Cross Teaching-Enhanced Multispectral Remote Sensing Object Detection With Transformer

Jiahe Zhu, Huan Zhang, Simin Li, Shengjin Wang, *Senior Member, IEEE*, and Hongbing Ma

***Abstract*—Remote sensing platforms are often equipped with sensors of multiple spectrums to capture the diverse reflective properties of ground areas, typically including the visible spectrum and the near infrared (NIR) spectrum. Moreover, thermal infrared (TIR) sensors capture the radiated heat of targets and are capable of all-day observation regardless of illumination conditions. By leveraging the complementary features of different spectrums, multispectral fusion techniques enhance the precision and robustness of remote sensing object detection methods. In this article, we present an object detection method for remote sensing imagery named multispectral detection transformer (multispectral DETR). The model fuses multispectral features with deformable attention and utilizes fused features for object detection. The multispectral deformable attention fusion block integrates the flexibility of dynamic weights with the principle of fusion based on local regions. Then, we propose a simple yet effective oriented object detection scheme based on angle prediction. Finally, we introduce a novel cross-teaching method between single-spectral and multispectral models, which alleviates the spectral interference issue caused by inconsistent target visibility. Experimental results demonstrate that multispectral DETR achieves state-of-the-art results on both the RGB-NIR VEDAI and the RGB-TIR DroneVehicle datasets.**

***Index Terms*—Detection transformer, feature fusion, knowledge distillation, multispectral remote sensing image, object detection.**

I. Introduction

OBJECT detection in remote sensing imagery is crucial in various applications, including urban planning, traffic management, search and rescue operations, as well as harbor and airport monitoring. Remote sensing platforms, such as satellites and drones, are often equipped with multispectral imaging sensors to capture ground areas with diverse reflective properties. The most commonly used spectrums in remote sensing platforms are the visible (RGB) spectrum and the infrared (IR) spectrum, particularly the near-infrared (NIR) spectrum [1], [2], [3]. Recently, the manufacturing of thermal infrared (TIR) cameras has become more feasible. TIR cameras can capture the radiated heat of objects in the 8–14 μm wavelength range, enabling all-day observation of ground scenes regardless of illumination conditions. This overcomes the imaging limitation of traditional remote sensing platforms, which are restricted to daytime observation. However, TIR images lack color and texture information, making it challenging for humans and computer vision algorithms to distinguish between objects of different categories. In contrast, RGB images capture rich color and texture information but are susceptible to illumination conditions. Samples of RGB, NIR, and TIR remote sensing images are shown in Fig. 1.

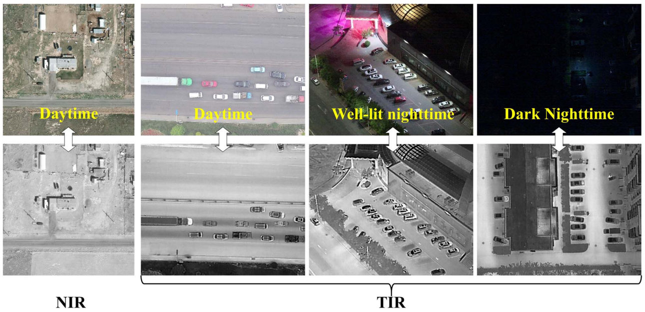

Fig. 1. Images captured with RGB, NIR, and TIR cameras on remote sensing platforms [4], [5] at different times.

Considering the complementary or enhancing features of different spectrums, an all-day remote sensing object detection method with high precision can be developed through fusion of multispectral information. The majority of previous studies have shown that feature fusion, also known as halfway fusion, outperforms input image fusion and decision fusion in deep learning models [6], [7]. Existing research focused on designing convolutional blocks for multispectral feature map fusion [8], [9], [10], a process that often involves extensive trial and error. Recently, some studies [11], [12] explored the use of transformers for feature fusion [13]. The transformer's ability to handle an arbitrary number of input tokens enables the joint processing of RGB and IR tokens, allowing the model to compute their relationships. However, the original self-attention mechanism of transformer can incorrectly fuse unrelated regions of the RGB-IR image pairs, such as fusing the upper left corner of the RGB image with the bottom right corner of the IR image.

In this work, we propose a multispectral fused oriented object detector based on the detection transformer (DETR) mechanism [14] and deformable attention [15]. Our proposed

Received 1 September 2024; revised 29 October 2024; accepted 15 November 2024. Date of publication 27 November 2024; date of current version 30 December 2024. This work was supported in part by the National Key Research and Development Program of China in the 14th Five-Year under Grant 2021YFFO602103 and Grant 2021YFF0602102, and in part by Shanghai Aerospace Science and Technology Innovation Foundation under Grant SAST2019-048. *(Corresponding author: Hongbing Ma.)*

Jiahe Zhu, Huan Zhang, Shengjin Wang, and Hongbing Ma are with the Department of Electrical Engineering, Tsinghua University, Beijing 100084, China (e-mail: hbma@tsinghua.edu.cn).

Simin Li is with the Institute of Electronic Engineering, China Academy of Engineering Physics, Mianyang 621900, China.

Digital Object Identifier 10.1109/JSTARS.2024.3504549

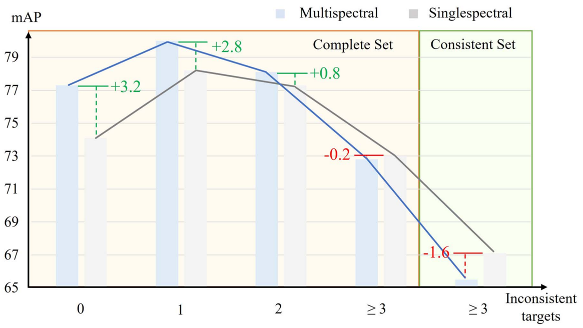

Fig. 2. Performance comparison between Single-spectral DETR and multi-spectral DETR with different numbers of inconsistent targets. The complete set is the original training set of DroneVehicle. The consistent set is a subset that excludes training images with more than two inconsistent targets.

Multi-spectral DETR) fuses RGB-IR features using an extended multispectral deformable attention (MDA) module, which samples RGB-IR feature tokens around the same coordinate with computed attention weights. The attention weights enable flexible feature fusion, while sampling around the same coordinate prevents fusion of unrelated regions. To deal with the arbitrary orientations of targets in remote sensing imagery, we also propose a simple yet effective oriented object detection scheme.

However, a fundamental assumption in multispectral fused object detection is that the same target should appear in the same position on both images, which may not hold true for RGB-TIR remote sensing image pairs captured at night. In extremely low-light conditions, many targets become invisible to human eyes in RGB images, as illustrated in the dark nighttime image pair in Fig. 1. We refer to these targets as inconsistent targets, as they do not consistently appear in both spectrums. The RGB features of the inconsistent targets might interfere with their TIR features, causing the fused features to have lower quality than TIR features alone.

To verify the feature interference problem in a multispectral model arising from inconsistent visibility, we compare the performance of multi-spectral DETR and its single-spectral counterpart on the DroneVehicle dataset [5], as shown in Fig. 2. Without losing generality, deformable DETR [15] trained and tested on TIR images is referred to as single-spectral DETR throughout this article. The test set of DroneVehicle is split according to the number of inconsistent objects. On the 6657 image pairs where all targets have consistent visibility, multispectral DETR outperforms single-spectral DETR by 3.2 in terms of mean average precision (mAP). However, as the number of inconsistent targets increases, the performance gap between multispectral DETR and single-spectral DETR narrows. On the 1456 image pairs containing at least three inconsistent objects, single-spectral DETR surpasses multispectral DETR by 0.2 in mAP. To emphasize the problem, we exclude images with no less than three inconsistent targets from the training set and retrain the single-spectral and multispectral models using the remaining consistent data. As shown in Fig. 2, the performance gap widens to 1.6. This finding confirms that, despite having access to the complete TIR image, the multispectral model is affected by the interference from invisible targets in the RGB spectrum, which degrades its performance. In other words, the multispectral model is lacking in robustness against inconsistent target invisibility. Therefore, a multispectral model needs to learn not only to fuse multispectral features but also to mitigate interference.

To address this issue, we further propose a knowledge distillation-based approach that encourages the multispectral model to mimic the single-spectral model, which is naturally unaffected by the interference problem, on inconsistent targets. Building upon this concept, we first use multispectral DETR to help train an enhanced single-spectral DETR, which is then used in return to guide multispectral DETR on inconsistent targets. In contrast to conventional unidirectional distillation methods [16] that transfer knowledge from a high-performance teacher to a student model, our proposed cross-teaching between single-spectral DETR and multispectral DETR (CT-SSMS) method combines the strengths of both single-spectral and multispectral models, thereby improving their performance simultaneously.

