

Article

DMAC-Net: Direction-Aware Multi-Granularity Enhancement with Asymmetric Context Guidance for Multimodal UAV-Based Small Object Detection

Qing Cheng ^{1,2,†}, Yan Jiang ^{1,†}, Yuan Gao ^{3,*}, Zeng Gao ³, Su Liu ¹ and Xiaoguang Tu ³

¹ College of Air Traffic Control, Civil Aviation Flight University of China, Chengdu 641400, China; chengqing@cafuc.edu.cn (Q.C.); yanj@cafuc.edu.cn (Y.J.)

² Low-Altitude Operation Intelligent Control Research Center, Chengdu 641400, China

³ College of Aviation Electronic and Electrical Engineering, Civil Aviation Flight University of China, Chengdu 641400, China

* Correspondence: gaoyuan@cafuc.edu.cn

† These authors contributed equally to this work.

Abstract

In complex UAV aerial scenes, small object detection tasks face challenges such as extremely low pixel occupancy, strong background interference, and sparse effective features, which are further compounded by environmental factors like low illumination. Consequently, single-modality detection algorithms are prone to severe target feature loss and missed detections. Multi-modal image fusion, which complements the texture details of visible light with the thermal radiation characteristics of infrared, is considered an effective approach to overcome the limitations of single physical imaging. However, conventional fusion mechanisms often suffer from semantic gaps when processing heterogeneous data, easily introducing redundant noise and background false alarms. To further improve the accuracy and robustness of small object detection in UAV aerial scenes, this paper proposes a multi-modal detection network that integrates direction-aware multi-granularity and asymmetric context guidance, termed DMAC-Net. Specifically, a Direction-Aware Granularity Enhancement (DAGE) module is first constructed for unified backbone feature extraction, which captures local directions and contour edges of small objects in UAV aerial images with high sensitivity, and expands the receptive field through a multi-granularity mechanism, effectively suppressing false positives induced by complex backgrounds while enhancing the recall of occluded and weakly featured targets. Additionally, the Asymmetric Context Guided Fusion (ACGF) module builds a spatial mechanism via asymmetric receptive fields and performs semantic soft alignment of cross-modal features with dynamic weight assignment, effectively filtering out artifacts and clutter from cross-modal interaction. Experimental results on multiple aerial datasets, including RGBTDronePerson, AVMS and LLVIP demonstrate that the proposed method outperforms existing mainstream models in terms of overall detection accuracy and missed-detection suppression, while exhibiting strong generalization capability and stability under complex lighting transitions and multi-scale variations in UAV monitoring environments.

Keywords: UAV multi-modal imagery; direction-aware; multi-granularity enhancement; asymmetric context guidance; small object detection



Academic Editor: Fabio D'Andreagiovanni

Received: 18 June 2026

Revised: 19 July 2026

Accepted: 30 July 2026

Published: 1 August 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.

1. Introduction

In recent years, the vigorous development of the low-altitude economy has driven the rapid popularization of UAV aerial imaging technology. UAV-based micro-object detection has been widely applied across various domains, including remote sensing mapping [1], agricultural plant protection [2], military reconnaissance [3], and inspection monitoring [4], making research on object detection algorithms from UAV perspectives a hot topic. Concurrently, integrating advanced artificial intelligence models into UAV onboard sensors has emerged as an important trend [5]. Recent studies have demonstrated that large language models and AI models exhibit transformative potential in empowering UAV applications, ranging from adaptive aerial decision-making in resource-constrained environments [6] to domain-specific visual perception tasks—for example, leveraging LLM-based agents to optimize advanced object detectors for complex UAV inspection missions. Inspired by these advances, efficient and robust perception models remain the cornerstone for tasks in dynamic low-altitude economy scenarios. Returning to the task itself, small object detection under UAV aerial scenes inherently encounters challenges such as large target quantities, complex background interference, and sparse effective features. Particularly under low illumination and high-noise conditions, existing segmentation and detection algorithms often suffer from severe target feature loss when performing remote sensing tasks from UAV perspectives. Moreover, the receptive field of small targets in UAV views tends to be more complex, and the pixel occupancy of targets is even smaller (typically less than 32×32 pixels), which further exacerbates the difficulty of feature extraction, leaving substantial room for improving the engineering recognition accuracy and robustness of current algorithms [7–10].

