IGIANet: Illumination Guided Implicit Alignment Network for Infrared–Visible UAV Detection

Xiangqi Chen[13*], Dawei Zhang[3*], Li Zhao[3*], Chengzhuan Yang[3], Zhongyu Chen[3], Jungang Lou[4], Zhonglong Zheng[123†], Sang-Woon Jeon[6], Hua Wang[5]

[1]Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University, Jinhua, Zhejiang 321004, China
[2]China-Mozambique "Belt and Road" Joint Laboratory on Smart Agriculture, Zhejiang Normal University, Jinhua, Zhejiang, 321004, China
[3]School of Computer Science and Technology, Zhejiang Normal University, Zhejiang, 321004, China
[4]College of Information Engineering, Huzhou University, Huzhou 313000, China
[5]Institute for Sustainable Industries and Liveable Cities, College of Engineering and Science, Victoria University, Melbourne, VIC, 8001, Australia
[6]Department of Electrical and Electronic Engineering, Hanyang University, Seoul,04763, South Korea
{zjnu_cxq, davidzhang, zhaoli2023, czyang, czy}@zjnu.edu.cn, ljg@zjhu.edu.cn, zhonglong@zjnu.edu.cn, sangwoonjeon@hanyang.ac.kr, hua.wang@vu.edu.au

Abstract

Visible-Infrared (RGB-IR) Unmanned Aerial Vehicle (UAV) object detection integrates complementary cues from visible and infrared sensors, offering broad application potential. However, due to sensor parallax, it still faces the challenge of weak spatial misalignment, which significantly limits its performance in UAV-based object detection. Existing methods emphasize strict alignment, overlooking spectral heterogeneity under varying illumination. To address these issues, we propose the **I**llumination **G**uided Implicit **A**lignment **Net**work **(IGIANet)** to mitigate modality heterogeneity without explicit alignment. Specifically, we integrate three novel modules. First, we propose an illumination-guided frequency modulation module that adaptively allocates fusion weights to visible and infrared features based on global illumination estimation, effectively alleviating modality imbalance under varying lighting conditions. Second, we introduce a frequency-guided cross-modality differential enhancement module, which computes differential cues across frequency domains to enhance complementary information and highlight weakly aligned and low-contrast regions. Finally, we introduce an implicit alignment-driven dynamic fusion module that actively estimates offsets and generates dynamic, position-adaptive fusion kernels to align and fuse modalities. Extensive experiments demonstrate that IGIANet outperforms state-of-the-art models on various benchmarks, achieving 80.9% mAP on DroneVehicle, 57.1% mAP on VEDAI, and 49.4% mAP on FLIR.

Introduction

Unmanned Aerial Vehicle (UAV) object detection has garnered increasing attention due to its wide range of applications, such as forest rescue (Yao, Qin, and Chen 2019; Fu et al. 2021), traffic surveillance (Shakhatreh et al. 2019; Chen et al. 2025b), and autonomous driving (Li et al. 2022b; Chen et al. 2024). However, most existing detectors are designed for visible (RGB) images, which, despite being rich in color and texture, often degrade significantly under low-light or poor visibility conditions. In contrast, infrared (IR) sensors are more robust to environmental variations and can operate effectively in complex scenarios. Consequently, UAV-based RGB-IR object detection has emerged as a promising approach for enhancing detection performance in challenging environments.

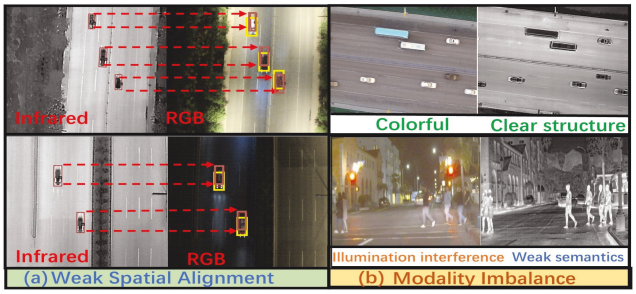

Figure 1: (a) illustrates the weak alignment issue: the red and yellow boxes denote ground-truth annotations in the IR and RGB images, showing a clear spatial misalignment. (b) highlights the modality imbalance: RGB images offer rich color and texture but degrade under poor illumination, while IR images lack color yet retain structural details, making targets clearer in low-light scenes.

In recent years, researchers have increasingly focused on combining the structural advantages of IR imagery in complex nighttime scenarios with the rich color and texture information from RGB images, aiming to achieve efficient multimodal fusion and improve RGB–IR object de-

*These authors contributed equally.
†Corresponding Author.

tection performance on UAV platforms. However, as illustrated in Figure 1, existing methods still face two major challenges (Zhou, Chen, and Cao 2020; Yuan and Wei 2024; Chen et al. 2023, 2025a): **1. Weak spatial alignment:** Due to the asynchronous imaging of RGB and IR sensors, the same object often appears in different spatial locations across the two modalities. This leads to alignment errors, which in turn impair the accurate localization of key object regions and compromise detection performance. **2. Modality imbalance:** While RGB images provide rich details under favorable lighting conditions, they are susceptible to illumination variations. In contrast, IR images are illumination-invariant and can consistently capture the structural characteristics of objects, yet they lack color and fine-grained texture (Li et al. 2022a). These inherent disparities and imbalances in information content pose significant challenges to effective multimodal object detection.

To address the above challenges, one of the most straightforward strategies is to perform pixel-wise alignment between RGB and IR images prior to detection (Chen, Liu, and Tan 2021), as illustrated in Figure 2 (a). While this approach explicitly considers the misalignment issue, it fails to fully exploit the complementary information between modalities during the subsequent feature fusion stage. Another line of work focuses on local feature alignment (Zhang et al. 2019b), as shown in Figure 2 (b), where the region of interest (ROI) from one modality is aligned to that of the other via random perturbations or jitters. Although this method shows some improvements, it suffers from low computational efficiency and may not be well-suited for real-time UAV-based applications.