This article is an extended version of [17] and a conference paper in IGARSS 2024. However, in the previous publications, multispectral fusion based on deformable attention was not compared with other fusion choices and the design of oriented object detection was not thoroughly justified. Therefore, this extended paper makes several new contributions, including a comparison between multispectral feature fusion with convolution and fusion with deformable attention in terms of computational budget and precision. Also, we conduct an ablation study on key choices in our oriented object detection scheme quantitatively as well as qualitatively with visualizations of detection results and sampling points. Finally, we extend experiments to oriented object detection on the VEDAI dataset [4]. This extension and previous research revolve around one purpose: realizing a high-precision multispectral fused object detection model suited to remote sensing imagery in both daytime and nighttime scenes. The overall contributions of this extended journal version are summarized as follows:

1) We introduce multispectral DETR, a novel multispectral remote sensing object detection model leveraging the deformable attention mechanism. We compare the multispectral deformable attention module with input fusion and convolutional fusion in terms of precision and computational efficiency. In addition, we design a simple yet effective oriented object detection scheme applicable to DETR-like models.
2) We propose cross-teaching between single-spectral and multispectral DETRs (CT-SSMS) to address the interference issue caused by inconsistent targets. CT-SSMS enhances single-spectral DETR with the high-performance multispectral DETR, while utilizing the enhanced single-spectral DETR to guide multispectral DETR on inconsistent targets.
3) Comprehensive experiments on DroneVehicle and VEDAI datasets demonstrate the state-of-the-art performance of multispectral DETR and the effectiveness of the CT-SSMS method.

II. Related Works

A multispectral fused remote sensing object detector comprises two essential components: multispectral fusion and oriented object detection. The former enables all-time detection, while the latter is necessitated by the characteristics of remote sensing platforms. This section reviews relevant research in general-purpose object detection, remote sensing object detection, and multispectral fused object detection, with a focus on transformer-based approaches.

A. General-Purpose Object Detection

In the deep learning era, object detection algorithms were initially categorized into two-stage models [18], [19], [20] and one-stage models [21], [22], [23]. The advent of transformer [13] led to the development of DETR [14], which introduced a query mechanism into the object detection task. The core of DETR is an encoder–decoder architecture consisting of transformer modules, where the encoder performs global feature attention and the decoder queries the image feature map to determine the presence of a target of interest in its corresponding region. By leveraging the Hungarian algorithm, DETR realizes one-to-one matching between object queries and ground-truth (GT) objects, eliminating the need for nonmaximum suppression (NMS). However, the original DETR model is plagued by slow training and low precision in detecting small targets.

Several subsequent studies [15], [24], [25], [26] proposed different solutions to address the two limitations of DETR. Specifically, deformable DETR [15] introduces the deformable transformer as an alternative module for the original transformer. The deformable transformer calculates sampling offsets relative to the query's reference position and their attention weights by passing the query vector through linear layers. This conversion from global to local attention introduces an inductive bias for object detection, facilitating model training. Also, the reduced computational budget enables the use of multiscale feature maps, leading to improved small target detection.

Although deformable DETR reduces computational costs in single-scale attention, it still struggles with a computational bottleneck of multiscale attention. Sparse DETR [27] observes that most regions of the feature maps do not require attention enhancement, especially those not queried by the decoder. Therefore, sparse DETR proposes to sparsify the encoder, which fuses local information in a subset of regions predicted by a scoring network. We adopt the sparsification introduced in sparse DETR into our single-spectral and multispectral DETRs due to its computational efficiency.

B. Remote Sensing Object Detection

Remote sensing platforms capture ground scenes from an overhead perspective, resulting in targets with random orientations and dense distributions. Oriented object detection was proposed to make bounding boxes align with the orientations of targets so that one bounding box does not contain multiple objects. A straightforward approach to oriented object detection is to predict an angle for each bounding box, yielding a five-parameter representation (x, y, w, h, θ), where (x, y) denote the center coordinates, (w, h) represent width and height of the bounding box, and θ is the orientation angle. However, the periodic nature of angles can lead to abrupt increases in loss values at boundary cases [28], [29], [30].

Previous studies have proposed solutions based on two main approaches. One approach is to transform the angle regression task into an angle classification task [28], [29], [31]. For instance, circular smooth loss [28] reformulates angle prediction as a 180-classification problem, with GT values in the range of $[-90^\circ, 90^\circ)$. The key is the use of smooth labels, which tolerates classifying one angle into its neighboring angles, including boundary cases. In specific, predicting -90° for an angle with GT value 89° will not incur a large loss value. Then, densely coded labels [29] furthers this approach by reducing the encoding length.

The second line of approaches uses loss functions that measure the Intersection-over-Union (IoU) between bounding boxes, rather than directly measuring parameter differences [32], [33]. In boundary cases, IoU-based losses only require high overlap between predicted and GT boxes, without directly comparing predicted and GT angles, thereby circumventing the periodicity issue. Building on this idea, the Rotated IoU Loss [34] measures the intersection area of two rotated rectangles with the Shoelace theorem [35]. Several studies transform oriented bounding boxes into Gaussian distributions and measure the distance between them using Gaussian Wasserstein Distance (GWD) Loss [32] and Kullback–Leibler Divergence (KLD) loss [33], respectively. KFIoU Loss [36] first obtains the overlapping distribution of the two Gaussians with Kalman filtering, and then converts the overlapping Gaussian into an oriented box, whose area approximates the overlap of the two oriented boxes.

Recently, efforts have been devoted to integrating the DETR method with oriented object detection. AO2 DETR [37] modifies different components of deformable DETR to accommodate oriented bounding box prediction, including an oriented proposal generation mechanism, an adaptive oriented proposal refinement module, a transformer decoder with angle prediction, and a set of rotation-aware matching costs. However, AO2 DETR is susceptible to the angle periodicity problem due to its use of angle regression with L1 Loss. ARS DETR [38] adopts an angle classification approach and introduces an aspect ratio-aware circular smooth label method. In addition, it proposes rotating the sampling points of deformable attention based on a coarse predicted angle.

In this work, we employ the Rotated IoU Loss [34] to supervise angle prediction. Within this framework, we investigate various choices of L1 Loss and sampling point rotation. Furthermore, we demonstrate that our proposed oriented object detection scheme for DETR achieves comparable performance to recent state-of-the-art methods [37], [38].

C. Multispectral Fused Object Detection

Given the complementary or enhancing features of RGB-IR images, fusing multispectral information can enhance the precision and robustness of downstream vision tasks. In the context of RGB-TIR fused pedestrian detection for autonomous driving, Liu et al. [6] compared various fusion stages, including early

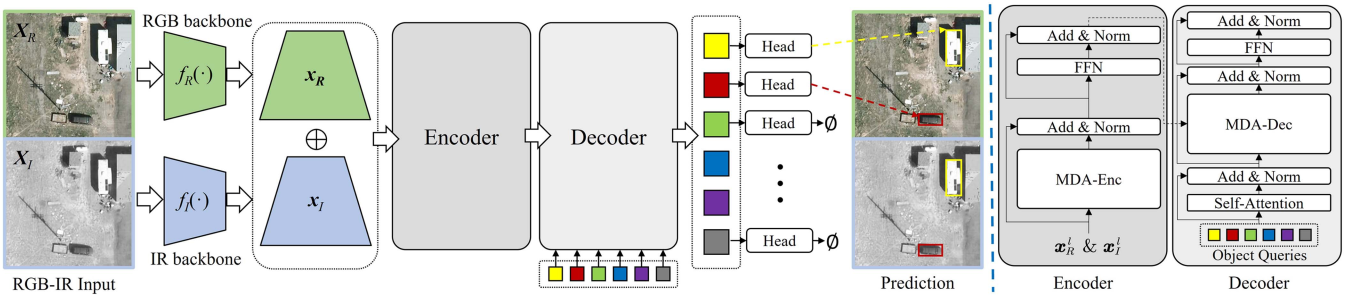

Fig. 3. Structure of multispectral DETR. There are four main components, namely, the RGB backbone, the IR backbone, the multispectral encoder, and the multispectral decoder. The RGB backbone and the IR backbone are separable and do not share weights.

fusion, halfway fusion, late fusion, and score fusion, and showed that halfway fusion yielded the best performance. Early studies of feature fusion [6], [39] combine multispectral feature maps into a single feature map and propagate it through subsequent convolutional layers. Later, Fusion CSPNet [7] reveals that it is more effective to propagate RGB-IR images through complete and independent backbones and then extract their feature maps from different stages for fusion.

In remote sensing imagery, such as satellites, uncrewed aerial vehicles (UAVs), and drones are often equipped with sensors of multiple spectrums, it is logical to develop multispectral fused object detection methods. On the vehicle detection in aerial imagery (VEDAI) dataset [4], which provides RGB-NIR images, YOLOrs [40] also compares input fusion with feature fusion and demonstrates the superiority of feature fusion. Fang et al. [10] proposed a cross-modality attentive feature fusion network designed to select shared features across modalities while amplifying modality-specific features.

With the DroneVehicle dataset [5] providing RGB-TIR image pairs, recent researches have also identified the problem of inconsistent visibility. C2Former [11] describes the inconsistent visibility problem as fusion imprecision and proposes a transformer-based fusion module for RGB-TIR fusion. CALNet [12], on the other hand, describes the same problem as semantic conflict. Their proposed solution is a conflict rectification module based on transformer, which incorporates K-nearest-neighbors (KNN) to prevent the fusion of conflicting features. Our work characterizes the same problem as interference, highlighting that it is the low-quality RGB features that cause this problem. We propose to address this issue with CT-SSMS, a training-time-only knowledge distillation method that does not change the model architecture.