Most of these detection networks are designed for single-modal imagery, such as visible or infrared [11–13]. Among existing single-modal imaging technologies, visible images have become the most widely used data source due to their rich color and detail information. However, their inherent imaging principle makes them sensitive to environmental conditions. In practical engineering scenarios such as UAV inspection and field reconnaissance, diurnal variations in illumination intensity and stochastic fluctuations in weather conditions (e.g., low light) are unavoidable. These factors can lead to overexposure, underexposure, blurring, and other issues in visible images, significantly reducing the reliability of object detection. Infrared thermal imaging technology, on the other hand, overcomes illumination dependency and enables all-weather target perception, but it can only reflect the thermal radiation distribution of targets and cannot meet the requirements of tasks that demand fine-grained classification and recognition. The texture, color, and detail features that are crucial in visible images are precisely what infrared images lack. Compared with single-modal image detection, infrared-visible fusion detection exhibits stronger robustness and detection accuracy in complex receptive fields, multi-scale targets, and dense multi-object scenarios [14].

At present, the vast majority of mainstream fusion studies adopt feature-level fusion strategies [15,16], mainly including decision-level fusion and simple feature fusion. Decision-level fusion only performs modal interaction at the decision stage, neglecting modal correlations during feature extraction, which results in substantial waste of valuable complementary information and leads to sharp performance degradation when a single modality fails. Although simple feature fusion advances modal interaction to the feature stage, it relies only on simple operations such as concatenation (Concat) or element-wise addition (Add). Concat indiscriminately preserves all features, introducing a large amount of task-irrelevant redundant noise. Add aligns the feature distributions of different modalities and emphasizes pixel-wise superposition of targets, which may disrupt the semantic independence of each modality and cause feature information loss. Both methods lack

a dynamic evaluation mechanism for modality importance and cannot adaptively fuse the most effective features according to different scenes. In addition, there is another category of detection methods that separate pixel-level fusion in the image preprocessing stage from the subsequent detection task, fusing multi-modal images into a single image before feeding them into the detector. Such fusion tasks may result in upstream fusion requirements that are not fully compatible with the downstream object detection task, and it is also difficult to achieve deep feature complementarity between modalities.

To address the challenges of sparse small-target features from UAV perspectives, limited single-modal perception boundaries, and the susceptibility to redundant noise from naive concatenation-based fusion, this paper proposes a multi-modal detection network termed Direction-aware Multi-granularity and Asymmetric Context Network (DMAC-Net). This network enhances detection robustness under complex lighting transitions and highly dynamic scenes by performing fine-grained directional enhancement and cross-scale semantic reconstruction of infrared and visible features. The detection performance of single-modal baseline models and the DMAC-Net multi-modal model under low-light conditions is illustrated in Figure 1.

