Regular Article

Efficient Multimodal Feature Refinement via Adaptive RGB-IR Interaction for Robust Drone Detection and Classification

Van-Phuc Hoang[1], Thien Huynh-The[2], Thanh-Dat Tran[2], Xuan Tinh Hoang[1]

[1] Institute of System Integration, Le Quy Don Technical University, Hanoi, Vietnam
[2] Department of Electronics and Information Engineering, HCM City University of Technology and Engineering (HCM-UTE)

Correspondence: Thien Huynh-The, thienht@hcmute.edu.vn
Communication: received 18 March 2026, revised 24 April 2026, accepted 30 April 2026
Online publication: 19 May 2026, Digital Object Identifier: 10.21553/rev-jec.451

Abstract– **The rapid proliferation of unmanned aerial vehicles (UAVs) has intensified the need for robust surveillance systems capable of distinguishing drones from biological entities like birds in unpredictable environments. While multispectral vision provides a resilient alternative to uni-modal sensors under adverse weather and lighting, existing architectures often struggle with cross-modal feature alignment and noise-induced spatial distortions. This paper proposes Multispectral Attention Context and Receptive-field Network (MACR-Net), an ultra-lightweight multimodal framework designed for high-precision drone detection. MACR-Net introduces a Global-Local Cross-Scale Interaction (GLCI) module to capture multi-scale semantic context and a Multimodal Spatial Cross-Perception (MSCP) mechanism to adaptively fuse RGB-IR streams while preserving target-specific thermal and structural signatures. Furthermore, we design an improved hybrid neck integrating Coordinate-Aware Attention (CAA) and Receptive Field Deformable (RFD) modules to anchor precise spatial coordinates and mitigate geometric distortions. Experimental results on the benchmark Multimodal Drone Detection Dataset demonstrate that MACR-Net outperforms state-of-the-art models, achieving a peak mAP_{50} of 91.13% and a significant mAP_{50-95} of 65.77%. Remarkably, the architecture maintains an extremely compact footprint with only 2.77M parameters and 0.77 GFLOPs, establishing an optimal balance between superior detection robustness and real-time feasibility for resource-constrained edge deployment.**

Keywords– **Deep learning, drone recognition, multimodal object detection, RGB-IR fusion.**

1 Introduction

The rapid evolution of unmanned aerial vehicles (UAVs) delivers numerous benefits across civil and industrial sectors, such as remote sensing, precision agriculture, disaster management, and infrastructure inspection [1–4]. However, their malicious deployment creates severe vulnerabilities for aviation security and public safety [5–7], making the establishment of robust detection architectures in unpredictable real-world environments an urgent research priority [8]. Traditional detection systems relying on visible-spectrum cameras alone frequently encounter severe limitations under adverse conditions such as low illumination, dense fog, or thick vegetation, where standard optical sensors lose spatial resolution and fail to separate micro aerial devices from biological entities such as birds [9–11].

To overcome these optical bottlenecks, multispectral fusion combining visible RGB and infrared IR sensors has emerged to exploit the complementary physical properties of each spectrum. RGB data provide detailed geometric textures and color profiles, while IR images capture thermal radiation to expose targets under visual camouflage or severe degradation conditions [12, 13]. Despite this synergy, current multispectral detection models still face significant barriers in cross-spectral alignment and feature fusion, particularly under fierce background noise or atmospheric distortion, where phase deviations and thermal overlaps obscure target features and yield high missed-detection rates and poor localization precision [14–17].

These challenges stem from architectural limitations that have persisted throughout the evolution of detection networks. Early two-stage methods such as Mask R-CNN [18] achieve strong spatial localization but are too computationally heavy for real-time detection of agile aerial vehicles. RetinaNet [19] pioneered single-stage detection via focal loss, yet its successive convolutions erode fine spatial detail, causing small aerial targets to dissolve into environmental noise. The YOLO family from v5 to v12 [20, 21] has continuously advanced multi-scale feature preservation and NMS-free inference, but its uni-modal design leaves even the latest versions blind under low-light or foggy conditions and unable to reliably distinguish UAVs from birds of similar wingspan. Furthermore, advanced Vision Transformer (ViT) architectures such as RT-DETR [22] achieve effective global context perception but require large parameter volumes and consume high computational power; this characteristic renders these models impractical for edge hardware.

The limitations of uni-modal systems have driven development of multispectral architectures. TSFADet [23] effectively exploits RGB–IR complementarity but at the cost of high computational volume and spatial misalignment risk. YOLOv10N-PRD [24] achieves

competitive frame rates but degrades under dense background noise. DVIF-Net [25] provides stable signal fusion yet produces imprecise bounding boxes in low-contrast scenes. GAANet [26] addresses spectral misalignment through graph aggregation but proves slow and memory-intensive on high-resolution inputs. MDFN [27] links multi-dimensional fusion channels but unintentionally amplifies thermal noise from cloudy environments, obscuring small targets. CADDN [28] reduces downsampling information loss but sacrifices spatial resolution for distant small targets. Most recently, the adaptive cross-modal fusion of Chen *et al.* [29] balances visible-thermal streams but incurs significant inference latency, degrading detection when UAVs perform rapid maneuvers or hide behind dense vegetation. These persistent limitations across spectral alignment, noise suppression, resolution preservation, and real-time performance collectively motivate the construction of a dedicated multispectral refinement architecture proposed in this work.

To resolve these challenges, this research proposes MACR-Net, an advanced multispectral object detection framework designed for accurate drone and bird identification in complex outdoor environments. The proposed architecture introduces a cross-spectral perception and multi-level context integration mechanism to fuse visible and infrared features, enabling the selective extraction of thermal and optical signals while preserving target information without amplifying background noise. In addition, a hybrid neck structure that combines coordinate attention and deformable convolution (conv) is employed to enhance multi-receptive-field feature extraction, improve spatial coordinate alignment, and adapt to geometric distortions caused by severe camera angles or adverse weather conditions. The conceptual novelty of the MACR-Net architecture resides not in the creation of a new solitary mathematical operator, but in the symbiotic design philosophy specific to the multispectral target detection domain. In complex outdoor environments, visual degradation and thermal radiation noise cause standard models to collapse. Distinct from the GAANet architecture, which relies on heavy graph structures prone to semantic drift, or the DVIF-Net framework, which frequently endures severe spatial phase deviations under fog conditions, the proposed design delivers specific conceptual breakthroughs to resolve these exact bottlenecks. Experimental results demonstrate the superiority of MACR-Net, achieving 91.13% mAP_{50} and 65.77% mAP_{50-95} while maintaining an ultra-lightweight design with 2.77M parameters and 0.77 GFLOPs, which makes the model suitable for real-time deployment on edge devices.