III. Proposed Method

In this section, we first introduce the overall architecture of multispectral DETR, focusing on its MDA block and oriented object detection scheme. Then, the proposed CT-SSMS method is described in detail.

A. Structure of Multispectral DETR

Multispectral DETR is a multispectral object detection model that leverages the deformable attention mechanism [15]. The architecture of multispectral DETR is illustrated in Fig. 3. The input to multispectral DETR is the RGB-IR aerial image pair $\{\boldsymbol{X}_R, \boldsymbol{X}_I\}$ that shows different information about the same scene. The two images go through their separate backbones. Multiscale image features are extracted as follows:

$$\boldsymbol{x}_R^l = f_R\left(\boldsymbol{X}_R\right), \boldsymbol{x}_I^l = f_I\left(\boldsymbol{X}_I\right), l = 1, 2, \ldots, L \qquad (1)$$

where f_R and f_I denote the RGB and IR backbones, and L is the number of feature levels. Then, multiscale feature maps from the two spectrums are concatenated along the level dimension, forming a $2L$-level feature pyramid. Note that the multispectral features are not fused within the two backbones, which have separate weights.

Subsequently, the RGB and IR feature maps, each comprising L levels, are fused in the encoder. In the decoder, object queries calculate attention with the fused feature maps. The output are decoder embeddings with the same dimension as the object queries, which will be processed with the prediction heads into class labels and bounding boxes. The specific structures of the encoder and decoder will be discussed in the next section. The GT value for each prediction is determined via the one-to-one matching mechanism of DETR. The optimal matching result is obtained by minimizing the matching cost

$$\hat{\sigma} = \arg\min_{\sigma} \sum_{i=1}^{N_q} \mathcal{L}_{\text{match}}(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_{\sigma(i)}) \qquad (2)$$

where $\boldsymbol{y}_i$ is the ith GT bounding box and $\hat{\boldsymbol{y}}_{\sigma(i)}$ is the predicted bounding box of the $\sigma(i)$th decoder embedding. In short, by minimizing the matching cost, each predicted bounding box is assigned to a GT box in a way that the overall distance between the predicted and GT boxes is the smallest.

Multispectral fused object detection assumes that every target appears in the same position of both spectrums. Therefore, given an aligned RGB-IR aerial image pair, multispectral DETR generates a single set of predictions that encompasses all objects, rather than producing separate predictions for each spectrum.

B. Multispectral Deformable Attention

To enable RGB-IR fusion, we extend the deformable attention block to a MDA block. The structure of the MDA block in the encoder (MDA-Enc) is shown in Fig. 4(a). In MDA-Enc, queries are feature tokens of the two spectrums. A query $\boldsymbol{q}^{l_0}$ is a token

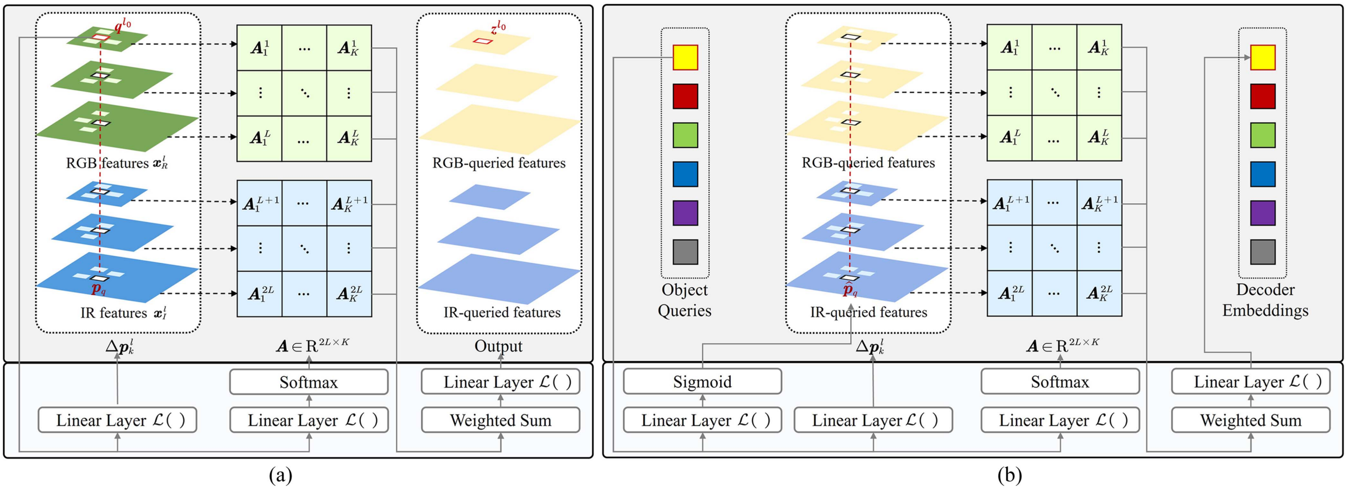

Fig. 4. Structures of the MDA blocks in the encoder and decoder of multispectral DETR. Queries are RGB or IR feature tokens in the encoder, which performs multispectral feature fusion. Queries are object queries in the decoder, where attention is calculated with multispectral fused features. (a) Multi-spectral deformable attention in encoder. (b) Multi-spectral deformable attention in decoder.

from the l_0th layer in the feature pyramid as

$$q^{l_0} = x^{l_0}\left(p_q\right), l_0 = 1, 2, \ldots, 2L \tag{3}$$

where p_q is the 2-D position of the query, which serves as the reference point and $x^{l_0}(p_q)$ means picking the query token from position p_q of the feature map x^{l_0}. As shown in Fig. 4(a), when $l_0 \leq L$, q^{l_0} is an RGB feature token. Conversely, when $L < l_0 \leq 2L$, q^{l_0} corresponds to an IR feature token. On each of the $2L$ levels, K sampling offsets Δp_k^l are calculated by passing query token q^{l_0} through a linear layer, yielding relative positions with respect to p_q. As a result, the sampled points on the first half of layers are RGB features, while those on the second half are IR features. Likewise, the attention weight matrix $A \in R^{2L \times K}$ is obtained by passing q^{l_0} through a linear layer followed by a softmax layer. Finally, all sampling points, processed with another linear layer, are multiplied by their corresponding attention weights and aggregated into an output token z^{l_0} at the same position as q^{l_0}

$$z^{l_0} = \mathcal{L}\left\{\sum_{l=1}^{L}\sum_{k=1}^{K}[A_k^l \times \mathcal{L}(x_R^l(p_q + \Delta p_k^l)) \right.$$
$$\left. + A_k^{l+L} \times \mathcal{L}(x_I^l(p_q + \Delta p_k^{l+L}))]\right\} \tag{4}$$

where $\mathcal{L}()$ denotes a linear layer. When $l_0 \leq L$, z^{l_0} is an RGB-queried multispectral fused token; otherwise, when $L < l_0 \leq 2L$, z^{l_0} is an IR-queried token.

As can be seen, the deformable attention block can inherently manage multispectral fusion. Compared to multispectral fusion based on global attention of the original transformer, which calculates attention between unrelated regions, the local attention of the deformable transformer performs fusion at the same positions across RGB-IR spectrums.

The structure of MDA in the decoder (MDA-Dec) is shown in Fig. 4(b). In MDA-Dec, queries are the N object queries. A reference point $\hat{p}_q$ is first calculated with the object query passing through a linear layer and a sigmoid function. Similarly, sampling offsets and attention weights are computed with the object queries. The weighted sum of the sampled feature tokens is the same as (4). By viewing RGB-queried and IR-queried features together, the object queries are updated into decoder embeddings, which contain features of both RGB and IR spectrums and information of target category and position. Therefore, subsequent prediction heads can process the decoder embeddings into bounding boxes.

While being reasonable, MDA offers a more flexible and generalizable alternative to traditional multispectral feature fusion methods, such as feature map addition or concatenation followed by a 1 × 1 convolution, as used in [6] and [7]. Notably, when L and K are set to 1, all Δp_k^l are 0, the feature token at the query position is summed with the corresponding feature token from the other spectrum, with their weights learned by the model. Under these restrictions, the MDA block resembles convolutional fusion.

The computational complexity of MDA fusion can also be compared with convolutional fusion. It is worth noting that convolutional fusion lacks the capabilities of multiscale fusion and spatial attention, requiring additional modules to achieve comparable effects. To simplify the calculation, we limit our discussion to fusion at a single scale. Assume the sizes of the RGB feature map and the NIR feature map are both $H \times W \times C$, where H, W, and C represent the height, width, and number of channels, respectively. Then:

(1) The computational complexity of convolutional fusion is $2HWC^2$ when the output feature map has C channels. When the output feature map has $2C$ channels, the computational complexity is increased to $4HWC^2$.

