Citation information: DOI: 10.1109/TGRS.2026.3674946,

BANet: Enhancing Weakly Aligned Multimodal Object Detection via Balanced Bidirectional Alignment Network

Yutian Shi, Guoquan Li, *Member, IEEE*, Zhilong Shen, Hongying Meng, *Senior Member, IEEE*, and Yu Pang, *Member, IEEE*

***Abstract*—Multimodal object detection in remote sensing imagery has achieved remarkable performance, primarily owing to its ability to exploit complementary information from multiple modalities. However, most existing methods often suffer from substantial performance degradation under weakly aligned conditions, primarily due to the asymmetric utilization of information across different modalities. Therefore, we propose a novel multimodal object detection network, termed BANet, which aims to improve detection accuracy in weakly aligned multimodal remote sensing imagery~~. BANet~~ adopts a dual-path architecture and incorporates a dedicated Weakly Aligned Module (WAM) to explicitly mitigate misalignment and enhance cross-modal feature interaction. WAM includes three cooperative components. ~~Specifically~~, the Adaptive Cross-Modal Correlation Module (ACMCM) is designed to establish semantic correspondence by jointly modeling global dependencies and local similarities in a bidirectional manner. Then, the Symmetric Offset Generator (SOG) adopts a coarse-to-fine strategy to produce stable and symmetric offsets, thereby enabling precise and robust spatial alignment. Finally, the Progressive Fusion Strategy (PFS) adaptively integrates the original and aligned features through learnable weighting, effectively preserving modality-specific characteristics while enhancing both spatial alignment and semantic consistency. Extensive experiments on the DroneVehicle and VEDAI multimodal remote sensing datasets demonstrate the superiority of the proposed method over advanced multimodal remote sensing object detectors. Notably, BANet performs best on the two datasets with only 8.8M parameters, highlighting its effectiveness and efficiency for real-time UAV applications.**

***Index Terms*—Multimodal object detection, Weak alignment, Feature alignment, Cross-modal fusion, Remote sensing.**

I. Introduction

UAV object detection aims to accurately locate and classify objects within aerial imagery and has been widely applied in various domains such as precision crop monitoring [1], infrastructure inspection [2], and wildlife conservation surveillance [3]. Most existing methods rely solely on RGB images, making their performance highly sensitive to imaging conditions [4]. Although RGB data can provide rich texture and structural information under favorable circumstances, it tends to degrade severely in challenging environments such as nighttime or foggy weather [5]. To address these limitations, researchers have introduced multimodal detection frameworks that combine visible and infrared information, leveraging their complementary characteristics to enhance robustness and generalization [6], [7]. However, effectively aligning and fusing heterogeneous features from visible and infrared modalities remains a challenging task.

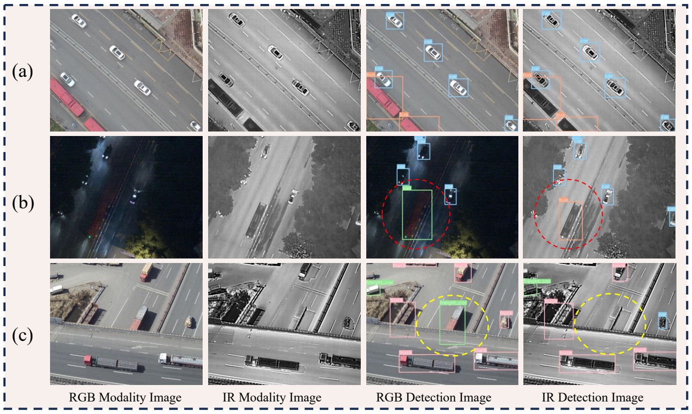

Fig. 1. Illustration of weak alignment in multimodal UAV imagery. (a) Multimodality well-aligned RGB-IR pair. (b) Semantic misalignment between RGB and IR images. The red circle indicates a case of a semantic weak misalignment. (c) Spatial misalignment between RGB and IR images. The yellow circle indicates a case of a spatial weak misalignment.

Current fusion strategies can be broadly categorized into feature-level and decision-level fusion. Feature-level fusion [8], [9] integrates information from multiple modalities within the intermediate layers of a network to enable joint representation learning. Decision-level fusion [10] combines the outputs of independent detection streams from each modality. However, both feature-level and decision-level fusion methods still face significant challenges in achieving effective cross-modal interaction and fully exploiting complementary information, which may lead to suboptimal performance in complex or highly dynamic scenes [11].

Furthermore, the limited cross-modal interaction in existing fusion frameworks mainly stems from the assumption of perfect spatial alignment between multimodal inputs. In real UAV-based sensing, this assumption rarely holds due to differences in sensor installation and viewpoint variations [12] during image acquisition. These factors often cause geometric deviation and semantic inconsistency between visible and infrared modalities, reducing the reliability of feature

Manuscript created November, 2025.

This work was supported by the National Natural Science Foundation of China under Grant U21A20447 and 12411530119. (Corresponding author: Guoquan Li.)

Yutian Shi, Guoquan Li and Zhilong Shen are with the School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: s230101127@stu.cqupt.edu.cn; ligq@cqupt.edu.cn; szldragon@foxmail.com).

Hongying Meng is with the Department of Electronic and Electrical Engineering, Brunel University London, London UB8 3PH, U.K. (e-mail: hongying.meng@brunel.ac.uk).

Yu Pang is with the School of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail:pangyu@cqupt.edu.cn).

correspondence and restricting effective joint learning. As a result, even advanced fusion strategies may fail to maintain stable and complementary feature relationships under weakly aligned conditions [13]. Beyond spatial misalignment, weakly aligned imagery introduces a deeper problem, which is the asymmetric utilization of cross-modal information. In practical scenarios, one modality usually dominates the fusion process, suppressing the subtle but valuable cues from the other [14]. This imbalance leads to biased alignment and incomplete semantic matching between modalities. Consequently, current frameworks [15] still find it difficult to achieve a balanced bidirectional exchange of semantic and geometric information, which limits their robustness and generalization under weak alignment conditions.

To bridge this gap, we propose a bidirectional alignment network (BANet) to enhance detection performance under weakly aligned multimodal conditions. BANet adopts a dual-path architecture and incorporates a dedicated Weakly Aligned Module (WAM) to explicitly mitigate misalignment and enhance the cross-modal feature interaction. The WAM consists of three cooperative components. First, the Adaptive Cross-Modal Correlation Module (ACMCM) establishes bidirectional semantic correspondence by jointly modeling global dependencies and local similarities, ensuring consistent and balanced feature representation across modalities. Second, the Symmetric Offset Generator (SOG) employs a coarse-to-fine strategy to generate stable and symmetric offsets, thereby achieving precise and robust spatial alignment. Finally, the Progressive Fusion Strategy (PFS) adaptively integrates the original, aligned, and re-fused features through learnable weighting, effectively preserving modality-specific characteristics while improving both spatial consistency and semantic complementarity. The main contributions of this study are summarized as follows:

1) We construct a multimodal object network (BANet) based on a dual-path framework for remote sensing in multinodal object detection, which can enhance detection accuracy and mitigate the asymmetric utilization of cross-modal information under weakly aligned conditions.
2) In BANet, we proposed WAM that integrates three cooperative components, ACMCM, SOG, and PFS. Specifically, ACMCM builds bidirectional semantic correspondence, SOG generates stable symmetric offsets for precise spatial alignment, and PFS adaptively fuses the original and aligned features through learnable weighting, preserving modality-specific information while enhancing semantic and spatial consistency.
3) Extensive experiments conducted on multimodal remote sensing datasets (DroneVehicle [16], VEDAI [5]) validate the effectiveness of BANet. The proposed model achieves state-of-the-art performance with lower computational complexity and parameter count, confirming its robustness and generalization in weakly aligned multimodal detection scenarios.

II. Related Works

This section briefly introduces related work, including multimodal remote sensing object detection. ~~And we~~ also introduce weakly aligned multimodal feature fusion.

A. Multimodal Remote Sensing Object Detection

Multimodal object detection, particularly through the fusion of visible (RGB) and infrared (IR) imagery, which become a fundamental research direction in remote sensing. This approach uses the complementary characteristics of different modalities, offering advantages over single-modality object detection [17], [18].

Early research in multimodal object detection primarily adapted single-modality detectors through direct feature-level fusion strategies. These methods typically integrated RGB and IR information using operations such as mid-level concatenation or channel-wise weighting, thereby improving detection performance in different conditions. However, such fusion strategies often fail to effectively distinguish complementary cross-modal information.

In recent years, several works have focused on constructing fusion blocks and multi-branch feature aggregation to preserve fine details [8], [10], [19] in remote sensing scenarios. Wang et al. present CM-YOLO [20], a lightweight object detector that enhances infrared feature awareness through a prior modality translator and infrared-visible gating modules, effectively addressing the limitations of RGB-dependent detection in complex environments. IN [6], LRAF-Net is designed, which introduces asymmetric data augmentation, cross-feature enhancement, and long-range dependence fusion to effectively capture and integrate global cross-modal contextual relationships for visible-infrared object detection.

Furthermore, other approaches address the challenge of input variability by employing explicitly designed components such as frequency filters or modality-aware training procedures [21], [22]. Sun et al. [21] proposed the LF-MDet, a low-rank multimodal detection method that integrates frequency filtering experts to achieve efficient and robust visible-infrared object detection. Liu et al. introduce MAFNet [22], an adaptive cross-modal object tracking framework that dynamically integrates RGB and NIR(near-infrared) images using a modality-aware weighting mechanism.

Moreover, recent contributions include transformer-based calibration with complementary learning for multimodal detection. Yuan et al. [11] introduce a transformer-based module that uses intermodality cross-attention to explicitly calibrate and complement RGB IR features while reducing computational cost via adaptive feature sampling.