Motivated by the above observations, we argue that an ideal RGB-IR fusion strategy for UAV-based object detection should: (1) simultaneously address both spatial misalignment and modality imbalance, and (2) avoid expensive explicit registration or complex geometric supervision. To this end, we introduce the Illumination Guided Implicit Alignment Network (IGIANet), as illustrated in Figure 2 (c). Specifically, we propose the illumination-guided frequency modulation module (IFMM), which estimates the global illumination coefficient from the RGB image to reweight the RGB and IR features adaptively. In poorly lit environments, this mechanism suppresses the unreliable modality (e.g., downweighting RGB at night) and emphasizes the modality that carries more discriminative cues under current lighting conditions. Secondly, we design the frequency-guided cross-modality differential enhancement module (FG-CMDEM) to highlight complementary structural details that are most vulnerable to spatial misalignment. It computes cross-modal differential cues in the frequency domain and selectively enhances high-frequency discrepancies while suppressing shared low-frequency components. Finally, we propose the implicit alignment-driven dynamic fusion module (IADDF), which predicts offset fields via global and local branches. IR features are then implicitly aligned to RGB features using grid sampling. Unlike prior works that treat alignment and fusion separately, IGIANet unifies illumination-aware feature balancing, spectral enhancement, and alignment into a single differentiable pipeline guided by global illumination.

In summary, our main contributions are as follows: (1) A principled illumination-guided spectral recalibration strategy that dynamically balances the importance of RGB and IR modalities based on scene brightness. (2) A frequency-guided differential enhancement mechanism designed to mine and amplify differential clues. (3) An implicit alignment and dynamic fusion strategy for effective modality alignment and fusion. (4) Extensive experiments on public datasets demonstrate the effectiveness and robustness of the proposed IGIANet model.

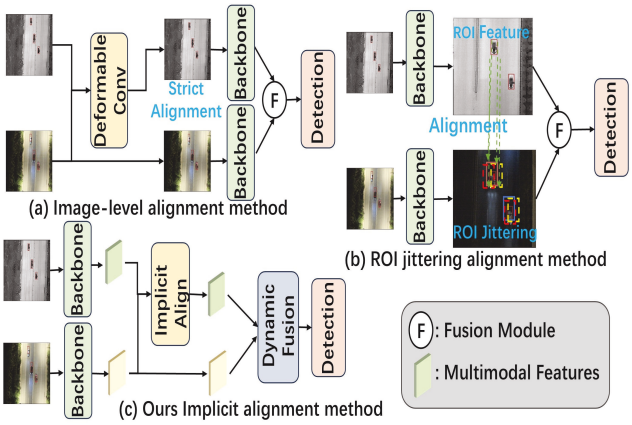

Figure 2: Illustration of different solutions for RGB-IR weak alignment: (a) adopts a global, rigid pixel-wise alignment strategy. (b) applies a local ROI jittering-based alignment method. (c) presents our implicit alignment approach combined with a dynamic convolution fusion strategy.

Proposed Method

Overall Architecure

As illustrated in Figure 3, we propose IGIANet, a unified framework that progressively enhances RGB–IR feature fusion via four tightly coupled modules, each designed to address a specific aspect of modality imbalance under illumination variation and spatial misalignment. First, the opponent-aware light illumination module (OLIM) derives a global illumination weight map from opponent color and luminance cues, which is then utilized by IFMM to balance spectral responses in the frequency domain. Next, FG-CMDEM complements and fuses cross-modal information by computing differential frequency cues. Finally, IADDF predicts and samples position-aware offsets via dual branches to align IR features with RGB ones, and generates dynamic, patch-wise kernels for adaptive fusion.

Opponent-aware Light Illumination Module

To robustly guide frequency modulation and alignment under varying illumination conditions, it is essential to capture color shifts and luminance variations. Color opponency highlights the color imbalance introduced by scene lighting, while perceived luminance encodes the overall brightness. Therefore, we design the OLIM to generate illumination-aware weights.

