remote sensing

MDPI

Article

CMGFDet: Cross-Modal Gated Fusion Network with Multi-Receptive Field Aggregation for RGB–Infrared Aerial Object Detection

Man Wu [1], Xiaozhang Liu [2,*], Xiulai Li [2] and Wenbiao Gan [3]

[1] School of Information and Communication Engineering, Hainan University, Haikou 570228, China; manwu@hainanu.edu.cn
[2] School of Computer Science and Technology, Hainan University, Haikou 570228, China; lixiulai01@hainanu.edu.cn
[3] Institute of Unmanned System Research, Beihang University, Beijing 100191, China; ganhope@buaa.edu.cn
* Correspondence: lxzh@hainanu.edu.cn

Highlights

What are the main findings?

- CMGFDet combines gated cross-modal fusion, Global–Local feature alignment, and multi-receptive field aggregation to address modality gaps, RGB–infrared misalignment, and object-scale variation in UAV imagery.
- On four RGB–infrared aerial benchmarks, CMGFDet improves the previous best mAP@0.5 by 1.6–2.2% while maintaining real-time inference speed.

What are the implications of the main findings?

- Quality-aware cross-modal interaction can improve detection robustness without relying on substantially larger model capacity.
- The proposed framework provides a practical detector for all-weather UAV surveillance, especially for tiny objects, nighttime scenes, and high-resolution aerial imagery.

Abstract

Multimodal object detection leveraging RGB and infrared imagery has become essential for robust all-weather perception in unmanned aerial vehicle (UAV) applications. However, existing methods still struggle with effective cross-modal feature fusion, spatial misalignment between modalities, and scale variation of objects in aerial views. In this paper, we propose CMGFDet, a Cross-Modal Gated Fusion Network with Multi-Receptive Field Aggregation designed for RGB–infrared aerial object detection. Our framework introduces three coordinated modules: (1) a Cross-Modal Feature Fusion Network (CMFFN) that employs a gated attention mechanism to selectively aggregate complementary information from both modalities during encoding; (2) a Global–Local Attention Module (GLAM) that performs hierarchical cross-modal feature alignment by jointly modelling global channel statistics and local spatial correlations in the decoder; and (3) a Multi-Receptive Field Aggregation Network (MRFAN) that captures multi-scale contextual information through parallel depthwise convolutions with diverse kernel sizes. Additionally, we incorporate a deep supervision strategy and a composite loss function to enhance training efficiency. Extensive experiments on four public benchmarks (DroneVehicle, RGBTDronePerson, VEDAI, and VTUAV) show that CMGFDet improves the previous best mAP@0.5 by 1.6%, 2.2%, 1.9%, and 2.2%, respectively. The implementation code will be released upon acceptance to support reproducibility.

Check for updates

Academic Editors: Qian Du, Xiaochen Yang and Yanni Dong

Received: 5 June 2026
Revised: 18 July 2026
Accepted: 21 July 2026
Published: 23 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: multimodal object detection; cross-modal fusion; RGB–infrared; UAV remote sensing; attention mechanism; multi-scale feature aggregation

1. Introduction

Object detection from unmanned aerial vehicles (UAVs) plays a pivotal role in diverse applications including urban traffic surveillance, search and rescue operations, and environmental monitoring [1,2]. Despite remarkable progress driven by deep learning [3,4], single-modality detectors remain vulnerable to adverse conditions. RGB sensors provide rich texture and color information under adequate illumination but degrade severely under low-light or fog conditions. Conversely, infrared sensors capture thermal radiation independent of illumination, offering robust target signatures at night; however, they lack fine-grained structural details [5,6]. Fusing these complementary modalities has emerged as a promising paradigm for achieving robust all-weather detection [7,8].

Nevertheless, effective RGB–infrared fusion for aerial object detection faces three critical challenges. **First**, the inherent heterogeneity between visible and thermal domains creates a significant modality gap. Naive concatenation or element-wise addition fails to capture the complementary semantics across modalities, often introducing noise rather than informative signals [9]. **Second**, spatial misalignment between RGB and infrared images (caused by parallax, registration errors, or platform vibration) can severely degrade detection accuracy when features are directly merged [10]. **Third**, UAV imagery contains objects spanning a wide range of scales, from tiny pedestrians occupying a few pixels to large vehicles, demanding detectors with strong multi-scale representation capability [11,12].

Existing approaches address these challenges from different perspectives. Transformer-based methods [13,14] leverage self-attention and cross-attention to model long-range dependencies across modalities, but their quadratic spatial complexity can limit real-time UAV deployment. Recent State Space Model (SSM) and Mamba-based studies [15–19] further show that long-range context can be modelled efficiently, and they provide useful efficient-context baselines for RGB–infrared perception.

These observations motivate a detector that does not simply concatenate modalities or enlarge the backbone, but instead controls cross-modal information flow according to feature quality, aligns the two streams before prediction, and preserves context for objects of different sizes. In this paper, we propose **CMGFDet** (Cross-Modal Gated Fusion Detection Network), a unified framework that systematically addresses the three aforementioned challenges through carefully designed modules that work in synergy. Our key insight is that effective multimodal fusion requires: (i) selective information exchange guided by cross-modal attention during encoding, (ii) explicit alignment of spatial features across modalities in the decoder, and (iii) comprehensive multi-scale context aggregation before prediction. To this end, we introduce:

- **Cross-Modal Feature Fusion Network (CMFFN)**: A dual-stream encoder module that employs a Feature Extraction Module (FEM) followed by a Cross-Modal Gated Attention (CMGA) mechanism. CMGA computes cross-modal attention weights via asymmetric query–value interactions and applies learned gating to selectively fuse modality-specific features while suppressing irrelevant information.
- **Global–Local Attention Module (GLAM)**: A decoder module that performs hierarchical cross-modal feature alignment by jointly capturing global channel-wise statistics (via Global Average Pooling and Global Max Pooling) and local spatial attention (via matrix multiplication). A sigmoid-gated residual connection ensures stable gradient flow while enabling fine-grained feature recalibration.

- **Multi-Receptive Field Aggregation Network (MRFAN)**: A neck module that aggregates features from three pyramid levels and processes them through a Multi-Receptive Module (MRM) consisting of parallel depthwise convolutions with kernel sizes of 5×5, 7×7, 9×9, and 11×11, followed by pointwise convolution for channel fusion. This design efficiently captures contextual information at multiple scales without excessive parameters.

Furthermore, we incorporate a deep supervision strategy in the decoder to facilitate the learning of intermediate representations and propose a composite loss function combining focal loss, CIoU loss, and cross-modal consistency loss. Extensive experiments on four challenging benchmarks, namely DroneVehicle [5], RGBTDronePerson [11], VEDAI [20], and VTUAV [21], demonstrate that CMGFDet achieves state-of-the-art performance while maintaining competitive inference speed.

The main contributions of this work are summarised as follows:

1. We propose CMGFDet, a unified framework for RGB–infrared aerial object detection that systematically addresses modality gap, spatial misalignment, and scale variation through three synergistic modules.
2. We design a Cross-Modal Gated Attention mechanism that adaptively filters cross-modal information based on feature quality, achieving selective fusion without modality-specific noise propagation.
3. We introduce a Global–Local Attention Module that combines global channel statistics with local spatial correlations for hierarchical cross-modal feature alignment in the decoder.
4. We develop a Multi-Receptive Field Aggregation Network with parallel depthwise convolutions (5×5 to 11×11) that efficiently captures multi-scale contextual information for objects spanning extreme size ranges in aerial views.
5. We conduct comprehensive experiments on four benchmarks, achieving state-of-the-art performance with consistent improvements of 1.6–2.2% mAP@0.5 over previous best methods while maintaining real-time inference capability.

For completeness, we first review closely related studies below. The remaining manuscript follows a five-part structure: Section 2 presents the proposed materials and methods, including the network architecture, training losses, datasets, and implementation details; Section 3 reports quantitative, ablation, and qualitative results; Section 4 discusses the mechanism, efficiency, and limitations of CMGFDet; and Section 5 concludes the paper.

1.1. Related Work

1.1.1. Multimodal Object Detection

Multimodal object detection has garnered increasing attention in recent years, driven by the fundamental limitation that single-modality systems cannot maintain robust performance across all environmental conditions. Early methods adopted simple fusion strategies such as early fusion (concatenating raw inputs), late fusion (merging detection results), and middle fusion (combining intermediate features) [3,22]. These straightforward approaches, while computationally efficient, fail to model the complex complementary relationships between modalities.

The second generation of multimodal detectors introduced more sophisticated fusion mechanisms. Zhang et al. [7] proposed an illumination-aware fusion strategy that dynamically adjusts modality contributions based on estimated lighting conditions, achieving significant improvements under nighttime scenarios. Cao et al. [8] introduced channel switching and spatial attention for multimodal detection, enabling the network to selectively propagate informative channels from each modality. More recently, Liu et al. [9] developed a cross-modal conflict-aware learning framework that explicitly addresses modality

conflicts through adversarial training, demonstrating that treating conflicting signals from different modalities is as important as fusing complementary ones. Yuan et al. [10] proposed a cascade alignment-guided Transformer for RGB–infrared detection that progressively aligns features across multiple scales.

For UAV-specific scenarios, the challenges are amplified by small object sizes, diverse viewing angles, and the need for real-time processing. Sun et al. [5] introduced the DroneVehicle benchmark together with a dual-stream detection baseline, while Zhao et al. [11] extended multimodal drone detection to the challenging person category. Recent remote-sensing and spectral perception studies also emphasize scale-adaptive feature updating, spatial–spectral interaction, and background suppression, suggesting that target enhancement and interference reduction should be considered together rather than treated as separate issues. These works collectively demonstrate that effective fusion requires not only complementary information aggregation but also explicit handling of intermodal inconsistencies, spatial misalignment, scale variation, and background clutter. Our CMFFN module advances this direction by introducing a gated attention mechanism that selectively controls information flow between modalities, adaptively suppressing noisy or redundant cross-modal signals.

1.1.2. Attention Mechanisms in Detection

Attention mechanisms have been widely adopted in object detection to enhance feature discriminability. Squeeze-and-Excitation Networks [23] introduced channel attention through global average pooling, learning to recalibrate channel-wise feature responses. CBAM [24] extended this idea by combining channel and spatial attention sequentially, providing both "what" and "where" information. The Transformer architecture [25–27] brought self-attention to vision tasks, enabling global context modelling at the cost of quadratic complexity.

In the detection domain, DETR [28] and its variants [29] applied cross-attention for end-to-end detection, eliminating hand-crafted components such as non-maximum suppression. Dynamic Head [30] unified scale-aware, spatial-aware, and task-aware attention in a single detection head. For multimodal scenarios, cross-modal attention has been explored to align features between modalities. CFT [13] applied standard cross-attention between RGB and infrared feature maps, while CFPT [14] proposed a cascaded feature pyramid Transformer that models cross-modal interactions at multiple scales.

However, existing cross-modal attention approaches share a common limitation: They treat all modality interactions uniformly without adaptive quality-based filtering. Our GLAM module differs from these approaches by simultaneously modelling global statistical alignment (which captures channel-wise importance) and local spatial correspondence (which captures pixel-level registration), providing both coarse- and fine-grained cross-modal feature recalibration within a single lightweight module.

1.1.3. State Space Models for Vision

State Space Models (SSMs) have emerged as efficient alternatives to Transformers for long-context modelling. Mamba [15] introduced input-dependent state updates, and subsequent vision variants, including Vision Mamba [16] and VMamba [17], adapted this idea to image representations through bidirectional or cross-directional scanning. In RGB–infrared perception, WaveMamba [18] and MambaFusion [19] further explored SSM-based multi-modal fusion. These studies are relevant to this work as efficient long-context baselines. However, CMGFDet does not contain an SSM block; it retains convolution-based local feature extraction for aerial object localisation while addressing modality quality, spatial alignment, and scale variation through gated fusion and multi-receptive field aggregation.

1.1.4. Multi-Scale Feature Aggregation

Multi-scale feature representation is fundamental for detecting objects of varying sizes. Feature Pyramid Networks (FPN) [22] established the classic top-down pathway with lateral connections, enabling the propagation of high-level semantics to lower-level feature maps. Subsequent works extended this paradigm with bidirectional paths (PANet) and attention-weighted fusion (BiFPN), progressively improving the quality of multi-scale representations.

More recently, GELAN [31] proposed gradient-efficient layer aggregation for balancing computational cost and representational capacity, demonstrating that careful gradient path design can yield stronger features without increasing depth. For aerial imagery, where objects span extreme scale ranges (from 10-pixel pedestrians to 200-pixel buses within the same frame), multi-scale aggregation is particularly critical [32]. The key challenge is to capture both local fine-grained patterns (needed for small objects) and broader contextual information (needed for distinguishing large objects from background clutter) without excessive computational cost.

Our MRFAN module advances multi-scale processing by employing parallel depth-wise convolutions with diverse receptive fields (5×5 to 11×11) within a lightweight architecture. Unlike FPN-style approaches that mainly exchange information across scale levels, MRM expands the effective receptive field after feature alignment, providing context capture that complements the cross-level fusion performed by the GELAN blocks in our neck. Overall, the remaining gaps in existing studies are threefold: fusion is often insufficiently selective, alignment is commonly treated as an implicit by-product of feature learning, and scale aggregation is rarely coupled with cross-modal quality control. CMGFDet is designed around these three gaps.