Nevertheless, most existing approaches assume perfectly aligned image pairs during training and inference. However, such approaches assume perfect spatial alignment between modalities and often fail to adequately address misalignment.

B. Weakly Aligned Multimodal Feature Fusion

In the past few years, some existing studies have focused on alignment learning for object detection tasks. We divide

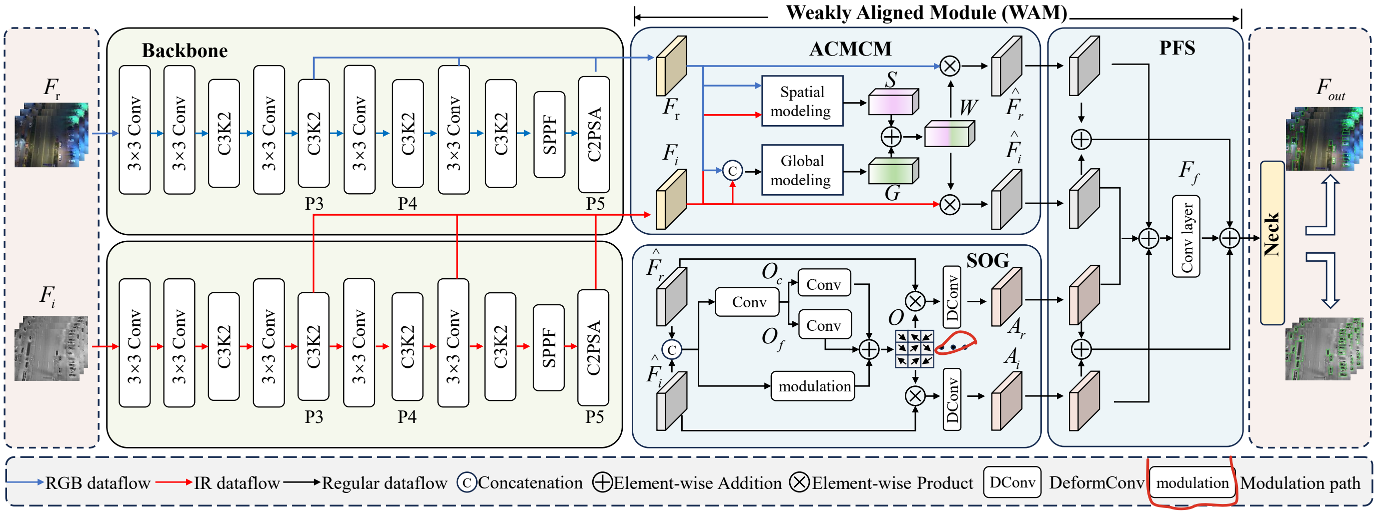

Fig. 2. The overall architecture of our proposed BANet. The light-green area is the two-stream backbone. The light-blue area is the Weakly Aligned Module (WAM). WAM includes three cooperative components. First, the Adaptive Cross-Modal Correlation Module (ACMCM) which consists of the global modeling and local modeling. Then, the Symmetric Offset Generator (SOG) employs a coarse-to-fine strategy and a modulation operation. Finally, the Progressive Fusion Strategy (PFS) integrates the original and aligned features through learnable weighting. WAM constructs symmetric semantic and spatial correspondence, and it also performs progressive feature integration. Furthermore, WAM explicitly mitigates misalignment and enhances cross-modal feature interaction.

these efforts into unimodal and multimodal. For unimodal, Han et al. proposed an S^2A-Net [23], which tackles the feature-anchor misalignment in oriented object detection through an anchor refinement network and alignment convolution. Song et al. [24] achieved precise cross-modal feature alignment for LiDAR-camera fusion through graph-based nearest-neighbor matching and an attention-weighted refinement mechanism. Wang et al. [25] introduced an unparameterized adjacent feature alignment module to dynamically integrate multiscale spatial features for salient object detection. Xie et al. [26] reintegrated decoupled features to generate shared offsets that enhance object detection.

In addition, in multimodal object detection, Zhang et al. [27] constructed a region feature alignment module with a similarity constraint to address positional misalignment in multimodal detection. To address the dual-perspective misalignment, Liu et al. [28] designed a cross-modal feature alignment module and a task-head alignment module, which can resolve spatial inconsistencies in anchor-free detection heads. Chen et al. [29] introduced an offset-guided adaptive feature alignment method that addresses weakly aligned RGB-IR images through cross-modality spatial offset modeling for precise offset estimation. [30] proposed a unified framework that addresses both semantic inconsistency and modality conflict in weakly aligned RGB-IR UAV detection through offset-guided semantic alignment. However, current existing works have only considered pixel or geometric misalignment; none of them have adequately considered the asymmetric utilization of cross-modal information. To bridge this gap, we propose a dual-path alignment framework.

III. PROPOSED METHOD

Our method aims to enhance the accuracy and robustness of the multimodal object detection under weakly misalignment conditions. In this section, we first present an overview of BANet and then introduce the proposed Weakly Aligned Module (WAM), which is composed of three key components: ACMCM, SOG, and PFS. Specifically, ACMCM establishes bidirectional semantic correspondence by jointly modeling global dependencies and local similarities. Then, SOG employs a coarse-to-fine strategy to generate stable and symmetric offsets. Finally, adaptively integrates original and aligned features through learnable weighting.

A. Overall architecture

To address the asymmetric utilization of cross-modal information under weakly aligned conditions in remote sensing scenarios, we propose a multimodal UAV object detection approach dubbed Bidirectional Alignment Network (BANet), as illustrated in Fig. 2. Firstly, a pair of weakly misaligned infrared and visible images is input into a two-stream backbone network, where multi-scale RGB and IR features are extracted independently. Then, the Weakly Aligned Module (WAM) is constructed to simultaneously address semantic inconsistency and spatial misalignment in weakly aligned inputs. The WAM comprises three core components. Within the WAM, the Adaptive Cross-Modal Correlation Module (ACMCM) serves as a semantic bridge by establishing bidirectional correlations, thereby enhancing semantic consistency and suppressing modality bias. Furthermore, a Symmetric Offset Generator (SOG) employs a coarse-to-fine strategy to predict symmetric and stable offsets, enabling more precise reconstruction of spatial correspondence under weak alignment. Furthermore, the Progressive Fusion Strategy (PFS) adaptively integrates the original and aligned features by combining them with learnable weights to preserve modality-specific characteristics.

B. Weakly Aligned Module (WAM)

In practical multimodal UAV images, RGB and IR images often exhibit weak alignment, which manifests not only as

Algorithm 1 Weak Alignment Module

```
Input:RGB feature $\mathbf{F}_\mathrm{r}$, IR feature $\mathbf{F}_\mathrm{i}$
Output:Aligned and fused representation $\mathbf{F}_\mathrm{out}$
WAM($\mathbf{F}_\mathrm{r}$,$\mathbf{F}_\mathrm{i}$):
Generate channel-wise weights $\mathbf{G}$ via Eq. 1.
if $\mathrm{m}$ is RGB then
    Calculate ($\mathbf{\Delta}_{\mathrm{r}\to\mathrm{i}}$, $\mathbf{\Pi}_{\mathrm{r}\to\mathrm{i}}$) via Eq. 2
    Obtain correlation weights $\mathbf{W}_{\mathrm{r}\to\mathrm{i}}$ via Eq. 5
    Calculate features $\hat{\mathbf{F}}_\mathrm{r}$ via Eq. 7
    Offset $\mathbf{O} \leftarrow (\hat{\mathbf{F}}_\mathrm{r}, \hat{\mathbf{F}}_\mathrm{i})$ via Eq. 11
    Aligned RGB features $\mathbf{A}_\mathrm{r} \leftarrow (\hat{\mathbf{F}}_\mathrm{r}, \mathbf{O})$ via Eq. 12
end if
if $\mathrm{m}$ is IR then
    Calculate ($\mathbf{\Delta}_{\mathrm{i}\to\mathrm{r}}$, $\mathbf{\Pi}_{\mathrm{i}\to\mathrm{r}}$) via Eq. 3
    Obtain correlation weights $\mathbf{W}_{\mathrm{i}\to\mathrm{r}}$ via Eq. 6
    Calculate features $\hat{\mathbf{F}}_\mathrm{i}$ via Eq. 7
    Offset $\mathbf{O} \leftarrow (\hat{\mathbf{F}}_\mathrm{r}, \hat{\mathbf{F}}_\mathrm{i})$ via Eq. 11
    Aligned IR features $\mathbf{A}_\mathrm{i} \leftarrow (\hat{\mathbf{F}}_\mathrm{i}, \mathbf{O})$ via Eq. 12
end if
$\mathbf{F}_\mathrm{f} \leftarrow \mathbf{F}_\mathrm{r}, \mathbf{F}_\mathrm{i}, \mathbf{A}_\mathrm{r}$ and $\mathbf{A}_\mathrm{i}$
$\mathbf{F}_\mathrm{out} \leftarrow (\mathbf{F}_\mathrm{r}, \mathbf{F}_\mathrm{i}, \mathbf{A}_\mathrm{r}, \mathbf{A}_\mathrm{i}, \mathbf{F}_\mathrm{f})$ via Eq. 14
Return $\mathbf{F}_\mathrm{out}$
```

spatial shifts but also generates semantic bias at the feature level. This often causes one modality to dominate the decision during fusion, thereby losing complementary information from the other modality. Most existing methods address alignment from a single dimension [29], [30], which makes it difficult to ensure both semantic consistency and spatial correspondence simultaneously. To address these limitations, we propose a novel Weakly Aligned Module (WAM). The WAM is structured with three dedicated components: Adaptive Cross-Modal Correlation Module (ACMCM), Symmetric Offset Generator (SOG), and Progressive Fusion Strategy (PFS). WAM explicitly maximizes bidirectional information transfer, constructs symmetric semantic and spatial correspondence, and it also performs progressive feature integration. Thus, it achieves robust multimodal fusion without relying on strictly registered image pairs.