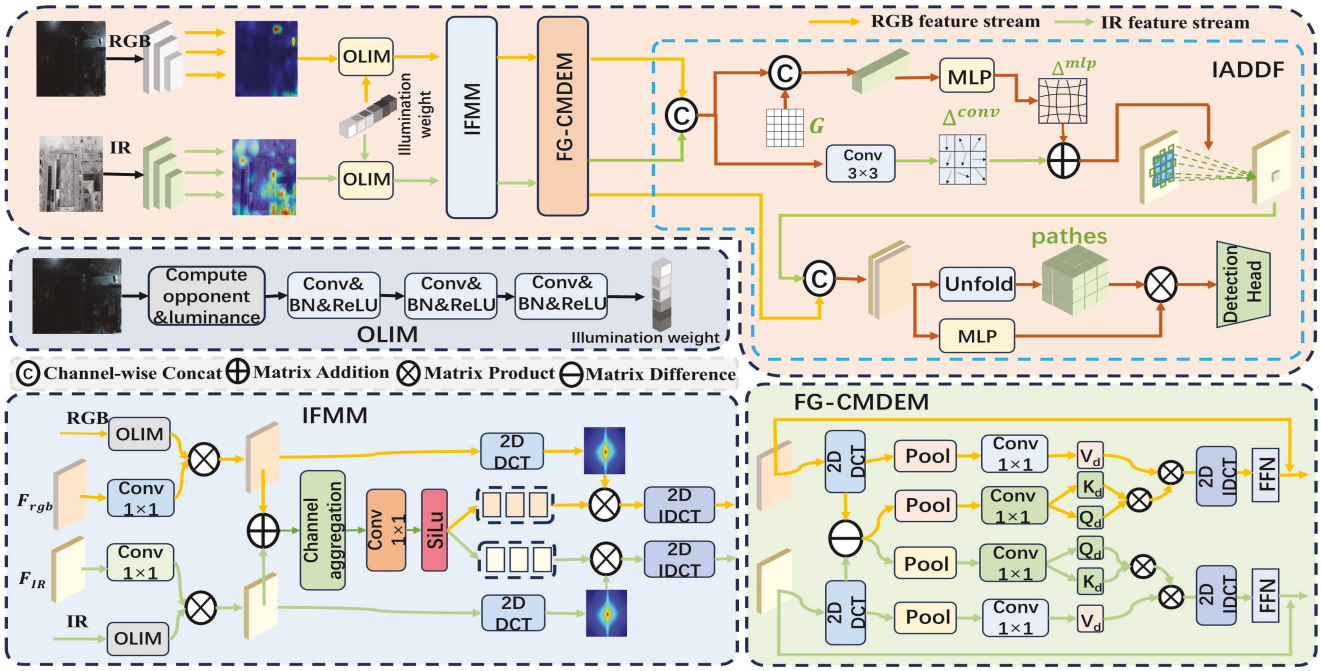

Figure 3: Overview of the IGIANet framework, which integrates an opponent-aware light illumination module (OLIM), illumination-guided frequency modulation module (IFMM), frequency-guided cross-modality differential enhancement module (FG-CMDEM), and implicit alignment-driven dynamic fusion module (IADDF) for UAV-based RGB-IR object detection.

Given an input RGB image $I \in \mathbb{R}^{3\times H\times W}$, H and W denote the height and width of the input image, we compute two opponent color channels, o_1 and o_2:

$$o_1 = R - G, o_2 = B - \tfrac{1}{2}(R+G), \tag{1}$$

where R, G, and B denote the red, green, and blue channels, respectively. Simultaneously, we convert the RGB image I into the LAB color space, normalize the L channel, and extract the perceived luminance channel L.

$$L = \frac{\mathrm{Lab}_L}{100}, \tag{2}$$

where $\mathrm{Lab}_L \in [0, 100]$ is the lightness component. These channels are concatenated to form the initial image $\hat{I}$, which effectively captures both color contrast and luminance distribution.

$$\hat{I} = [o_1;\, o_2;\, L] \in \mathbb{R}^{3\times H\times W}, \tag{3}$$

where $[;]$ is the concation operation. Subsequently, as shown in equation 4, we apply several convolutional blocks for downsampling and feature extraction, and finally use a sigmoid activation to produce the illumination weight map W_{illum}.

$$W_{illum} = \sigma(ConvBlock(\hat{I})), \tag{4}$$

where σ is the sigmoid operation.

Illumination-Guided Frequency Modulation Module

The IFMM adaptively balances and enhances RGB and IR features in the frequency domain under the guidance of illumination. Given the input RGB features F_{rgb}, IR features F_{ir}, and the per-pixel illumination weight map W_{illum}, we first normalize and smooth W_{illum} to ensure it lies within the range $\left[0.5 - \frac{\beta}{2},\ 0.5 + \frac{\beta}{2}\right]$. Then, we compute the complementary weights:

$$W_{\text{rgb}} = W_{\text{illum}}, \quad W_{\text{ir}} = 1 - W_{\text{illum}}, \tag{5}$$

$$\widetilde{F}_{\text{rgb}} = \mathrm{Conv}^{\text{g}}_{1\times1}(F_{\text{rgb}} \otimes W_{\text{rgb}}), \tag{6}$$

$$\widetilde{F}_{\text{ir}} = \mathrm{Conv}^{\text{g}}_{1\times1}(F_{\text{ir}} \otimes W_{\text{ir}}), \tag{7}$$

where Conv^{g} denotes group convolution with C groups, $\otimes$ denotes the matrix product. These weights are broadcast along the channel dimension C and applied to the features via grouped 1×1 convolutions. Subsequently, the weighted $\widetilde{F}_{\text{rgb}}$ and $\widetilde{F}_{\text{ir}}$ features are combined via element-wise addition to obtain F_{sum}:

$$F_{\text{sum}} = \widetilde{F}_{\text{rgb}} + \widetilde{F}_{\text{ir}}. \tag{8}$$

We then apply channel attention (Hu, Shen, and Sun 2018) to F_{sum} to perform channel-wise feature aggregation, resulting in the aggregated feature Z, which captures global semantic information across modalities:

$$Z = Attn(F_{\text{sum}}) \otimes F_{\text{sum}}, \tag{9}$$

where $Attn$ denotes the channel atention. Next, we apply a 1×1 convolution followed by a SiLU activation to Z, generating two sets of frequency modulation weights, A_{rgb} and A_{ir}, which are used to enhance the frequency components of the F_{rgb} and F_{ir} features:

$$[A_{rgb}; A_{ir}] = \mathrm{SiLU}\big(\mathrm{Conv}_{1\times1}(Z)\big), \tag{10}$$

$$A_{rgb}, A_{ir} \leftarrow \mathrm{split}(A_{rgb}; A_{ir}). \tag{11}$$

Specifically, the modulated features $\tilde{F}_{\text{RGB}}$ and $\tilde{F}_{\text{IR}}$ are fed into dynamic filters along with their corresponding weights A_{rgb} and A_{ir}, respectively. These filters perform weighted 2D DFT–IDFT operations in the frequency domain, enabling position-adaptive spectral reconstruction to dynamically compensate for illumination-induced response discrepancies:

$$\widehat{F}_m(x,y) = \mathcal{F}^{-1}\Big(A_m(u,v) \otimes \mathcal{F}\{\widetilde{F}_m(x,y)\}\Big), \quad m \in \{\text{rgb}, \text{ir}\} \tag{12}$$

where $\mathcal{F}\{\cdot\}$ and $\mathcal{F}^{-1}\{\cdot\}$ denote the 2D Discrete Cosine Transform (2D DCT) and its inverse (2D IDCT), respectively; $A_m(u,v)$ is the corresponding frequency-domain modulation weight map. This makes the frequency modulation more targeted and more effective at emphasizing discriminative information. By embedding illumination correction into the generation of frequency-domain weights, IFMM not only rectifies illumination bias in the spatial domain but also ensures that the spectral reconstruction process focuses on low-light regions based on lighting conditions, thereby significantly enhancing feature robustness.

Frequency-Guided Cross-Modality Differential Enhancement Module

The FG-CMDEM is designed to complement and enhance cross-modality differential information between frequency-modulated RGB and IR features, as illustrated in Figure 3. Unlike prior methods that apply frequency-based enhancement globally, our FG-CMDEM uniquely computes cross-modal differential cues in the frequency domain, enabling targeted structural enhancement.