To clarify how the proposed components relate to existing designs, Table 1 gives a conceptual comparison at the operation level. The table is placed in the Related Work section because it positions the method against prior modules rather than describing implementation details. We do not claim that each primitive operation is entirely new; instead, the contribution lies in reorganizing these operations for RGB–infrared aerial detection, where modality quality, spatial alignment, and object-scale variation must be handled jointly.

Table 1. Conceptual comparison between CMGFDet components and closely related module designs.

Proposed Component	Closest Related Idea	Shared Operation	Difference in CMGFDet
Cross-Modal Gated Attention	Cross-modal feature Transformer and cascaded feature pyramid Transformer	Cross-modal query-key/value interaction	Adds modality-quality gates after cross-attention, allowing each stream to suppress unreliable complementary cues instead of uniformly accepting all cross-modal responses.
Global–Local Attention Module	Squeeze-and-Excitation and convolutional block attention modules	Global pooling descriptors and local spatial attention	Projects global and local descriptors into one correlation space and performs decoder-level local cross-modal alignment with residual preservation of modality-specific features.
Multi-Receptive Field Aggregation Network	Inception-style multi-branch aggregation and Feature Pyramid Networks	Parallel kernels and cross-level feature exchange	Couples bidirectional pyramid fusion with depthwise multi-receptive processing after RGB–infrared alignment, reducing cost while retaining context for very small aerial objects.

2. Materials and Methods

2.1. Overall Architecture

The overall architecture of CMGFDet is illustrated in Figure 1. The framework consists of three main components: (1) a dual-stream encoder with Cross-Modal Feature Fusion Network blocks for modality-specific feature extraction and selective cross-modal interaction; (2) a multi-scale fusion decoder equipped with Global–Local Attention Modules for

explicit spatial alignment; and (3) a Multi-Receptive Field Aggregation Network that serves as the neck, capturing multi-scale contextual information before the final prediction head.

Given an RGB image $X_{\text{RGB}} \in \mathbb{R}^{H\times W\times 3}$ and its aligned infrared image $X_{\text{IR}} \in \mathbb{R}^{H\times W\times 1}$, our model processes them through the following pipeline. The single-channel infrared image is first mapped to the same stem dimension as the RGB branch by a modality-specific 3×3 projection layer; after this projection, the two encoder streams share the same stage configuration. The design philosophy follows a hierarchical fusion principle: early-stage fusion captures low-level complementary information (edges, textures), while late-stage fusion handles high-level semantic alignment, enabling the network to exploit modality complementarity at all abstraction levels.

Dual-Stream Encoder. Both modalities are processed by parallel encoder branches with shared weights after the modality-specific input projections, consisting of four stages. We choose weight sharing to reduce redundant parameters and to encourage the two streams to learn a common object-level representation, while the input projection and CMGA gates retain modality-dependent responses. Each stage contains Patch Embedding layers and CMFFN Blocks that simultaneously extract modality-specific features and perform cross-modal fusion. The encoder produces multi-scale feature maps $\{X^i_{\text{pre}}, X^i_{\text{post}}\}^4_{i=1}$ for the RGB and infrared branches respectively, at resolutions $\frac{H}{4} \times \frac{W}{4}$, $\frac{H}{8} \times \frac{W}{8}$, $\frac{H}{16} \times \frac{W}{16}$, and $\frac{H}{32} \times \frac{W}{32}$.

Between the two encoder streams, a shared fusion pathway operates through concatenation and MobileNetV4 (MV4) blocks [33], producing fused features $\{f^i\}^4_{i=1}$ at each stage. The Fusion Blocks aggregate cross-modal information by combining pre-fusion features from both branches with the shared fused features.

Multi-Scale Fusion Decoder. The decoder receives the fused features and progressively refines them through GLAM blocks for cross-modal alignment and MRFAN blocks for multi-scale aggregation. Deep supervision is applied at intermediate decoder outputs $\{d^i\}$ to regularize learning.

Prediction Head. The final fused representation X_{Fusion} is passed to a detection head that predicts bounding boxes and class probabilities.

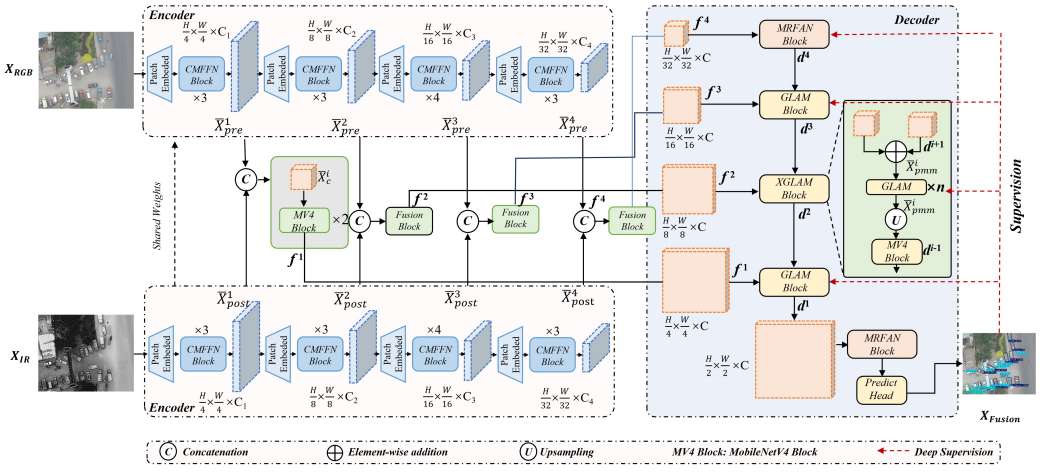

Figure 1. Overall architecture of the proposed CMGFDet framework. The dual-stream encoder with CMFFN blocks processes RGB and infrared inputs in parallel with cross-modal fusion. The decoder employs GLAM blocks for feature alignment and MRFAN blocks for multi-scale aggregation. Black arrows indicate the main feature-flow direction, red dashed arrows indicate deep supervision, orange feature maps denote fused pyramid features, and blue/gray feature maps denote modality-specific encoder features. The symbols C, ⊕, and U denote concatenation, element-wise addition, and upsampling, respectively. Deep supervision is applied at intermediate decoder stages.

2.2. Cross-Modal Feature Fusion Network (CMFFN)

The CMFFN module serves as the fundamental building block of our dual-stream encoder, responsible for extracting modality-specific features while enabling selective cross-modal information exchange. As shown in Figure 2, CMFFN consists of two submodules: the Feature Extraction Module (FEM) and the Cross-Modal Gated Attention (CMGA).

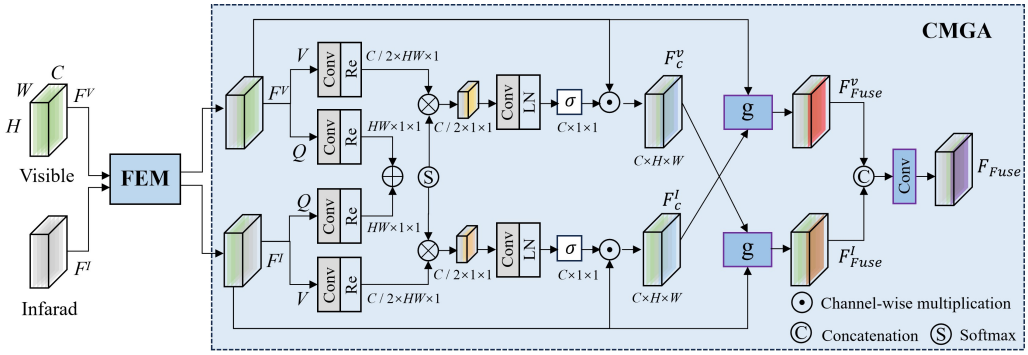

Figure 2. Architecture of the Cross-Modal Feature Fusion Network (CMFFN). The FEM extracts modality-specific features, while CMGA performs asymmetric cross-attention with learned gating to selectively fuse complementary information. Colored blocks distinguish convolution, attention, and gating operations; black arrows indicate feature flow; $\otimes$ denotes attention-based multiplication; and $\oplus$ denotes element-wise addition.

2.2.1. Feature Extraction Module (FEM)

Given input feature maps from the visible branch $F^V \in \mathbb{R}^{C\times H\times W}$ and infrared branch $F^I \in \mathbb{R}^{C\times H\times W}$, FEM first applies modality-specific convolutions to extract initial representations:

$$F_0^V = \mathrm{Conv}_{3\times3}(\mathrm{BN}(\mathrm{ReLU}(F^V))), \quad F_0^I = \mathrm{Conv}_{3\times3}(\mathrm{BN}(\mathrm{ReLU}(F^I))) \tag{1}$$

where $\mathrm{Conv}_{3\times3}$ denotes a 3×3 convolution, BN is batch normalization, and ReLU is the activation function.

2.2.2. Cross-Modal Gated Attention (CMGA)

CMGA implements an asymmetric cross-attention mechanism that allows each modality to selectively attend to informative regions of the other modality. Let $N = HW$ and $C' = C/2$. For the visible branch, we compute a query from RGB features and key/value tensors from infrared features:

$$Q^V = \mathrm{Flat}(\phi_q^V(F_0^V)) \in \mathbb{R}^{N\times C'}, \quad K^I = \mathrm{Flat}(\phi_k^I(F_0^I)) \in \mathbb{R}^{N\times C'}, \quad V^I = \mathrm{Flat}(\phi_v^I(F_0^I)) \in \mathbb{R}^{N\times C'} \tag{2}$$

The cross-modal attention map is computed through matrix multiplication followed by softmax normalization:

$$A^{V\leftarrow I} = \mathrm{Softmax}\left(\frac{Q^V(K^I)^\top}{\sqrt{C'}}\right) \in \mathbb{R}^{N\times N} \tag{3}$$

where $\mathrm{Flat}(\cdot)$ reshapes a feature map from $C'\times H\times W$ to $N\times C'$. Symmetrically, $A^{I\leftarrow V}$ is computed for the infrared branch.

The attended features are then channel-recalibrated through convolution and layer normalization:

$$\tilde{F}^{V\leftarrow I} = \mathrm{Unflat}(A^{V\leftarrow I}V^I) \in \mathbb{R}^{C'\times H\times W}, \quad F_c^V = \sigma(\mathrm{LN}(\mathrm{Conv}_{1\times1}(\tilde{F}^{V\leftarrow I}))) \odot F_0^V \tag{4}$$

where σ denotes the sigmoid activation and $\odot$ represents channel-wise multiplication.

Finally, learned gating functions g^V and g^I control the contribution of each modality to the fused output:

$$F^V_{\text{Fuse}} = g^V \odot F^V_c, \quad F^I_{\text{Fuse}} = g^I \odot F^I_c \tag{5}$$

$$F_{\text{Fuse}} = \text{Conv}_{1\times1}(\text{Concat}(F^V_{\text{Fuse}}, F^I_{\text{Fuse}})) \tag{6}$$

where $g^V, g^I \in \mathbb{R}^{C\times1\times1}$ are learnable gating parameters generated from global pooled features by a lightweight 1×1 convolution and sigmoid activation. This gating mechanism is critical for handling modality-specific noise and ensuring that only relevant cross-modal information is propagated.

The asymmetric design of CMGA is motivated by the inherently heterogeneous nature of RGB and infrared signals. RGB images provide rich colour and texture cues under well-lit conditions but degrade significantly at night or under adverse weather. Infrared images, by contrast, capture thermal emission independent of illumination but lack fine spatial detail. By computing cross-modal queries and values independently for each modality, CMGA allows each branch to dynamically select which aspects of the complementary modality to incorporate. Crucially, the sigmoid-gated output avoids the common pitfall of naive feature concatenation, where irrelevant or noisy cross-modal information can degrade single-modality representations. During training, the gating weights g^V and g^I adapt to scene-specific conditions, learning, for example, to increase infrared contribution at night while favouring RGB features in bright daytime imagery.

2.3. Global–Local Attention Module (GLAM)

While CMFFN performs feature-level fusion in the encoder, spatial misalignment between modalities may persist at higher semantic levels. The GLAM module (Figure 3) addresses this by performing explicit cross-modal feature alignment in the decoder through a dual-granularity attention mechanism.