(2) According to [15], the computational complexity of the deformable attention module is given by

$$\Omega(N_q, K, H, W, C) = 2N_qC^2 + \min(HWC^2, N_qKC^2) \tag{5}$$

where N_q is the number of query vectors, and K is the number of sampling offsets. In the MDA block, due to the introduction of the encoder sparsification method [27], N_q is $\rho \times H \times W$, where ρ is the update ratio (set to 30% according to the default choice of [27]), and K is 2×4. Since the RGB feature points and IR feature points alternately serve as query, the computation in (5) is repeated twice, resulting in the final computational complexity of

$$\Omega(0.3HW, 8, H, W, C) = 3.2HWC^2. \tag{6}$$

The module's output has 2C channels.

From the analysis above, it is evident that the MDA block has a comparable computational complexity ($3.2HWC^2$) to convolutional fusion ($4HWC^2$) due to encoder sparsification [27]. The module also includes spatial attention, making it an efficient multispectral fusion block.

C. Oriented Object Detection With DETR

DETR and its variants are typically designed for horizontal object detection, rendering them inadequate for handling targets of arbitrary orientations in remote sensing imagery. To address this limitation, we propose a simple yet effective solution to enable oriented object detection with DETR. We adopt a basic angle-based representation, where the output of the bounding box regression branch is modified to predict an angle θ alongside the (x, y, w, h) parameters. Matching cost and loss functions associated with the bounding boxes are also adapted, where Rotated IoU Loss [34] serves as the primary guide for oriented bounding box matching and regression

$$\mathcal{L}_{\text{RIoU}} = 1 - \frac{\boldsymbol{A}_P(B_1 \cap B_2)}{\boldsymbol{A}_R(B_1) + \boldsymbol{A}_R(B_2) - \boldsymbol{A}_P(B_1 \cap B_2)} \tag{7}$$

where B_1 and B_2 are the two oriented bounding boxes, $\boldsymbol{A}_R()$ is the function to calculate the area of a rectangle, and $\boldsymbol{A}_P()$ is the function to calculate the area of a polygon. Notably, the loss function in (7) does not directly involve the angles of B_1 and B_2, thereby avoiding the periodicity problem.

L1 loss is another loss function used in original DETR, which directly compares estimated parameters against GT parameters. Contrary to the original design, we do not impose constraints on w, h, or θ to circumvent the angle periodicity problem [28]. Center point coordinates (x, y) remain in L1 loss to facilitate convergence of DETR. With these modifications, DETR gains the capability to effectively detect oriented objects. The effectiveness of these design choices will be evaluated through experiments.

D. Cross Teaching Between Single-Spectral DETR and Multispectral DETR

In this section, we first specify how to determine consistent and inconsistent objects. The DroneVehicle dataset provides separate annotations for RGB and TIR images. Following [5], TIR annotations are considered GT as TIR images are not affected by varying lighting conditions. RGB annotations reflect whether the annotator can identify the target in the RGB image, i.e., whether the target is visible. Therefore, objects are deemed consistent if they possess annotations in both RGB and TIR images.

Since vehicles do not overlap in remote sensing imagery, consistency can be determined by calculating the IoU between the ith TIR bounding box and each RGB bounding box. The ith TIR object is labeled as a consistent object if there exists an RGB bounding box that satisfies

$$\text{IoU}(B_i^{\text{TIR}}, B^{\text{RGB}}) \geq T \tag{8}$$

where IoU() is the IoU calculator and T is the IoU threshold. Conversely, objects annotated exclusively in the TIR images are considered inconsistent. This means that if no RGB bounding box satisfies (8), the ith TIR bounding object is labeled as inconsistent. Taking the misalignment [41], [42] problem into consideration, threshold T is set to 0.3. It turns out that, while the majority of objects captured during daytime hours or well-lit nighttime conditions exhibit consistency, a notable proportion of targets captured in dark nighttime scenes are inconsistent. For example, Fig. 5(a) shows a pair of RGB-TIR images with a total of ten targets identified from the TIR image, 7 of which are consistent. The remaining three targets are not visible in the RGB image, thus labeled as inconsistent.

The overview of CT-SSMS is illustrated in Fig. 5(b) with consistency determined. To solve the interference problem, the knowledge of single-spectral DETR, which is unaffected by RGB features, is utilized to teach multispectral DETR. On inconsistent targets, multispectral DETR is trained to mimic the output of single-spectral DETR with distillation losses. Also, it has been demonstrated that a single-spectral model can learn from a multispectral model [43], despite having no access to the second spectrum. Therefore, the knowledge of multispectral DETR can also be utilized to enhance the overall performance of single-spectral DETR. Based on the finding that multispectral DETR performs worse on inconsistent targets, distillation loss from multispectral DETR to single-spectral DETR is only calculated on consistent targets. When every iteration of training contains knowledge distillation of the two directions, the cross teaching cycle is formed, which is a virtuous circle that combines the strengths of both multispectral and single-spectral DETRs.

For the proposed fine-grained knowledge distillation based on visibility consistency, the one-to-one matching mechanism of DETR [14] is considered. In each iteration, the bipartite matching results of single-spectral DETR and multispectral DETR are first obtained with (2). Then, knowledge is distilled between object queries that match the same GT object. For the ith GT object, if it is consistent, the matched single-spectral object query Q_i^{SS} learns from the multispectral object query Q_i^{MS}. Conversely, if the object is inconsistent, Q_i^{MS} learns from Q_i^{SS}. The learning directions in CT-SSMS are illustrated in Fig. 5(b). For the seven consistent targets, the learning direction is from multispectral DETR to single-spectral DETR (MS2SS). For the three inconsistent targets, reverse distillation is performed from single-spectral DETR to multispectral DETR (SS2MS).

The knowledge distillation loss is computed based on detected bounding boxes and includes both classification and regression parts. KLD loss is applied to the classification part based on the

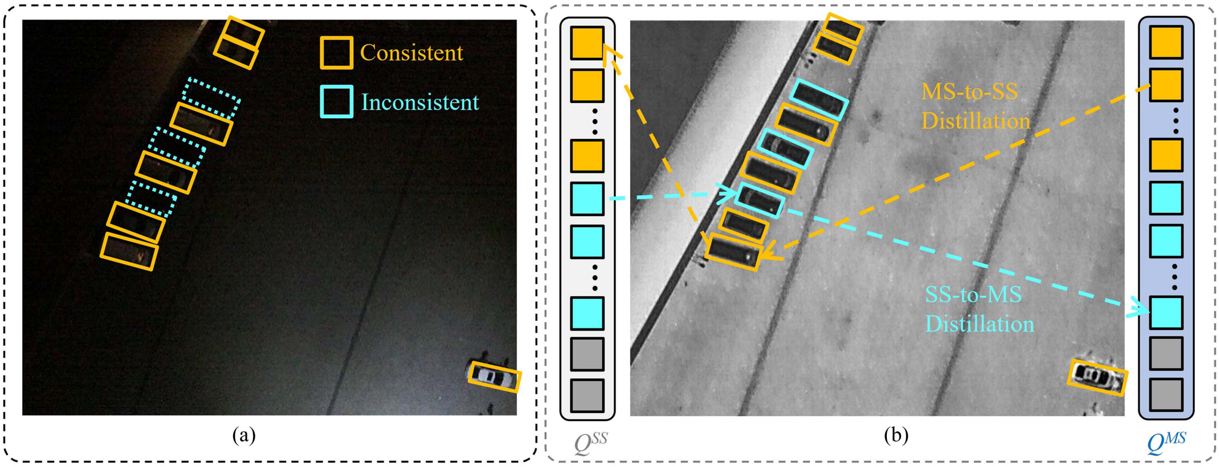

Fig. 5. Demonstration of the proposed CT-SSMS method. Note that knowledge distillation is done on all pairs of object queries that match a GT target. For clarity of demonstration, only one arrow is shown for each distillation direction. (a) Consistent and inconsistent objects. (b) Cross teaching directions.

determined learning directions

$$\mathcal{L}_{KLD} = \sum_{i=1}^{N_q} \Bigg(\mathbb{I}_{\text{Consistent}}(i) \times \hat{\boldsymbol{y}}^{cls*}_{\hat{\sigma}^{MS}(i)} \log \frac{\hat{\boldsymbol{y}}^{cls*}_{\hat{\sigma}^{MS}(i)}}{\hat{\boldsymbol{y}}^{cls}_{\hat{\sigma}^{SS}(i)}} + \mathbb{I}_{\text{Inconsistent}}(i) \times \hat{\boldsymbol{y}}^{cls*}_{\hat{\sigma}^{SS}(i)} \log \frac{\hat{\boldsymbol{y}}^{cls*}_{\hat{\sigma}^{SS}(i)}}{\hat{\boldsymbol{y}}^{cls}_{\hat{\sigma}^{MS}(i)}} \Bigg) \tag{9}$$

where $\mathbb{I}$ is the indicator function, $\hat{\boldsymbol{y}}_{\hat{\sigma}^{MS}(i)}$ and $\hat{\boldsymbol{y}}_{\hat{\sigma}^{SS}(i)}$ are the predictions of multispectral DETR and single-spectral DETR that match the same GT box, superscript cls denotes the classification part of the prediction, and * indicates that gradient is not propagated through this vector. For the regression part, mean square error (MSE) loss is used

$$\mathcal{L}_{MSE} = \sum_{i=1}^{N_q} \Bigg(\mathbb{I}_{\text{Consistent}}(i) \times \left\| \hat{\boldsymbol{y}}^{reg*}_{\hat{\sigma}^{MS}(i)} - \hat{\boldsymbol{y}}^{reg}_{\hat{\sigma}^{SS}(i)} \right\|^2 + \mathbb{I}_{\text{Inconsistent}}(i) \times \left\| \hat{\boldsymbol{y}}^{reg*}_{\hat{\sigma}^{SS}(i)} - \hat{\boldsymbol{y}}^{reg}_{\hat{\sigma}^{MS}(i)} \right\|^2 \Bigg) \tag{10}$$

where superscript reg denotes the regression part of the prediction.