2 Methodology

2.1 MACR-Net: Multispectral Attention Context and Receptive-field Network

In this paper, this research proposes MACR-Net (**M**ultispectral **A**ttention **C**ontext and **R**eceptive-field **Net**work), an architecture with emphasis on adaptive multimodal feature refinement through the close coordination of specialized modules. Instead of passive data fusion, MACR-Net establishes an active feature interaction mechanism through a hierarchical structure as shown in Figure 1. The architecture commences with a dual-stream Backbone in Figure 1a to extract multi-scale features and narrow the spectral gap. These representations flow into a specialized Neck where coordinate-aware attention and deformable refinement modules anchor precise localization and adapt to geometric distortions. Finally, the Detection Head performs multi-scale regression and classification. This rigorous coordination from backbone to head achieves an optimal balance between superior accuracy and real-time operational capability.

Global-Local Cross-Scale Interaction (GLCI): Extreme variation in object scale represents the most significant challenge in drone detection from aerial imagery. To resolve this small-target problem, this research proposes the GLCI module as shown in Figure 1b. Unlike traditional static pooling mechanisms, GLCI establishes a multi-scale feature interaction process through three parallel branches to optimize the simultaneous extraction of fine-grained features and semantic context. First, the global context branch performs channel-wise feature calibration. For an input feature $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, a 1×1 conv layer reduces the channel dimension to $C' = C/2$ to create a condensed representation. This mechanism leverages the global average pooling operation $\mathcal{G}_{\text{avg}}$ to compress spatial information directly into a global context vector of size $\mathbb{R}^{1\times 1\times C'}$. This representation passes through a calibration conv layer to learn spectral correlations and generate a dynamic weight map. The output of this branch follows an element-wise multiplication

$$\mathbf{X}_{\text{global}} = \mathbf{X} \odot \sigma(\mathcal{C}_{1\times 1}\left(\mathcal{G}_{\text{avg}}\left(\mathbf{X}\right)\right)), \tag{1}$$

where σ denotes the Sigmoid activation function, $\odot$ represents element-wise multiplication, and $\mathcal{C}_{n\times n}$ signifies a regular convolution operation with an $n\times n$ kernel. This function establishes a channel attention mechanism which allows the neural network to prioritize feature channels with critical information based on wide-area context. Consequently, the model rebalances features to adapt effectively to changing background environments.

Second, the local difference branch acts as a detail-oriented filter to emphasize high-frequency structures such as edges and small drone patterns while suppressing low-frequency background noise. The feature map $\mathbf{X}$ is first processed by a depthwise conv followed by a sigmoid activation to generate an attenuation map. The differential operation is then defined as

$$\mathbf{X}_{\text{diff}} = \mathbf{X} - \left(\mathbf{X} \odot \sigma\left(\mathcal{DW}_{3\times 3}\left(\mathbf{X}\right)\right)\right), \tag{2}$$

where $\mathcal{DW}(\cdot)$ denotes the depthwise conv operation. The subtraction effectively removes the smoothed low-frequency component, functioning as a dynamic high-pass filter that highlights weak edge structures of dis-

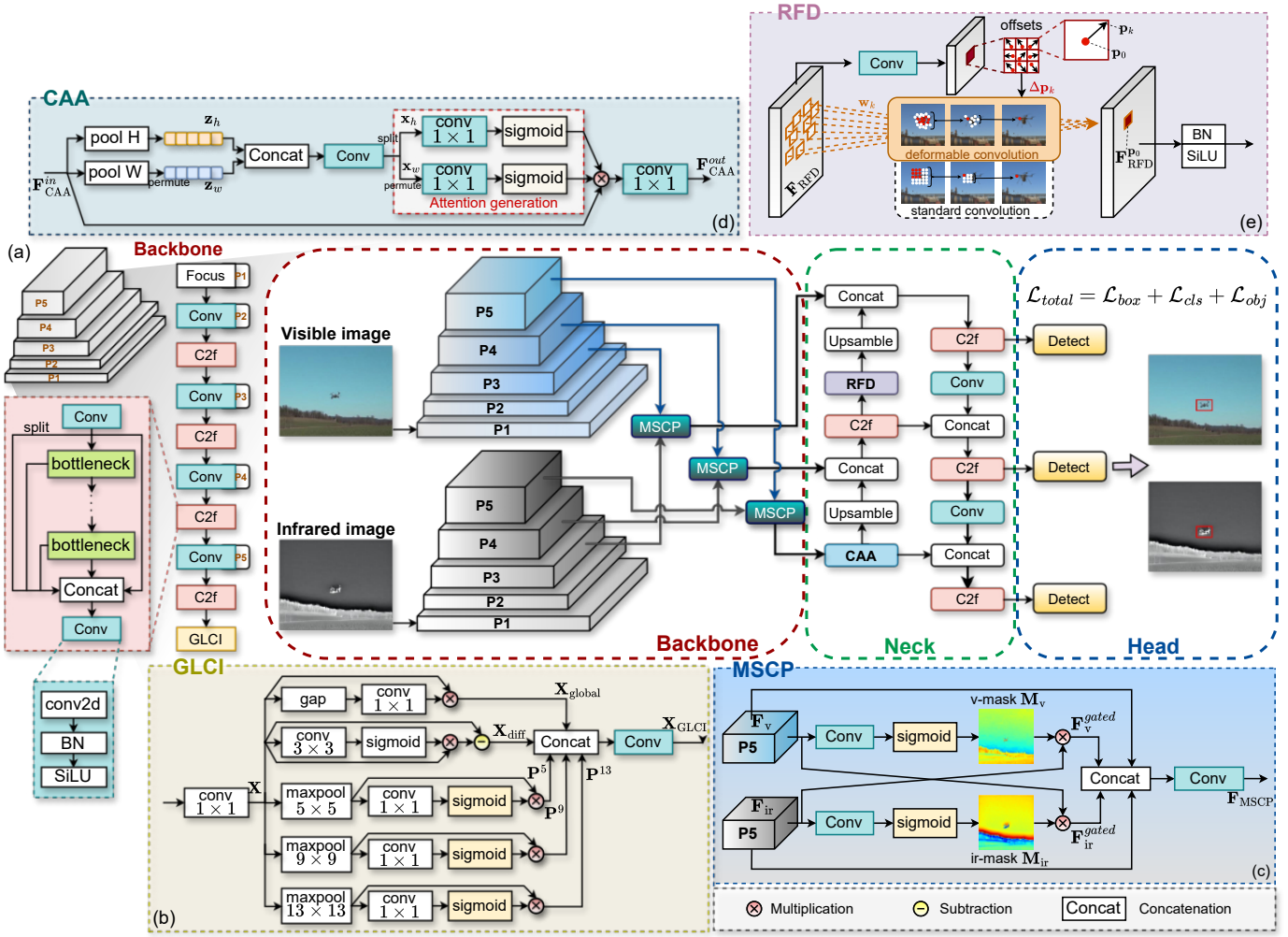

Figure 1. Architecture of the proposed MACR-Net, including (a) detailed stream-wise backbone, (b) GLCI module, (c) MSCP module, and an improved neck with the CAA (d) and RFD (e) modules.

tant targets while suppressing homogeneous regions such as sky or clouds.