In WAM, the RGB and infrared modality feature maps are denoted as $\mathbf{F}_\mathrm{r}$ and $\mathbf{F}_\mathrm{i}$, respectively. The WAM first employs the ACMCM to estimate bidirectional correlation weights $\mathbf{W}_{\mathrm{r}\to\mathrm{i}}$ and $\mathbf{W}_{\mathrm{i}\to\mathrm{r}}$, which jointly capture global dependencies and local variations to emphasize modality-consistent semantic regions. These weights are subsequently applied to the original features to generate refined representations, $\hat{\mathbf{F}}_\mathrm{r}$ and $\hat{\mathbf{F}}_\mathrm{i}$, which exhibit enhanced semantic alignment. Then, the SOG predicts symmetric coarse-to-fine offsets $\mathbf{O}$ with modulation and normalization, which guide deformable convolutions to compensate for spatial misalignments between modalities, thereby producing the aligned features $\mathbf{A}_\mathrm{r}$ and $\mathbf{A}_\mathrm{i}$. Moreover, the re-fused features $\mathbf{F}_\mathrm{f}$ are progressively integrated with the original and aligned feature streams through learnable balancing weights. This strategy enables the network to preserve modality-specific information while ensuring spatial consistency and semantic integrity. Finally, we get the final output features $\mathbf{F}_\mathrm{out}$.

In conclusion, the WAM offers an effective solution for multimodal remote sensing detection under weak alignment conditions. It explicitly mitigates asymmetric cross-modal information utilization by incorporating balanced correlation modeling, stable offset learning, and adaptive feature integration within a unified framework.

C. Adaptive Cross-Modal Correlation Module (ACMCM)

In weakly aligned multimodal scenarios, simple feature concatenation or one-way correlation often results in semantic bias, where one modality dominates the shared representation. Traditional cross-modal correlation methods rely mainly on local feature similarity, ignoring the asymmetric reliability between RGB and IR modalities. Although some studies [29], [30] introduce global attention to strengthen cross-modal interaction, they still exploit either global correlations or local spatial similarities, which often result in unbalanced and semantically inconsistent representations across modalities. To overcome this asymmetry, we construct an Adaptive Cross-Modal Correlation Module (ACMCM), which shares models' global dependencies and local spatial information in a unified correlation estimator. This design enables balanced and bidirectional information exchange between modalities, improving semantic consistency across them. It also reduces redundant modality-specific responses and provides a stable correlation prior for the subsequent alignment process. Fig. 3 depicts the ACMCM structure.

ACMCM computes correlation weights in three steps. Firstly, we aggregate global dependencies by concatenating the two modalities and applying global average pooling (GAP). This process can be expressed as

$$\mathbf{G} = \sigma\left(\mathbf{W}_\mathrm{g} \cdot \mathrm{GAP}([\mathbf{F}_\mathrm{r}; \mathbf{F}_\mathrm{i}])\right), \tag{1}$$

where GAP is global average pooling, $[\,;\,]$ indicates channel concatenation, $\mathbf{W}_\mathrm{g}$ is a two-layer 1×1 convolution, and σ is the sigmoid activation. This operation generates channel-wise weights $\mathbf{G} \in \mathbb{R}^{\mathrm{B}\times\mathrm{C}\times 1\times 1}$ that emphasize modality-shared semantics.

Then, to complement the global view, we further extract bidirectional local similarity and difference, which can be expressed as

$$\mathbf{\Delta}_{\mathrm{r}\to\mathrm{i}} = |\mathbf{F}_\mathrm{r} - \mathbf{F}_\mathrm{i}| \quad \mathbf{\Pi}_{\mathrm{r}\to\mathrm{i}} = \mathbf{F}_\mathrm{r} \odot \mathbf{F}_\mathrm{i} \tag{2}$$

$$\mathbf{\Delta}_{\mathrm{i}\to\mathrm{r}} = |\mathbf{F}_\mathrm{i} - \mathbf{F}_\mathrm{r}| \quad \mathbf{\Pi}_{\mathrm{i}\to\mathrm{r}} = \mathbf{F}_\mathrm{i} \odot \mathbf{F}_\mathrm{r}, \tag{3}$$

where $\mathbf{\Delta}$ encodes inter-modal differences, while $\mathbf{\Pi}$ captures correlated activations. $\odot$ denotes element-wise multiplication. Moreover, we compute the mean and max pooling operation over both $\mathbf{\Delta}$ and $\mathbf{\Pi}$ and concatenate them, which is shown as

$$\mathbf{S} = \sigma\left(\mathbf{f}_\mathrm{s}\left([\mathrm{mean}(\mathbf{\Delta}),\, \max(\mathbf{\Delta}),\, \mathrm{mean}(\mathbf{\Pi}),\, \max(\mathbf{\Pi})]\right)\right), \tag{4}$$

where $\mathbf{S} \in \mathbb{R}^{\mathrm{B}\times 1\times\mathrm{H}\times\mathrm{W}}$ highlights spatially consistent regions while suppressing noisy responses. $\mathbf{f}_\mathrm{s}$ denotes a lightweight convolutional operation consisting of two successive convolutional layers with non-linear activation. In addition, we simultaneously compute $\mathbf{S}_{\mathrm{r}\to\mathrm{i}}$ and $\mathbf{S}_{\mathrm{i}\to\mathrm{r}}$.

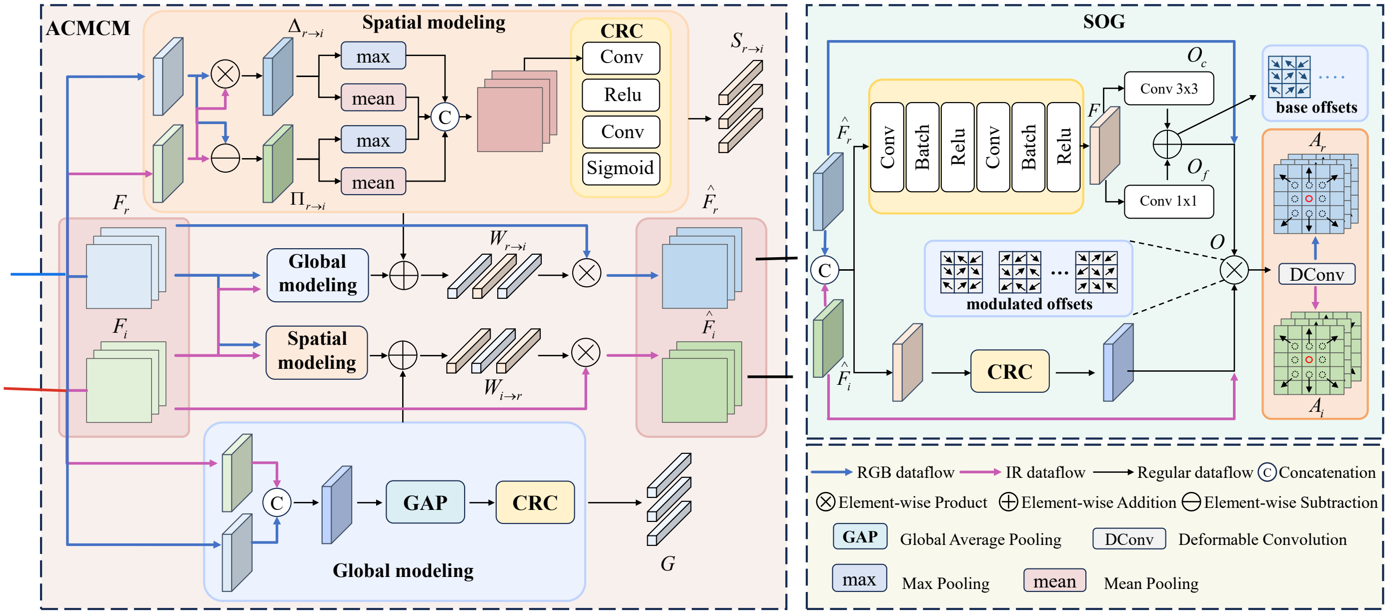

Fig. 3. The overall structure of ACMCM and SOG. In ACMCM, RGB and IR modality features are input into a unified space, establishing bidirectional correlations for semantic correction. Subsequently, output semantically aligned features to guide spatial alignment in SOG. In SOG, the features from ACMCM are used to generate symmetric coarse-to-fine offsets for both modalities. These offsets progressively correct spatial misalignment through deformable convolution, ensuring consistent structural correspondence across RGB and IR features.

Finally, the global and spatial branches are combined through a learnable balance vector $\beta = [\beta_1, \beta_2]$ with softmax normalization

$$\mathbf{W}_{\mathrm{r}\to\mathrm{i}} = \beta_1 \cdot \mathbf{G} + \beta_2 \cdot \mathrm{expand}(\mathbf{S}_{\mathrm{r}\to\mathrm{i}}) \tag{5}$$

$$\mathbf{W}_{\mathrm{i}\to\mathrm{r}} = \beta_1 \cdot \mathbf{G} + \beta_2 \cdot \mathrm{expand}(\mathbf{S}_{\mathrm{i}\to\mathrm{r}}), \tag{6}$$

where expand$(\cdot)$ presents the spatial map across channels. This adaptive weighting enables the network to dynamically emphasize global or local cues depending on the input characteristics. The final correlation weights are calculated as

$$\hat{\mathbf{F}}_{\mathrm{r}} = \mathbf{F}_{\mathrm{r}} \cdot (1 + \lambda \mathbf{W}_{\mathrm{r}\to\mathrm{i}}), \quad \hat{\mathbf{F}}_{\mathrm{i}} = \mathbf{F}_{\mathrm{i}} \cdot (1 + \lambda \mathbf{W}_{\mathrm{i}\to\mathrm{r}}), \tag{7}$$

where $\lambda = 0.5$ ensures moderate feature amplification. After performing bidirectional modal correlation calculations, the semantically aligned features $\hat{\mathbf{F}}_{\mathrm{r}}$ and $\hat{\mathbf{F}}_{\mathrm{i}}$ are prepared for subsequent spatial alignment.