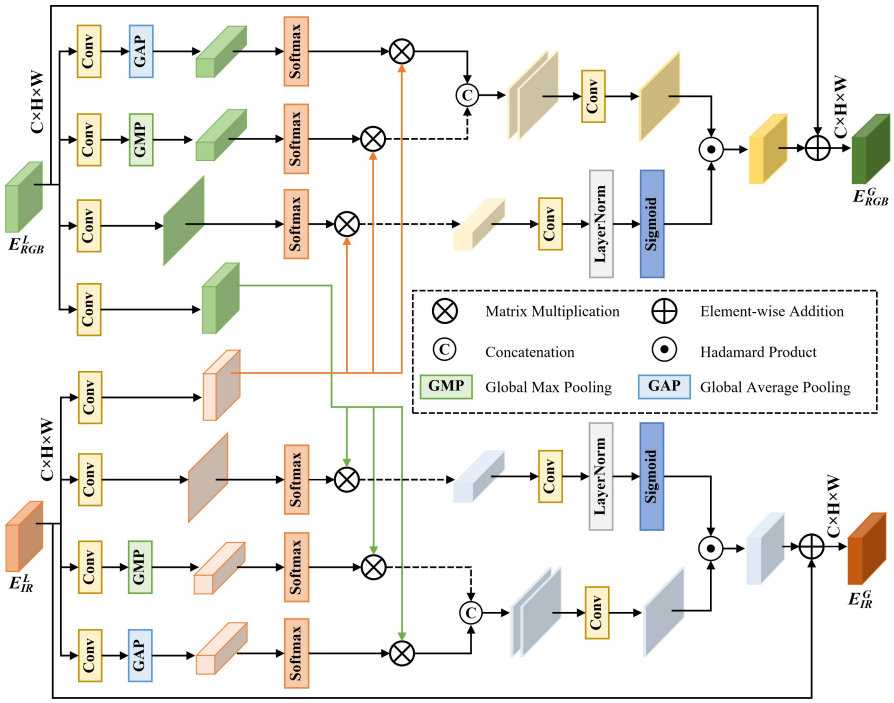

Figure 3. Architecture of the Global–Local Attention Module (GLAM) for cross-modal feature alignment. Three parallel branches (GAP, GMP, and local convolution) capture multi-granularity attention. Different colors distinguish the global pooling, local convolution, and fusion/gating branches,

while black arrows indicate feature flow. Cross-modal matrix multiplication aligns spatial features, while sigmoid gating with residual connections produce the final aligned representations.

Given encoder features from both branches $E^l_{\mathrm{RGB}}, E^l_{\mathrm{IR}} \in \mathbb{R}^{C\times H\times W}$ at decoder level l, GLAM computes aligned outputs $E^G_{\mathrm{RGB}}, E^G_{\mathrm{IR}}$ through three parallel attention branches. Before the three branches are combined, each branch is projected to a shared embedding dimension and resized to the same spatial grid, so the subsequent element-wise and matrix operations are performed on comparable representations rather than on raw heterogeneous descriptors.

2.3.1. Global Statistical Attention

For each modality, we derive three attention descriptors using different pooling strategies:

$$\alpha_{\mathrm{gap}} = \mathrm{Softmax}(\mathrm{Conv}_{1\times1}(\mathrm{GAP}(E^l_{\mathrm{RGB}}))), \quad \alpha_{\mathrm{gmp}} = \mathrm{Softmax}(\mathrm{Conv}_{1\times1}(\mathrm{GMP}(E^l_{\mathrm{RGB}}))) \tag{7}$$

where GAP and GMP denote Global Average Pooling and Global Max Pooling, respectively. GAP captures the average response distribution, while GMP emphasizes the most salient features. Their outputs are channel descriptors and are broadcast back to the $H \times W$ grid after 1×1 projection. A third branch applies spatial convolution followed by softmax to generate local attention weights α_{local}. We then concatenate the projected descriptors and use a 1×1 convolution to obtain α_{RGB} and α_{IR} with identical dimensions.

2.3.2. Cross-Modal Spatial Correlation

The unified attention maps from both modalities are combined through cross-modal matrix multiplication:

$$M_{\mathrm{cross}} = \mathrm{Softmax}(\alpha_{\mathrm{RGB}} \otimes \alpha_{\mathrm{IR}}^{\top}) \tag{8}$$

This operation computes pairwise spatial correlations between the two modalities, enabling explicit alignment of corresponding regions regardless of minor spatial shifts. To keep the cost practical, the operation is applied only on decoder feature maps after downsampling to pyramid levels P_3–P_5; in our implementation, local correlation is further computed within a $k \times k$ neighborhood around each position ($k = 7$) rather than over the full input-resolution image. Therefore, the practical complexity is $O(HWk^2)$ per level instead of $O(H^2W^2)$ on the original 1920×1080 VTUAV frames.

2.3.3. Gated Fusion with Residual Connection

The cross-modal correlation features are concatenated and processed through a global branch:

$$W_{\mathrm{global}} = \mathrm{Sigmoid}(\mathrm{LN}(\mathrm{Conv}_{1\times1}(\mathrm{Concat}(M^V_{\mathrm{cross}}, M^I_{\mathrm{cross}})))) \tag{9}$$

The final aligned features are obtained through Hadamard product and element-wise addition:

$$E^G_{\mathrm{RGB}} = W_{\mathrm{global}} \odot \mathrm{Conv}(M^V_{\mathrm{cross}}) \oplus E^l_{\mathrm{RGB}} \tag{10}$$

$$E^G_{\mathrm{IR}} = (1 - W_{\mathrm{global}}) \odot \mathrm{Conv}(M^I_{\mathrm{cross}}) \oplus E^l_{\mathrm{IR}} \tag{11}$$

where $\oplus$ denotes element-wise addition. The residual connection preserves the original modality-specific information while incorporating the aligned cross-modal features.

The design of GLAM with three parallel attention branches serves a clear functional purpose. Global Average Pooling provides a holistic statistical summary of the entire feature map, making it robust to local noise and effective at capturing the overall scene

distribution. Global Max Pooling, in contrast, identifies the single most salient response at each channel, which is particularly useful for detecting rare or small objects against cluttered backgrounds. The local convolution branch retains spatial resolution throughout, preserving fine-grained positional cues that would be lost by global pooling. Because all three branches are first projected to a common feature space, the subsequent multiplication is not a direct pixel-wise operation between incompatible GAP, GMP, and local convolution outputs. Instead, it measures cross-modal correspondence after descriptor normalization and projection, enabling robust alignment even when objects appear at slightly different pixel positions across the two imaging modalities.

2.4. Multi-Receptive Field Aggregation Network (MRFAN)

After cross-modal alignment, features at different scales need to be effectively aggregated for accurate detection of objects with diverse sizes. The MRFAN module (Figure 4) serves as the neck component that bridges the decoder and the prediction head.

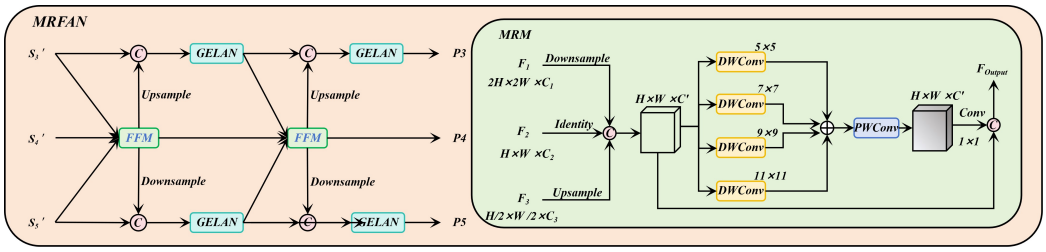

Figure 4. Architecture of the Multi-Receptive Field Aggregation Network (MRFAN). (**Left**): The bidirectional feature fusion structure with GELAN blocks. (**Right**): The Multi-Receptive Module (MRM) with parallel depthwise convolutions at kernel sizes 5, 7, 9, and 11. Black arrows indicate feature flow; the symbols C and $\oplus$ denote concatenation and element-wise addition, respectively; and the colored backgrounds distinguish the MRFAN fusion structure from the MRM submodule.

2.4.1. Multi-Scale Feature Fusion

MRFAN is the complete aggregation network and receives three feature maps $\{S_3', S_4', S_5'\}$ from different decoder levels. The MRM described below is a submodule inside MRFAN and is applied after the three levels have been brought to a compatible resolution. This distinction avoids ambiguity with Algorithm 1, where MRFAN denotes the multi-input neck rather than a single-feature operation. A bidirectional feature fusion structure connects these levels through upsampling and downsampling operations, with GELAN blocks [31] and Feature Fusion Modules (FFMs) facilitating efficient information exchange:

$$P_3 = \text{GELAN}(\text{Concat}(\text{GELAN}(S_3'), \text{Up}(\text{FFM}(S_4', S_5')))) \tag{12}$$

$$P_5 = \text{GELAN}(\text{Concat}(\text{GELAN}(S_5'), \text{Down}(\text{FFM}(S_3', S_4')))) \tag{13}$$

$$P_4 = \text{Up}(\text{FFM}(P_3)) + \text{Down}(\text{FFM}(P_5)) \tag{14}$$

2.4.2. Multi-Receptive Module (MRM)

The core of MRFAN is the Multi-Receptive Module, which captures contextual information at multiple scales simultaneously. Given features from three scales, MRM first aligns them to a common resolution through downsampling, identity mapping, and upsampling:

$$F_1 = \text{Down}(P_3) \in \mathbb{R}^{H\times W\times C_1}, \quad F_2 = P_4 \in \mathbb{R}^{H\times W\times C_2}, \quad F_3 = \text{Up}(P_5) \in \mathbb{R}^{H\times W\times C_3} \tag{15}$$

The aligned features are concatenated and processed by four parallel depthwise separable convolutions [34] with varying kernel sizes:

$$F_{\text{cat}} = \text{Concat}(F_1, F_2, F_3) \in \mathbb{R}^{H \times W \times (C_1 + C_2 + C_3)} \tag{16}$$

$$F_{\text{multi}} = \text{Concat}(\text{DWConv}_5(F_{\text{cat}}), \text{DWConv}_7(F_{\text{cat}}), \text{DWConv}_9(F_{\text{cat}}), \text{DWConv}_{11}(F_{\text{cat}})) \tag{17}$$

Finally, a pointwise convolution followed by a 1×1 convolution produces the output:

$$F_{\text{output}} = \text{Conv}_{1\times1}(\text{PWConv}(F_{\text{multi}})) \in \mathbb{R}^{H \times W \times C'} \tag{18}$$

The use of depthwise separable convolutions ensures parameter efficiency while the diverse kernel sizes (5×5 to 11×11) provide receptive fields ranging from local texture to broader contextual information, which is crucial for detecting both small and large objects in aerial imagery.

Algorithm 1 Training Procedure of CMGFDet

Require: Training dataset $\mathcal{D} = \{(X^n_{\text{RGB}}, X^n_{\text{IR}}, Y^n)\}_{n=1}^N$, learning rate η, epochs T, loss weights μ_1, μ_2, deep supervision weights $\{\lambda_i\}$
Ensure: Trained model parameters Θ

Initialize model parameters Θ
for $t = 1$ to T **do**
  for each mini-batch $(X_{\text{RGB}}, X_{\text{IR}}, Y)$ from $\mathcal{D}$ **do**
    // Dual-Stream Encoding with CMFFN
    $\{X^i_{\text{pre}}\}_{i=1}^4 \leftarrow \text{Encoder}_{\text{RGB}}(X_{\text{RGB}})$ with CMFFN blocks
    $\{X^i_{\text{post}}\}_{i=1}^4 \leftarrow \text{Encoder}_{\text{IR}}(X_{\text{IR}})$ with CMFFN blocks
    $\{f^i\}_{i=1}^4 \leftarrow \text{FusionBlocks}(\{X^i_{\text{pre}}\}, \{X^i_{\text{post}}\})$
    // Decoder with GLAM Alignment
    $d^4 \leftarrow \text{MV4Block}(f^4)$
    for $l = 3$ to 1 **do**
      $E^G_{\text{RGB}}, E^G_{\text{IR}} \leftarrow \text{GLAM}(f^l_{\text{RGB}}, f^l_{\text{IR}})$
      $d^l \leftarrow \text{MV4Block}(\text{Up}(d^{l+1}), E^G_{\text{RGB}}, E^G_{\text{IR}})$
    end for
    // Final Prediction
    $X_{\text{Fusion}} \leftarrow \text{MRFAN}(d^1, d^2, d^3)$
    $\hat{Y} \leftarrow \text{PredictHead}(X_{\text{Fusion}})$
    // Compute Loss
    $\mathcal{L}_{\text{cls}} \leftarrow \text{FL}(\hat{Y}, Y)$
    $\mathcal{L}_{\text{reg}} \leftarrow \text{CIoU}(\hat{Y}, Y)$
    $\mathcal{L}_{\text{cm}} \leftarrow \text{CrossModalConsistency}(F^V_{\text{Fuse}}, F^I_{\text{Fuse}})$
    $\mathcal{L}_{\text{deep}} \leftarrow \sum_{i=1}^3 \lambda_i \mathcal{L}_{\text{det}}(d^i, Y)$
    $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{cls}} + \mu_1 \mathcal{L}_{\text{reg}} + \mu_2 \mathcal{L}_{\text{cm}} + \mathcal{L}_{\text{deep}}$
    $\Theta \leftarrow \Theta - \eta \nabla_\Theta \mathcal{L}_{\text{total}}$
  end for
end for
return Θ

The bidirectional feature fusion structure in MRFAN is designed to overcome a key limitation of standard top-down Feature Pyramid Networks, which propagate high-level semantics downwards but lose precise low-level spatial detail during upsampling. By incorporating both top-down and bottom-up pathways, MRFAN ensures that each output scale P_3, P_4, P_5 is informed by features from all other scales simultaneously. This is particularly important for aerial datasets where object scale variation is extreme: a vehicle may span fewer than 20 pixels in high-altitude VTUAV frames but exceed 100 pixels in low-altitude DroneVehicle imagery. The GELAN blocks within the fusion paths apply gradient-efficient aggregation to maintain strong gradient flow across all levels during

training, while the Feature Fusion Modules provide lightweight channel recalibration before each scale connection.