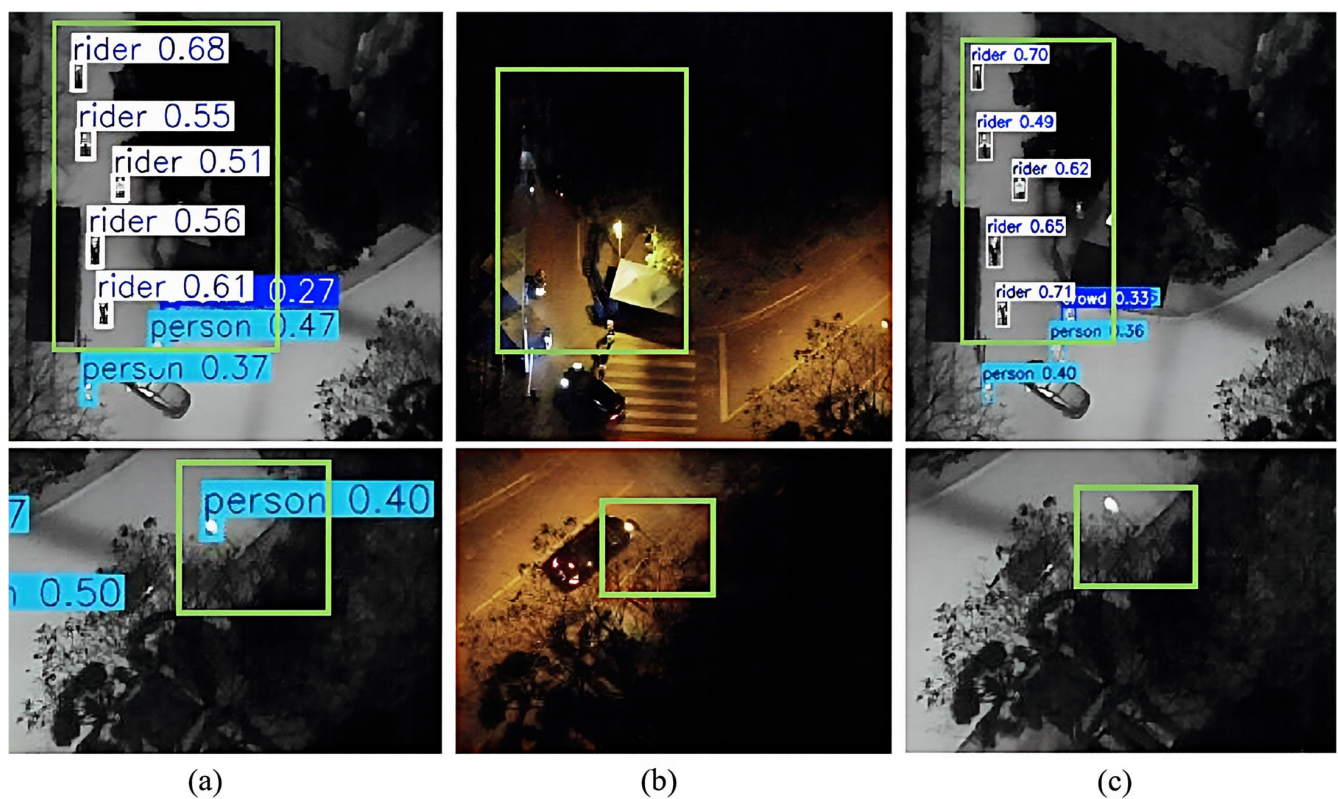


Figure 1. Comparison of low-light detection performance between single-modality baselines and the proposed DMAC-Net multimodal model. For each scene, columns (a,b) show the detection results of the infrared and visible modalities, respectively, using the YOLOv11 single-modality branch. Column (c) presents the results of the DMAC-Net multimodal fusion model proposed in this work. The green box indicates the main target area for comparison. The modalities are arranged by scene, demonstrating mutual complementarity—infrared supplies thermal signatures when visible fails, and visible provides spatial details to suppress thermal-induced false alarms.

As depicted in the figure, under complex aerial environments with low illumination and high noise, the infrared and visible modalities demonstrate distinct yet complementary characteristics. Infrared is capable of recovering targets missed by the visible spectrum through thermal radiation, whereas visible light can rectify false positives arising from thermal source interference via texture and detail information. Consequently, for UAV aerial scenarios characterized by low light and strong interference, the multi-source visible-infrared data fusion achieved by DMAC-Net—which is specifically redesigned and enhanced based on task characteristics—provides a reasonable and favorable solution for overcoming the inherent detection bottlenecks.

The main contributions of this paper are summarized as follows:

1. To address the unique physical challenges inherent in complex aerial environments, we construct a multimodal detection network that integrates direction-aware multi-granularity and asymmetric context guidance, termed DMAC-Net, which provides a deep coupling paradigm for small object detection in all-weather scenarios.
2. A Direction-Aware Multi-Granularity Enhancement (DAGE) module is designed for unified backbone feature extraction. By capturing local orientations and edge contours of targets to suppress background false alarms, and enlarging the receptive field via a multi-granularity mechanism, it effectively improves the detection performance of dense and feature-sparse small targets.
3. The Asymmetric Context Guided Fusion (ACGF) module is constructed to achieve semantic alignment and complementarity between infrared thermal radiation priors and visible geometric details via asymmetric feature guidance and dynamic weight assignment, thereby enabling efficient cross-modal feature interaction.
4. The effectiveness of the proposed model is validated across multiple multimodal datasets. Extensive experiments show that DMAC-Net achieves superior performance not only on UAV-view tasks including RGBTDronePerson and AVMS, but also maintains high-precision localization and robustness in extreme lighting and low-altitude near-ground wide-area surveillance scenarios, as demonstrated on the LLVIP dataset.