In summary, the ACMCM establishes a more equitable semantic foundation for subsequent spatial alignment and feature fusion by balancing global dependencies and local variations.

D. Symmetric Offset Generator (SOG)

Although ACMCM mitigates semantic inconsistencies, cross-modal features still exhibit spatial discrepancies due to sensor misalignment, parallax, or viewpoint shifts. Such spatial discrepancies lead to inconsistent receptive regions and degraded correspondence, especially in small or distant objects. Conventional offset prediction methods [29], [30] often estimate deformations in a single direction or at a single scale, leading to unstable alignment and making one modality dominate. To overcome this, we proposed Symmetric Offset Generation (SOG), which primarily consists of a coarse-to-fine offset estimation strategy followed by a precise modulation stage. The coarse branch handles large displacements, and the fine branch refines local misalignments. Together, the symmetric multi-scale design restores positional correspondence and reduces modality bias, producing more balanced spatial alignment. The architecture of SOG is shown in Fig. 3.

Firstly, the SOG concatenates modality-specific features with their corresponding correlation priors and processes the combined input through a shared convolutional block

$$\mathrm{F} = \mathbf{f}_{\mathrm{shared}}([\hat{\mathbf{F}}_{\mathrm{r}}; \hat{\mathbf{F}}_{\mathrm{i}}]), \tag{8}$$

where $\mathbf{f}_{\mathrm{shared}}(\cdot)$ denotes a convolutional block composed of a 1×1 convolution for channel projection, followed by a 3×3 convolution with ReLU activation to capture local cross-modal context. The resulting feature map F encodes the joint cross-modal representation.

Subsequently, two complementary offset branches are generated for coarse-to-fine alignment. The first branch employs a convolutional predictor $\mathbf{f}_{\mathrm{coarse}}(\cdot)$ to estimate coarse offsets $\mathbf{O}_{\mathrm{c}}$, which primarily compensate for large spatial displacements caused by sensor misalignment or viewpoint variation. However, such coarse offsets primarily provide global corrections and may be insufficient to resolve subtle structural inconsistencies. Therefore, the second branch employs a lightweight convolutional predictor $\mathbf{f}_{\mathrm{fine}}(\cdot)$ to generate fine offsets $\mathbf{O}_{\mathrm{f}}$, thereby refining alignment in regions with localized variations and detailed structures. By combining these two components, the final offset estimation jointly incorporates global displacement correction and local refinement, expressed as

$$\mathbf{O}_{\mathrm{c}} = \mathbf{f}_{\mathrm{coarse}}(\mathrm{F}), \quad \mathbf{O}_{\mathrm{f}} = \mathbf{f}_{\mathrm{fine}}(\mathrm{F}), \tag{9}$$

where $\mathbf{O}_\mathrm{c}$ and $\mathbf{O}_\mathrm{f}$ denote the coarse and fine offsets predicted from the shared feature map H, respectively. The aggregated offset is computed as

$$\mathbf{O}_\mathrm{b} = \mathbf{O}_\mathrm{c} + \lambda \cdot \mathbf{O}_\mathrm{f}, \tag{10}$$

where λ is a scaling factor empirically set to 0.5 to balance the contribution of local refinement.

Then, we introduce a unified step to enhance the stability of offset estimation, which includes the integration modulation and normalization. Specifically, a modulation operation is first predicted from the concatenated features $[\hat{\mathbf{F}}_\mathrm{r}; \hat{\mathbf{F}}_\mathrm{i}]$ via a convolutional predictor $\mathbf{f}_\mathrm{mod}(\cdot)$ with sigmoid activation $\sigma(\cdot)$. This operation is applied element-wise to the base offset $\mathbf{O}_\mathrm{b}$ and serves to suppress spatially unstable regions. Furthermore, to prevent excessive deformation, the modulated offset is subsequently normalized and scaled using a hyperbolic tangent function. The final refined offset is computed as

$$\mathbf{O} = \tanh\left(\frac{\mathbf{O}_\mathrm{b} \odot \sigma(\mathbf{f}_\mathrm{mod}([\hat{\mathbf{F}}_\mathrm{r}; \hat{\mathbf{F}}_\mathrm{i}]))}{\mathrm{std}(\mathbf{O}_\mathrm{b})} \right) \cdot \mathrm{std}(\mathbf{O}_\mathrm{b}) \cdot 2 \tag{11}$$

where $\mathrm{std}(\cdot)$ denotes the standard deviation computed across the spatial dimensions for each sample.

Finally, the refined offsets are employed to guide bidirectional deformable convolutions, warping the enhanced feature maps to produce spatially aligned representations

$$\mathbf{A}_\mathrm{r} = \mathrm{DConv}(\hat{\mathbf{F}}_\mathrm{r}, \mathbf{O}), \quad \mathbf{A}_\mathrm{i} = \mathrm{DConv}(\hat{\mathbf{F}}_\mathrm{i}, \mathbf{O}), \tag{12}$$

DConv refers to the deformable convolution operator [31]. This step generates aligned RGB features $\mathbf{A}_\mathrm{r}$ and IR features $\mathbf{A}_\mathrm{i}$. By explicitly coupling offset prediction with deformable alignment, SOG achieves symmetric spatial alignment across modalities, thereby supplying reliable inputs for subsequent fusion.

Therefore, the SOG enables robust multimodal alignment by jointly estimating symmetric coarse-to-fine offsets, which not only stabilize deformable convolution but also mitigate asymmetric dependence on a single modality. This design facilitates more accurate correction of spatial misalignment while preserving a balanced cross-modal information flow.

E. Progressive Fusion Strategy (PFS)

Even after semantic–geometric alignment is achieved, cross-modal fusion remains challenging due to the inherently imbalanced reliability between RGB and IR modalities. Differences in illumination, texture, and thermal response can cause the network to over-rely on one modality, leading to asymmetric information utilization where the dominant modality suppresses complementary information. Existing fusion methods [29], [30] either directly replace the original features, resulting in the loss of modality-specific details, or simply concatenate them, which introduces redundancy and noise. To address these limitations, the proposed Progressive Fusion Selection (PFS) hierarchically integrates both original and aligned features through an adaptive gating mechanism. This progressive design gradually balances modality contributions, enhances spatial coherence and semantic completeness, yielding a balanced and discriminative representation for multimodal detection under weakly misalignment conditions.

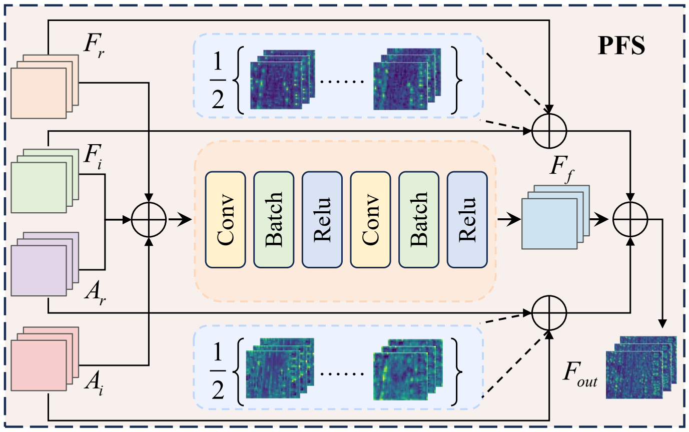

Fig. 4. Overview of the proposed PFS. In PFS, the aligned and original features are progressively fused through multi-level interaction. A learnable gating mechanism adaptively balances modality contributions, preserving complementary information while suppressing redundancy to generate the final representation.

As shown in Fig. 4, the PFS hierarchically integrates multiple feature streams. Specifically, given the original features $\mathbf{F}_\mathrm{r}$ and $\mathbf{F}_\mathrm{i}$, along with the aligned features $\mathbf{A}_\mathrm{r}$ and $\mathbf{A}_\mathrm{i}$ as inputs, the fusion process is carried out in two successive stages. In the first stage, a compact re-fused representation $\mathbf{F}_\mathrm{f}$ is constructed by concatenating all feature streams, followed by a projection through a transformation block

$$\mathbf{F}_\mathrm{f} = \mathbf{f}_\mathrm{proj}([\mathbf{F}_\mathrm{r}; \mathbf{F}_\mathrm{i}; \mathbf{A}_\mathrm{r}; \mathbf{A}_\mathrm{i}]), \tag{13}$$

where $\mathbf{f}_\mathrm{proj}(\cdot)$ consists of two successive convolutional layers with batch normalization and ReLU activation. This step reduces redundancy while enhancing cross-modal interactions.

In the second stage, the network adaptively balances the contributions of three complementary feature streams using a learnable weight vector $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \gamma_3]$, which is normalized via a softmax function

$$\mathbf{F}_\mathrm{out} = \gamma_1 \cdot \tfrac{1}{2}(\mathbf{F}_\mathrm{r} + \mathbf{F}_\mathrm{i}) \; + \; \gamma_2 \cdot \tfrac{1}{2}(\mathbf{A}_\mathrm{r} + \mathbf{A}_\mathrm{i}) \; + \; \gamma_3 \cdot \mathbf{F}_\mathrm{f}. \tag{14}$$

The first term retains modality-specific information from the original feature streams, the second enforces spatial consistency via the aligned representations, and the third introduces complementary cues from the re-fused features. By progressively integrating these components, the final output $\mathbf{F}_\mathrm{out}$ yields a balanced representation with enhanced discriminability and robustness.

To sum up, this method employs hierarchical integration of original, aligned, and re-fused streams with adaptive weighting, ensuring that neither modality dominates the fusion process. By preserving modality-specific cues while enforcing cross-modal consistency, the progressive design mitigates asymmetric information utilization and yields robust multimodal representations with strong generalization under weak alignment conditions.