Let $\widehat{F}_{\text{rgb}}, \widehat{F}_{\text{ir}} \in \mathbb{R}^{C\times H\times W}$ be the feature maps from the RGB and infrared branches after IFMM processing. To effectively capture illumination-driven differences between modalities, we first transform both modality features into the frequency domain using the 2D DCT, as follows:

$$\bar{F}_{\text{rgb}} = \mathcal{F}(\widehat{F}_{\text{rgb}}), \quad \bar{F}_{\text{ir}} = \mathcal{F}(\widehat{F}_{\text{ir}}). \tag{13}$$

For each modality $m \in \{\text{rgb}, \text{ir}\}$, we then compute the element-wise difference in the frequency domain:

$$F_d^m = \bar{F}_m - \bar{F}_{\bar{m}} \in \mathbb{R}^{C\times H\times W}, \tag{14}$$

where $\bar{m}$ denotes the opposite modality. Next, both $\bar{F}_m$ and F_d^m are downsampled by a factor of s using average pooling:

$$\bar{F}'_{\text{m}}, {F'_d}^m \in \mathbb{R}^{C\times H'\times W'}, \quad H' = \tfrac{H}{s}, W' = \tfrac{W}{s}. \tag{15}$$

To generate attention primitives, the difference map ${F'_d}^m$ is passed through two 1×1 convolutions that reduce its channel dimension from C to $C_q = C/s$, yielding the query $Q_d^m \in \mathbb{R}^{C_q\times H'\times W'}$ and key $K_d^m \in \mathbb{R}^{C_q\times H'\times W'}$. Meanwhile, the value $V_d^m \in \mathbb{R}^{C_q\times H'\times W'}$ is generated by applying a 1×1 convolution to $\bar{F}'_m$. All these are reshaped into sequences of length $N = H'\times W'$ and used for attention computation:

$$F_m^{att} = \mathcal{F}^{-1}\left(\text{FFN}\left(\text{softmax}\left(\frac{Q_d^m K_d^m}{\sqrt{d_k}}\right)V_d^m\right)\right) \oplus \widehat{F}_{\text{m}}, \tag{16}$$

where FFN denotes a feed-forward network, and $\oplus$ denotes residual addition with the original spatial-domain feature. The proposed FG-CMDEM is a cross-modal attention mechanism that fuses different modalities through explicit difference-driven fusion, as RGB-IR images have lighting differences that cannot be solved by simple addition.

Implicit Alignment-driven Dynamic Fusion Module

The proposed IADDF, as illustrated in Figure 3, is primarily designed to address the weak alignment issue between RGB and IR images. It first corrects IR features spatially using pixel-wise implicit offsets, then generates dynamic convolutional kernels to adaptively fuse the aligned bimodal features, and finally projects the fused output back to the original channel dimension. To robustly estimate spatial alignment offsets, we employ a dual-branch offset estimation module, consisting of an MLP-based global branch and a convolutional local branch. Specifically, given the input attention features $F_m^{\text{att}} \in \mathbb{R}^{C\times H\times W}$, we first generate a normalized coordinate map $\mathbf{G} \in [-1,1]^{H\times W\times 2}$ and concatenate each pixel's spatial coordinates (x_i, y_i) with its RGB and IR features:

$$\mathbf{z} = \big[\mathbf{G},\, F_m^{\text{att}}\big] \in \mathbb{R}^{(2+2C)\times H\times W}. \tag{17}$$

This tensor is passed through a two-layer MLP to estimate a pixel-wise offset:

$$\Delta^{\text{mlp}} = \text{MLP}(\mathbf{z}) \in \mathbb{R}^{2\times H\times W}. \tag{18}$$

Meanwhile, in the convolutional branch, we concatenate the RGB and IR features and apply a 3×3 convolution:

$$\Delta^{\text{conv}} = \text{Conv}_{3\times3}\big([F_{\text{rgb}}^{\text{att}},\, F_{\text{ir}}^{\text{att}}]\big) \in \mathbb{R}^{2\times H\times W}. \tag{19}$$

The two offset maps are summed to obtain the final displacement field:

$$\Delta = \Delta^{\text{mlp}} + \Delta^{\text{conv}}, \tag{20}$$

which is then used in bilinear sampling (GridSample) to align the IR features:

$$F'_{\text{ir}} = \text{GridSample}\big(F_{\text{ir}}^{\text{att}},\, \Delta\big). \tag{21}$$

Next, we concatenate the aligned IR and original RGB features:

$$F^{\text{cat}} = \big[F_{\text{rgb}}^{\text{att}},\, F'_{\text{ir}}\big] \in \mathbb{R}^{2C\times H\times W}. \tag{22}$$

For each spatial location i, its concatenated feature vector $F_i^{\text{cat}} \in \mathbb{R}^{2C}$ is fed into another MLP to predict dynamic convolution kernel weights:

$$\mathbf{w}_i = \text{MLP}_{\text{kernel}}\big(F_i^{\text{cat}}\big) \in \mathbb{R}^{k^2}. \tag{23}$$

After batch normalization and activation, these weights are reshaped into a position-specific kernel map $\mathcal{W} \in \mathbb{R}^{1\times H\times W\times k\times k}$. Simultaneously, F^{cat} is unfolded into local patches $\mathcal{P} \in \mathbb{R}^{2C\times H\times W\times k\times k}$. Each location i is convolved with its corresponding kernel:

$$F_i^{\text{dyn}} = \sum_{u,v} \mathcal{P}_i \otimes \mathcal{W}_i, \quad F^{\text{dyn}} \in \mathbb{R}^{2C\times H\times W}. \tag{24}$$