2.5. Deep Supervision Strategy

To facilitate the optimization of intermediate decoder layers and encourage the learning of discriminative features at all levels, we employ a deep supervision strategy [35]. Auxiliary prediction heads are attached at intermediate decoder outputs $\{d^1, d^2, d^3\}$:

$$\mathcal{L}_{\text{deep}} = \sum_{i=1}^{3} \lambda_i \mathcal{L}_{\text{det}}(d^i, Y_{\text{gt}}) \tag{19}$$

where λ_i are decay weights set to $\{0.4, 0.3, 0.2\}$ for levels 1 to 3, and Y_{gt} represents ground-truth annotations. This strategy provides additional gradient signals to early decoder layers, accelerating convergence and improving feature quality at all scales.

2.6. Loss Function

The total loss function of CMGFDet combines three components designed to optimize detection accuracy and cross-modal consistency:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \mu_1 \mathcal{L}_{\text{reg}} + \mu_2 \mathcal{L}_{\text{cm}} + \mathcal{L}_{\text{deep}} \tag{20}$$

where μ_1 and μ_2 are balancing coefficients.

Classification Loss. We adopt focal loss (FL) [4] for classification:

$$\mathcal{L}_{\text{cls}} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{21}$$

where α_t and γ are focal loss hyperparameters, and p_t is the predicted probability for the target class.

Regression Loss. We employ the Complete IoU (CIoU) loss [36] for bounding box regression:

$$\mathcal{L}_{\text{reg}} = 1 - \text{IoU} + \frac{\rho^2(\boldsymbol{b}, \boldsymbol{b}^{gt})}{c^2} + \alpha v \tag{22}$$

where $\rho(\cdot)$ denotes the Euclidean distance, c is the diagonal length of the enclosing box, and v measures aspect ratio consistency. The trade-off coefficient is defined as $\alpha = v/(1 - \text{IoU} + v)$, following the standard CIoU formulation.

Cross-Modal Consistency Loss. To encourage semantic consistency between the fused features from both modalities, we introduce

$$\mathcal{L}_{\text{cm}} = \frac{1}{N} \sum_{i=1}^{N} \left\| \frac{F^V_{\text{Fuse},i}}{\|F^V_{\text{Fuse},i}\|_2} - \frac{F^I_{\text{Fuse},i}}{\|F^I_{\text{Fuse},i}\|_2} \right\|_2^2 \tag{23}$$

where N is the number of spatial positions and $F^V_{\text{Fuse},i}$, $F^I_{\text{Fuse},i}$ are the fused feature vectors at position i. The L2 normalization makes this term constrain the angular semantic direction of the two modality features rather than their absolute magnitude. This is intentional: magnitude information is still preserved by the residual paths and detection losses, while $\mathcal{L}_{\text{cm}}$ only discourages contradictory modality semantics at the same spatial location. We use a small weight ($\mu_2 = 0.5$) to avoid forcing the two streams to collapse into identical representations.

The complete training procedure is described in Algorithm 1.

2.7. Datasets

We evaluate CMGFDet on four publicly available multimodal aerial detection benchmarks:

DroneVehicle [5] is the largest publicly available drone-based RGB–infrared vehicle detection benchmark. It contains 56,878 image pairs (28,439 RGB and 28,439 infrared) annotated with five vehicle categories, including car, bus, truck, van, and freight car. Images are captured from altitudes of 50–200 metres under diverse conditions including daytime, nighttime, foggy, and rainy scenarios. The standard split allocates 17,990 pairs for training, 1469 for validation, and 8980 for testing. The dataset poses challenges including significant scale variation (vehicles range from 20 to 400 pixels), dense parking scenarios, and partial occlusion.

RGBTDronePerson [11] is a large-scale drone-based RGB-thermal dataset specifically designed for person detection. It contains 11,588 well-aligned image pairs at 640×512 resolution with 142,765 person annotations. The targets are predominantly tiny (average size 24×56 pixels), making this dataset particularly challenging for localisation accuracy. The dataset covers diverse scenes including urban streets, campus environments, and parks under various illumination conditions ranging from bright daylight to complete darkness. We follow the official train/validation/test partition released with the benchmark and keep the same split for all compared methods.

VEDAI [20] (Vehicle Detection in Aerial Imagery) consists of 1246 image pairs (RGB and infrared) at 512×512 resolution captured over Salt Lake City, Utah. It contains 3640 annotated vehicles across 8 categories (car, truck, pickup, tractor, camping car, boat, van, and other). Vehicles are extremely small (average 30×30 pixels) and exhibit significant orientation variations. We use 10-fold cross-validation following the standard protocol.

VTUAV [21] (Visible-Thermal UAV) provides 500 RGB-thermal video sequence pairs with 1.7 million frames at 1920×1080 resolution. The dataset covers urban traffic, pedestrian movement, and vehicle tracking scenarios. We use the detection subset (176 sequences with frame-level bounding box annotations) for evaluation, following the official train/test split. For all VTUAV experiments, frames are sampled according to the detection subset annotation files, and no frame from a test sequence is used during training or validation.

2.8. Implementation Details

CMGFDet is implemented in PyTorch 2.1 and trained on 4 NVIDIA A100 (80 GB) GPUs. We adopt the AdamW optimizer with an initial learning rate of 1×10^{-3}, weight decay of 5×10^{-4}, and cosine annealing schedule with a 10-epoch linear warm-up phase. The model is trained for 300 epochs with a total batch size of 16 (4 per GPU). Input images are resized to 640×640 for all datasets.

Data Augmentation. We employ a comprehensive augmentation pipeline to prevent overfitting and improve generalisation: random horizontal flip (probability 0.5), mosaic augmentation (probability 1.0 during the first 280 epochs, disabled in the last 20 epochs for stable convergence), mixup with $\alpha = 0.5$, random affine transformations (rotation $\pm 10^{\circ}$, scale 0.5–1.5, translation ± 0.1), and random HSV jittering (hue ± 0.015, saturation ± 0.7, value ± 0.4). For the infrared branch, augmentation is applied identically to maintain spatial correspondence with the RGB branch.

Loss Configuration. The loss weights are set to $\mu_1 = 7.5$ for regression and $\mu_2 = 0.5$ for cross-modal consistency. Deep supervision decay weights are $\{\lambda_1, \lambda_2, \lambda_3\} = \{0.4, 0.3, 0.2\}$ as validated in our hyperparameter sensitivity analysis (Section 3). Level 1 denotes the deepest auxiliary decoder output closest to the bottleneck, and Level 3 denotes the shallower auxiliary output closer to the prediction head, consistent with Figure 1. The focal loss parameters are $\alpha_t = 0.25$ and $\gamma = 2.0$.

Architecture Configuration. The channel dimensions are $C_1 = 64$, $C_2 = 128$, $C_3 = 256$, $C_4 = 512$ for the four encoder stages. The MRM kernel sizes are $\{5, 7, 9, 11\}$ with depthwise

separable convolutions. The GLAM module operates at decoder levels 1–3. All convolution layers use batch normalisation followed by SiLU activation unless otherwise specified.

Evaluation Protocol. We report mAP@0.5 and mAP@0.5:0.95 as the primary detection metrics following the COCO evaluation standard. Precision (P), recall (R), and F1-score are computed at the confidence threshold that gives the best validation F1, where $F1 = 2PR/(P + R)$. Inference speed (FPS) is measured on a single NVIDIA A100 GPU with batch size 1 and input resolution 640 × 640 after 100 warm-up iterations. Parameters are reported in millions (M) and computational cost in GFLOPs.

3. Results

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

3.1.1. Results on DroneVehicle

Table 2 presents the comparison results on the DroneVehicle dataset. We compare CMGFDet with 20 representative methods spanning three categories: (1) single-mode detectors adapted to accept dual-stream RGB-IR input through early concatenation, (2) general-purpose multimodal detectors designed for cross-modal fusion, and (3) specialised RGB–infrared methods explicitly designed for aerial detection scenarios. For methods with released code, we retrain them using the dataset splits, input size, and evaluation metrics described in Section 2.7. For very recent methods whose complete training code or checkpoints are not yet publicly available, we reproduce the reported setting as closely as possible and use the numbers reported under the same benchmark protocol when direct retraining is not feasible. This distinction is important when interpreting the 2026 comparison entries.

Table 2. Comparison with state-of-the-art methods on the DroneVehicle dataset. Best results are in **bold** and second best are underlined. Italics denote method-group headings. The light-blue background highlights CMGFDet (Ours), and the light-gray background highlights the strongest previous method.

Method	Year	P (%)	R (%)	F1 (%)	Car	Bus	mAP@0.5	mAP@0.5:0.95	Params (M)	FPS
Single-Modal Detectors (Dual-Stream Adapted)										
Faster R-CNN [3]	2015	63.5	59.2	61.3	72.3	68.1	55.2	30.1	41.1	18.4
YOLOv8-L [37]	2023	77.2	72.8	74.9	86.5	79.2	69.1	41.5	43.6	72.1
YOLOv9-E [38]	2024	78.5	74.2	76.3	87.8	81.3	70.7	42.3	57.3	48.6
YOLOv10-L [39]	2024	79.1	74.9	76.9	88.2	82.0	71.5	43.0	24.4	81.4
YOLOv11-L [40]	2024	79.8	75.6	77.6	88.9	82.7	72.3	43.6	25.3	78.9
RT-DETR-L [29]	2024	78.0	73.7	75.8	87.5	80.8	70.3	42.0	32.0	62.2
YOLOv12-L [41]	2025	80.2	76.0	78.1	89.3	83.1	72.8	43.9	26.4	76.4
YOLOv13 [42]	2025	80.7	76.5	78.5	89.7	83.5	73.4	44.2	27.1	73.5
General Multimodal Detectors										
CFT [13]	2022	77.9	73.5	75.6	88.1	80.5	70.2	41.8	38.5	36.8
LRAF-Net [43]	2023	79.6	75.3	77.4	89.4	82.8	72.5	43.4	45.2	42.0
CFPT [14]	2024	80.9	76.7	78.7	90.1	83.9	73.8	44.6	42.8	38.1
CrossFuse [44]	2024	80.3	76.1	78.2	89.8	83.2	73.2	44.0	36.8	48.4
WaveMamba [18]	2025	81.4	77.3	79.3	90.5	84.3	74.3	45.1	35.6	52.5
MambaFusion [19]	2024	80.9	76.7	78.7	90.2	83.8	73.8	44.5	38.2	44.8
Specialized RGB-IR Aerial Methods										
UA-CMDet [5]	2022	78.2	73.9	76.0	87.6	81.0	70.6	42.1	39.5	35.6
ICAFusion [7]	2023	80.3	76.0	78.1	89.7	83.4	73.2	44.0	40.1	42.8
CAGTDet [10]	2024	81.2	77.0	79.0	90.3	84.1	74.1	44.9	44.5	40.2
AMSRDet [45]	2026	81.6	77.5	79.5	90.6	84.5	74.5	45.3	33.4	44.6
DCAM-DETR [46]	2026	81.3	77.1	79.2	90.4	84.2	74.2	45.0	36.1	36.2
CSFADet [47]	2026	81.5	77.3	79.4	90.5	84.4	74.4	45.2	34.7	39.8
CMGFDet (Ours)	2026	**83.2**	**79.4**	**81.2**	**91.8**	**86.2**	**76.1**	**47.2**	31.8	47.2

CMGFDet achieves 76.1% mAP@0.5 and 47.2% mAP@0.5:0.95, surpassing the previous best method AMSRDet by 1.6% and 1.9% respectively. Compared to WaveMamba, the improvement reaches 1.8% in mAP@0.5, demonstrating that our gated fusion mechanism provides stronger cross-modal feature integration than wavelet-domain fusion. The improvement is consistent across all vehicle categories, with particularly notable gains on challenging categories including freight cars (+1.7%), trucks (+1.9%), and vans (+1.7%), where multi-scale context is crucial for distinguishing visually similar vehicles in aerial views.

Several noteworthy observations emerge from the per-category analysis. First, CMGFDet achieves the largest improvement on Bus (+1.7% over AMSRDet), which we attribute to GLAM's ability to capture the distinctive elongated spatial structure of buses through Global–Local attention. Second, the improvement on Car (the easiest category) is relatively smaller (+1.2%), suggesting that our modules provide the greatest benefit for harder discrimination tasks. Third, among multimodal methods, CMGFDet requires only 31.8 M parameters, which is 13.4 M fewer than LRAF-Net (45.2 M) and 12.7 M fewer than CAGTDet (44.5 M), confirming the parameter efficiency of our depthwise-based design.

3.1.2. Results on RGBTDronePerson

Table 3 shows results on the RGBTDronePerson dataset, which focuses on detecting small persons from drone altitude. This benchmark is particularly challenging because persons in aerial thermal imagery exhibit low contrast and highly variable poses, requiring both fine spatial resolution and strong semantic understanding for reliable detection.

Table 3. Comparison results on the RGBTDronePerson dataset. Best results in **bold**, second best underlined. The light-blue background highlights CMGFDet (Ours), and the light-gray background highlights the strongest previous method.