TABLE I
PERFORMANCE COMPARISON OF UNIMODAL OBJECT DETECTION METHODS ON THE DRONEVEHICLE DATASET. BOLDING INDICATES THE BEST RESULTS

Methods	Modality	Heads	Car	Truck	Freight-car	Bus	Van	mAP50(%)	Params(M)	GFLOPs(G)
RetinaNet [32]	RGB	OBB	78.5	34.4	24.1	69.8	28.8	47.1	36.5	129.0
Faster R-CNN [33]	RGB	OBB	79.0	49.0	37.2	77.0	37.0	55.9	41.1	133.6
Oriented R-CNN [34]	RGB	OBB	80.1	53.8	41.6	85.4	43.3	60.8	-	-
S^{2}A-Net [23]	RGB	OBB	80.0	54.2	42.2	84.9	43.8	61.0	38.6	93.0
RoITransformer [35]	RGB	OBB	61.6	55.1	42.3	**85.5**	**44.8**	**61.6**	-	-
R^{3}Det [36]	RGB	OBB	79.3	42.2	24.5	76.0	28.5	50.1	37.0	107.6
YOLOv5s [37]	RGB	HBB	96.3	76.9	58.3	95.4	56.5	76.7	**7.0**	**15.8**
YOLOv8s [38]	RGB	HBB	96.8	77.7	58.0	95.5	59.6	77.5	11.1	28.4
YOLO11s [39]	RGB	HBB	**96.8**	**77.7**	**58.3**	95.5	**59.6**	**77.5**	9.4	21.3
RetinaNet [32]	IR	OBB	88.8	35.4	39.5	76.5	32.1	54.5	36.5	129.0
Faster R-CNN [33]	IR	OBB	89.4	53.5	48.3	87.0	42.6	64.2	41.1	133.6
Oriented R-CNN [34]	IR	OBB	89.8	57.4	53.1	89.3	45.4	67.0	-	-
S^{2}A-Net [23]	IR	OBB	89.9	54.5	55.8	88.9	48.4	67.5	38.6	93.0
RoITransformer [35]	IR	OBB	90.1	60.4	58.9	89.7	52.2	70.3	-	-
R^{3}Det [36]	IR	OBB	89.5	48.3	16.6	87.1	39.9	62.3	37.0	107.6
YOLOv5s [37]	IR	HBB	98.1	79.3	68.7	95.3	60.6	80.4	**7.0**	**15.8**
YOLOv8s [38]	IR	HBB	98.3	79.8	67.9	95.8	61.9	80.7	11.1	28.4
YOLO11s [39]	IR	HBB	**98.4**	80.5	68.6	**96.0**	**64.1**	**81.5**	9.4	21.3
Ours	RGB+IR	HBB	97.7	**82.3**	**69.9**	95.6	**65.1**	**82.1**	8.8	40.6

TABLE II
PERFORMANCE COMPARISON OF UNIMODAL OBJECT DETECTION METHODS ON THE VEDAI DATASET. BOLDING INDICATES THE BEST RESULTS

Methods	Modality	Car	Pickup	Camping	Truck	Other	Tractor	Boat	Van	mAP50(%)	Params(M)	GFLOPs(G)
YOLOrs [40]	RGB	85.2	72.9	70.3	50.6	42.6	76.7	18.6	38.9	57.0	-	-
RetinaNet [32]	RGB	91.3	82.4	79.7	78.1	60.8	75.3	69.5	69.0	75.8	36.5	212.3
SuperYOLO [10]	RGB	81.6	80.2	83.6	64.0	47.9	56.7	22.1	69.8	63.2	**7.71**	-
FCOS [41]	RGB	88.9	82.4	**87.5**	**86.3**	58.6	68.4	**78.0**	71.4	**77.7**	32.1	206.5
YOLOv5x [37]	RGB	92.3	84.4	72.4	75.1	53.9	73.5	59.4	70.1	70.0	97.21	246.9
YOLOv8x [38]	RGB	**93.4**	84.4	73.2	70.2	55.5	**83.3**	50.3	69.4	72.5	68.16	258.1
YOLOv10x [42]	RGB	92.9	**84.8**	73.6	73.7	**63.8**	68.3	51.7	66.1	71.9	31.66	**171.1**
YOLOrs [40]	IR	82.0	73.9	63.8	54.2	43.9	54.3	21.9	43.3	54.7	-	-
RetinaNet [32]	IR	88.6	84.5	87.0	81.3	40.1	51.7	71.3	78.7	72.9	36.5	212.3
SuperYOLO [10]	IR	82.2	78.2	78.2	55.3	24.4	45.1	31.2	77.1	59.0	**7.71**	-
FCOS [41]	IR	82.1	75.5	**89.0**	**92.7**	56.1	59.2	**73.7**	74.1	**75.3**	32.1	206.5
YOLOv5x [37]	IR	89.7	80.6	72.2	80.5	44.8	64.0	52.7	**75.4**	70.0	97.21	246.9
YOLOv8x [38]	IR	88.5	79.6	74.8	81.5	37.9	**67.4**	61.5	68.7	70.0	68.16	258.1
YOLOv10x [42]	IR	**90.7**	79.7	74.1	72.1	33.3	61.9	51.5	63.2	65.8	31.66	**171.1**
Ours	RGB+IR	88.8	**85.9**	75.6	74.1	**70.5**	**77.4**	**67.9**	**82.4**	**77.8**	**8.8**	**40.6**

IV. EXPERIMENTAL RESULTS

This section presents experimental evaluations of the proposed approach. We first describe the implementation details, including the datasets, evaluation metrics, and parameter settings. Then, we compare our BANet with several state-of-the-art methods. In addition, the visualization results are provided to demonstrate the effectiveness of the proposed model.

A. Datasets

In the experiment, we utilize two public drone multimodal remote sensing benchmark datasets, DroneVehicle [16] and VEDAI [5] (Vehicle Detection in Aerial Imagery), to verify the robustness of the proposed BANet. These two datasets include different numbers of image pairs, varying image resolutions, diverse detected object categories, distinct object annotation forms, varied annotation results for each modality, and various lighting conditions.

1) DroneVehicle: The DroneVehicle [16] dataset is a large-scale drone-captured benchmark for RGB-Infrared cross-modal vehicle detection. It comprises a total of 56,878 images, evenly distributed between RGB and infrared modalities. The dataset encompasses a variety of complex aerial imaging scenarios and includes five categories of vehicles: cars, trucks, buses, vans, and freight cars. In this work, all images are uniformly resized to 640 × 640 pixels. For robust evaluation, the dataset is divided into 17,990 training images, 1,469 validation images, and 8,980 test images.

2) VEDAI: The VEDAI [5] dataset consists of 1246 pairs of RGB and IR aerial images, sourced from the Utah Auto-

TABLE III
PERFORMANCE COMPARISON OF MULTIMODAL OBJECT DETECTION METHODS ON THE DRONEVEHICLE DATASET. BOLDING INDICATES THE BEST RESULTS

Methods	Car	Truck	Freight-car	Bus	Van	mAP50(%)	Params(M)	GFLOPs(G)
CSL [43]	88.4	22.4	32.1	60.0	22.5	45.0	37.1	109.6
S^2ANet [23]	89.8	45.2	44.6	88.8	39.4	61.5	38.6	93.0
DEYOLO [44]	97.4	60.5	52.4	94.7	49.1	70.8	78.3	-
CFT [45]	96.0	56.8	51.8	93.5	48.9	69.4	206.2	403.9
ICAFusion [46]	81.6	56.0	33.3	85.7	31.8	57.7	120.2	180.0
C^2Former [11]	83.1	69.6	60.5	88.9	55.7	71.6	79.0	224.7
SuperYOLO [10]	87.5	46.8	60.7	87.1	38.0	64.0	138.7	-
LF-MDet [21]	82.2	73.6	59.6	86.6	57.0	71.8	38.7	77.7
MROD-YOLO [47]	96.7	63.2	47.3	93.0	51.9	70.4	45.3	227.3
DPAL-R [28]	95.3	72.1	56.1	93.8	45.3	72.6	-	-
DPAL-P [28]	95.2	74.8	58.6	94.1	51.6	74.9	-	-
DMM [48]	90.4	79.8	63.0	89.9	68.6	79.4	87.97	-
OAFA [29]	90.3	76.8	**73.3**	90.3	66.0	79.4	-	-
Ours	**97.7**	**82.3**	69.9	**95.6**	**69.0**	**82.1**	**8.8**	**40.6**

TABLE IV
PERFORMANCE COMPARISON OF MULTIMODAL OBJECT DETECTION METHODS ON THE VEDAI DATASET. BOLDING INDICATES THE BEST RESULTS

Methods	Car	Pickup	Camping	Truck	Other	Tractor	Boat	Van	mAP50(%)	Params(M)	GFLOPs(G)
ICAFusion [46]	83.7	77.6	69.9	48.6	45.7	71.3	45.9	67.0	63.7	120.2	311.5
SuperYOLO [10]	90.6	83.3	77.7	68.5	52.4	82.1	62.2	69.4	73.3	**7.0**	81.5
FFCA [49]	89.6	85.7	78.7	**85.7**	48.6	81.8	61.5	67.0	74.8	-	-
AFD [50]	88.5	85.5	71.4	73.3	58.7	**89.1**	59.8	80.5	75.9	-	-
CFT [45]	**93.6**	83.3	73.9	74.8	51.0	75.9	64.7	72.1	73.7	206.2	-
YOLOv5x [37]	93.2	84.8	69.9	75.8	63.3	76.8	51.3	68.5	73.0	97.4	247.1
YOLOv8x [38]	91.3	85.8	68.6	77.4	59.1	72.2	61.0	78.5	74.2	68.2	258.3
YOLOv10x [42]	91.7	81.0	67.9	77.4	63.6	76.9	59.3	71.4	73.7	31.7	171.3
DEYOLO [44]	93.3	81.7	74.9	77.6	68.3	73.9	50.9	68.6	73.7	78.3	-
C^2Former [11]	87.2	80.7	**82.7**	77.4	58.4	72.9	**71.4**	75.2	75.7	101.0	430.5
Ours	88.8	**85.9**	75.6	74.1	**70.5**	77.4	67.9	**82.4**	**77.8**	8.8	**40.6**

mated Geographic Reference Center (AGRC) database. It is widely adopted as a benchmark for vehicle detection in high-resolution remote sensing imagery. Each original image has dimensions of $16,000 \times 16,000$ pixels, with a ground resolution of approximately 12.5 cm per pixel. In our experiments, we use the 1024×1024 image size. The detection task involves eight object categories: car, pickup, camping car, truck, other, tractor, boat, and van. The dataset is split into 1,089 image pairs for training and 121 pairs for testing.