IV. Experimental Settings

A. Datasets

Currently, two public datasets are available for the task of multispectral fused remote sensing object detection: the VEDAI dataset [4] with RGB-NIR spectrums and the DroneVehicle dataset [5] with RGB-TIR spectrums.

The DroneVehicle dataset [5] was collected using a DJI M200 drone with a Zenmuse XT 2 imaging device, with the TIR camera operating in the 7.5–13.5 μm wavelength range. The dataset contains 28 439 pairs of 640 × 512 resolution RGB-TIR images, annotated with five categories of vehicles: car, truck, bus, freight-car, and van. The DroneVehicle dataset is divided into 17 990 pairs of training images, 1469 pairs of validation images, and 8980 pairs of test images. Most experiments in this section adhere to this division. For experiments on knowledge distillation, we also use subsets of the test set with 0 and ≥ 3 inconsistent targets, respectively.

The VEDAI dataset [4] was built on the high-resolution satellite images from the Utah Automated Geographic Reference Center in June 2012. The dataset contains 1251 pairs of fully aligned RGB and NIR images at a resolution of 1024 × 1024. The images in the VEDAI dataset are divided into ten folds, each containing 1089 training images and 121 test images. The dataset includes annotations for a total of nine categories: car, truck, small truck, van, camper, tractor, plane, boat, and other. The ablation experiments in this article are conducted on the first fold of the VEDAI dataset, while the comparative experiments with existing research use tenfold cross-validation.

B. Implementation Details

The multispectral DETR model is implemented with the Pytorch framework [44]. The AdamW [45] optimizer is used with an initial learning rate set to 10^{-4}. Swin transformer [46] is selected as RGB and IR backbones.

On the DroneVehicle dataset, the model is trained for 60 epochs, with the learning rate reduced by a factor of 0.1 after the 40th epoch. On the VEDAI dataset, due to its smaller number of samples, the training schedule is extended to 120 epochs, with the learning rate decayed after the 100th epoch.

In CT-SSMS, both single-spectral DETR and multispectral DETR serve as teacher models and are initially trained separately for 50 epochs using default hyperparameter settings. Then, CT-SSMS is carried out for ten epochs with a learning rate of $1e-5$. Original detection losses are included during CT-SSMS.

The comparisons on the DroneVehicle and VEDAI datasets use mAP under IoU threshold 0.5 as the evaluation metric.

V. Discussion

In this section, we first discuss the design choices of multispectral DETR through ablation studies, focusing on the MDA

TABLE I
COMPARISON OF DIFFERENT FUSION METHODS

Method	GFLOPs	Params(M)	DroneVehicle	VEDAI
RGB-only	51.83	41.57	67.4	75.6
IR-only	51.83	41.57	74.5	73.8
Input Fusion	51.86	41.57	75.3	76.4
Convolutional Fusion	104.12	74.18	76.3	77.0
MDA (Ours)	103.66	72.36	**76.9**	**77.3**

The bold values represent highest score among different methods.

TABLE II
ABLATION STUDY ON ORIENTED OBJECT DETECTION

L1 Loss	RDA	DroneVehicle	VEDAI
(x, y, w, h, θ)		67.3	72.6
–	–	76.7	77.2
(x, y)		**76.9**	**77.3**
(x, y)	From scratch	76.0	76.5
	Finetune	76.6	77.1

The bold values represent highest score among different methods.

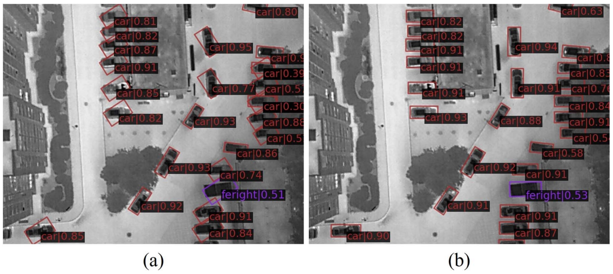

Fig. 6. Predicted oriented bounding boxes of multispectral DETR. (a) Including all (x, y, w, h, θ) parameters in L1 Loss. (b) Including only (x, y) in L1 Loss.

block and the proposed oriented object detection scheme. Then, several sets of experiments on knowledge distillation are conducted to provide a detailed discussion on the effectiveness of CT-SSMS. Further, we compare our method with state-of-the-art remote sensing object detection methods on both RGB-NIR and RGB-TIR datasets. Finally, by analyzing limitations and challenges of the proposed methods, we suggest several directions for future research.

A. Ablation Studies

In Table I, different fusion methods are compared with single-spectral models on the two datasets. Computational budget measured in giga floating point operations (GFLOPs) and network parameters are also compared. GFLOPs is measured assuming input resolution 640 × 512 of the DroneVehicle dataset. The RGB-only and IR-only models are sparse DETRs trained and tested exclusively on RGB and IR images, respectively. The input fusion model, treating the IR image as a fourth channel concatenated to the RGB channels, can be implemented with Sparse DETR by modifying the first layer of the backbone. Convolutional Fusion concatenates different stages of feature maps along the channel dimension and fuses them with a 1×1 convolution before feeding them to the deformable attention blocks.

As shown in Table I, the IR-only model outperforms the RGB-only model by a large margin on DroneVehicle due to the invisibility of RGB targets in nighttime scenes. By contrast, the RGB-only model has a higher mAP on VEDAI, which contains daytime scenes only. Therefore, subsequent experiments use the TIR spectrum for DroneVehicle and the RGB spectrum for VEDAI when evaluating single-spectral models.

As for multispectral fusion, on both datasets, incorporating multispectral information improves the accuracy of remote sensing detection. Like existing works [6], [7], it verifies the effectiveness of fusing the complementary features of RGB and IR images. Among the three compared fusion methods, Input Fusion shows slightly better results than the single-spectral models. The convolutional fusion model demonstrates some improvement over the input fusion model, which is consistent with most studies [6], [7]. Among the fusion methods, MDA achieves the highest results, with an mAP of 76.9 on DroneVehicle and 77.3 on VEDAI. In terms of computational budget and network parameters, while the proposed multispectral DETR with MDA as the fusion method has higher performance than the convolutional fusion model, it is also more lightweight.

Next, in Table II, we compare different choices of L1 Loss and rotated deformable attention (RDA) proposed in [38]. Rotated IoU Loss is consistently used in all experiments. It turns out that simply adding an angle parameter θ in model prediction and L1 Loss falls into the angle periodicity problem. As visualized in Fig. 6(a), when a target has an angle with the image boundary, the predicted oriented bounding box is accurate. However, when a target is nearly horizontal, the predicted bounding box remains rotated. The reason can be that, in the training phase, the model is often penalized for predicting nearly horizontal boxes, as boundary cases lead to a large L1 Loss. On the other hand, Rotated IoU requires that the predicted boxes overlap with the GT boxes. Consequently, the model finds a compromise by predicting a rotated box for every target, which is a local minimum where neither L1 Loss nor rotated IoU Loss is excessively large.

As parameters (x, y) are not influenced by the periodicity problem, while parameters (w, h, θ) have boundary cases, experiments were conducted by removing (w, h, θ) from L1 Loss and by removing L1 loss altogether. It can be seen from Fig. 6(b) that multispectral DETR with the modified L1 Loss can accurately predict oriented boxes, indicating alleviation of the periodicity problem. Table II shows that, eliminating the periodicity problem significantly improves the performance of multispectral DETR. In addition, keeping (x, y) in L1 Loss provides a slight advantage on the datasets.

Under this scheme, the sampling points of MDA are visualized in Fig. 7(d). On a shallow, high-resolution feature map layer, the sampling points are searching for the edges of targets, with points on the edge having larger attention weights. On a deep, low-resolution feature map layer, the sampling points form three crosses, which is the result of the eight attention heads sampling in eight different directions. This regular pattern of sampling points indicate that the DETR model is well-trained.