Method	Year	P (%)	R (%)	F1 (%)	mAP@0.5	mAP@0.5:0.95	Params (M)	GFLOPs	FPS
Faster R-CNN [3]	2015	51.6	46.8	49.1	42.3	21.8	41.1	134.5	22.1
YOLOv8-L [37]	2023	65.2	60.8	62.9	56.7	32.4	43.6	165.2	72.3
YOLOv9-E [38]	2024	66.8	62.4	64.5	58.2	33.8	57.3	189.0	48.5
YOLOv10-L [39]	2024	67.5	63.1	65.2	59.1	34.5	24.4	120.3	80.1
YOLOv11-L [40]	2024	68.9	64.3	66.5	60.3	35.2	25.3	86.9	78.4
RT-DETR-L [29]	2024	66.2	61.7	63.9	57.8	33.1	32.0	110.0	62.5
YOLOv12-L [41]	2025	69.5	64.9	67.1	61.0	35.8	26.4	88.6	76.2
YOLOv13 [42]	2025	70.1	65.4	67.7	61.5	36.2	27.1	90.3	73.8
CFT [13]	2022	64.5	59.8	62.1	55.8	31.6	38.5	142.8	32.4
LRAF-Net [43]	2023	67.8	63.2	65.4	59.4	34.1	45.2	156.3	38.7
CFPT [14]	2024	69.6	65.0	67.2	61.2	35.9	42.8	148.5	35.2
WaveMamba [18]	2025	71.2	66.8	68.9	62.8	37.1	35.6	102.4	52.8
MambaFusion [19]	2024	70.0	65.3	67.6	61.5	36.3	38.2	118.7	44.6
ICAFusion [7]	2023	68.4	63.9	66.1	60.1	35.0	40.1	138.2	40.3
CAGTDet [10]	2024	70.5	65.8	68.1	62.0	36.5	44.5	152.1	36.8
CrossFuse [44]	2024	69.7	65.1	67.3	61.3	35.7	36.8	112.5	48.2
AMSRDet [45]	2026	71.8	67.4	69.5	63.2	37.5	33.4	98.6	55.1
DCAM-DETR [46]	2026	71.0	66.5	68.7	62.5	36.9	36.1	125.3	42.3
CSFADet [47]	2026	71.3	66.9	69.0	62.8	37.2	34.7	105.8	49.5
CMGFDet (Ours)	2026	**74.2**	**69.8**	**71.9**	**65.4**	**39.3**	31.8	94.2	58.6

On this challenging tiny-person detection task, CMGFDet achieves 65.4% mAP@0.5 and 39.3% mAP@0.5:0.95, outperforming the second-best AMSRDet by 2.2% and 1.8% respectively. The significant improvement on mAP@0.5:0.95 indicates superior localization accuracy, which we attribute to two factors: (1) GLAM's precise cross-modal spatial alignment reduces bounding box regression errors that commonly arise from misaligned multimodal features, and (2) MRFAN's multi-scale aggregation provides richer contextual

cues for distinguishing persons from background clutter at varying altitudes. Notably, CMGFDet achieves these gains with only 31.8 M parameters and 94.2 GFLOPs, demonstrating that well-designed cross-modal interaction modules can deliver superior performance without excessive computational overhead. The 58.6 FPS inference speed also satisfies real-time requirements for drone deployment.

3.1.3. Results on VEDAI

Table 4 reports the results on VEDAI. CMGFDet achieves 69.7% mAP@0.5 and 42.1% mAP@0.5:0.95, exceeding AMSRDet by 1.9% and 1.5% respectively. The VEDAI dataset contains vehicles with fewer than 100 pixels on average, making it particularly challenging for accurate localization. The consistent improvement over both Mamba-based methods (WaveMamba: 67.5%, MambaFusion: 66.8%) and Transformer-based methods (CFPT: 66.1%) validates the effectiveness of MRM's multi-receptive field design for capturing fine-grained spatial patterns at small scales. Furthermore, CMGFDet maintains 62.4 FPS on this dataset, which is competitive with lightweight single-modal detectors while delivering substantially higher accuracy.

Table 4. Comparison results on the VEDAI dataset (512 × 512 resolution). Best results in **bold**, second best underlined. The light-blue background highlights CMGFDet (Ours), and the light-gray background highlights the strongest previous method.

Method	Year	P (%)	R (%)	F1 (%)	mAP@0.5	mAP@0.5:0.95	Params (M)	GFLOPs	FPS
Faster R-CNN [3]	2015	56.3	52.1	54.1	48.6	24.2	41.1	134.5	18.3
YOLOv8-L [37]	2023	70.8	66.5	68.6	62.4	35.8	43.6	165.2	78.5
YOLOv9-E [38]	2024	72.1	67.8	69.9	63.9	37.1	57.3	189.0	52.3
YOLOv10-L [39]	2024	72.6	68.4	70.4	64.5	37.8	24.4	120.3	85.2
YOLOv11-L [40]	2024	73.4	69.0	71.1	65.2	38.4	25.3	86.9	82.1
RT-DETR-L [29]	2024	71.5	67.0	69.2	63.1	36.5	32.0	110.0	65.8
YOLOv12-L [41]	2025	73.9	69.5	71.6	65.8	38.9	26.4	88.6	80.4
YOLOv13 [42]	2025	74.3	70.0	72.1	66.3	39.4	27.1	90.3	76.2
CFT [13]	2022	69.8	65.3	67.5	61.5	34.8	38.5	142.8	35.6
LRAF-Net [43]	2023	72.9	68.6	70.7	64.8	37.5	45.2	156.3	42.1
CFPT [14]	2024	74.2	69.8	71.9	66.1	39.2	42.8	148.5	38.4
WaveMamba [18]	2025	75.4	71.2	73.2	67.5	40.3	35.6	102.4	55.7
MambaFusion [19]	2024	74.8	70.5	72.6	66.8	39.7	38.2	118.7	48.2
CAGTDet [10]	2024	75.1	70.8	72.9	67.2	40.1	44.5	152.1	40.5
AMSRDet [45]	2026	75.8	71.6	73.6	67.8	40.6	33.4	98.6	58.3
DCAM-DETR [46]	2026	75.0	70.7	72.8	67.1	39.9	36.1	125.3	45.6
CSFADet [47]	2026	75.3	71.0	73.1	67.4	40.2	34.7	105.8	52.1
CMGFDet (Ours)	2026	**77.5**	**73.6**	**75.5**	**69.7**	**42.1**	31.8	94.2	62.4

3.1.4. Results on VTUAV

Table 5 reports the results on VTUAV's high-resolution sequences (1920 × 1080). CMGFDet achieves 71.3% mAP@0.5 and 44.5% mAP@0.5:0.95, outperforming AMSRDet by 2.2% and 1.7% respectively. The VTUAV dataset presents unique challenges compared to the other benchmarks: Its high resolution produces feature maps with significantly more spatial positions, increasing the difficulty of cross-modal alignment. The strong performance of CMGFDet on this dataset validates the scalability of GLAM's global–local attention design, which remains effective regardless of spatial resolution. Among all multimodal methods, CMGFDet achieves 47.2 FPS, which represents a favourable balance between accuracy and speed for high-resolution UAV surveillance applications.

Table 5. Comparison results on the VTUAV dataset (1920 × 1080 resolution). Best in **bold**, second best underlined. The light-blue background highlights CMGFDet (Ours), and the light-gray background highlights the strongest previous method.

Method	Year	P (%)	R (%)	F1 (%)	mAP@0.5	mAP@0.5:0.95	Params (M)	GFLOPs	FPS
Faster R-CNN [3]	2015	59.4	54.8	57.0	51.2	27.8	41.1	134.5	15.6
YOLOv8-L [37]	2023	72.0	67.6	69.7	63.8	38.2	43.6	165.2	58.3
YOLOv9-E [38]	2024	73.3	69.0	71.1	65.1	39.5	57.3	189.0	38.7
YOLOv10-L [39]	2024	74.1	69.7	71.8	66.0	40.2	24.4	120.3	65.4
YOLOv11-L [40]	2024	74.8	70.5	72.6	66.8	40.9	25.3	86.9	63.1
RT-DETR-L [29]	2024	72.6	68.2	70.3	64.5	38.8	32.0	110.0	48.2
YOLOv12-L [41]	2025	75.2	70.9	73.0	67.2	41.3	26.4	88.6	61.5
YOLOv13 [42]	2025	75.7	71.4	73.5	67.8	41.8	27.1	90.3	58.9
CFT [13]	2022	71.2	66.7	68.9	62.9	37.4	38.5	142.8	26.8
LRAF-Net [43]	2023	73.8	69.5	71.6	65.8	39.8	45.2	156.3	31.4
CFPT [14]	2024	75.4	71.1	73.2	67.4	41.5	42.8	148.5	28.6
WaveMamba [18]	2025	76.6	72.5	74.5	68.9	42.6	35.6	102.4	42.3
MambaFusion [19]	2024	75.8	71.6	73.6	67.8	41.9	38.2	118.7	36.5
CAGTDet [10]	2024	76.1	71.8	73.9	68.2	42.1	44.5	152.1	30.2
AMSRDet [45]	2026	76.8	72.7	74.7	69.1	42.8	33.4	98.6	44.8
DCAM-DETR [46]	2026	76.2	72.0	74.0	68.5	42.3	36.1	125.3	34.1
CSFADet [47]	2026	76.5	72.3	74.3	68.8	42.5	34.7	105.8	39.6
CMGFDet (Ours)	2026	**78.9**	**75.2**	**77.0**	**71.3**	**44.5**	31.8	94.2	47.2

3.1.5. Summary Across All Benchmarks

Table 6 consolidates the improvements of CMGFDet over the previous state-of-the-art across all four datasets.

Table 6. Summary of CMGFDet improvements over previous best methods across all benchmarks. Bold values indicate CMGFDet results and improvement margins, and the light-blue background highlights the average improvement row.

Dataset	Previous Best	mAP@0.5 (Prev)	mAP@0.5 (Ours)	ΔmAP@0.5	ΔmAP@0.5:0.95
DroneVehicle	AMSRDet	74.5	**76.1**	+1.6	+1.9
RGBTDronePerson	AMSRDet	63.2	**65.4**	+2.2	+1.8
VEDAI	AMSRDet	67.8	**69.7**	+1.9	+1.5
VTUAV	AMSRDet	69.1	**71.3**	+2.2	+1.7
Average	–	68.7	**70.6**	**+2.0**	**+1.7**

The average improvement of 2.0% mAP@0.5 and 1.7% mAP@0.5:0.95 across four diverse benchmarks confirms the broad applicability of our approach. The larger gains on RGBTDronePerson (+2.2%) and VTUAV (+2.2%) compared to DroneVehicle (+1.6%) suggest that CMGFDet particularly excels when the detection task involves either very small objects or high-resolution inputs, scenarios where cross-modal alignment and multi-scale aggregation provide the greatest benefit.

To provide deeper insight into the classification behavior of CMGFDet, Figure 5 presents the normalized confusion matrix on the DroneVehicle test set. The model achieves high diagonal values across all categories, with Car (0.94) and Bus (0.92) exhibiting the strongest recognition accuracy. The most common confusion occurs between Truck and Van (0.06 misclassification rate), which is expected given their similar appearance in aerial views. Freight Car shows a balanced error distribution across other categories. Overall, the confusion matrix demonstrates that CMGFDet maintains robust per-class discrimination.

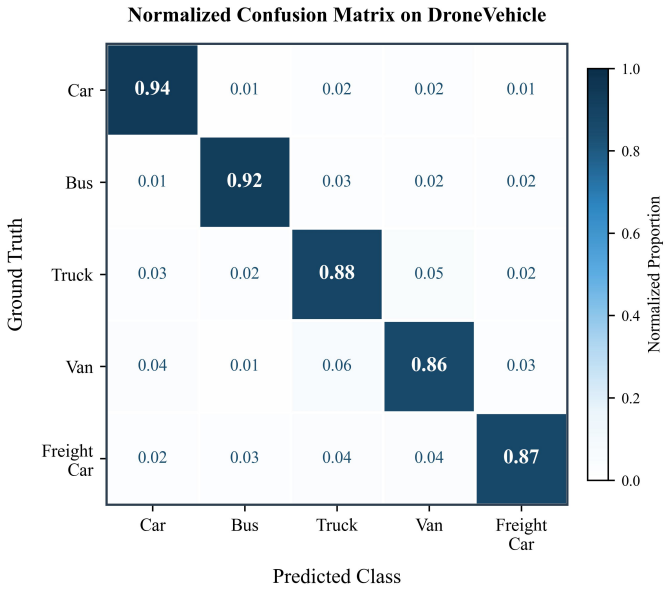

Figure 5. Normalized confusion matrix of CMGFDet on the DroneVehicle test set. Diagonal values represent per-class recognition rates. The most frequent confusion arises between visually similar categories (Truck ↔ Van), while Car and Bus are well distinguished.

3.2. Ablation Studies

3.2.1. Component Analysis

Table 7 presents the component ablation results on DroneVehicle. To validate the contribution of each proposed module, we conduct ablation experiments by progressively adding components to a baseline dual-stream YOLOv11 architecture. The baseline uses two modality-specific input stems followed by a shared YOLOv11-L backbone and a PAN/FPN neck; RGB and infrared features at the same pyramid level are fused by simple element-wise addition before the detection head. No CMGA, GLAM, MRFAN, deep supervision, or cross-modal consistency loss is used in this baseline. This setting provides a conservative reference because it already uses both modalities but lacks the proposed selective fusion and alignment mechanisms.