B. Implementation details

1) Parameter settings: All experiments were implemented using the PyTorch framework and conducted on an NVIDIA RTX A5000 GPU. The software environment included Python 3.9, PyTorch 2.0.1, and CUDA 12.2. The network was trained using Stochastic Gradient Descent (SGD) with an initial learning rate of 0.01, a momentum of 0.937, and a weight decay of 0.0005. Training was performed with a batch size of 16 for 150 epochs.

2) Evaluation metrics: In object detection, Precision, Recall, mAP50, and mAP50:95 are widely adopted as fundamental performance metrics. The definitions of them are as follows

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{15}$$

$$\text{AP} = \int_0^1 P(R)dR \qquad \text{mAP} = \frac{1}{C}\sum_{i=1}^{C} \text{AP}_i, \tag{16}$$

where TP represents the number of positive samples correctly detected. FP denotes the number of negative samples incorrectly detected as positive. FN indicates the number of positive samples that were not detected. AP is the average precision of each class, while mAP represents the average of AP values across all C classes. Specifically, mAP50 means the mAP calculated at an IoU threshold of 0.5. mAP50:95 refers to the mAP calculated over multiple IoU thresholds ranging from 0.5 to 0.95 with a step size of 0.05. As for model efficiency, the floating point operations (GFLOPs) and the number of parameters (Params) are used.

C. Comparison Experiment

In this part, we compare the proposed approach with several popular object detection methods on two benchmark datasets. We present the effectiveness of our model through both tabular and figure representations.

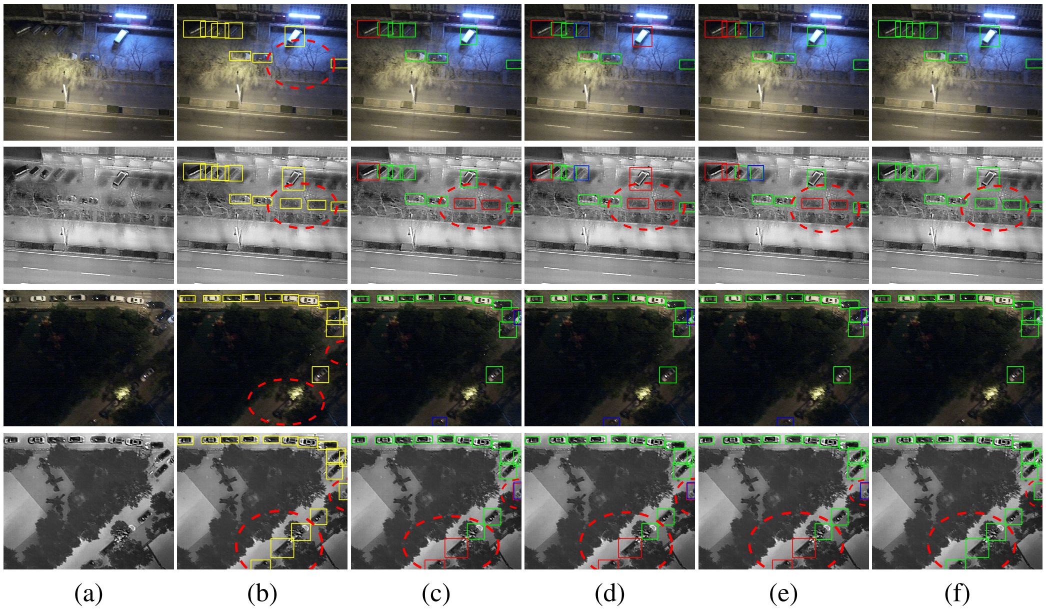

Fig. 5. Visualization on the DroneVehicle dataset under weakly misalignment conditions. (a) Base images. (b) Ground truth. (c) YOLOv5. (d) YOLOv8. (e) YOLOv11. (f) BANet. The red circle indicates a weak misalignment of the mode at this location. Green boxes indicate correctly detected, blue boxes indicate incorrectly detected, and red boxes indicate missed detections.

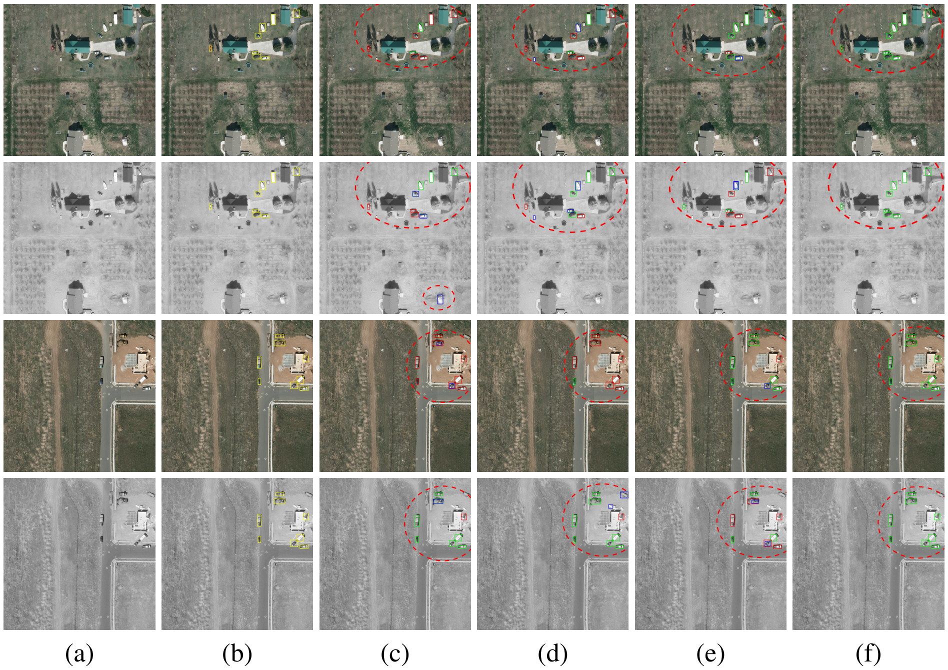

Fig. 6. Visualization on the VEDAI dataset under weakly misalignment conditions. (a) Base images. (b) Ground truth. (c) YOLOv5. (d) YOLOv8. (e) YOLOv11. (f) BANet. The red circle indicates a weak misalignment of the mode at this location. Green boxes indicate correctly detected, blue boxes indicate incorrectly detected, and red boxes indicate missed detections.

1) Comparison on VEDAI and DroneVehicle: The DroneVehicle [16] dataset contains both RGB and IR images captured under complex aerial conditions. The compared methods are divided into two groups. The first group is compared with unimodal detectors, we compare the proposed approach with several representative methods, including RetinaNet [32], Faster R-CNN [33], Oriented R-CNN [34], S^2A-Net [23], RoITransformer [35], R^3Det [36], YOLOv5s [37], YOLOv8s [38], and YOLO11s [39]. Each model is independently trained and tested on RGB and IR modalities to assess robustness under different imaging conditions. As shown in Table I, our BANet achieves the highest mAP50 of 82.1%, outperforming all unimodal methods. The IR-based YOLO11s (81.5%) and IR-based YOLOv8s (80.7%) perform slightly lower. Meanwhile, our network remains lightweight, with only 8.8 M parameters, much smaller than most traditional detectors. The second group is compared with multimodal detectors. We compare the proposed approach with several state-of-the-art multimodal methods, including DMM [48], OAFA [29], CFT [45], and C^2Former [11]. As shown in Table III, our approach achieves an mAP50 of 82.1%, surpassing DMM and OAFA (both 79.4%). In addition, it requires only 8.8M parameters and 40.6 GFLOPs, whereas CFT and C^{2}Former need 206.2M and 79.0M parameters, respectively. These results indicate that the proposed approach achieves a favorable balance between detection accuracy and computational efficiency on the DroneVehicle dataset.

The VEDAI [5] dataset includes Car, Pickup, Camping, Truck, Other, Tractor, Boat, and Van. In these categories, different types of images have different resolutions and object sizes. The detection performance compared with unimodal detectors is shown in Table II. We compare our approach with YOLOrs [40], RetinaNet [32], SuperYOLO [10], FCOS [41], YOLOv5x [37], YOLOv8x [38], and YOLOv10x [39]. Our BANet, which fuses RGB and IR features, achieves a mAP50 of 77.8%, exceeding all unimodal methods with only 8.8M parameters and 40.6GFLOPs. It also attains strong category-level results, such as Other (70.5%), Van (82.4%), and Tractor (77.4%). For multimodal comparison on the VEDAI dataset, as shown in Table IV. Our approach consistently outperforms existing multimodal methods, achieving a mAP50 of 77.8% compared to 75.9% for AFD [50] and 75.7% for C^{2}Former [11]. Although these methods use more complex models, their performance is still lower than our BANet. Even with a lightweight design, our approach surpasses large-scale detectors such as YOLOv8x (68.2M, 258.3 GFLOPs) and YOLOv5x (97.4M, 247.1 GFLOPs). These results demonstrate that the proposed approach can efficiently integrate multimodal information while maintaining strong detection performance on the VEDAI dataset.