The problem is that the sampling points that match a target do not automatically align with the direction of the target. The RDA mechanism [38] was proposed to rotate the sampling points to a roughly predicted angle. As shown in Fig. 7(a), the sampling points of ARS DETR align with the targets. When adopted in

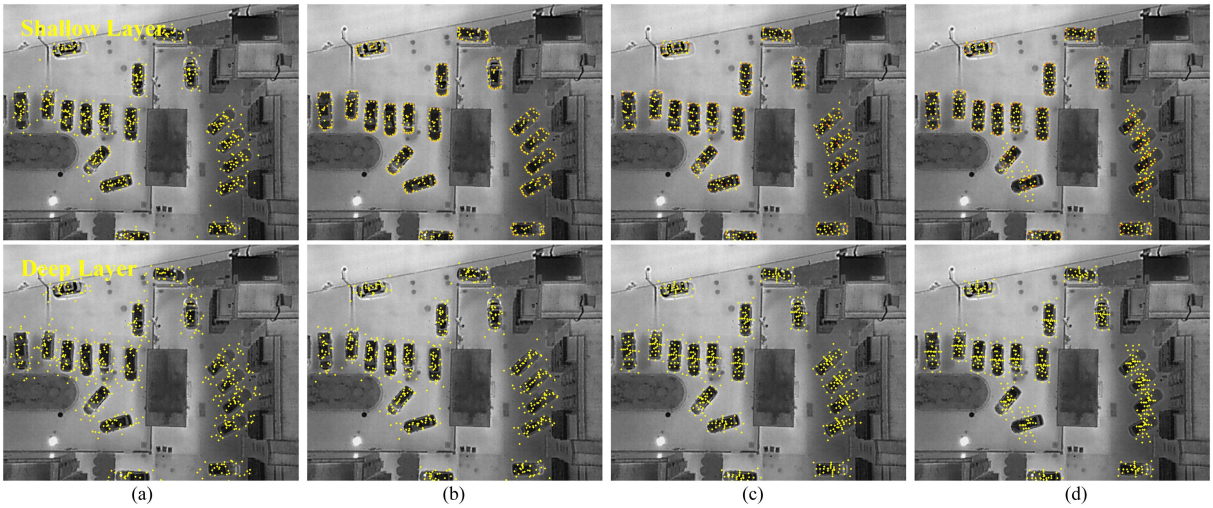

Fig. 7. Sampling points of deformable attention on shallow and deep feature map layers. (a) ARS DETR [38] with RDA. (b) RDA from scratch. (c) RDA finetune. (d) Ours.

TABLE III
RESULTS OF UNIDIRECTIONAL KNOWLEDGE DISTILLATION AND CT-SSMS

Tested Model	Distillation	Test set		
		Complete	0 Inconsistent	$\geq$ 3 Inconsistent
Single-spectral DETR	–	74.5	74.1	73.0
	Unidirectional MS2SS (All targets)	74.7(+0.2)	**74.7(+0.6)**	73.6(+0.6)
	Unidirectional MS2SS (Inconsistent Targets)	**75.2(+0.7)**	74.5(+0.4)	**74.4(+1.4)**
	–	74.5	74.1	73.0
	CT-SSMS (Time)	76.4(+1.9)	75.4(+1.3)	75.0(+2.0)
	CT-SSMS (Consistency)	**76.6(+2.1)**	**75.9(+1.8)**	**75.3(+2.3)**
Multi-spectral DETR	–	76.9	77.3	72.8
	Unidirectional SS2MS (All targets)	77.0(+0.1)	77.6(+0.3)	72.8(+0.0)
	Unidirectional SS2MS (Inconsistent Targets)	**77.2(+0.3)**	**77.7(+0.4)**	**73.3(+0.5)**
	–	76.9	77.3	72.8
	CT-SSMS (Time)	77.5(+0.6)	77.9(+0.6)	**73.7(+0.9)**
	CT-SSMS (Consistency)	**77.7(+0.8)**	**78.2(+0.9)**	73.6(+0.8)

The bold values represent highest score among different methods.

multispectral DETR, as shown in Fig. 7(b), the RDA mechanism also rotates the sampling points to align with the targets. However, the RDA mechanism disrupts the regular pattern and leads to a 0.9-point decrease in the performance of multispectral DETR, as shown in Table II.

To simultaneously achieve the regular pattern and aligned sampling points, we test a finetune strategy that rotates the sampling points of a well-trained multispectral DETR. Meanwhile, the learning rate of the decoder, i.e., the linear layers that calculate sampling offsets, is decreased to 1e−6 to maintain the regular pattern of sampling points. As shown in Fig. 7(c) and Table II, although the finetuning strategy maintains the regular sampling pattern while rotating the sampling points, it does not improve model performance. Therefore, it can be concluded that the RDA module does not work on multispectral DETR with angle regression.

B. Results and Analysis of CT-SSMS

In this section, the proposed CT-SSMS method is evaluated on the RGB-TIR DroneVehicle dataset. Note that the interference problem is unique to RGB-TIR nighttime scenes. Thus, CT-SSMS is not tested on VEDAI, which does not have inconsistent targets.

Table III presents the results of unidirectional knowledge distillation and CT-SSMS. The performance is measured on the complete test set of DroneVehicle as well as the two representative subsets with 0 and ≥ 3 inconsistent targets, which is aligned with Fig. 2. On the complete test set, single-spectral DETR and multispectral DETR without distillation achieve mAP scores of 74.5 and 76.9, respectively. It is worth noting that, for fair comparison, the training schedule of the two baseline models have the same number of epochs as those with knowledge distillation as reported in Section IV-B. On the ≥ 3 inconsistent subset, though the model size of multispectral DETR is twice that of single-spectral DETR, its mAP is 0.2 lower.

Before proceeding to CT-SSMS, it is necessary to first analyze unidirectional knowledge distillation to show the importance of selecting consistent or inconsistent targets for distillation of the two directions. Table III compares two different unidirectional knowledge distillation schemes: one with all targets and one with selected consistent or inconsistent targets. When all targets are used for MS2SS knowledge distillation with KLD and MSE Losses, the performance improvement of single-spectral DETR is minimal. When consistent targets are selected for MS2SS distillation, however, the improvement becomes more pronounced, especially on the ≥ 3 inconsistent subset, where single-spectral DETR gains 1.4 mAP. Experimental results indicate that the knowledge of multispectral DETR is generally beneficial for single-spectral DETR. However, the output of multispectral DETR on inconsistent targets exhibits low precision, which has an adverse effect when used in distillation. Keeping 10% inconsistent targets from MS2SS distillation already brings noticeable improvements.

A similar trend is observed in SS2MS distillation. That is, using all targets for SS2MS distillation leads to minimal improvements for multispectral DETR (0.1 on the whole test and 0.0 on the ≥ 3 inconsistent subset), whereas selecting inconsistent targets for SS2MS distillation yields gains of 0.3 and 0.5 mAP, respectively. It may seem counterintuitive that, despite its lower performance, single-spectral DETR can still contribute to improving multispectral DETR when all targets are used for distillation. We hypothesize that the knowledge distillation loss serves as a regularization term for multispectral DETR. On the other hand, single-spectral DETR has an advantage over multispectral DETR in scenarios with numerous inconsistent targets. Thus, leveraging its output on inconsistent targets for SS2MS distillation has a positive impact on multispectral DETR.

In addition to our consistency-based CT-SSMS, we introduce a variant of CT-SSMS based on image capturing time, referred to as CT-SSMS (Time). We categorize images into day, well-lit night, and dark night categories, following [5]. As the name suggests, CT-SSMS (Time) utilizes all targets in day and well-lit night images for MS2SS distillation and reserves dark night targets for SS2MS distillation. As can be seen in Table III, both time-based and consistency-based CT-SSMS yield larger improvements over unidirectional distillation. Thus, the core competence of CT-SSMS lies in its bidirectional knowledge distillation, where the two models mutually reinforce each other, creating a positive feedback loop.

When comparing time-based and consistency-based CT-SSMS, it becomes apparent that the latter achieves greater improvements. The key of CT-SSMS is to use the better predictions of single-spectral DETR on certain targets to teach multispectral DETR, and conversely use the better predictions of multispectral DETR on other targets to teach single-spectral DETR. Accurate separation of the two sets of targets can lead to better effect of CT-SSMS. Therefore, this experimental result validates that consistency, rather than image capturing time, better separates targets on which single-spectral DETR performs better. However, in case separate annotations for the two spectrums are not available, CT-SSMS based on image capturing time serves as a viable alternative.

Finally, with consistency-based CT-SSMS, the mAP scores of single-spectral DETR and multispectral DETR are elevated to 76.6 and 77.7, respectively. Through CT-SSMS, the performance of single-spectral DETR approaches the multispectral baseline. The mAP of multispectral DETR on the ≥ 3 inconsistent subset (73.6) surpasses the baseline single-spectral DETR (73.0) by 0.6, indicating that the interference problem has been substantially mitigated and that the robustness of multispectral DETR is enhanced.

C. Comparison With Existing Methods

In this section, we first compare with existing methods for oriented object detection based on DETR in both single-spectral and multispectral settings. The results are summarized in Table IV. ARS DETR [38] and AO2 DETR [37] can be expanded to multispectral models with improved performance using the MDA block in Fig. 4, which verifies that the proposed MDA block can generalize to other methods. As shown in Table IV, our angle regression-based approach outperforms AO2 DETR, which is hindered by the angle periodicity issue, and performs on par with ARS DETR, which is based on angle regression, across the two datasets. It is worth noting that the DroneVehicle dataset has a much larger number of testing images than VDEAI, which better reveals the precision and generalizability of the model. Therefore, it is demonstrated that our proposed oriented object detection scheme for DETR-like models is effective.