Third, the attentive multi-scale pooling branch enhances feature extraction across different receptive fields. Unlike traditional pooling methods that aggregate all information within the kernel [30], our design introduces an independent spatial attention module for each max pooling layer $\mathcal{P}^k_{\max}$ with kernel sizes $k \in \{5, 9, 13\}$. The selection of these specific kernel sizes derives from a statistical analysis through the K-means cluster method on the target boundary box scale distribution within the dataset. Specifically, the 5×5 kernel maps to the cluster of micro targets at long distances. This size provides a proper local receptive field to encapsulate micro-features without the accumulation of heavy background noise. The 9×9 kernel is configured to encompass the structural morphology of medium-scale aerial vehicles. Meanwhile, the 13×13 kernel creates a broad receptive field to absorb the complex thermal distribution of larger biological entities, such as birds, along with their environmental context. This broad perception assists the network to filter out avian entities. For the pooled feature $\mathcal{P}^k_{\max}(\mathbf{X})$, an attention map is generated using a 1×1 conv and a sigmoid function

$$\mathbf{P}^k = \mathcal{P}^k_{\max}(\mathbf{X}) \odot \sigma\left(\mathcal{C}_{1\times1}\left(\mathcal{P}^k_{\max}(\mathbf{X})\right)\right), \quad (3)$$

where $\mathcal{P}^k_{\max}(\cdot)$ denotes the max pooling operation with kernel size $k \times k$. This mechanism allows the network to emphasize informative regions while suppressing irrelevant background signals at each scale.

Finally, the outputs of all branches–global context, local difference, and attentive multi-scale pooling–are fused through a conv layer

$$\mathbf{X}_{\text{GLCI}} = \mathcal{F}\left(\texttt{concat}\left[\mathbf{X}_{\text{global}}, \mathbf{X}_{\text{diff}}, \mathbf{P}^5, \mathbf{P}^9, \mathbf{P}^{13}\right]\right), \quad (4)$$

where $\texttt{concat}[\cdot]$ denotes channel-wise concatenation and $\mathcal{F}$ represents a composite block including a conv layer, Batch Normalization (BN), and a Sigmoid Linear Unit (SiLU) activation. This fusion integrates multi-branch features into a compact, discriminative representation for subsequent detection.

Multimodal Spatial Cross-Perception: In dual-sensor object detection systems, the spectral modality gap between visible and infrared images often causes information interference upon fusion. To thoroughly resolve this issue from the initial stage, we propose

the MSCP module shown in Figure 1(c). This module establishes a cross-modal interaction mechanism, wherein the spatial feature of one modality calibrates and guides the opposite modality. With the integration of the P3, P4, and P5 blocks, consider two input feature streams from the P5 blocks: the visible stream $\mathbf{F}_{\mathrm{v}}$ and the infrared stream $\mathbf{F}_{\mathrm{ir}}$. First, the system initializes two independent spatial mask maps through a mask-initialization conv layer and a sigmoid activation function

$$\mathbf{M}_{\mathrm{v}} = \sigma\left(\mathcal{F}_{\mathrm{v}}\left(\mathbf{F}_{\mathrm{v}}\right)\right), \quad \mathbf{M}_{\mathrm{ir}} = \sigma\left(\mathcal{F}_{\mathrm{ir}}\left(\mathbf{F}_{\mathrm{ir}}\right)\right), \tag{5}$$

the masks $\mathbf{M}_{\mathrm{v}}$ and $\mathbf{M}_{\mathrm{ir}}$ act as attention filters. These filters highlight regions with high activation levels characteristic of each spectrum, such as detailed structures in RGB images and thermal signals in IR images. The breakthrough of MSCP lies in the cross-calibration mechanism. This module transcends passive data combination to establish a bidirectional cross-calibration mechanism. Within this mechanism, thermal radiation from the infrared stream rescues obscured structural boundaries in the visible stream, while sharp geometric textures from the visible spectrum reshape blurred thermal signals. This mutual feature calibration represents a fundamental paradigm shift from conventional fusion methods. Instead of self-attention, the visible mask filters the infrared stream, and conversely, the infrared mask filters the visible stream. The point-wise interaction operation occurs as follows

$$\mathbf{F}_{\mathrm{v}}^{gated} = \mathbf{F}_{\mathrm{v}} \odot \mathbf{M}_{\mathrm{ir}}, \quad \mathbf{F}_{\mathrm{ir}}^{gated} = \mathbf{F}_{\mathrm{ir}} \odot \mathbf{M}_{\mathrm{v}}. \tag{6}$$

This mechanism ensures that strong thermal signals from $\mathbf{M}_{\mathrm{ir}}$ rescue blurred details in the $\mathbf{F}_{\mathrm{v}}$ stream, such as a drone covered by fog, while sharp structural boundaries from $\mathbf{M}_{\mathrm{v}}$ reshape the typically blurred thermal signals of $\mathbf{F}_{\mathrm{ir}}$. This cross-calibration mechanism specifically addresses spatial misalignment and asynchronous capture delays. Standard networks collapse under these conditions due to rigid point-to-point aggregation. However, the sigmoid activation establishes a continuous gradient of attention weights rather than a rigid binary threshold. This mathematical property generates a spatial buffer zone around the primary thermal signature. If an aerial vehicle causes the thermal peak to shift by several pixels, the gradual weight decay ensures that displaced geometric features still receive sufficient activation. Through the neutralization of these coordinate discrepancies at the initial stage, the MSCP module prevents the downstream propagation of alignment errors and guarantees a coherent feature representation for subsequent computation layers.

Finally, to preserve the integrity of the original data while the network integrates the cross-calibrated streams, MSCP concatenates the four feature components and passes these components through a final fusion conv layer

$$\mathbf{F}_{\mathrm{MSCP}} = \mathcal{F}\left(\texttt{concat}\left[\mathbf{F}_{\mathrm{v}}, \mathbf{F}_{\mathrm{ir}}^{gated}, \mathbf{F}_{\mathrm{v}}^{gated}, \mathbf{F}_{\mathrm{ir}}\right]\right). \tag{7}$$

This architecture eliminates the spectral phase shift and provides an information-rich feature foundation for subsequent context computation layers.