2) Visualization Results: To give a clearer view of the detection performance, we select three methods and our BANet for qualitative comparison. The visualization results on the DroneVehicle [16] and VEDAI [5] datasets are shown in Fig. 5 and Fig. 6, respectively. It can be observed that the compared detectors exhibit limited performance under weakly misalignment conditions. As shown in Fig. 5, the contrast methods fail to detect the weakly misaligned objects in the IR image, whereas our proposed approach successfully detects all objects. Apart from this, except for our BANet, other methods usually miss or mistakenly localize some objects. Whereas, with smaller objects as shown in Fig. 6, none of the detectors except our BANet fully detected all the objects. Under weakly misalignment conditions, our method accurately identifies inconspicuous objects and maintains robust detection performance.

V. Ablation Studies

Here, we conduct ablation experiments on the DroneVehicle [16] dataset to evaluate the contribution of each proposed component. The quantitative results are summarized in Table V.

1) Baseline Setup: When adopting a single modality, the baseline achieves 76.8% mAP50 with the RGB modality and 80.7% with the IR modality, indicating that the IR modality provides more discriminative representations for the DroneVehicle [16] dataset.

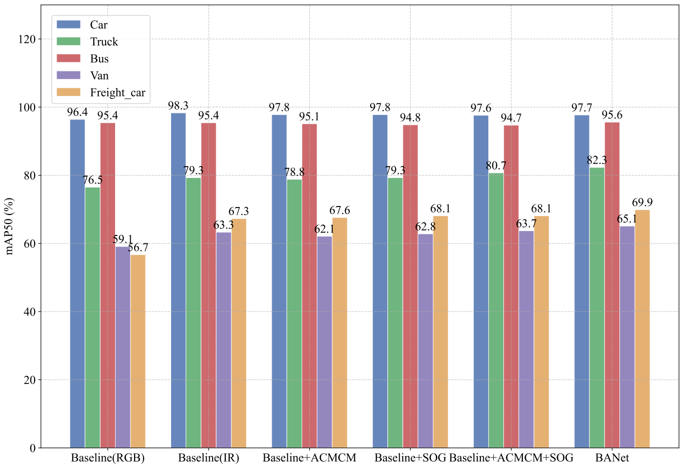

Fig. 7. MAP50 ablation experiments across different detection categories on the DroneVehicle.

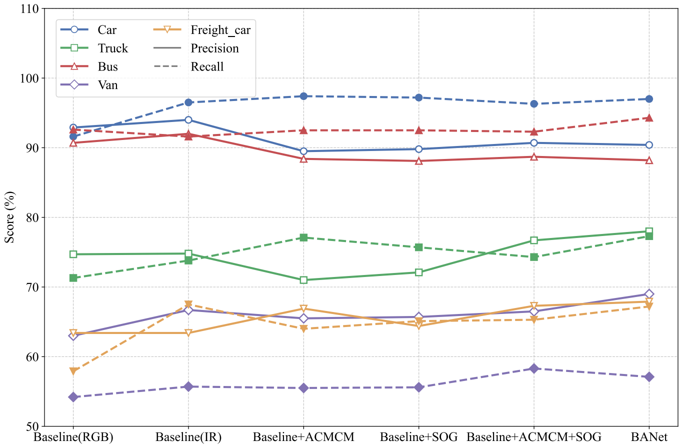

Fig. 8. Precision and Recall ablation experiments across different detection categories on the DroneVehicle.

2) Effects of the ACMCM: To explore the effect of ACMCM on BANet, we conducted experiments by adding ACMCM alone. Adding ACMCM after the backbone achieves better metrics. As shown in Table V, introducing only the ACMCM module achieves 80.3% mAP50, 72.7% mAP75, 76.3% Precision, 77.3% Recall, and 60.4% mAP50:95 on the DroneVehicle [16] dataset. According to Table VI, either global or spatial modeling alone improves performance, but combining both achieves the best results, demonstrating their complementary effect. These results confirm that correlation modeling enhances semantic interaction across modalities and reduces modality-specific inconsistencies. Furthermore, this design also improves weakly aligned multimodal feature learning.

3) Effects of the SOG: To verify the effectiveness of the proposed SOG, we conduct experiments on the DroneVehicle [16] dataset. The result is shown in Table V, after incorporating this module, achieves 80.6% mAP50 and 73.7% mAP75, outperforming both RGB and IR baselines. Table VII shows that adding either the coarse or fine branch alone brings performance gains. Furthermore, adding both branches with

TABLE V
RESULTS OF THE ABLATION STUDY FOR DIFFERENT ESSENTIAL COMPONENTS IN BANET. BOLDING INDICATES THE BEST RESULTS

Method	mAP50	mAP75	mAP50:95	Pre	Rec	Params(M)	GFLOPs(G)
Baseline(RGB)	76.8	65.6	55.0	75.3	73.5	**2.6**	**6.3**
Baseline(IR)	80.7	73.3	60.7	78.2	77.0	2.6	6.3
Baseline+ACMCM	80.3	72.7	60.4	76.3	77.3	3.9	10.4
Baseline+SOG	80.6	73.7	60.2	76.0	77.2	6.3	27.6
Baseline+ACMCM+SOG	81.0	73.5	60.8	78.0	77.3	6.6	29.0
Our BANet	**82.1**	**75.1**	**62.2**	**78.7**	**78.6**	8.8	40.6

TABLE VI
ABLATION STUDY FOR THE ACMCM ON THE DRONEVEHICLE TEST DATASET

ACMCM		mAP50	mAP75	Params(M)	GFLOPs(G)
Global modeling	Spatial modeling				
×	×	76.8	65.6	2.6	6.3
✓	×	79.5	70.1	3.3	8.4
×	✓	79.8	70.4	3.5	9.0
✓	✓	80.3	72.7	3.9	10.4

a modulation operation further enhances accuracy, indicating that the complete coarse-to-fine design stabilizes alignment under weak registration. These results validate that the SOG achieves stable and reliable feature alignment across diverse conditions. By generating both coarse-grained and fine-grained offsets through shared features and adaptive modulation, it effectively enhances alignment robustness.

TABLE VII
ABLATION STUDY FOR THE SOG ON THE DRONEVEHICLE TEST DATASET

SOG			mAP50	mAP75	Params(M)	GFLOPs(G)
Coarse	Fine	Modulation				
×	×	×	76.8	65.6	2.6	6.3
✓	×	×	79.2	70.0	5.4	23.6
✓	✓	×	79.8	70.8	5.8	25.1
✓	✓	✓	80.6	73.7	6.3	27.6

4) Effects of the PFS: In Table V, we evaluate the impact of PFS on the final prediction results. Overall, with the complete design, our BANet framework achieves the best performance on DroneVehicle [16], reaching 82.1% mAP50, 75.1% mAP75, and 62.2% mAP50:95. These results clearly demonstrate that this structure adaptively controls the contribution of each feature stream, suppressing redundancy while preserving complementary information. In addition, this fusion strategy maximizes the complementary advantages of multimodal images under weak alignment conditions.

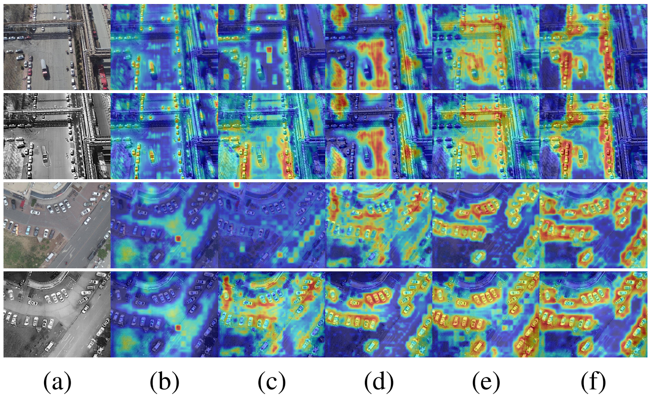

Fig. 9. Heatmap visualisation results on the DroneVehicle dataset. (a) Base images. (b) Baseline. (c) Baseline+ACMCM. (d) Baseline+SOG. (e) Baseline+ACMCM+SOG. (f) BANet.

5) Visualization Results: Furthermore, to validate the effectiveness of each proposed component, we conduct a comprehensive visualization analysis. Fig. 7 and 8 present the mAP50 performance, precision, and recall scores across different categories through progressive integration of our modules. The results demonstrate that each component contributes significantly to the final performance, with the complete BANet framework achieving the best results across all categories. Moreover, Fig. 9 provides thermal map comparisons. In contrast, the ACMCM enhances semantic consistency between modalities and reduces background interference. The SOG further improves spatial alignment by correcting positional deviations between RGB and IR features, which leads to more accurate and stable attention distributions. Furthermore, PFS helps the network effectively integrate the original and aligned features. Most notably, the complete BANet framework produces the most discriminative thermal responses, with sharp activations centered on target objects and minimal background interference. These visualizations quantitatively verify that our method successfully addresses the challenges of weakly aligned multimodal images.

VI. CONCLUSION

In this ~~article,~~ we propose a detection framework, BANet, designed for weakly aligned multimodal remote sensing object detection. The network effectively mitigates asymmetric cross-modal information utilization and spatial misalignment through the proposed WAM, which consists of the ACMCM, SOG, and a PFS. These components work collaboratively to enhance semantic interaction, achieve stable spatial alignment, and adaptively integrate multimodal features. Experimental results on the DroneVehicle [16] and VEDAI [5] datasets demonstrate that BANet achieves superior detection accuracy with lower computational complexity and fewer parameters compared to existing methods. ~~The proposed approach~~ exhibits strong robustness and generalization in challenging, weakly aligned scenarios. Future work will focus on further improving cross-modal representation learning, exploring more efficient fusion mechanisms, and extending the framework to more complex multimodal object detection tasks.