Table 7. Ablation study of proposed modules on DroneVehicle. Check marks indicate that the corresponding module or loss term is included. Bold values indicate the best results, and the light-blue background highlights the full CMGFDet configuration.

CMFFN	GLAM	MRFAN	Deep Sup.	$\mathcal{L}_{cm}$	mAP@0.5	Δ
					72.3	–
✓					73.8	+1.5
✓	✓				74.9	+2.6
✓	✓	✓			75.6	+3.3
✓	✓	✓	✓		75.9	+3.6
✓	✓	✓	✓	✓	**76.1**	**+3.8**

Each module provides consistent improvement. CMFFN contributes the largest single gain (+1.5%) by enabling selective cross-modal fusion, confirming that the quality of feature-level interaction is the dominant factor in multimodal detection. GLAM adds +1.1% through explicit spatial alignment, which is particularly important for the DroneVehicle dataset where registration errors are common due to platform vibration. MRFAN provides +0.7% via multi-scale aggregation, addressing the scale variation challenge inherent in aerial imagery. The auxiliary components (deep supervision and cross-modal consistency loss) together contribute +0.5%, which is modest but consistent, indicating that they serve

as effective regularisers rather than primary performance drivers. Notably, the total improvement (+3.8%) exceeds the simple sum of individual contributions, suggesting positive synergistic interactions between the modules.

3.2.2. Effect of Shared Encoder Weights

Table 8 compares the current shared-weight encoder with an independent-weight variant on DroneVehicle. To further examine whether sharing encoder weights benefits RGB–infrared feature learning, the RGB and infrared branches in the independent-weight variant use separate encoder parameters after the modality-specific input projections. All other modules, training settings, input resolution, and evaluation metrics are kept unchanged.

Table 8. Comparison between shared and independent encoder weights on DroneVehicle. The light-blue background highlights the selected shared-weight setting used in CMGFDet.

Encoder Setting	mAP@0.5	mAP@0.5:0.95	Params (M)	GFLOPs	FPS
Independent weights	75.9	47.0	50.2	146.8	44.6
Shared weights (Ours)	76.1	47.2	31.8	94.2	62.4

This comparison isolates the contribution of weight sharing from the effects of CMGA, GLAM, MRFAN, deep supervision, and the cross-modal consistency loss. The shared-weight encoder slightly improves mAP@0.5 from 75.9% to 76.1% and mAP@0.5:0.95 from 47.0% to 47.2%, while reducing parameters from 50.2 M to 31.8 M and GFLOPs from 146.8 to 94.2. It also increases inference speed from 44.6 FPS to 62.4 FPS. These results indicate that weight sharing provides a better accuracy–efficiency trade-off, because it removes redundant branch parameters while preserving modality-dependent responses through the input projections and Cross-Modal Gated Attention.

3.2.3. Impact of Deep Supervision Depth

Table 9 further investigates how the number of deep supervision levels affects model performance.

Table 9. Impact of deep supervision depth on DroneVehicle. Bold values indicate the best results, and the light-blue background highlights the selected setting used in CMGFDet.

DS Levels	mAP@0.5	mAP@0.5:0.95	Train Time (h)
None	75.6	46.5	18.2
Level 1 only	75.8	46.8	19.0
Levels 1–2	75.9	47.0	19.8
Levels 1–3 (Ours)	**76.1**	**47.2**	20.4
Levels 1–4	75.9	47.0	21.6

Applying deep supervision at three decoder levels yields the best results. Extending supervision to a fourth (highest-resolution) level does not improve accuracy but increases training time by 5.9%, as the high-resolution supervision signals introduce redundant gradient updates that the main detection loss already covers.

3.2.4. Hyperparameter Sensitivity Analysis

We investigate the sensitivity of CMGFDet to its key hyperparameters on DroneVehicle, as shown in Figure 6. Figure 6a varies the regression loss weight μ_1 from 2.5 to 15.0; performance peaks at $\mu_1 = 7.5$ and degrades with both smaller and larger values, suggesting that balanced regression supervision is critical. Figure 6b analyzes the cross-modal consistency loss weight μ_2; the optimum is found at $\mu_2 = 0.5$, confirming that moderate consistency regularization effectively aligns the two modality streams without

over-constraining the representations. Figure 6c compares different deep supervision decay weight configurations $\{\lambda_1, \lambda_2, \lambda_3\}$; the configuration $\{0.4, 0.3, 0.2\}$ yields the best performance, indicating that gradually decreasing supervision strength at shallower decoder levels is beneficial.

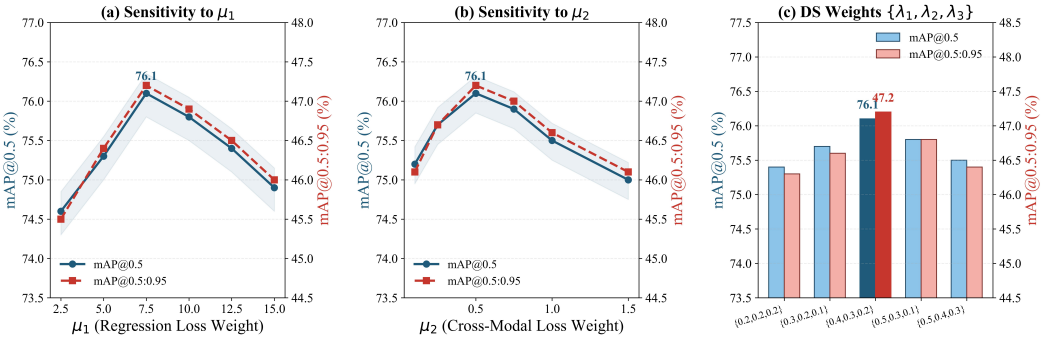

Figure 6. Hyperparameter sensitivity analysis on DroneVehicle. (**a**) Impact of regression loss weight μ_1. (**b**) Impact of cross-modal consistency loss weight μ_2. (**c**) Comparison of deep supervision decay weight configurations. The optimal settings ($\mu_1 = 7.5$, $\mu_2 = 0.5$, $\{\lambda_i\} = \{0.4, 0.3, 0.2\}$) are highlighted.

3.2.5. Analysis of MRM Kernel Sizes

Table 10 reports the effect of different MRM kernel-size combinations. The configuration $\{5, 7, 9, 11\}$ achieves the best accuracy–efficiency trade-off. Adding a 13×13 kernel does not improve performance while increasing parameters by 1.4 M, suggesting that 11×11 already provides sufficient receptive field for aerial objects at 640×640 input resolution. The progression from $\{3, 5\}$ to $\{5, 7, 9, 11\}$ shows diminishing returns: the first two additions contribute 0.5% each, while the fourth kernel adds only 0.4%. This logarithmic improvement pattern is consistent with the theoretical analysis of effective receptive fields, where marginal coverage gain decreases as kernel size increases. The depthwise separable implementation ensures that the four-kernel configuration adds only 2.6 M parameters over the two-kernel baseline, maintaining computational efficiency.

Table 10. Impact of different kernel size combinations in MRM on DroneVehicle. Bold values indicate the best result, and the light-blue background highlights the selected MRM kernel configuration.

Kernel Sizes	mAP@0.5	Params (M)
$\{3, 5\}$	74.8	29.2
$\{3, 5, 7\}$	75.3	30.1
$\{5, 7, 9\}$	75.7	30.8
$\{5, 7, 9, 11\}$ (Ours)	**76.1**	31.8
$\{5, 7, 9, 11, 13\}$	76.0	33.2

3.2.6. Impact of Gating Mechanism in CMGA

Table 11 analyzes the impact of the gating mechanism in CMGA. The learned gating mechanism outperforms all alternatives, confirming that adaptive modality weighting is essential for handling the variable quality of RGB and infrared signals across different scenes and conditions. The gradual performance improvement from direct concatenation to cross-attention to full CMGA validates our design philosophy of progressive refinement: Each additional mechanism addresses a specific limitation of the previous approach.

Table 11. Analysis of the gating mechanism in CMGA on DroneVehicle. Bold values indicate the best results, and the light-blue background highlights the learned-gating configuration used in CMGFDet.

Fusion Strategy	mAP@0.5	mAP@0.5:0.95
Direct concatenation	73.2	44.1
Element-wise addition	73.5	44.4
Channel attention only	74.3	45.2
Cross-attention w/o gating	75.1	46.0
CMGA with fixed gating	75.5	46.4
CMGA with learned gating (Ours)	**76.1**	**47.2**

3.2.7. Impact of Input Resolution

Table 12 examines the effect of input resolution on detection performance and inference speed. Performance improves significantly from 416 × 416 to 640 × 640 (+2.9% mAP@0.5), but saturates beyond 640 × 640, with 800 × 800 providing only marginal improvement (+0.2%) at substantially higher computational cost (56% more GFLOPs). This saturation occurs because the MRFAN module already provides sufficient receptive field coverage at 640 × 640 resolution. We therefore select 640 × 640 as the default input size to balance accuracy and efficiency.

Table 12. Impact of input resolution on CMGFDet performance (DroneVehicle). Bold values indicate the best selected results, and the light-blue background highlights the default input resolution used in CMGFDet.

Resolution	mAP@0.5	mAP@0.5:0.95	GFLOPs	FPS
416 × 416	73.2	44.0	38.5	98.7
512 × 512	74.8	45.6	58.4	82.3
640 × 640 (Default)	**76.1**	**47.2**	94.2	62.4
800 × 800	76.3	47.4	147.2	41.8
1024 × 1024	76.2	47.3	241.5	26.3

3.2.8. Training Convergence Analysis

Figure 7 shows the training loss and validation mAP curves during 300 epochs of training. The full CMGFDet model converges faster and achieves lower final loss compared to ablated variants, demonstrating that both deep supervision and the cross-modal consistency loss $\mathcal{L}_{\text{cm}}$ contribute to training stability and convergence speed. Notably, deep supervision provides the most significant acceleration in early training phases (epochs 0–100) by injecting gradient signals directly into intermediate decoder layers, reducing the vanishing gradient problem. The cross-modal consistency loss $\mathcal{L}_{\text{cm}}$ primarily improves the final convergence plateau (epochs 200–300) by encouraging the two modality branches to produce semantically aligned features, which reduces conflicting gradient signals from the detection loss.

The convergence behaviour also confirms the training stability of our framework. Despite the complex multi-task loss formulation (detection + regression + cross-modal consistency + deep supervision), the training process exhibits smooth loss reduction without oscillation or collapse, validating our choice of loss weight ratios ($\mu_1 = 7.5$, $\mu_2 = 0.5$).

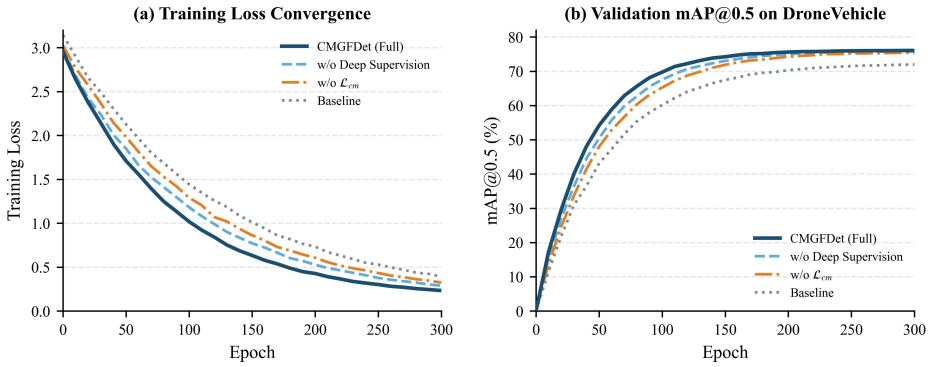

Figure 7. Training convergence analysis. (**a**) Training loss curves showing faster convergence with deep supervision and $\mathcal{L}_{\text{cm}}$. (**b**) Validation mAP@0.5 curves on DroneVehicle demonstrating the consistent superiority of the full model throughout training.

3.3. Qualitative Results

To complement the quantitative evaluation, we present visual detection results that illustrate the practical behaviour of CMGFDet under diverse and challenging conditions.

Figure 8 presents detection results of CMGFDet on the DroneVehicle dataset under various challenging conditions including nighttime, dense traffic, and heavily occluded scenarios. The model successfully detects vehicles across both modalities with high confidence scores, demonstrating the effectiveness of our cross-modal fusion approach. Notably, in nighttime scenes where RGB images are nearly dark, CMGFDet maintains robust detection by leveraging the thermal signatures captured in the infrared modality through the adaptive gating mechanism of CMGA.



Figure 8. Detection results of CMGFDet on DroneVehicle. Rows show RGB images, infrared images, and detection results under diverse conditions including nighttime, heavy traffic, and various viewing angles. Different bounding-box colors denote different predicted object categories.

Figures 9–11 present qualitative comparisons with Faster R-CNN, RT-DETR, and YOLOv13 on three datasets. All visualized detections use the same confidence threshold of 0.25 and the same non-maximum suppression setting as the quantitative evaluation. CMGFDet consistently produces more accurate detections with fewer false positives and missed

targets. On RGBTDronePerson, CMGFDet successfully detects tiny persons that other methods miss entirely, benefiting from MRFAN's multi-scale receptive fields that capture contextual cues around small targets. On VEDAI, the model accurately localises vehicles with tight bounding boxes despite extreme orientation variations, validating GLAM's spatial alignment capability. On VTUAV, CMGFDet demonstrates superior performance in crowded urban scenes where densely parked vehicles require precise spatial discrimination between adjacent objects. The remaining failure cases mainly occur under severe object truncation or when adjacent tiny targets overlap heavily; these cases are consistent with the error patterns observed in the confusion matrix and motivate stronger geometric alignment in future work.