Attentive Receptive-field Hybrid Neck: Subsequent to the MSCP feature fusion, the Neck refines representations before the prediction head. Successive backbone convs and downsampling operations severely degrade spatial coordinates and distort the geometries of tiny UAVs. To resolve this, we reconstruct the Neck using two sequential refinement modules. First, the Coordinate-Aware Attentive (CAA) module performs spatial factorization to anchor precise target coordinates, as illustrated in Figure 1(d). Next, the Receptive Field Deformable (RFD) module adapts this coordinate-embedded stream to complex geometric distortions, as shown in Figure 1(e). The CAA module mitigates spatial degradation through the direct embedding of coordinates into the channel attention mechanism. During coordinate information encoding, the input $\mathbf{F}_{\mathrm{CAA}}^{in}$ undergoes independent 1D spatial pooling along the vertical (pool H) and horizontal (pool W) directions to extract two directional context vectors

$$\mathbf{z}^{h} \in \mathbb{R}^{H\times 1\times C}, \quad \mathbf{z}^{w} \in \mathbb{R}^{1\times W\times C}. \tag{8}$$

A permutation on $\mathbf{z}^{w}$ precedes concatenation with $\mathbf{z}^{h}$. A shared conv layer compresses channels and promotes cross-interaction to generate an intermediate representation $\mathbf{f}$. In the attention generation stage, the system splits $\mathbf{f}$ into two streams $\mathbf{x}_h$ and $\mathbf{x}_w$. Each stream passes through a 1×1 conv and a sigmoid function to initialize spatial attention maps. The module multiplies the original input $\mathbf{F}_{\mathrm{CAA}}^{in}$ with these maps, and a final 1×1 conv aggregates cross-channel features to produce the ultimate output

$$\mathbf{F}_{\mathrm{CAA}}^{out} = \mathcal{C}_{1\times 1}(\mathbf{F}_{\mathrm{CAA}}^{in} \odot \sigma(\mathcal{C}_{1\times 1}(\mathbf{x}_h)) \odot \sigma(\mathcal{C}_{1\times 1}(\mathbf{x}_w))). \tag{9}$$

The $\mathbf{F}_{\mathrm{CAA}}^{out}$ feature carries highly precise coordinate information, whereas UAV shapes frequently vary due to rotation angles and flight perspectives. The RFD module addresses the limitations of a fixed convolution grid through a dynamic sampling mechanism. More importantly, the joint design of the CAA and RFD modules resolves a critical weakness of standard deformable convolutions. Under micro-target conditions, conventional dynamic sampling grids tend to disperse into background noise, which leads to unstable spatial localization. The proposed architecture enforces strict coordinate anchoring via the CAA module before the RFD module performs geometric adaptation. This coordinated mechanism ensures that dynamic sampling remains tightly aligned with target structures rather than drifting into irrelevant regions. Consequently, the model achieves highly reliable spatial precision for aerial targets at extreme distances. Such structural synergy establishes a clear performance advantage, where MACR-Net transcends simple component integration and forms a coherent solution for practical aerial observation tasks. A conv layer applies to the input to predict a 2D offset map $\Delta\mathbf{p}_k$ for each pixel. Instead of sampling on a rigid square grid, the neural network automatically adjusts the sampling points to closely adhere to the actual shape of the drone. The output value at position

$\mathbf{p}_0$ is calculated as follows

$$\mathbf{F}_{\mathrm{RFD}}^{\mathbf{p}_0} = \sum_{\mathbf{p}_k \in \mathcal{R}} \mathbf{w}_k \, \mathbf{F}_{\mathrm{RFD}}(\mathbf{p}_0 + \mathbf{p}_k + \Delta \mathbf{p}_k), \tag{10}$$

where $\mathcal{R}$ denotes the standard receptive grid, $\mathbf{w}_k$ signifies the conv weight, and $\Delta \mathbf{p}_k$ represents the dynamic offset learned from data to determine input feature map $\mathbf{x}$ at sampled coordinates via bilinear interpolation. Finally, the spatial stream passes through BN và SiLU activation to stabilize data distribution, which produces an invariant representation to bound tiny drone targets.

2.2 Loss Function

To optimize the network for tiny UAV detection, the objective function $\mathcal{L}_{total}$ integrates three distinct components: localization loss $\mathcal{L}_{box}$, classification loss $\mathcal{L}_{cls}$, and confidence loss $\mathcal{L}_{obj}$. The total loss is defined as a weighted sum of these individual terms

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{box} + \lambda_2 \mathcal{L}_{cls} + \lambda_3 \mathcal{L}_{obj} \tag{11}$$

where $\lambda_1, \lambda_2, \lambda_3$ represent the trade-off coefficients. Through empirical optimization, the default configuration establishes $\lambda_1 = 0.05$, $\lambda_2 = 0.5$, and $\lambda_3 = 1.0$. For the localization term $\mathcal{L}_{box}$, we employ the CIoU to accelerate bounding box regression, especially for objects with extreme scale variations. This metric accounts for the overlap area, central point distance, and aspect ratio to ensure precise spatial alignment. The classification and confidence terms utilize Binary Cross-Entropy to distinguish drones from complex backgrounds effectively. This composite loss strategy guides the model to achieve high localization accuracy and a low false-positive rate in challenging flight environments.

3 Simulations and Results

3.1 Dataset and Training Configurations

To evaluate the performance of the MACR-Net architecture, experiments are conducted on the Multimodal Drone Detection Dataset provided by the IEEE Signal Processing Society. This dataset consists of simultaneous RGB-IR image pairs with a 320 × 256 resolution and focuses on drone and bird categories. The split comprises 45.000 training pairs, 6.500 validation pairs, and 13.000 test samples. Targets appear in challenging environments such as mountainous terrain or dense forests under degraded visibility from cloud and fog. Scenarios also feature distant targets and complex flocking. Furthermore, the data incorporates signal distortions such as speckle noise, additive white Gaussian noise, and motion blur to evaluate model robustness. Figure 2 illustrates example scenes of these environmental conditions.

Training is performed on a workstation equipped with a 3.00 GHz CPU, 16 GB RAM, and an NVIDIA GeForce RTX 3060Ti GPU. MACR-Net is trained for 300 epochs with a batch size of 32, using the AdamW optimizer with an initial learning rate of 0.01 and a weight decay of 0.0005. To ensure fairness, all comparative experiments are conducted using the same test set. Object detection performance is evaluated using average precision (AP), mean average precision (mAP), precision (Pr), recall (Re), and the total number of trainable parameters. AP quantifies accuracy through the area under the precision–recall curve, while mAP represents the mean AP across categories. Specifically, mAP_{50} uses a 50% intersection over union (IoU) threshold, whereas mAP_{50-95} denotes the average AP over IoU thresholds from 50% to 95% with 5% increments.

Table I
Quantitative impact of multispectral fusion on detection accuracy.

Modality	Pr(%)	Re(%)	mAP_{50}(%)	mAP_{50-95}(%)
RGB	89.56	83.22	86.45	57.77
IR	87.12	85.23	88.04	58.47
RGB + IR	93.32	90.51	91.13	65.77

3.2 Results and Discussions

Modality Evaluation: To evaluate the contribution of each sensor type, the first experiment compares single-modality models (RGB-only or IR-only) with the proposed multispectral fusion method (RGB+IR). Based on the experimental data in Table II, a significant improvement in detection performance is observed when transitioning from single-modality systems to the proposed multispectral approach. Specifically, models using only RGB or IR data achieve mAP_{50} levels of 86.45% and 88.04%, respectively, which indicates certain limitations of each sensor modality when operating independently in cloudy or forested environments. Notably, with the integration of the MACR-Net fusion mechanism, the mAP_{50} index increases to 91.13%, which confirms superior object detection capability through information synergy between thermal and visual streams. Furthermore, the mAP_{50-95} index–representing localization accuracy and bounding box quality–records an impressive increase from below 59% in single-modality models to 65.77%. These results demonstrate that multispectral integration not only assists the system in detecting targets more easily under misty conditions but also provides detailed morphological features for absolute precision in drone localization, even against complex hilly terrains.