Method	Pub. + Year	RGB	Infrared	Car	Freight	Truck	Bus	Van	mAP (%)
RetinaNet	CVPR 2017	✓	✗	78.5	34.4	24.1	69.8	28.8	47.1
R3Det	AAAI 2021	✓	✗	80.3	56.1	42.7	80.2	44.4	60.8
S2ANet	TGRS 2021	✓	✗	80.0	54.2	42.2	84.9	43.8	61.0
Faster R-CNN	TPAMI 2016	✓	✗	79.0	49.0	37.2	77.0	37.0	55.9
RoITransformer	CVPR 2019	✓	✗	61.6	55.1	42.3	85.5	44.8	61.6
Oriented R-CNN	ICCV 2021	✓	✗	80.1	53.8	41.6	85.4	43.3	60.8
RetinaNet	CVPR 2017	✗	✓	88.8	35.4	39.5	76.5	32.1	54.5
R3Det	AAAI 2021	✗	✓	89.5	48.3	16.6	87.1	39.9	62.3
S2ANet	TGRS 2021	✗	✓	89.9	54.5	55.8	88.9	48.4	67.5
Faster R-CNN	TIPAMI 2016	✗	✓	89.4	53.5	48.3	87.0	42.6	64.2
RoITransformer	CVPR 2019	✗	✓	89.6	51.0	53.4	88.9	44.5	65.5
Oriented R-CNN	ICCV 2021	✗	✓	89.8	57.4	53.1	89.3	45.4	67.0
CIAN	IF 2019	✓	✓	90.1	63.8	60.7	89.1	50.3	70.8
MBNet	ECCV 2020	✓	✓	90.1	64.4	62.4	88.8	53.6	71.9
TSFADet	ECCV 2022	✓	✓	90.0	69.2	65.5	89.7	55.2	73.9
SLBAF-Net	MTA 2023	✓	✓	90.2	<u>68.6</u>	72.0	<u>89.9</u>	59.9	76.1
C2Former	TGRS 2024	✓	✓	90.2	68.3	64.4	89.8	58.5	74.2
DMM +Faster R-CNN	TGRS 2025	✓	✓	<u>90.4</u>	63.0	77.8	88.7	<u>66.0</u>	77.2
CoDAF	Arxiv 2025	✓	✓	90.3	66.2	<u>77.9</u>	80.5	65.0	<u>78.6</u>
IGIANet(Ours)	-	✓	✓	**90.5**	**76.2**	**78.2**	**90.3**	**69.4**	**80.9**

Table 1: Comparison on the DroneVehicle dataset. We compare our IGIANet method with single-modal and multimodal methods, both using Oriented Bounding Boxes (OBB) as the detection head. The best-performing results are highlighted in bold, and the second-best results are <u>underlined</u>.

Finally, a 1×1 convolution reduces the channel dimension:

$$F^{\text{out}} = \text{Conv}_{1\times1}\big(F^{\text{dyn}}\big) \in \mathbb{R}^{C\times H\times W}. \quad (25)$$

We adopt implicit alignment with two offset estimation branches: a global MLP branch capturing large-scale alignment deviations, and a local convolutional branch ensuring smooth, detail-consistent offsets. Their combination enables coarse global correction and fine local adjustment for accurate RGB–IR alignment.

Experiments

Experimental Settings

Datasets. We evaluate our method on three RGB-IR object detection benchmark datasets, including two UAV-based datasets:

(1) DroneVehicle (Sun et al. 2022) is a multimodal UAV dataset for vehicle detection and recognition, containing 28,439 paired RGB–IR images across diverse scenarios such as urban roads, highways, and parking lots. It includes five categories—car, truck, bus, van, and freight_car—with image pairs at a resolution of 650×512. The dataset is split into 17,990 training, 8,980 testing, and 1,469 validation samples.

(2) VEDAI (Razakarivony and Jurie 2016) provides RGB–IR imagery from suburban and rural scenes with varied ground textures. It includes two resolutions (1024×1025 and 512×512); we use the higher-resolution set of 1,246 image pairs. The dataset covers nine categories (e.g., car, truck, bus, van, airplane, ship), and we follow the same train/test split as in (Shen et al. 2024).

(3) FLIR (Zhang et al. 2020) is a challenging multimodal dataset comprising 5,142 RGB-IR object instances, with 4,129 images used for training and 1,013 for testing. It contains three object categories: person, bicycle, and car.

Implementation Details. We follow the same data processing protocol as previous work (Shen et al. 2024). The proposed IGIANet is built upon the YOLOv8 detector, implemented within the MMYOLO framework (Contributors 2022). We report mean Average Precision (mAP) consistent with prior methods, based on the IoU metric to evaluate detection performance. Specifically, mAP_{50} refers to the average precision at an IoU threshold of 0.5, while mAP denotes the average precision over a range of IoU thresholds from 0.5 to 0.95. Our model is optimized using Stochastic Gradient Descent (SGD) (Robbins and Monro 1951) with an initial learning rate of 0.000125 and a momentum of 0.937. The training is conducted over 36 epochs.

Main Results

We compare our proposed method, IGIANet, with several SOTA approaches on the DroneVehicle, VEDAI, and FLIR datasets. The corresponding results are presented in Tables 1, 2 and 3 respectively.

Comparison Results on DroneVehicle. As shown in Tables 1, our proposed IGIANet effectively addresses the weak

alignment problem and achieves efficient fusion of RGB and IR features, significantly improving detection performance. We compare our method with a variety of single-modality object detection methods, including R3Det (Yang et al. 2021), S2ANet (Han et al. 2021), Faster R-CNN (Ren et al. 2016), RetinaNet (Lin et al. 2017), and RoITransformer (Ding et al. 2019). IGIANet consistently outperforms these single-modality detectors and exceeds the best-performing RGB or IR-based methods by 11.5% and 17.3% in mAP. Compared with various advanced multimodal methods, including DMM+Faster R-CNN (Zhou et al. 2025), C2Former (Yuan and Wei 2024), SLBAF-Net (Cheng et al. 2023), CoDAF (Zongzhen et al. 2025), MBNet (Zhou, Chen, and Cao 2020), CIAN (Zhang et al. 2019a), and TS-FADet (Yuan, Wang, and Wei 2022), our method achieves superior performance. Specifically, IGIANet outperforms DMM+Faster R-CNN and CoDAF by 3.7% and 2.3% in mAP, respectively. These results demonstrate the effectiveness of our proposed method and also highlight the limitations of single-modality approaches in multimodal object detection tasks.