TABLE IV
COMPARISON WITH OTHER ORIENTED OBJECT DETECTION METHODS FOR DETRS

Spectrum	Models	DroneVehicle	VEDAI
Single-spectral	AO2 DETR [37]	73.8	69.9
	ARS DETR [38]	72.6	**76.5**
	Sparse DETR [27] + Ours	**74.5**	75.6
Multi-spectral	AO2 DETR [37]	76.0	74.7
	ARS DETR [38]	75.6	**78.7**
	Multi-spectral DETR(Ours)	**76.9**	77.3

The bold values represent highest score among different methods.

TABLE V
COMPARISON ON THE VEDAI DATASET

Models	Year	Oriented	Horizontal
RetinaNet [23]	2017	51.9	59.9
EfficientDet [47]	2020	52.6	–
YOLOrs [40]	2020	59.7	–
YOLOrs-Lite [48]	2021	62.2	–
SuperYOLO [49]	2023	–	75.1
HyperYOLO [50]	2024	–	76.7
CrossYOLO [51]	2024	–	79.8
ICAFusion [52]	2024	–	76.6
CMAFF [10]	2022	–	78.6
CFFIM [51]	2023	–	79.8
Multi-spectral DETR(Ours)	–	**69.8**	**82.2**

All methods are multispectral models.
The bold values represent highest score among different methods.

Next, we compare the performance of multispectral DETR with other multispectral models on the VEDAI dataset, as presented in Table V. Since many existing works [10], [49], [50], [51], [51], [52] perform horizontal object detection on VEDAI, we also report horizontal object detection results in Table V for a comprehensive comparison. For horizontal object detection, the performance of multispectral DETR without angle prediction is measured. In both oriented and horizontal object detection settings, multispectral DETR achieves state-of-the-art results.

Finally, we compare the performance of CT-SSMS-enhanced multispectral DETR with recent state-of-the-art methods for RGB-TIR remote sensing object detection, as shown in Table VI. MKD [43] represents a unidirectional knowledge distillation method from a multispectral GGHL [53] to a single-spectral GGHL. As evident from the results, CT-SSMS not only has a higher mAP than MKD, but also brings larger relative improvement to the single-spectral model. Among RGB-TIR models, CT-SSMS-enhanced multispectral DETR also attains the highest mAP score. Therefore, our proposed CT-SSMS method sets a new state-of-the-art benchmark for both TIR and RGB-TIR object detection in remote sensing imagery.

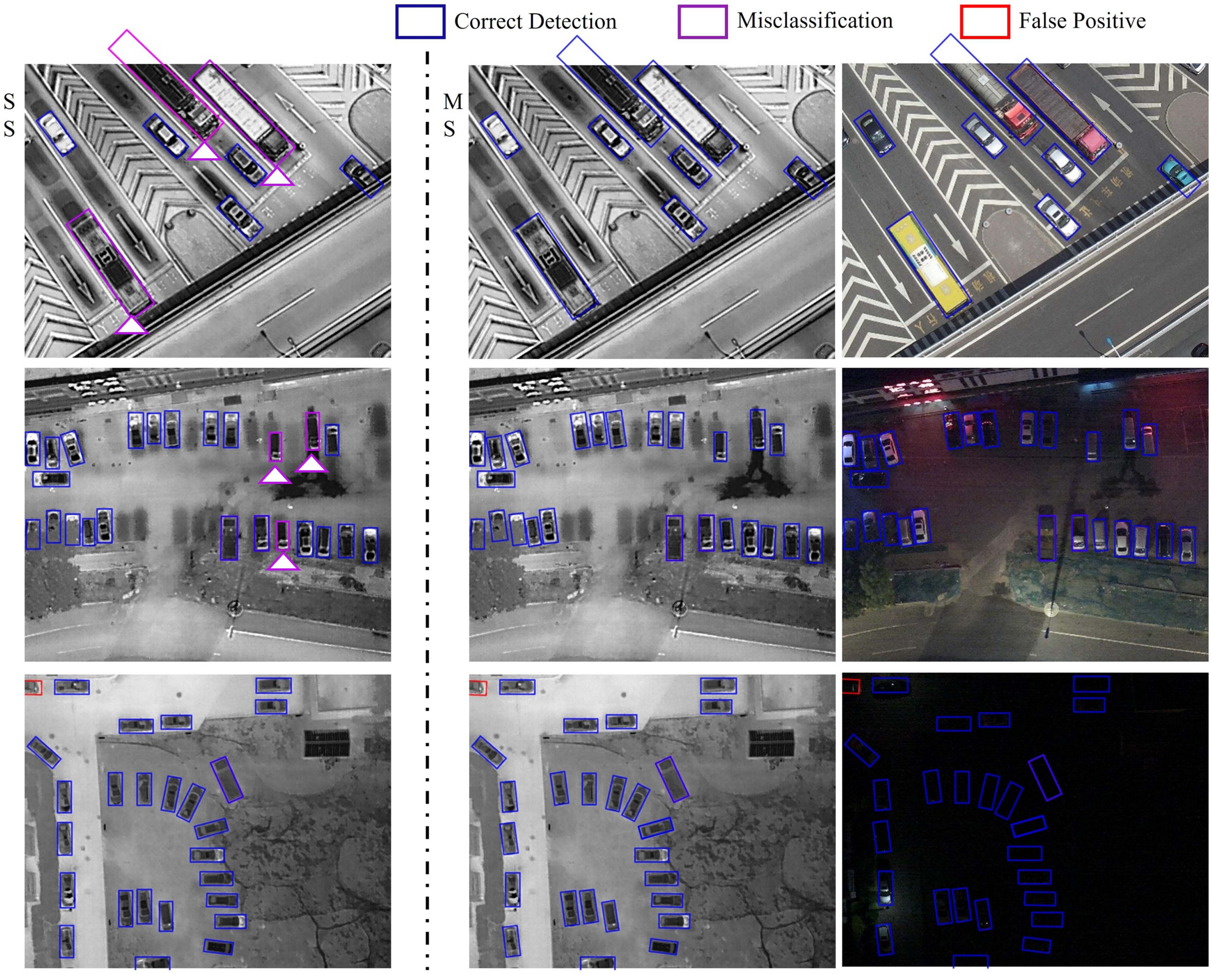

Fig. 8. Detection results of single-spectral DETR and multispectral DETR on daytime and nighttime scenes. Both methods are enhanced with CT-SSMS.

TABLE VI
COMPARISON ON THE DRONEVEHICLE DATASET

Method	Year	Spectrum	mAP
GGHL [53]	2022	TIR	67.7
MKD [43]	2023		69.0
Single-spectral DETR	2022		74.5
+ CT-SSMS (Ours)	–		**76.6**
UA-CMDet [5]	2022	RGB-TIR	64.0
GLFNet [54]	2024		71.4
Cascade-TSFADet [42]	2022		73.9
C2Former [11]	2024		74.2
RIFuse [55]	2023		75.4
CALNet [12]	2023		75.4
FFODNet [56]	2024		76.9
Multi-spectral DETR(Ours)	–		76.9
+ CT-SSMS (Ours)	–		**77.7**

The bold values represent highest score among different methods.

Fig. 8 shows the detection results of single-spectral DETR and multispectral DETR, both enhanced with CT-SSMS, in daytime and nighttime scenes. The bounding boxes are color-coded to indicate different detection outcomes: blue boxes for correct detections, purple boxes for misclassifications, and red boxes for false positives. Overall, the majority of targets are correctly detected by both single-spectral DETR and multispectral DETR. In daytime scenes, RGB images offer rich color and texture information about targets, which cannot be obtained from TIR images. As a result, while single-spectral DETR has several misclassifications, multispectral can correctly classify the targets, distinguishing among buses, trucks, and freight-cars, which demonstrates the advantage of multispectral fused object detection. In well-lit nighttime scenes, RGB information remains useful in preventing misclassifications. In dark nighttime scenes, the detection results of multispectral DETR are similar to those of single-spectral DETR. Note that the red box is an unlabeled target at the image corner, which should not be considered a false positive.

D. Limitations

Despite the promising results, several limitations in our approach warrant further investigation. While our feature fusion mechanism with the MDA block effectively leverages multispectral information to enhance detection performance, it introduces additional computational overhead by approximately doubling the network complexity. Furthermore, the classification accuracy of our model, which relies on direct classification of decoder embeddings, leaves room for improvement.

To address these limitations, future research could explore two primary directions. First, the development of tailored knowledge distillation methods could potentially compress the multispectral knowledge into a more efficient architecture while maintaining detection accuracy. Second, alternative classification schemes, such as incorporating prototype-based classifiers [57], could be investigated to replace the current direct classification of decoder embeddings, potentially leading to more robust and accurate object categorization. These improvements would contribute to both the computational efficiency and detection performance of object detection systems on remote sensing platforms.