Ablation Study: In the second experiment, an ablation study is conducted on the multispectral dataset to evaluate the contribution of each proposed module, with results summarized in Table III. The baseline model, which uses the original architecture without optimization modules, only achieves an mAP_{50} of 83.27% and an mAP_{50-95} of 53.11%, indicating the limitations of conventional feature extractors in handling visual camouflage and noise under challenging environmental conditions. When the refinement module cluster in the neck (configuration $\mathcal{M}_3$, without CAA+RFD) is removed, the mAP_{50-95} drops significantly from 65.77% to 60.42%, demonstrating the critical role of coordinate attention and deformable convolution in improving bounding box localization for ultra-small drones. Similarly, removing the MSCP module (configuration $\mathcal{M}_2$)

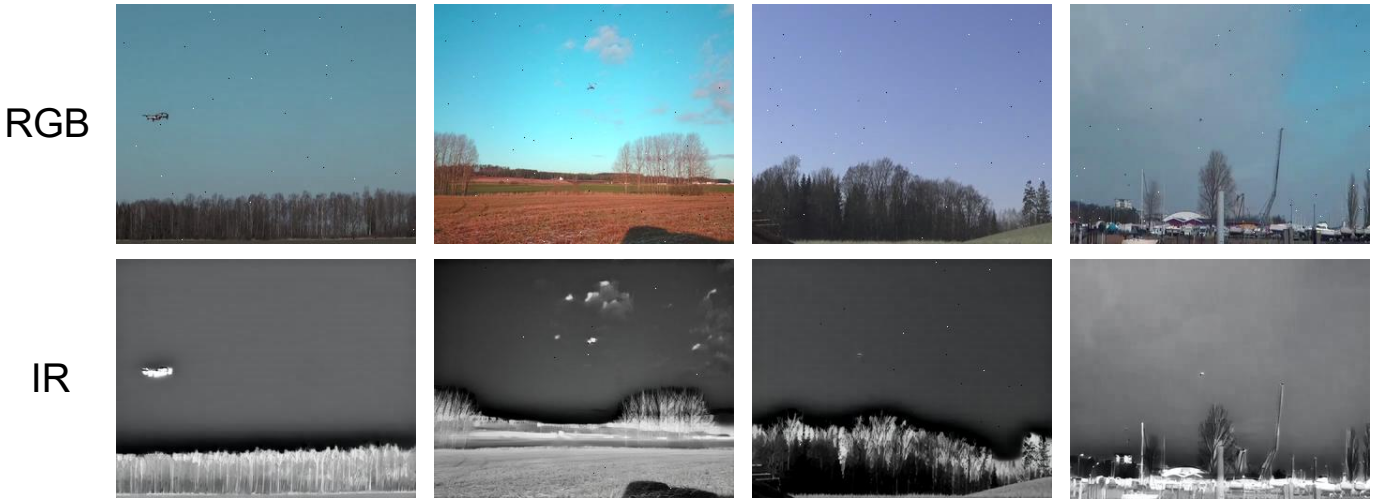

Figure 2. Representative visible-infrared image pairs from the benchmark dataset.

Table II
Sensitivity Analysis of the Localization Loss Coefficient (λ_1)

Loc (λ_1)	Cls (λ_2)	Obj (λ_3)	Pr (%)	Re (%)	mAP$_{50}$(%)	mAP$_{50-95}$(%)
0.03	0.5	1.0	90.12	85.97	89.65	64.21
0.04	0.5	1.0	91.56	87.11	90.32	64.98
0.05	**0.5**	**1.0**	**93.32**	**90.51**	**91.13**	**65.77**
0.06	0.5	1.0	92.10	89.45	90.54	65.12
0.07	0.5	1.0	91.05	88.90	89.88	64.55

Table III
Performance comparison of MACR-Net model configurations.

Model	Configurations			Pr	Re	mAP$_{50}$	mAP$_{50-95}$	Params
	GLCI	MSCP	CAA+RFD	(%)	(%)	(%)	(%)	(M)
Baseline	✗	✗	✗	85.12	81.45	83.27	53.11	0.93
$\mathcal{M}_1$	✗	✓	✓	91.56	88.62	90.18	62.88	1.14
$\mathcal{M}_2$	✓	✗	✓	90.11	87.25	89.34	61.55	1.37
$\mathcal{M}_3$	✓	✓	✗	89.45	86.30	88.15	60.42	2.51
MACR-Net	✓	✓	✓	**93.32**	**90.51**	**91.13**	**65.77**	**2.77**

reduces the mAP$_{50}$ to 89.34%, revealing the importance of effective cross-modal feature fusion between thermal and visible streams. Configuration $\mathcal{M}_1$ (without GLCI) also results in a performance decline, confirming the necessity of global context modeling at the backbone stage. Overall, integrating all three components enables MACR-Net to achieve substantial improvements of nearly 8% in mAP$_{50}$ and over 12% in mAP$_{50-95}$ compared to the baseline, while maintaining a lightweight design with only 2.77M parameters.

Subsequent to the module ablation, this research presents a sensitivity analysis on the localization loss coefficient λ_1 to evaluate architectural robustness, as reported in Table II. The experiment varies λ_1 from 0.03 to 0.07 while the configuration holds λ_2 and λ_3 constant at 0.5 and 1.0, respectively. The results demonstrate that the detection metrics experience minimal variance across all tested values. The model achieves peak performance at $\lambda_1 = 0.05$, where the network records 91.13% mAP$_{50}$ and 65.77% mAP$_{50-95}$. The marginal performance drop at extreme values proves that the composite loss strategy possesses strong hyperparameter tolerance. The stable detection accuracy across various parameter configurations validates the structural superiority of the proposed framework.

Method comparison: To assert the superiority of the proposed method, in the third experiment, MACR-Net is compared with a series of current state-of-the-art (SOTA) models is presented in Table IV. First, it is clear that traditional two-stage architectures such as Faster R-CNN [18] or RetinaNet [19] reveal significant bulkiness with massive capacities (from 58 to 145M parameters) and extremely large computational volumes, yet their mAP$_{50-95}$ performance only stagnates at 52-55% due to a lack of adaptability to complex environmental noise. Similarly, the advanced ViT model RT-DETR [22] requires 42M parameters and 136 GFLOPs, but this model encounters difficulties

Table IV
Performance comparison with state-of-the-art models.