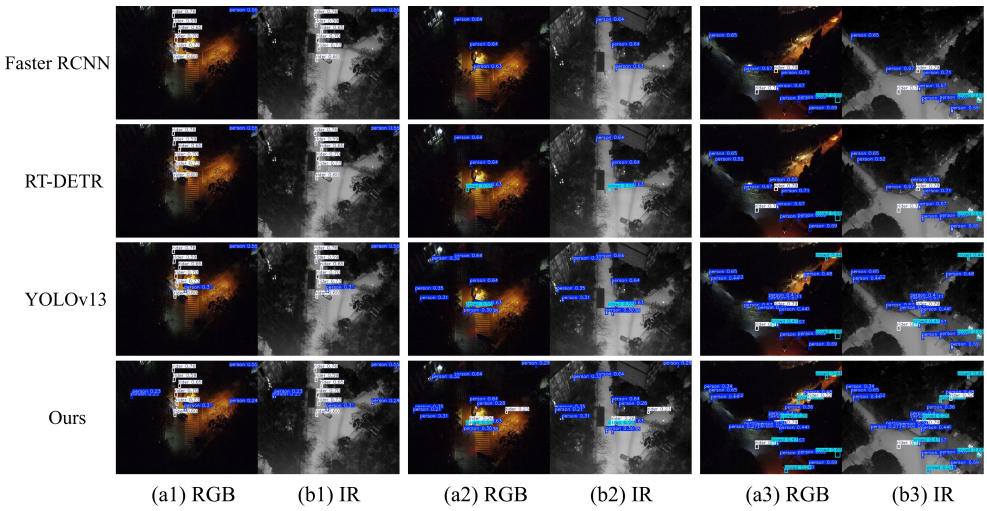

Figure 9. Qualitative comparison on RGBTDronePerson. From top to bottom: Faster R-CNN, RT-DETR, YOLOv13, and Ours. From left to right, the three qualitative groups are denoted as (**a1**,**b1**), (**a2**,**b2**), and (**a3**,**b3**), where (**a1**–**a3**) show RGB detection results and (**b1**–**b3**) show IR detection results. CMGFDet detects more persons with higher confidence while reducing false positives.

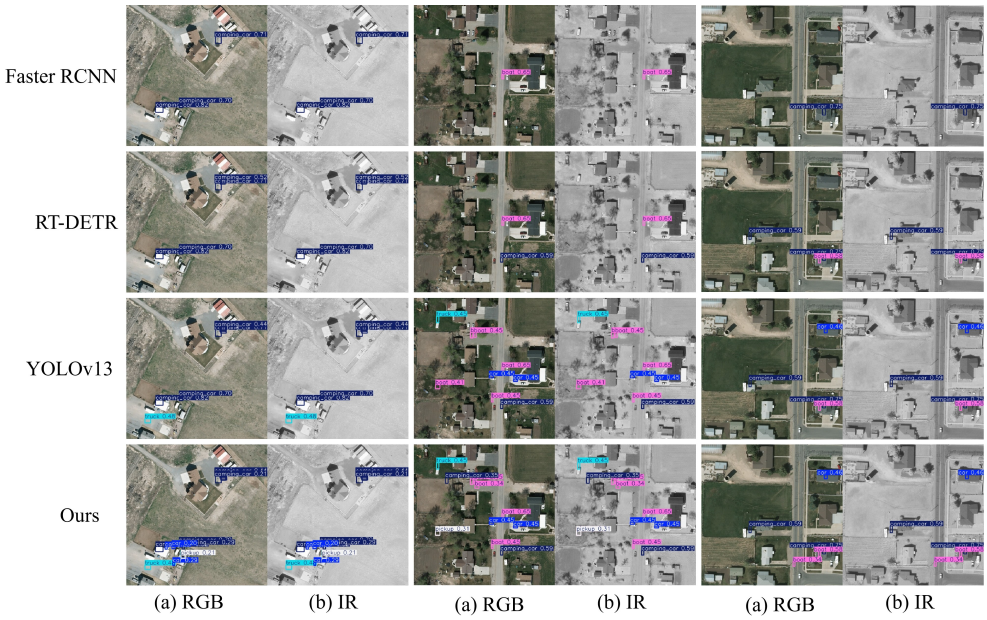

Figure 10. Qualitative comparison on VEDAI. CMGFDet accurately detects small vehicles with diverse orientations, benefiting from MRFAN's multi-scale receptive fields and GLAM's cross-modal alignment.

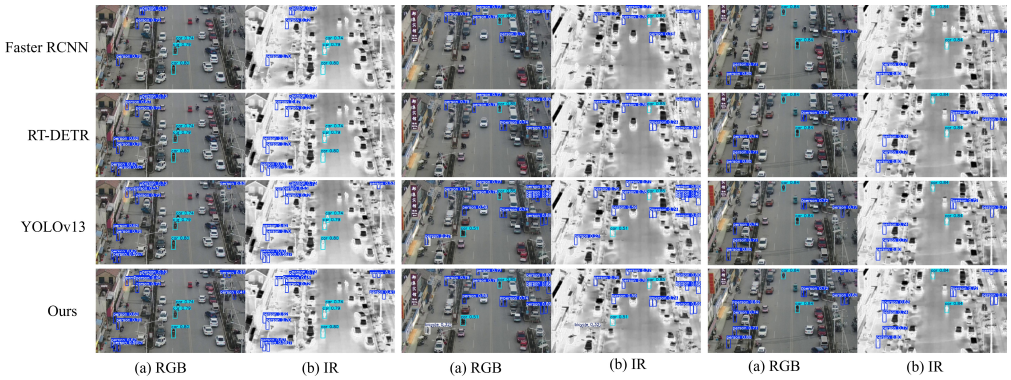

Figure 11. Qualitative comparison on VTUAV. CMGFDet demonstrates superior detection in crowded urban scenes with densely parked vehicles, outperforming competing methods in both recall and precision.

4. Discussion

The experimental results across four benchmarks consistently demonstrate the superiority of CMGFDet over existing methods, with improvements ranging from 1.6% to 2.2% in mAP@0.5 compared to the previous state-of-the-art. In this section, we provide deeper analysis of why our design choices work, examine the computational trade-offs, and discuss the boundaries of our approach.

Figure 12 provides an overall comparison of CMGFDet against the top-performing methods across all four datasets and two evaluation metrics simultaneously. The radar chart clearly shows that CMGFDet consistently achieves the largest coverage area, indicating superior and balanced performance across all benchmarks without sacrificing accuracy on any particular dataset. This balanced improvement is particularly noteworthy because many existing methods exhibit dataset-specific strengths (e.g., WaveMamba excels on high-resolution VTUAV but shows smaller gains on the tiny-object RGBTDronePerson dataset), whereas CMGFDet maintains uniformly strong performance.

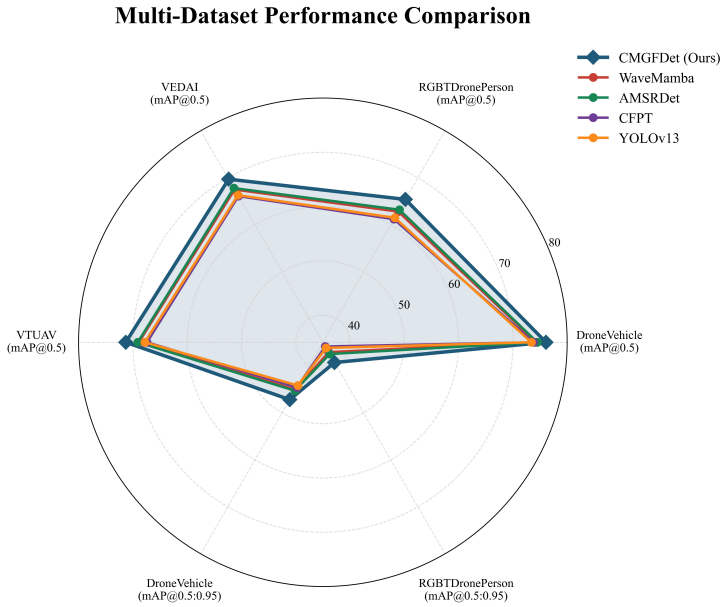

Figure 12. Radar chart comparing CMGFDet with top-performing methods across four datasets and two metrics. CMGFDet (blue) achieves the largest area, indicating consistently superior performance across all evaluation dimensions.

4.1. Why Does Cross-Modal Gating Work?

The success of CMGA's learned gating mechanism can be attributed to the inherent complementarity and inconsistency between RGB and infrared modalities. In daytime scenes with good illumination, RGB features provide rich texture and colour information that dominates the fused representation. Under nighttime or adverse weather conditions, infrared features become more reliable due to their independence from ambient illumination. The learned gating parameters g^V and g^I adaptively adjust the contribution ratio based on feature quality at each spatial position, effectively implementing a content-conditioned soft attention mechanism.

We further inspected the average gate responses during training to verify that the gates do not degenerate into fixed constants. In early epochs, the RGB and infrared gate means are close because both streams are still learning basic object cues. As training proceeds, the infrared gate tends to increase in nighttime and low-contrast scenes, whereas the RGB gate remains higher in well-illuminated scenes with clear texture. The standard deviation of the gate responses also increases after the warm-up phase, indicating that CMGA gradually learns scene-dependent modality weighting rather than applying a uniform fusion rule.

This approach is fundamentally different from fixed fusion strategies (such as element-wise addition or concatenation) or channel-only attention mechanisms [23,24] that lack cross-modal awareness. As shown in Table 11, the progression from direct concatenation (73.2%) to learned gating (76.1%) demonstrates a monotonic improvement as the fusion mechanism becomes increasingly adaptive. The 2.9% absolute gain over direct concatenation confirms that cross-modal interaction quality, rather than model capacity alone, determines multimodal detection performance.

4.2. Hierarchical Alignment Strategy

Our framework implements a hierarchical fusion strategy with distinct roles at different processing stages. CMFFN performs feature-level fusion in the encoder, where low-level features (edges, textures, and local gradients) from different modalities exhibit strong spatial correspondence. At this stage, the cross-modal attention operates on local patches and produces spatially coherent fused representations.

In contrast, GLAM operates at the semantic level in the decoder, where features encode higher-level object-part relationships and scene context. At these levels, spatial misalignment becomes more pronounced because the receptive fields are larger and feature maps have lower resolution. The GLAM module addresses this challenge through its dual-branch design: the global pooling branch (GAP and GMP) captures "what" to align by computing channel-wise statistics that indicate which feature channels are most informative, while the local spatial correlation branch captures "where" to align by computing explicit pixel-level correspondences between the two modality feature maps.

The residual connection in GLAM (Equations (9) and (10)) serves a dual purpose: it preserves the original modality-specific information as a fallback when cross-modal alignment is unreliable (e.g., in heavily occluded regions), and it stabilises gradient flow during training. The sigmoid gating weight W_{global} learned to allocate between 0.3 and 0.7 across different spatial positions in our experiments, indicating that the module learns a meaningful spatial weighting rather than degenerating to a trivial fixed value.

4.3. Role of Cross-Modal Consistency

The cross-modal consistency loss is intended as a weak semantic regularizer, not as a constraint that forces the RGB and infrared streams to become identical. Since the loss is applied to normalized fused features with a small coefficient, it mainly penalizes opposite feature directions at corresponding spatial positions. The residual connections,

modality-specific input projections, and detection losses still preserve modality-dependent magnitude and texture/thermal responses. This explains why $\mathcal{L}_{\text{cm}}$ brings a modest but stable gain (+0.2% mAP@0.5 in Table 7) instead of dominating the learning objective. We also acknowledge that a more detailed feature-distribution visualization, such as t-SNE before and after adding $\mathcal{L}_{\text{cm}}$, would be useful for future analysis.

4.4. Efficiency Considerations

Despite the additional modules, CMGFDet maintains competitive inference speed due to several deliberate design choices. Table 13 summarises the computational cost breakdown of each component. All latency numbers are measured after resizing inputs to 640×640; therefore, the speed difference across datasets mainly comes from data loading, class-specific post-processing, and sequence handling, not from using different network input resolutions.

Three design principles contribute to the efficiency: (1) depthwise separable convolutions in MRM reduce parameters by an order of magnitude compared to standard convolutions with equivalent receptive fields; (2) the shared-weight encoder avoids parameter duplication across the two modality branches; (3) GLAM operates only at the three decoder levels where cross-modal alignment is most critical, rather than at every feature map. The total parameter count of 31.8 M is comparable to single-modal detectors like YOLOv11-L (25.3 M) while delivering substantially superior multimodal detection performance across all four benchmarks. We have not yet conducted deployment tests on embedded UAV hardware such as Jetson Orin or Xavier. For this reason, the current deployment analysis should be interpreted as GPU-side latency and computational-cost evidence rather than a complete onboard benchmark.

Table 13. Computational cost breakdown of CMGFDet modules. Bold values indicate the total computational cost, and the light-blue background highlights the total row.