VI. Conclusion

In this article, we present a multispectral fused remote sensing object detection method based on transformer. The multispectral DETR model fuses RGB-IR feature maps with the multispectral deformable attention module and is capable of oriented object detection. Further, we examine the interference problem caused by inconsistent targets and introduce a cross teaching approach between singlespectral DETR and multispectral DETR. The CT-SSMS method utilizes the respective strengths of single-spectral and multispectral DETR to enhance each other, especially improving the performance of multispectral DETR on inconsistent targets. Experimental results show the effectiveness of multispectral deformable attention and the oriented object detection scheme. While experiments on unidirectional knowledge distillation reveal the importance of selecting consistent and inconsistent targets for distillation of the two directions, the bidirectional CT-SSMS method outperforms unidirectional distillation. With CT-SSMS, multispectral DETR achieves state-of-the-art results on RGB-IR fused remote sensing object detection tasks.

References

[1] X. Tang and X. Zhu, "The geometric calibration and validation for the ZY3-02 satellite optical image," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. 42, pp. 587–593, 2017.

[2] H. Lu, T. Fan, P. Ghimire, and L. Deng, "Experimental evaluation and consistency comparison of UAV multispectral minisensors," *Remote Sens.*, vol. 12, no. 16, pp. 2542–2560, 2020.

[3] D. P. Roy et al., "Landsat-8: Science and product vision for terrestrial global change research," *Remote Sens. Environ.*, vol. 145, pp. 154–172, 2014.

[4] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery: A small target detection benchmark," *J. Vis. Commun. Image Representation*, vol. 34, pp. 187–203, 2016.

[5] Y. Sun, B. Cao, P. Zhu, and Q. Hu, "Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 10, pp. 6700–6713, Oct. 2022.

[6] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 1–13.

[7] A. Wolpert, M. Teutsch, M. S. Sarfraz, and R. Stiefelhagen, "Anchor-free small-scale multispectral pedestrian detection," in *Proc. Brit. Mach. Vis. Conf.*, 2020.

[8] K. Zhou, L. Chen, and X. Cao, "Improving multispectral pedestrian detection by addressing modality imbalance problems," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 787–803.

[9] H. Zhang, E. Fromont, S. Lefèvre, and B. Avignon, "Guided attentive feature fusion for multispectral pedestrian detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 72–80.

[10] F. Qingyun and W. Zhaokui, "Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery," *Pattern Recognit.*, vol. 130, pp. 108786–108800, 2022.

[11] M. Yuan and X. Wei, "C^2 Former: Calibrated and complementary transformer for RGB-infrared object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5403712.

[12] X. He, C. Tang, X. Zou, and W. Zhang, "Multispectral object detection via cross-modal conflict-aware learning," in *Proc. ACM Int. Conf. Multimedia*, 2023, pp. 1465–1474.

[13] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[14] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[15] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–11.

[16] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[17] 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, 2023, Art. no. 5001405.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

[19] R. Girshick, "Fast R-CNN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1440–1448.

[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[22] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[23] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[24] S. Liu et al., "Dab-detr: Dynamic anchor boxes are better queries for detr," in *Proc. Int. Conf. Learn. Representation*, 2022.

[25] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "DN-DETR: Accelerate DETR training by introducing query denoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 13619–13627.

[26] H. Zhang et al., "DINO: DETR with improved denoising anchor boxes for end-to-end object detection," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–12.

[27] B. Roh, J. Shin, W. Shin, and S. Kim, "Sparse DETR: Efficient end-to-end object detection with learnable sparsity," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–11.

[28] X. Yang and J. Yan, "Arbitrary-oriented object detection with circular smooth label," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 677–694.

[29] X. Yang, L. Hou, Y. Zhou, W. Wang, and J. Yan, "Dense label encoding for boundary discontinuity free rotation detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15819–15829.

[30] D. Lu, D. Li, Y. Li, and S. Wang, "OSKDet: Orientation-sensitive keypoint localization for rotated object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1182–1192.

[31] X. Yang and J. Yan, "On the arbitrary-oriented object detection: Classification based approaches revisited," *Int. J. Comput. Vis.*, vol. 130, no. 5, pp. 1340–1365, 2022.

[32] X. Yang, J. Yan, Q. Ming, W. Wang, X. Zhang, and Q. Tian, "Rethinking rotated object detection with Gaussian Wasserstein distance loss," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2021, pp. 11830–11841.

[33] X. Yang et al., "Learning high-precision bounding box for rotated object detection via Kullback-Leibler divergence," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 18381–18394.

[34] D. Zhou et al., "IoU loss for 2D/3D object detection," in *Proc. Int. Conf. 3D Vis*, 2019, pp. 85–94.

[35] Y. Lee and W. Lim, "Shoelace formula: Connecting the area of a polygon and the vector cross product," *Math. Teacher*, vol. 110, no. 8, pp. 631–636, 2017.

[36] X. Yang et al., "The KFIoU loss for rotated object detection," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–13.

[37] L. Dai, H. Liu, H. Tang, Z. Wu, and P. Song, "AO2-DETR: Arbitrary-oriented object detection transformer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 5, pp. 2342–2356, May 2023.

[38] Y. Zeng, Y. Chen, X. Yang, Q. Li, and J. Yan, "ARS-DETR: Aspect ratio-sensitive detection transformer for aerial oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5610315.

[39] C. Li, D. Song, R. Tong, and M. Tang, "Multispectral pedestrian detection via simultaneous detection and segmentation," in *Proc. Brit. Mach. Vis. Conf.*, 2018, pp. 1–12.

[40] M. Sharma et al., "YOLOrs: Object detection in multimodal remote sensing imagery," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1497–1508, 2020.

[41] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, and Z. Liu, "Weakly aligned cross-modal learning for multispectral pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5127–5137.

[42] M. Yuan, Y. Wang, and X. Wei, "Translation, scale and rotation: Cross-modal alignment meets RGB-infrared vehicle detection," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 509–525.

[43] Z. Huang, W. Li, and R. Tao, "Multimodal knowledge distillation for arbitrary-oriented object detection in aerial images," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.

[44] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 8026–8037.

[45] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2018.

[46] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[47] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10781–10790.

[48] M. Sharma, P. P. Markopoulos, and E. Saber, "YOLOrs-lite: A lightweight CNN for real-time object detection in remote-sensing," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, IEEE, 2021, pp. 2604–2607.

[49] 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, 2023, Art. no. 5605415.

[50] G. Nan, Y. Zhao, L. Fu, and Q. Ye, "Object detection by channel and spatial exchange for multimodal remote sensing imagery," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 8581–8593, 2024.

[51] 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, 2024, Art. no. 5000405.

[52] 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, pp. 109913–109925, 2024.

[53] Z. Huang, W. Li, X.-G. Xia, and R. Tao, "A general Gaussian heatmap label assignment for arbitrary-oriented object detection," *IEEE Trans. Image Process.*, vol. 31, pp. 1895–1910, 2022.

[54] X. Kang, H. Yin, and P. Duan, "Global–local feature fusion network for visible–infrared vehicle detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 6005805.

[55] T. Tian, J. Cai, Y. Xu, Z. Wu, Z. Wei, and J. Chanussot, "RGB-infrared multi-modal remote sensing object detection using CNN and transformer based feature fusion," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2023, pp. 5728–5731.

[56] J. Wang et al., "Multi-modal object detection of UAV remote sensing based on joint representation optimization and specific information enhancement," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 1016–1025, 2024.

[57] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 4080–4090.

Jiahe Zhu received the B.S. degree in electronic information science and technology from the School of Electronic Science and Engineering, Nanjing University, Nanjing, China, in 2021, and the M.E. degree in electronic information engineering from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2024.

His research interests include computer vision, remote sensing object detection, multispectral fusion, medical image processing, and deep learning.

Huan Zhang received the B.S. and M.S. degrees from Beihang University, Beijing, China, in 2016 and 2019, respectively, and the Ph.D. degree from Tsinghua University, Beijing, China, in 2022.

She is currently a Postdoctoral Researcher with the Department of Electronic Engineering, Tsinghua University. Her research interests include deep learning, computer vision, and remote sensing information processing.

Simin Li received the Ph.D. degree in electronical science and technology from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2020.

She is currently an Assistant Researcher with the Institute of Electronic Engineering, China Academy of Engineering Physics, Mianyang, China. Her research interests lie in deep learning based signal processing, image classification and reconstruction, and intelligent control.

Shengjin Wang (Senior Member, IEEE) received the B.E. degree in electronic engineering from Tsinghua University, Beijing, China, in 1985 and the Ph.D. degree in electronic engineering from the Tokyo Institute of Technology, Tokyo, Japan, in 1997.

From May 1997 to August 2003, he was a member of the research Director in the Internet System Research Laboratories, NEC Corporation, Japan. Since September 2003, he has been a Professor with the Department of Electronic Engineering, Tsinghua University, where he is currently also the Director of the Research Center of Media Intelligence and Autonomous Systems. He has published more than 100 papers and is the holder of 20 patents. His current research interests include computer vision, machine learning, intelligent video analysis, person re-identification, and multimodal cooperative robotics.

Hongbing Ma received the B.E. degree in electronic engineering from Hebei Normal University, Shijiazhuang, China, in 1985 and the M.E. and Ph.D. degrees in electronic engineering from Peking University, Beijing, China, in 1996 and 1999, respectively.

He is currently a Professor with the Department of Electronic Engineering, Tsinghua University, Beijing, China. His research interests include image processing and pattern recognition.