Method	**Type**	mAP_{50}	mAP
Faster R-CNN	RGB	78.6	47.2
YOLOv5	RGB	74.3	38
FCOS	RGB	63.3	37.8
DINO	RGB	73.8	48.3
Faster R-CNN	IR	75.5	44.1
YOLOv5	IR	74	35.1
FCOS	IR	56.2	33.3
DINO	IR	68.8	45.4
YOLO Fusion	RGB+IR	78.6	49.1
ADMPF	RGB+IR	81.6	-
TINet	RGB+IR	82.6	44.1
CFT	RGB+IR	<u>85.3</u>	56.0
ICAFusion	RGB+IR	84.8	<u>56.6</u>
IGIANet (Ours)	RGB+IR	**88.9**	**57.1**

Table 2: Comparisons on the VEDAI dataset.

Comparison Results on VEDAI. As shown in Table 2, our proposed IGIANet demonstrates strong performance on the VEDAI dataset, achieving an mAP_{50} of 88.9%. Compared with various single-modality detectors, including Faster R-CNN, YOLOv5 (Jia et al. 2021), FCOS (Tian et al. 2020), ADMPF (Liu, Li, and Peng 2025), and DINO (Zhang et al. 2022), our detector surpasses the best IR- and RGB-based detectors by 13.4% and 10.3% in mAP_{50}, respectively. When compared with multimodal detectors such as YOLO Fusion (Qingyun and Zhaokui 2022), TINet (Zhang et al. 2023), CFT (Qingyun, Dapeng, and Zhaokui 2021), and ICAFusion (Shen et al. 2024), our method outperforms ICAFusion by 4.1% in mAP_{50} and 0.5% in mAP. These results demonstrate that our method effectively fuses feature information from both IR and RGB images, leading to superior detection performance.

Comparison Results on FLIR. As shown in Table 3, our IGIANet demonstrates excellent performance on the FLIR dataset. Compared with various multimodal detectors, including GAFF (Zhang et al. 2021), SMPD (Li et al. 2023), CFT (Qingyun, Dapeng, and Zhaokui 2021), ICAFusion (Shen et al. 2024), UniRGB (Yuan et al. 2024), MFPT (Zhu et al. 2023), LRAF-Net (Fu et al. 2024), and DAMSDet (Guo et al. 2025), our detector achieves noticeably better results in mAP_{75}, surpassing DAMSDet by 0.7% in this metric. Although IGIANet is slightly lower than DAMSDet in mAP_{50}, it still significantly outperforms LRAF-Net by 4.5% in mAP_{50}, and achieves a higher overall mAP, exceeding DAMSDet by 0.1%. These results indicate that our detector maintains superior performance even on non-UAV multimodal datasets.

Method	**Type**	mAP_{50}	mAP_{75}	mAP
YOLOv5	IR	80.1	–	42.4
YOLOv5	RGB	67.8	25.9	31.8
DINO	IR	80.6	42.7	44.8
DINO	RGB	70.9	25.9	–
GAFF	IR+RGB	72.9	32.9	36.6
SMPD	IR+RGB	73.6	–	–
CFT	IR+RGB	78.7	35.5	40.2
ICAFusion	IR+RGB	79.2	36.9	41.4
MFPT	IR+RGB	80.0	–	–
LRAF-Net	IR+RGB	80.5	–	42.8
UniRGB-IR	IR+RGB	83.9	40.1	43.8
DAMSDet	IR+RGB	**86.6**	<u>48.1</u>	<u>49.3</u>
Ours	IR+RGB	<u>85.0</u>	**48.9**	**49.4**

Table 3: Comparison on the FLIR dataset.

Ablation Study

Ablation Study on Each Component. As shown in Table 4, we conduct a detailed ablation study by decomposing the proposed IGIANet and evaluating the relative effectiveness of each component. Compared to the baseline, introducing the IFMM, FG-CMDEM, and IADDF modules improves the mAP by 0.3%, 1.6%, and 0.7%, respectively. This indicates that FG-CMDEM effectively captures and enhances the differential information between the two modalities, achieving complementary feature enhancement. In addition, incorporating the OLIM module further enhances the representational capacity of IFMM, increasing mAP from 45.8% to 46.2%, demonstrating that illumination-aware weighting can effectively modulate frequency features. Furthermore, by integrating the IADDF module, the mAP improves from 48.8% to 49.3%, indicating that the proposed method enables effective fusion of the two modalities.

Study on the Effectiveness of the Hyperparameter β To investigate the impact of the parameter β on the range of illumination weights $\left[0.5 - \frac{\beta}{2},\ 0.5 + \frac{\beta}{2}\right]$, we conduct an ablation study by varying the value of β. As shown in Figure 4, the best detection performance is achieved when $\beta = 0.3$. However, when β exceeds 0.7, the detection performance degrades significantly.

OLIM	IFMM	DEM	IADDF	mAP_{50}	mAP
-	-	-	-	81.2	45.3
-	✓	-	-	81.7	45.8
✓	✓	-	-	82.5	46.2
-	-	✓	-	82.6	46.9
-	-	-	✓	81.8	46
-	✓	✓	-	83.7	48.1
✓	✓	✓	-	84.2	48.8
✓	✓	✓	✓	85.0	49.4

Table 4: Ablation study of each proposed module on the FLIR dataset, where DEM denotes FG-CMDEM, "–" indicates module disabled; "✓" indicates enabled.

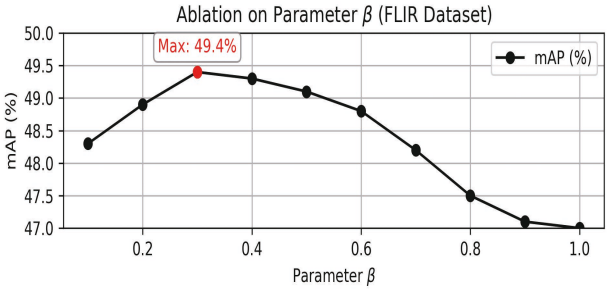

Figure 4: Ablation on the parameter β on the FLIR Dataset.