Model	Pr (%)	Re (%)	mAP_{50} (%)	mAP_{50-95} (%)	Params (M)	FLOPs (G)	Latency (ms)
RetinaNet	84.42	80.15	81.21	52.77	145	99.9	119.36
Faster R-CNN	86.08	84.02	84.98	55.47	58.30	169.5	73
RT-DETR	86.48	83.57	85.71	56.23	42	136	9.4
YOLOv5n	85.29	84.62	86.67	56.89	2.50	7.2	**6.3**
YOLOv8n	88.37	86.16	88.68	60.61	3.01	8.1	8.1
YOLOv10n	89.86	87.56	89.54	61.62	**2.26**	6.5	6.8
YOLOv11n	90.29	88.33	90.85	62.14	2.58	6.3	6.5
YOLOv12n	90.79	89.23	90.82	63.15	2.51	6.6	6.7
YOLOv10N-PRD	92.54	89.62	90.05	62.15	4.7	12.8	28.45
DVIF-Net	90.87	**90.56**	90.24	62.38	2.49	8.2	22.73
MACR-Net	**93.32**	90.51	**91.13**	**65.77**	2.77	**0.77**	12.24

with micro-target distortion and achieves a mAP_{50-95} of 56.23%. Turning to the most advanced one-stage detection models in the YOLO family (from YOLOv5n to YOLOv12n), although they are excellently optimized in terms of capacity (fluctuating around 2.5M parameters), the highest recorded performance only stops at an mAP_{50} of 90.82% and an mAP_{50-95} of 63.15% (YOLOv12n). Even when compared with recently proposed specialized architectures such as YOLOv10N-PRD [24] or DVIF-Net [25], MACR-Net still establishes a new performance limit. Although DVIF-Net has a very small marginal lead in the recall index (Re reaches 90.56%), MACR-Net completely dominates in comprehensive detection capability with a peak mAP_{50} of 91.13% and the highest mAP_{50-95} at 65.77%. The superior gap of over 2.6% in the mAP_{50-95} index compared to YOLOv12n is undeniable evidence that the spatial attention and multispectral fusion mechanisms of MACR-Net thoroughly handle the problem of small target distortion in hilly regions. The most decisive highlight of MACR-Net lies in computational efficiency: despite processing dual-stream data, the model still maintains a very modest parameter count of 2.77M, while the computational volume is compressed to a minimum of only 0.77G–dozens of times lower than its competitors. Furthermore, hardware-level evaluations verify that this structural economy translates into a remarkably low inference latency of 12.24 ms per frame. This rapid execution capability confirms that MACR-Net is not only the most powerful solution to penetrate fog and forest foliage but also a perfect architecture for real-time deployment on edge devices carried on UAVs with limited power resources.

Robustness Evaluation: In the final experiments, the confusion matrix is analyzed on the test set to evaluate the classification capability as shown in Figure 3. The results indicate that the standard YOLOv10n and YOLOv10n-PRD face difficulties in separating objects from the background with drone omission rates of 5.8% and 6.7% respectively. In contrast, MACR-Net demonstrates excellent background suppression capability as it reduces the drone omission rate to 3.2% and achieves an accuracy of 95.6% for this class. Additionally, the confusion between birds and drones is significantly reduced with bird class accuracy increasing from 82.1% to 87.4%. This confirms that the multispectral fusion mechanism of MACR-Net sharply distinguishes between mechanical and biological structures even under fluctuating image quality. The practical performance of the models is visualized across complex environmental scenarios in Figure 4. Under thick forest cover and cloudy skies, the baseline models such as YOLOv10n and YOLOv10n-PRD often miss targets or produce bounding boxes with low confidence scores due to visual camouflage. In hilly regions and foggy conditions where contrast is weak and objects are distorted, DVIF-Net still fails to optimize bounding box boundaries effectively. Conversely, MACR-Net maintains superior robustness through the extraction of information from dual-image streams which achieves high confidence scores ranging from 0.79 to 0.92. The bounding boxes predicted by MACR-Net align precisely with the actual boundaries of the drones and birds which serves as clear evidence for the capability of the CAA+RFD module in coordinate refinement and flexible adaptation to geometric distortions.

Despite its robust performance, MACR-Net faces certain physical boundaries, as illustrated in Figure 5. The first failure mode relates to simultaneous thermal crossover and visual camouflage (Figure 5(a)), where a bird is misclassified as a drone because both its thermal signature and visual textures blend into the background, which neutralizes the MSCP module. The second limitation is extreme physical occlusion (Figure 5(b)), where dense obstacles completely block both RGB and IR signals, which inevitably causes missed detections. These single-frame spatial limits motivate our future direction to integrate spatial-temporal memory and cross-frame kinematic modeling to maintain trajectory prediction under severe interference conditions.

4 Conclusions

This research proposes MACR-Net, a robust multispectral object detection framework designed for drone and bird identification in complex outdoor environments. The architecture effectively integrates information from RGB and IR sensors through the proposed

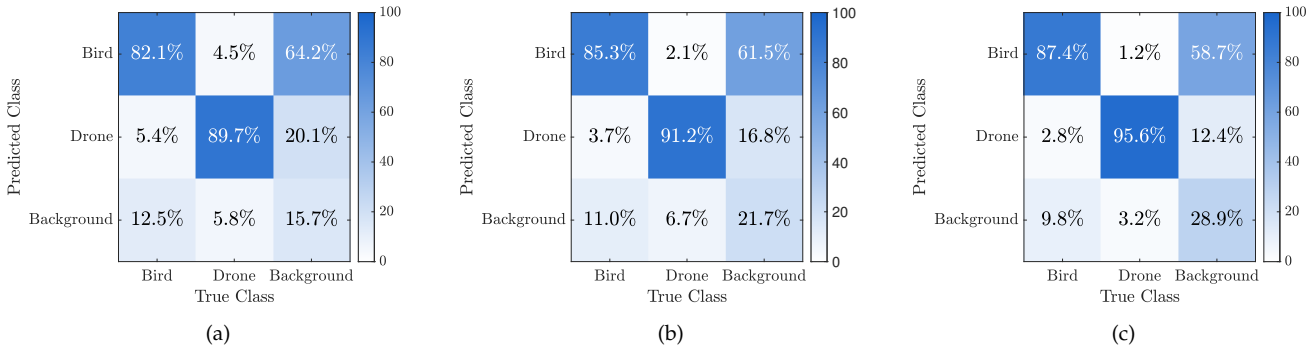

Figure 3. Confusion matrix: (a) YOLOv10n; (b) YOLOv10N-PRD; (c) MACR-Net.