References

[1] J. Su, D. Yi, B. Su, Z. Mi, C. Liu, X. Hu, X. Xu, L. Guo, and W.-H. Chen, "Aerial visual perception in smart farming: Field study of wheat yellow rust monitoring," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 3, pp. 2242–2249, 2021.

[2] W. Sakla, G. Konjevod, and T. N. Mundhenk, "Deep multi-modal vehicle detection in aerial isr imagery," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017, pp. 916–923.

[3] K. Telegraph and C. Kyrkou, "Spatiotemporal object detection for improved aerial vehicle detection in traffic monitoring," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 12, pp. 6159–6171, 2024.

[4] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1037–1045.

[5] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery: A small target detection benchmark," *Journal of Visual Communication and Image Representation*, vol. 34, pp. 187–203, 2016.

[6] H. Fu, S. Wang, P. Duan, C. Xiao, R. Dian, S. Li, and Z. Li, "Lraf-net: Long-range attention fusion network for visible–infrared object detection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 10, pp. 13 232–13 245, 2023.

[7] Y. Zhang, Y. Zhang, Z. Shi, R. Fu, D. Liu, Y. Zhang, and J. Du, "Enhanced cross-domain dim and small infrared target detection via content-decoupled feature alignment," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[8] C. Jiang, H. Ren, H. Yang, H. Huo, P. Zhu, Z. Yao, J. Li, M. Sun, and S. Yang, "M2fnet: Multi-modal fusion network for object detection from visible and thermal infrared images," *International Journal of Applied Earth Observation and Geoinformation*, vol. 130, p. 103918, 2024.

[9] M. Liang, J. Hu, C. Bao, H. Feng, F. Deng, and T. L. Lam, "Explicit attention-enhanced fusion for rgb-thermal perception tasks," *IEEE Robotics and Automation Letters*, vol. 8, no. 7, pp. 4060–4067, 2023.

[10] 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 Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[11] M. Yuan and X. Wei, "C^{2}former: Calibrated and complementary transformer for rgb-infrared object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–12, 2024.

[12] P. Lyu, P.-H. Yeung, X. Yu, X. Cheng, C. Wu, and J. C. Rajapakse, "Efficient fourier filtering network with contrastive learning for aav-based unaligned bimodal salient object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–12, 2025.

[13] X. Bi, R. Qie, C. Tao, Z. Zhang, and Y. Xu, "Unsupervised multimodal uav image registration via style transfer and cascade network," *Remote Sensing*, vol. 17, no. 13, p. 2160, 2025.

[14] Z. Wang and Q. Zhang, "Real-time aerial multispectral object detection with dynamic modality-balanced pixel-level fusion," *Sensors*, vol. 25, no. 10, p. 3039, 2025.

[15] R. Li, J. Xiang, F. Sun, Y. Yuan, L. Yuan, and S. Gou, "Multiscale cross-modal homogeneity enhancement and confidence-aware fusion for multispectral pedestrian detection," *IEEE Transactions on Multimedia*, vol. 26, pp. 852–863, 2023.

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

[17] S. K. Roy, A. Deria, D. Hong, B. Rasti, A. Plaza, and J. Chanussot, "Multimodal fusion transformer for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–20, 2023.

[18] C. Zhang, H. Wang, Y. Cai, L. Chen, Y. Li, M. A. Sotelo, and Z. Li, "Robust-fusionnet: Deep multimodal sensor fusion for 3-d object detection under severe weather conditions," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–13, 2022.

[19] 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 *IGARSS 2023-2023 IEEE international geoscience and remote sensing symposium*. IEEE, 2023, pp. 5728–5731.

[20] Z. Wang, S. Li, and K. Huang, "Cross-modal adaptation for object detection in infrared remote sensing imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 22, pp. 1–5, 2025.

[21] X. Sun, Y. Yu, and Q. Cheng, "Low-rank multimodal remote sensing object detection with frequency filtering experts," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.

[22] L. Liu, M. Zhang, C. Li, C. Li, and J. Tang, "Cross-modal object tracking via modality-aware fusion network and a large-scale dataset," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 4, pp. 6981–6994, 2025.

[23] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *IEEE transactions on geoscience and remote sensing*, vol. 60, pp. 1–11, 2021.

[24] 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 Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 4, pp. 2619–2632, 2024.

[25] Q. Wang, Y. Liu, Z. Xiong, and Y. Yuan, "Hybrid feature aligned network for salient object detection in optical remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[26] X. Xie, C. Lang, S. Miao, G. Cheng, K. Li, and J. Han, "Mutual-assistance learning for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 15 171–15 184, 2023.

[27] L. Zhang, Z. Liu, X. Zhu, Z. Song, X. Yang, Z. Lei, and H. Qiao, "Weakly aligned feature fusion for multimodal object detection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 3, pp. 4145–4159, 2025.

[28] Y. Liu, W. Guo, C. Yao, and L. Zhang, "Dual-perspective alignment learning for multimodal remote sensing object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–15, 2025.

[29] C. Chen, J. Qi, X. Liu, K. Bin, R. Fu, X. Hu, and P. Zhong, "Weakly misalignment-free adaptive feature alignment for uavs-based multimodal object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 836–26 845.

[30] L. Zongzhen, L. Hui, W. Zhixing, W. Yuxing, Z. Haorui, and Z. Jianlin, "Cross-modal offset-guided dynamic alignment and fusion for weakly aligned uav object detection," *arXiv preprint arXiv:2506.16737*, 2025.

[31] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[32] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[33] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[34] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented r-cnn for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3520–3529.

[35] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning roi transformer for oriented object detection in aerial images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2849–2858.

[36] X. Yang, J. Yan, Z. Feng, and T. He, "R3det: Refined single-stage detector with feature refinement for rotating object," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 4, 2021, pp. 3163–3171.

[37] G. Jocher, "Ultralytics yolov5," 2020. [Online]. Available: https://github.com/ultralytics/yolov5

[38] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolov8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[39] G. Jocher and J. Qiu, "Ultralytics yolo11," 2024. [Online]. Available: https://github.com/ultralytics/ultralytics

[40] M. Sharma, M. Dhanaraj, S. Karnam, D. G. Chachlakis, R. Ptucha, P. P. Markopoulos, and E. Saber, "Yolors: Object detection in multimodal remote sensing imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 1497–1508, 2020.

[41] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: A simple and strong anchor-free object detector," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 4, pp. 1922–1933, 2020.

[42] L. L. e. a. Ao Wang, Hui Chen, "Yolov10: Real-time end-to-end object detection," *arXiv preprint arXiv:2405.14458*, 2024.

[43] X. Yang and J. Yan, "Arbitrary-oriented object detection with circular smooth label," in *European conference on computer vision*. Springer, 2020, pp. 677–694.

[44] Y. Chen, B. Wang, X. Guo, W. Zhu, J. He, X. Liu, and J. Yuan, "Deyolo: Dual-feature-enhancement yolo for cross-modality object detection," in *International Conference on Pattern Recognition*. Springer, 2024, pp. 236–252.

[45] F. Qingyun, H. Dapeng, and W. Zhaokui, "Cross-modality fusion transformer for multispectral object detection," *arXiv preprint arXiv:2111.00273*, 2021.

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

[47] S. Wang, X. Yang, R. Lu, D. Zhang, W. Xie, S. Su, and Z. Zhang, "Mrod-yolo: Multimodal joint representation for small object detection in remote sensing imagery via multi-scale iterative aggregation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–14, 2025.

[48] M. Zhou, T. Li, C. Qiao, D. Xie, G. Wang, N. Ruan, L. Mei, Y. Yang, and H. T. Shen, "Dmm: Disparity-guided multispectral mamba for oriented object detection in remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[49] Y. Zhang, M. Ye, G. Zhu, Y. Liu, P. Guo, and J. Yan, "Ffca-yolo for small object detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.

[50] M. Ji, B. Heo, and S. Park, "Show, attend and distill: Knowledge distillation via attention-based feature matching," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 9, 2021, pp. 7945–7952.

Yutian Shi received his B.A degree from Chongqing University of Posts and Telecommunications in 2023. He is currently pursuing the M.S. degree with the School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China. His research interests include object detection and deep learning.

Guoquan Li (Member, IEEE) received the M.S. and Ph.D. degree in circuits and systems from Chongqing University, Chongqing, China, in 2006 and 2012, respectively. From 2009 to 2010, he was a Visiting Ph.D. Student with the Department of Electrical and Computer Engineering, University of Delaware, Newark, DE, USA. He is currently a Professor with the School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications (CQUPT), Chongqing. His current research interests include image processing, machine learning, and MIMO wireless networks.

Zhilong Shen received the M.S. degree from Chongqing University of Science and Technology, in 2023. He is currently pursuing the Ph.D. degree with the School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China. His research interests include object detection and image recognition.

Hongying Meng (Senior Member, IEEE) received his Ph.D.degree in Communication and Electronic Systems from Xi'an Jiaotong University, Xi'an, China. He is an associate editor for IEEE Transactions on Circuits and Systems for Videos Technology (TCSVT) and IEEE Transactions on Cognitive and Developmental Systems (TCDS). He has authored over 200 publications including IEEE TIP, TCYB, TFS, TAC, TCSVT, TBE, TCDS, ICASSP and CVPR. He is currently a ~~Reader~~ at the Department of Electronic and Electrical Engineering, Brunel University London, U.K. His research interests include digital signal processing, affective computing, machine learning, human computer interaction, and computer vision.

Yu Pang (Member, IEEE) received his Ph.D. degree from McGill University in Canada in 2010. He is currently a Professor with the School of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing. His current research interests include integrated circuit design, artificial intelligence, and the development of digital medical devices.