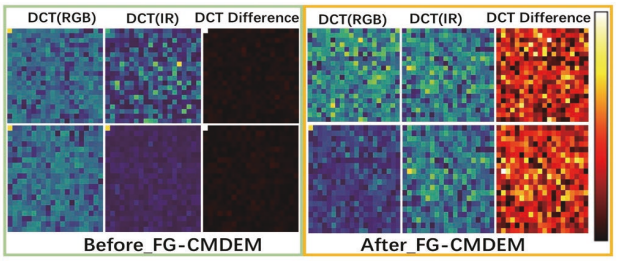

Figure 5: Visualization of the frequency maps before and after applying FG-CMDEM. The module selectively enhances cross-modal frequency differences.

Figure 6: Visualizing the illumination weight in Equation 4 reveals that different weights are dynamically assigned to different lighting conditions.

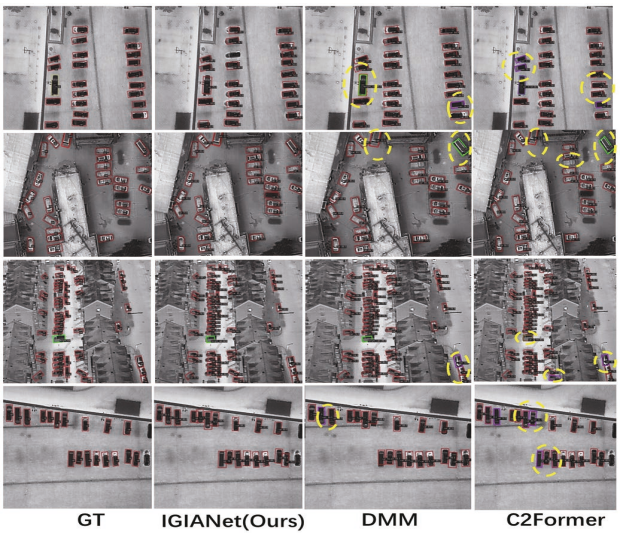

Figure 7: Comparison of different detection results, where the yellow dashed line represents detection errors.

Visualization Results Comparison. We provide a visual comparison to demonstrate the effectiveness of IGIANet. As shown in Figure 7, detection results from different detectors are displayed, with yellow ellipses marking incorrect detections. Previous methods are more prone to errors, especially in scenes with multiple similar objects. In contrast, IGIANet captures more discriminative information from both RGB and IR modalities, enabling more accurate and robust detection. As shown in Figure 5, we compare frequency differences before and after applying FG-CMDEM, observing a significantly greater distinction between RGB and IR modalities after. By computing cross-modal frequency differences, FG-CMDEM selectively enhances modality-specific information and effectively facilitates the integration of complementary features across modalities. As shown in Figure 6, our module perceives varying illumination conditions and dynamically assigns weights W_{illum}. The visualization shows that regions with high-quality illumination in the RGB image receive higher weights, enabling effective feature complementation.

Conclusion

In this paper, we propose an Illumination Guided Implicit Alignment Network (IGIANet) for robust UAV-based RGB-IR object detection. The proposed method effectively addresses the weak alignment problem through implicit alignment and dynamic convolution. In addition, under the guidance of illumination-aware weights, it performs spectral modulation on RGB-IR features and applies frequency-based differential enhancement, effectively leveraging both modality-specific discrepancies and cross-modality complementary information. Extensive experiments demonstrate the effectiveness of IGIANet in UAV-based RGB-IR object detection. In the future, we plan to extend our method to other multimodal tasks.

Acknowledgments

This work was supported in part by the Project of China-Mozambique "Belt and Road" Joint Laboratory on Smart Agriculture under Grant 2024YFE0214000, in part by the National Natural Science Foundation of China under Grant 62272419, and Grant 62402449.

References

Chen, N.; Xie, J.; Nie, J.; Cao, J.; Shao, Z.; and Pang, Y. 2023. Attentive alignment network for multispectral pedestrian detection. In *Proceedings of the 31st ACM international conference on multimedia*, 3787–3795.

Chen, X.; Jin, S.; Zhao, L.; Yang, C.; Zhang, D.; Wang, X.; He, X.; Wang, H.; Chen, Z.; and Zheng, Z. 2025a. Mask-Guided Frequency Feature Fusion for Visible–Infrared Remote Sensing Object Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–15.

Chen, X.; Liu, L.; and Tan, X. 2021. Robust pedestrian detection based on multi-spectral image fusion and convolutional neural networks. *Electronics*, 11(1): 1.

Chen, X.; Yang, C.; Mo, J.; Sun, Y.; Karmouni, H.; Jiang, Y.; and Zheng, Z. 2024. CSPNeXt: A new efficient token hybrid backbone. *Engineering Applications of Artificial Intelligence*, 132: 107886.

Chen, X.; Yang, C.; Mo, J.; Sun, Y.; Zhao, L.; Chen, H.; and Zheng, Z. 2025b. A Cross-domain Feature Fusion Network for Nighttime Drone-view Object Detection. *Pattern Recognition*, 112635.

Cheng, X.; Geng, K.; Wang, Z.; Wang, J.; Sun, Y.; and Ding, P. 2023. SLBAF-Net: Super-Lightweight bimodal adaptive fusion network for UAV detection in low recognition environment. *Multimedia Tools and Applications*, 82(30): 47773–47792.

Contributors, M. 2022. MMYOLO: OpenMMLab YOLO series toolbox and benchmark. https://github.com/open-mmlab/mmyolo.

Ding, J.; Xue, N.; Long, Y.; Xia, G.-S.; and Lu, Q. 2019. Learning RoI transformer for oriented object detection in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2849–2858.

Fu, H.; Wang, S.; Duan, P.; Xiao, C.; Dian, R.; Li, S.; and Li, Z. 2024. LRAF-Net: Long-Range Attention Fusion Network for Visible–Infrared Object Detection. *IEEE Transactions on Neural Networks and Learning Systems*, 35(10): 13232–13245.