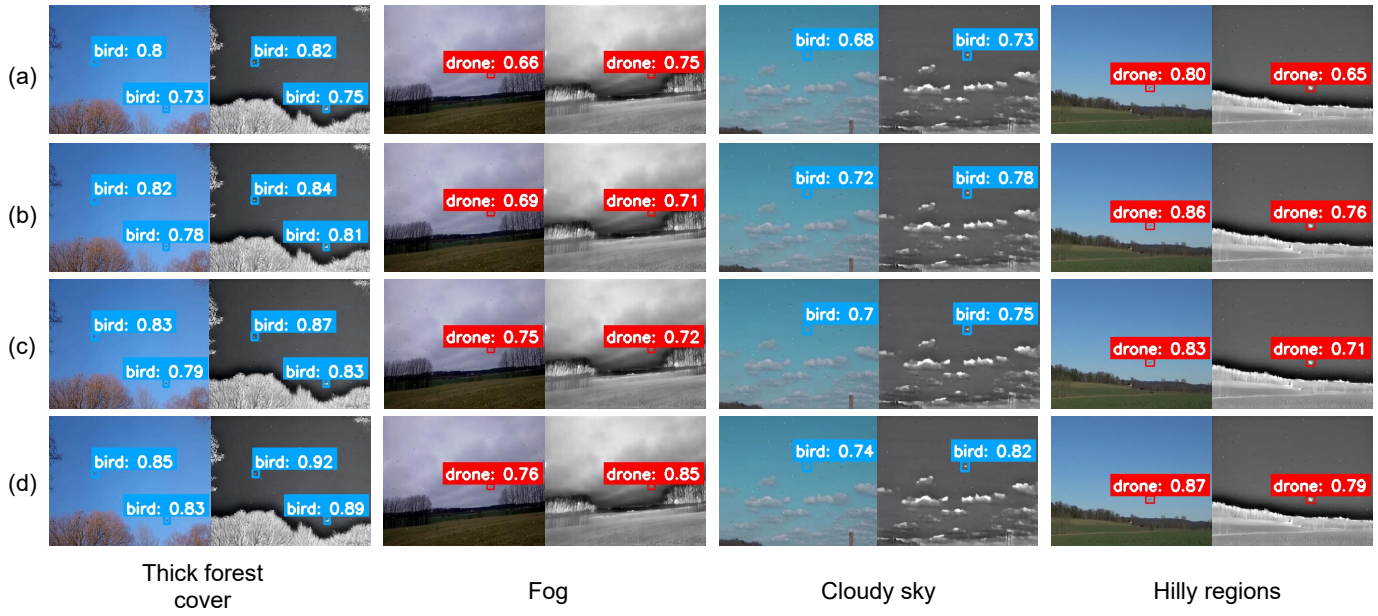

Figure 4. Comparative detection performance under varied outdoor complexities: (a) YOLOv10n; (b) YOLOv10n-PRD; (c) DVIF-Net; (d) MACR-Net.

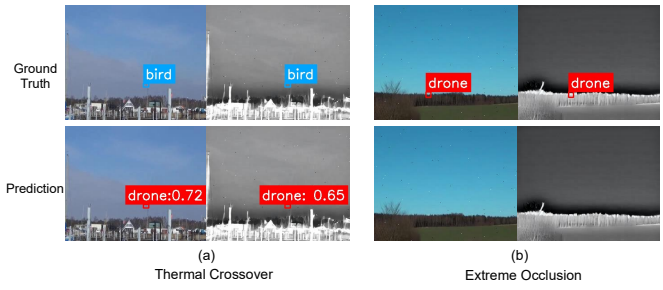

Figure 5. Typical failure scenarios: (a) False detection caused by thermal crossover, and (b) Missed detection due to extreme occlusion.

modules including MSCP and GLCI. Additionally, the combination of coordinate attention mechanism and deformable convolution at the neck stage ensures high precision in bounding box refinement and adaptability to geometric distortions. Experimental results demonstrate the superiority of MACR-Net over several SOTA models. The proposed method achieves 91.13% mAP_{50} and 65.77% mAP_{50-95}, while maintaining high efficiency with only 2.77M parameters and 0.77 GFLOPs. Hardware-level evaluations further verify this efficiency with a remarkably low inference latency of 12.73 ms, confirming the practical readiness of MACR-Net for real-time deployment on edge devices. Although the framework establishes robust defense mechanisms against environmental variations, future research will prioritize explicit cross-domain evaluations to validate its generalization capacity beyond benchmark datasets. Furthermore, to overcome the inherent physical limits of single-frame detection under extreme occlusion and thermal crossover, subsequent studies will integrate spatial-temporal memory and cross-frame kinematic modeling to maintain continuous trajectory prediction under severe interference.

Acknowledgment

This research is funded by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.02-2023.06.

References

[1] D. M.-T. Nguyen and T. Huynh-The, "Xsnet: Lightweight object detection model using x-shaped architecture in remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 22, pp. 1–5, 2025.

[2] T. Huynh-The, S. N. Truong, and G.-V. Nguyen, "Hb-senet: A hybrid bilateral network for accurate semantic segmentation of remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 14 179–14 193, 2024.

[3] G.-V. Nguyen and T. Huynh-The, "Enhancing aerial semantic segmentation with feature aggregation network for deeplabv3+," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024.

[4] G.-V. Nguyen, N. Huynh-The, and et al., "Prune and quantize semantic segmentation network for aerial objects recognition," *REV Journal on Electronics and Communications*, vol. 14, no. 3-4, pp. 18–29, 2024.

[5] G. Ding, Y. Ren, Y. Liu, Q. Zhao, and S. Li, "Vision-based anti-unmanned aerial technology: Opportunities and challenges," *IEEE Geoscience and Remote Sensing Magazine*, vol. 13, no. 4, pp. 382–405, Dec. 2025.

[6] I. Anagnostis, P. Kotzanikolaou, and C. Douligeris, "Understanding and securing the risks of uncrewed aerial vehicle services," *IEEE Access*, vol. 13, pp. 47 955–47 995, Mar. 2025.

[7] U. Seidaliyeva, L. Ilipbayeva, K. Taissariyeva, N. Smailov, and E. Matson, "Advances and challenges in drone detection and classification techniques: A state-of-the-art review," *Sensors*, vol. 24, p. 125, 12 2023.

[8] T. S. Gunawan, M. H. Aiman Mazlan, M. Kartiwi, and N. M. Yusoff, "Robust drone detection in adverse weather using YOLOv11 with synthetic rain augmentation," in *Proceedings of the 2025 IEEE 11th International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA)*, 2025, pp. 159–164.

[9] G. Gutierrez, J. P. Llerena, L. Usero, and M. A. Patricio, "A comparative study of convolutional neural network and transformer architectures for drone detection in thermal images," *Applied Sciences*, vol. 15, no. 1, 2025.

[10] A. Coluccia, A. Fascista, A. Dimou, D. Zarpalas, L. Sommer, A. Schumann, and E. Mele, "The drone-vs-bird detection grand challenge at IJCNN 2025," in *Proceedings of the 2025 International Joint Conference on Neural Networks (IJCNN)*, Rome, Italy, 2025, pp. 1–8.