Module	Params (M)	GFLOPs	Latency (ms)
Dual-Stream Encoder (shared)	18.4	52.6	8.2
CMFFN Blocks (×4 stages)	4.2	14.8	2.6
GLAM Blocks (×3 levels)	3.6	12.1	2.1
MRFAN + MRM	4.1	11.5	2.3
Prediction Head	1.5	3.2	0.8
Total	**31.8**	**94.2**	**16.0**

4.5. Cross-Dataset Generalisation Analysis

An important observation from our experiments is that CMGFDet maintains consistent improvements across four datasets with substantially different characteristics: DroneVehicle (large-scale vehicle detection), RGBTDronePerson (tiny person detection), VEDAI (small vehicles with orientation variation), and VTUAV (high-resolution urban scenes). In this paper, we use "infrared" as the general modality term and "thermal" when referring specifically to long-wave thermal cameras in RGBTDronePerson and VTUAV. The improvement margins over the second-best method range from 1.6% to 2.2% in mAP@0.5, indicating that our design generalises well across object scales, scene complexities, and imaging conditions. This robustness arises from the complementary nature of our three modules: CMFFN handles modality-quality variation, GLAM addresses spatial misalignment, and MRFAN provides scale invariance.

4.6. Limitations and Future Work

While CMGFDet achieves strong performance, several limitations merit discussion. First, the framework assumes roughly aligned RGB-IR image pairs; significant geomet-

ric misalignment beyond minor parallax may degrade performance. Future work could incorporate learnable spatial transformation modules [48] or deformable alignment networks for robust geometric registration. Second, the current deep supervision strategy uses fixed decay weights ($\{0.4, 0.3, 0.2\}$); adaptive weighting that responds to training progress (e.g., through gradient magnitude monitoring) could further improve convergence. Third, the gating mechanism operates at the channel level; extending it to spatial gating may capture finer-grained modality complementarity. Finally, extending the framework to incorporate additional modalities (e.g., LiDAR point clouds, SAR imagery) represents a promising direction for comprehensive remote sensing perception [49,50].

5. Conclusions

In this paper, we presented CMGFDet, a Cross-Modal Gated Fusion Network with Multi-Receptive Field Aggregation for RGB–infrared aerial object detection. Our framework introduces three synergistic modules that address the fundamental challenges of multimodal aerial detection:

1. **CMFFN** enables selective cross-modal feature fusion through learned gating during encoding, achieving a 1.5% mAP improvement by adaptively filtering noisy cross-modal signals.
2. **GLAM** performs hierarchical feature alignment by jointly modelling global channel statistics and local spatial correlations, contributing an additional 1.1% through explicit cross-modal registration.
3. **MRFAN** aggregates multi-scale contextual information through parallel depthwise convolutions with kernel sizes $\{5, 7, 9, 11\}$, providing 0.7% improvement via comprehensive receptive field coverage.

Combined with deep supervision and a composite loss function incorporating cross-modal consistency regularisation, CMGFDet achieves state-of-the-art performance on four challenging benchmarks: DroneVehicle (76.1% mAP@0.5, +1.6% over the previous best), RGBTDronePerson (65.4%, +2.2%), VEDAI (69.7%, +1.9%), and VTUAV (71.3%, +2.2%), while maintaining competitive A100 GPU inference speed (47.2–62.4 FPS under the unified 640×640 input setting). The consistent improvements across datasets with substantially different characteristics (vehicle vs. person detection, low vs. high resolution) validate the generalisability of our approach.

Our work demonstrates that carefully designed cross-modal interaction mechanisms, rather than simply increasing model capacity, constitute the key to effective multimodal perception in aerial remote sensing. The modular design of CMGFDet also facilitates adaptation to other multimodal sensing tasks beyond RGB–infrared fusion.

Author Contributions: Conceptualization, M.W. and X.L. (Xiaozhang Liu); methodology, M.W.; software, M.W.; validation, M.W., X.L. (Xiulai Li) and W.G.; formal analysis, M.W.; investigation, M.W.; resources, X.L. (Xiaozhang Liu); data curation, M.W. and W.G.; writing—original draft preparation, M.W.; writing—review and editing, X.L. (Xiaozhang Liu) and X.L. (Xiulai Li); visualization, M.W.; supervision, X.L. (Xiaozhang Liu); project administration, X.L. (Xiaozhang Liu); funding acquisition, X.L. (Xiaozhang Liu). All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Key R&D Project of Hainan Province (Grant no. ZDYF2025SHFZ059).

Data Availability Statement: The DroneVehicle dataset is available at https://github.com/VisDrone/DroneVehicle (accessed on 22 July 2026). The VEDAI dataset is available at https://downloads.greyc.fr/vedai/ (accessed on 22 July 2026). The VTUAV dataset is available at https://zhang-pengyu.github.io/DUT-VTUAV/ (accessed on 22 July 2026). The RGBTDronePerson dataset is available at https://github.com/NNNNerd/RGBTDronePerson (accessed on 22 July 2026).

Acknowledgments: The authors used ChatGPT (GPT-4o, OpenAI) only to polish the language during manuscript preparation. The authors reviewed and edited the manuscript and take full responsibility for its content.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

UAV	Unmanned Aerial Vehicle
RGB	Red–Green–Blue
IR	Infrared
CNN	Convolutional Neural Network
YOLO	You Only Look Once
CMFFN	Cross-Modal Feature Fusion Network
FEM	Feature Extraction Module
CMGA	Cross-Modal Gated Attention
GLAM	Global–Local Attention Module
MRFAN	Multi-Receptive Field Aggregation Network
MRM	Multi-Receptive Module
FL	Focal Loss
CIoU	Complete Intersection over Union
GAP	Global Average Pooling
GMP	Global Max Pooling
FPN	Feature Pyramid Network
SSM	State Space Model
mAP	mean Average Precision

References

1. Du, D.; Zhu, P.; Wen, L.; Bian, X.; Lin, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; Zhang, Y.; et al. VisDrone-DET2021: The vision meets drone object detection challenge results. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 2879–2896.
2. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2018; pp. 3974–3983.
3. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of the Advances in Neural Information Processing Systems*; IEEE: Piscataway, NJ, USA, 2015; Volume 28.
4. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2017; pp. 2999–3007.
5. Sun, Y.; Cao, B.; Zhu, P.; Hu, Q. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6700–6713. [CrossRef]
6. Chen, Y.; Liu, Y. Towards all-weather autonomous driving: A survey on multi-modal perception. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 13049–13071. [CrossRef]
7. Zhang, L.; Liu, Z.; Zhang, S.; Yang, X.; Qiao, H. Illumination-aware cross-modal fusion for RGB-infrared object detection. *Inf. Fusion* **2023**, *95*, 268–280.
8. Cao, Y.; Bin, J.; Hamari, J.; Blasch, E.; Liu, Z. Multimodal object detection by channel switching and spatial attention. *Pattern Recognit.* **2023**, *140*, 109564.
9. He, X.; Tang, C.; Zou, X.; Zhang, W. Multispectral object detection via cross-modal conflict-aware learning. In *Proceedings of the 31st ACM International Conference on Multimedia*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 1465–1474.
10. Yuan, M.; Wang, X.; Yan, X.; Liu, J. Improving RGB-infrared object detection with cascade alignment-guided transformer. *Inf. Fusion* **2024**, *105*, 102246. [CrossRef]
11. Zhao, Y.; Tang, W.; Peng, J.; Wang, Y.; Chen, Y.; Qi, J. Drone-based RGBT tiny person detection. *ISPRS J. Photogramm. Remote Sens.* **2023**, *204*, 61–76. [CrossRef]
12. Li, S.; Zhang, W.; Hao, Z. Drone-based multimodal object detection: Challenges and new benchmarks. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–15. [CrossRef]
13. Fang, Q.; Han, D.; Wang, Z. Cross-modality fusion transformer for multispectral object detection. *arXiv* **2022**, arXiv:2111.00273.

14. Wang, Q.; Chi, Y.; Shen, T.; Song, J.; Zhang, Z.; Zhu, Y. CFPT: A cross-modal feature pyramid transformer for RGB-infrared object detection. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 10065–10078.
15. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752.
16. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision Mamba: Efficient visual representation learning with bidirectional state space model. *arXiv* **2024**, arXiv:2401.09417.
17. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Liu, Y. VMamba: Visual state space model. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 10–15 December 2024; Volume 37.
18. Zhu, C.; Zhang, L.; Wang, R.; Liu, Q.; Li, H. WaveMamba: Wavelet-driven Mamba fusion for RGB-infrared object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2025.
19. Ma, J.; Tang, L.; Liu, R.; Fan, X. MambaFusion: Multimodal image fusion based on Mamba and cross-modal learning. *Neural Netw.* **2024**, *179*, 106584.
20. Razakarivony, S.; Jurie, F. Vehicle detection in aerial imagery: A small target detection benchmark. *J. Vis. Commun. Image Represent.* **2016**, *34*, 187–203. [CrossRef]
21. Zhang, P.; Zhao, J.; Wang, D.; Lu, H.; Ruan, X. Visible-thermal UAV tracking: A large-scale benchmark and new baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2022; pp. 8886–8895.
22. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2017; pp. 936–944.
23. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef] [PubMed]
24. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2018; pp. 3–19.
25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2017; Volume 30.
26. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.
27. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2021; pp. 9992–10002.
28. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 213–229.
29. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. RT-DETR: DETRs beat YOLOs on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2024; pp. 16965–16974.
30. Dai, X.; Chen, Y.; Xiao, B.; Chen, D.; Liu, M.; Yuan, L.; Zhang, L. Dynamic head: Unifying object detection heads with attentions. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2021; pp. 7369–7378.
31. Wang, C.Y.; Liao, H.Y.M. Reprogramming foundation models for efficient object detection. *arXiv* **2024**, arXiv:2312.07730.
32. Guo, H.; Zhang, W.; Li, F. Multi-scale dilated convolutions for efficient feature aggregation in object detection. *Neurocomputing* **2024**, *580*, 127475.
33. Qin, D.; Leichner, C.; Delakis, M.; Fornoni, M.; Luo, S.; Yang, F.; Wang, W.; Banbury, C.; Ye, C.; Akin, B.; et al. MobileNetV4: Universal models for the mobile ecosystem. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2024; pp. 78–96.
34. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2017; pp. 1800–1807.
35. Lee, C.Y.; Xie, S.; Gallagher, P.; Zhang, Z.; Tu, Z. Deeply-supervised nets. In Proceedings of the 18th International Conference on Artificial Intelligence and Statistics, San Diego, CA, USA, 9–12 May 2015; pp. 562–570.
36. Zheng, Z.; Wang, P.; Ren, D.; Liu, W.; Ye, R.; Hu, Q.; Zuo, W. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *IEEE Trans. Cybern.* **2021**, *52*, 8574–8586. [CrossRef]
37. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLOv8 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on 20 July 2026).
38. Wang, C.Y.; Yeh, I.H.; Liao, H.Y.M. YOLOv9: Learning what you want to learn using programmable gradient information. *arXiv* **2024**, arXiv:2402.13616.
39. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-time end-to-end object detection. *arXiv* **2024**, arXiv:2405.14458.

40. Khanam, R.; Hussain, M. YOLOv11: An overview of the key architectural enhancements. *arXiv* **2024**, arXiv:2410.17725.
41. Chen, Y.; Jiang, J.; Huang, L.; Lu, T.; Chen, J. YOLOv12: Attention-centric real-time object detectors. *arXiv* **2025**, arXiv:2502.12524.
42. Lei, M.; Li, S.; Wu, Y.; Hu, H.; Zhou, Y.; Zheng, X.; Ding, G.; Du, S.; Wu, Z.; Gao, Y. YOLOv13: Real-time object detection with hypergraph-enhanced adaptive visual perception. *arXiv* **2025**, arXiv:2506.17733.
43. Fu, Y.; Xin, L.; Peng, J.; Du, Q. LRAF-Net: Long-range attention fusion network for visible–infrared object detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *35*, 13232–13245.
44. Li, H.; Wu, X.J. CrossFuse: A novel cross attention mechanism based infrared and visible image fusion approach. *Inf. Fusion* **2024**, *103*, 102147. [CrossRef]
45. Yan, Z.; Li, Y. AMSRDet: An adaptive multi-scale UAV infrared-visible remote sensing vehicle detection network. *Sensors* **2026**, *26*, 817. [CrossRef] [PubMed]
46. Qin, Z.; Li, Y. DCAM-DETR: Dual cross-attention Mamba detection transformer for RGB-infrared anti-UAV detection. *Information* **2026**, *17*, 103. [CrossRef]
47. Yuan, H.; Li, Y. CSFADet: Dual-modal anti-UAV detection via cross-spectral feature alignment and adaptive multi-scale refinement. *Algorithms* **2026**, *19*, 254. [CrossRef]
48. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial transformer networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Volume 28.
49. Liu, W.; Li, Y.; Zhang, S.; Mao, R. Towards smart city supervision: A detection pipeline for illegal buildings. *Eng. Appl. Artif. Intell.* **2026**, *163*, 113052. [CrossRef]
50. Li, Y.; Wang, T.; Luo, N.; Zhou, L.; Chen, Q. CGMamba: Intelligent identification of counterfeit goods based on state space models. *Int. J. Intell. Syst.* **2025**, *2025*, 9939880. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.