Fu, T.; Li, Y.; Ye, X.; Tan, X.; Sun, H.; Shen, F.; and Ding, E. 2021. Lifting the veil of frequency in joint segmentation and depth estimation. In *Proceedings of the 29th ACM International Conference on Multimedia*, 944–952.

Guo, J.; Gao, C.; Liu, F.; Meng, D.; and Gao, X. 2025. DAMSDet: Dynamic adaptive multispectral detection transformer with competitive query selection and adaptive feature fusion. In *Proceedings of the European Conference on Computer Vision*, 464–481. Springer.

Han, J.; Ding, J.; Li, J.; and Xia, G.-S. 2021. Align deep features for oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–11.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141.

Jia, X.; Zhu, C.; Li, M.; Tang, W.; and Zhou, W. 2021. LLVIP: A Visible-Infrared Paired Dataset for Low-Light Vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3496–3504.

Li, Q.; Zhang, C.; Hu, Q.; Fu, H.; and Zhu, P. 2022a. Confidence-aware fusion using dempster-shafer theory for multispectral pedestrian detection. *IEEE Transactions on Multimedia*, 25: 3420–3431.

Li, Q.; Zhang, C.; Hu, Q.; Zhu, P.; Fu, H.; and Chen, L. 2023. Stabilizing multispectral pedestrian detection with evidential hybrid fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4): 3017–3029.

Li, Y.; Yu, A. W.; Meng, T.; Caine, B.; Ngiam, J.; Peng, D.; Shen, J.; Lu, Y.; Zhou, D.; Le, Q. V.; et al. 2022b. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17182–17191.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2980–2988.

Liu, K.; Li, T.; and Peng, D. 2025. Aerial Image Object Detection Based on RGB-Infrared Multibranch Progressive Fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–14.

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

Qingyun, F.; and Zhaokui, W. 2022. Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. *Pattern Recognition*, 130: 108786.

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

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6): 1137–1149.

Robbins, H.; and Monro, S. 1951. A stochastic approximation method. *The annals of mathematical statistics*, 400–407.

Shakhatreh, H.; Sawalmeh, A. H.; Al-Fuqaha, A.; Dou, Z.; Almaita, E.; Khalil, I.; Othman, N. S.; Khreishah, A.; and Guizani, M. 2019. Unmanned aerial vehicles (UAVs): A survey on civil applications and key research challenges. *IEEE Access*, 7: 48572–48634.

Shen, J.; Chen, Y.; Liu, Y.; Zuo, X.; Fan, H.; and Yang, W. 2024. ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognition*, 145: 109913.

Sun, Y.; Cao, B.; Zhu, P.; and Hu, Q. 2022. Drone-Based RGB-Infrared Cross-Modality Vehicle Detection via Uncertainty-Aware Learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10): 6700–6713.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2020. FCOS: A simple and strong anchor-free object detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4): 1922–1933.

Yang, X.; Yan, J.; Feng, Z.; and He, T. 2021. R3Det: Refined single-stage detector with feature refinement for rotating object. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 3163–3171.

Yao, H.; Qin, R.; and Chen, X. 2019. Unmanned aerial vehicle for remote sensing applications—A review. *Remote Sensing*, 11(12): 1443.

Yuan, M.; Cui, B.; Zhao, T.; Wang, J.; Fu, S.; Yang, X.; and Wei, X. 2024. UniRGB-IR: A Unified Framework for Visible-Infrared Semantic Tasks via Adapter Tuning. *arXiv preprint arXiv:2404.17360*.

Yuan, M.; Wang, Y.; and Wei, X. 2022. Translation, scale and rotation: cross-modal alignment meets RGB-infrared vehicle detection. In *Proceedings of the European Conference on Computer Vision*, 509–525. Springer.

Yuan, M.; and Wei, X. 2024. C2Former: Calibrated and complementary transformer for rgb-infrared object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–12.

Zhang, H.; Fromont, E.; Lefevre, S.; and Avignon, B. 2020. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In *2020 IEEE International Conference on Image Processing*, 276–280. IEEE.

Zhang, H.; Fromont, E.; Lefèvre, S.; and Avignon, B. 2021. Guided attentive feature fusion for multispectral pedestrian detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 72–80.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022. DINO: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

Zhang, L.; Liu, Z.; Zhang, S.; Yang, X.; Qiao, H.; Huang, K.; and Hussain, A. 2019a. Cross-modality interactive attention network for multispectral pedestrian detection. *Information Fusion*, 50: 20–29.

Zhang, L.; Zhu, X.; Chen, X.; Yang, X.; Lei, Z.; and Liu, Z. 2019b. Weakly aligned cross-modal learning for multispectral pedestrian detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5127–5137.

Zhang, Y.; Yu, H.; He, Y.; Wang, X.; and Yang, W. 2023. Illumination-Guided RGBT Object Detection With Inter- and Intra-Modality Fusion. *IEEE Transactions on Instrumentation and Measurement*, 72: 1–13.

Zhou, K.; Chen, L.; and Cao, X. 2020. Improving multispectral pedestrian detection by addressing modality imbalance problems. In *Proceedings of the European Conference on Computer Vision*, 787–803. Springer.

Zhou, M.; Li, T.; Qiao, C.; Xie, D.; Wang, G.; Ruan, N.; Mei, L.; Yang, Y.; and Shen, H. T. 2025. DMM: Disparity-guided multispectral mamba for oriented object detection in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 1–1.

Zhu, Y.; Sun, X.; Wang, M.; and Huang, H. 2023. Multi-modal feature pyramid transformer for rgb-infrared object detection. *IEEE Transactions on Intelligent Transportation Systems*, 24(9): 9984–9995.

Zongzhen, L.; Hui, L.; Zhixing, W.; Yuxing, W.; Haorui, Z.; and Jianlin, Z. 2025. Cross-modal Offset-guided Dynamic Alignment and Fusion for Weakly Aligned UAV Object Detection. *arXiv preprint arXiv:2506.16737*.