[11] P. Suo, J. Zhu, Q. Zhou, W. Kou, X. Wang, and W. Suo, "YOLOv9-AAG: Distinguishing birds and drones in infrared and visible light scenarios," *IEEE Access*, vol. 13, pp. 76 609–76 619, Apr. 2025.

[12] F. Xu, Z. Lin, C. Chen, R. Yang, and H. Zhai, "Anti-drone detection in aerial multispectral images," in *Proceedings of the 2024 IEEE International Geoscience and Remote Sensing Symposium*, Athens, Greece, 2024, pp. 9388–9391.

[13] D. Tavaris, A. d. Zan, A. Toma, G. Luca Foresti, I. Scagnetto, and N. Martinel, "Multi-perspective RGB and infrared dataset for UAV detection," *IEEE Access*, vol. 13, pp. 168 792–168 803, Sep. 2025.

[14] S. Sarkar, M. A. Hasan, K. A. Shahriar, F. Labiba, N. Tasnim, and S. A. H. Fattah, "EGD-YOLO: A lightweight multimodal framework for robust drone-bird discrimination via ghost-enhanced YOLOv8n and EMA attention under adverse condition," *ArXiv*, vol. abs/2510.10765, 2025.

[15] Y.-Y. Chen, S.-Y. Jhong, and Y.-J. Lo, "Reinforcement-and-alignment multispectral object detection using visible–thermal vision sensors in intelligent vehicles," *IEEE Sensors Journal*, vol. 23, no. 21, pp. 26 873–26 886, Nov. 2023.

[16] F. Meng, A. Hong, H. Tang, and G. Tong, "FQDNet: A fusion-enhanced quad-head network for RGB-infrared object detection," *Remote Sensing*, vol. 17, no. 6, 2025.

[17] Z. Zhuo, R. Lu, Y. Yao, S. Wang, Z. Zheng, J. Zhang, and X. Yang, "TAF-YOLO: A small-object detection network for UAV aerial imagery via visible and infrared adaptive fusion," *Remote Sensing*, vol. 17, no. 24, 2025.

[18] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the 2017 IEEE International Conference on Computer Vision*, Dec. 2017, pp. 2980–2988.

[19] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the 2017 IEEE International Conference on Computer Vision*, 2017, pp. 2999–3007.

[20] J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González, "A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS," *Machine Learning and Knowledge Extraction*, vol. 5, no. 4, pp. 1680–1716, 2023.

[21] C.-Y. Wang, I.-H. Yeh, and H.-Y. Mark Liao, "YOLOv9: Learning what you want to learn using programmable gradient information," in *Proceedings of the 2024 European Conference on Computer Vision (ECCV 2024)*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Cham: Springer Nature Switzerland, 2025, pp. 1–21.

[22] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "DETRs beat YOLOs on real-time object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 965–16 974.

[23] M. Yuan, Y. Wang, and X. Wei, "Translation, scale and rotation: Cross-modal alignment meets RGB-infrared vehicle detection," in *Proceedings of the 2022 European Conference on Computer Vision (ECCV 2022)*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 509–525.

[24] J. Zhu, J. Rong, W. Kou, Q. Zhou, and P. Suo, "Robust real-time recognition of drones and birds in complex scenarios: a multimodal data fusion recognize approach," *Complex & Intelligent Systems*, vol. 12, no. 1, p. 8, Nov. 2025.

[25] X. Zhao, H. Zhang, C. Li, K. Wang, and Z. Zhang, "DVIF-Net: A small-target detection network for UAV aerial images based on visible and infrared fusion," *Remote Sensing*, vol. 17, no. 20, Oct. 2025.

[26] Z. Zheng, M. Zhou, Z. Shang, X. Wei, H. Pu, J. Luo, and W. Jia, "GAANet: Graph aggregation alignment feature fusion for multispectral object detection," *IEEE Transactions on Industrial Informatics*, vol. 21, no. 11, pp. 8282–8292, Nov. 2025.

[27] F. Yang, B. Liang, W. Li, and J. Zhang, "Multidimensional fusion network for multispectral object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 1, pp. 547–560, Jan. 2025.

[28] L. Zhang, Y. Liu, X. Wang, Y. He, G. Li, Y. Zhang, C. Liu, Z. Jiang, and Y. Liu, "CADDN: A content-aware downsampling-based detection method for small objects in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–17, 2025.

[29] Y.-Y. Chen, S.-Y. Jhong, H.-C. Lin, and Y.-C. Wu, "Vision–language-guided adaptive cross-modal fusion for multispectral object detection under adverse weather conditions," *IEEE MultiMedia*, vol. 32, no. 2, pp. 22–32, Jun. 2025.

[30] Y. Qiu, Y. Liu, Y. Chen, J. Zhang, J. Zhu, and J. Xu, "A2SPPNet: Attentive atrous spatial pyramid pooling network for salient object detection," *IEEE Transactions on Multimedia*, vol. 25, pp. 1991–2006, Jan. 2023.

Van-Phuc Hoang received PhD degree in Electronic Engineering from The University of Electro-Communications, Tokyo, Japan in 2012. He has worked as postdoc researcher, visiting scholar at The University of Electro-Communications, Tokyo, Japan, Telecom Paris, France and University of Strathclyde, Glasgow, UK during the period of 2012-2018. He is working as an Associate Professor, Director with Institute of System Integration, Le Quy Don Technical University, Hanoi, Vietnam. His research interests include digital signal processing, hardware security, VLSI design and intelligent systems for Internet of Things. He was the Technical Program Chair of several IEEE international conferences such as ICDV 2017, MCSoC 2018, SigTelCom 2019, APCCAS 2020 and ATC 2020. He is a member of IEEE.

Thien Huynh-The (IEEE Senior Member) received the Ph.D. degree in Computer Science and Engineering from Kyung Hee University (KHU), South Korea, in 2018. He was a recipient of the Superior Thesis Prize awarded by KHU. From March 2018 to August 2018, he was a Postdoctoral Researcher with Ubiquitous Computing Laboratory, KHU. From September 2018 to May 2022, he was a Postdoctoral Researcher with ICT Convergence Research Center, Kumoh National Institute of Technology, South Korea. He is currently a Lecturer of Ho Chi Minh City University of Technology and Engineering (HCM-UTE), Vietnam. He was a recipient of Golden Globe Award 2020 for Vietnamese Young Scientist and IEEE ATC Best Paper Award in 2023, 2024, and 2025. His research interests include digital image processing, radio signal processing, computer vision, wireless communications, IoT applications, machine learning, and DL. He is currently serving as an Editor of IEEE COMML.

Thanh-Dat Tran is currently pursuing a B.S. degree in Computer Engineering Technology with the Department of Computer and Communication Engineering, Ho Chi Minh City University of Technology and Engineering (HCM-UTE), Vietnam. His current research interests include digital image processing, radio signal processing, computer vision, machine learning, and deep learning.