2. Related Work

As one of the core tasks in computer vision and remote sensing, UAV-based aerial small object detection aims to accurately locate and identify target entities with low pixel occupancy from images characterized by complex backgrounds, illumination variations, and severe occlusion. Small objects often lack sufficient shape, texture, and other detailed features, making detection performance highly sensitive to environmental noise. Therefore, efficient small object detection holds irreplaceable practical significance in real-world applications such as UAV monitoring, natural exploration, traffic surveillance, and autonomous driving [17–19].

In recent years, extensive research has been devoted to breaking through the bottlenecks of single-modal aerial small object detection, particularly in mitigating feature degradation and addressing physical perception blind spots. To tackle the challenge of weak small-object features, early studies primarily relied on data augmentation and multi-scale feature fusion. The “Copy-Paste” strategy proposed by Kisantal et al. [20] enriches feature information by copying and pasting small objects at multiple locations in images to increase their training frequency. However, such stitching operations, if applied indiscriminately without directional awareness, are prone to introducing edge artifacts that interfere with detection performance. Building on this, the Stitcher mechanism [21] dynamically stitches images of different scales to balance the network’s response to multi-scale targets, which poses greater challenges for optimizing data distribution and balancing feature extraction capability for weak targets. Combining high-resolution shallow spatial details

with deep high-level semantics remains the most mainstream solution. The classic FPN [22] and its subsequent variants are widely used to construct feature pyramids. However, most conventional extraction networks lack targeted perception of local directions, and during the lengthy hierarchical propagation process, the already weak semantic information of single-modal small objects is easily overwhelmed or diluted by cumulatively increasing background noise. To address the high computational cost of small object detection in high-resolution images, Yang et al. proposed QueryDet [23], which designs a cascaded sparse query mechanism to significantly accelerate inference, though its capability in capturing fine-grained edge features still leaves room for improvement. Methods based on context modeling and attention mechanisms [24] strive to suppress complex background noise and enhance feature representation of weak targets. SMA-YOLO [25] introduces a non-semantic sparse attention (NSSA) module and a bidirectional multi-branch auxiliary feature pyramid to extract low-level non-semantic features by suppressing global over-semanticization, thereby improving model sensitivity to small objects. However, as pointed out in that study, single-stage networks such as MS-YOLOv7 and Drone-YOLO, when adding extra detection heads or complex feature layers, tend to encounter information interaction conflicts in dense scenes. Particularly under extreme conditions such as night or heavy fog, detection relying solely on the visible modality is prone to near-failure due to the lack of low-level signals.

To break the perceptual blind spots of single modalities, recent research has gradually shifted toward multi-modal fusion. LF-DETR [26] achieves high-precision dual-modal target localization through the DETR architecture, but its large parameter count and high computational overhead make it unsuitable for deployment on UAV platforms, and its robustness in handling spatially misaligned multi-modal sensors still requires substantial improvement. Building on this, DEYOLO [27] designs a dual-enhancement mechanism and a bidirectional decoupled focus module within the CNN framework to integrate complex feature fusion mechanisms for enhanced cross-modal representation. Notably, most mainstream feature-level fusion networks, including DEYOLO, lack effective direction-aware and enhancement mechanisms specifically optimized for infrared thermal contours and visible edge textures of tiny targets from UAV perspectives. Such feature extraction mechanisms remain susceptible to cross-feature interference under complex backgrounds. Meanwhile, the successful application of multi-branch YOLO frameworks further demonstrates that the highly modular design of this detection series is well-suited for extending to multi-modal inputs. In exploring more efficient cross-modal interactions, to alleviate the high spatial complexity caused by stacking multiple Transformer blocks, ICAFusion [28] proposes an iterative cross-attention guided feature enhancement (ICFE) module that leverages shared parameters between blocks and incorporates global information to enhance modal features. However, when handling aerial tiny objects, its preceding spatial feature contraction operation tends to compress feature maps prematurely, leading to loss of the already weak geometric details of small objects and posing challenges for maintaining modal features in deeper networks. Furthermore, SuperYOLO [29], which adopts pixel-level compact fusion, and PSFusion [30], which serves high-level tasks through progressive semantic injection, both employ symmetric equal-weight fusion strategies. This raises the question of decoupling the asymmetric relationship between the detail advantages of visible light and the contour advantages of infrared in terms of their physical properties. While intuitive, such strategies struggle to fully accommodate the asymmetric differences between heterogeneous modalities in physical attributes (e.g., visible-light detail advantages versus infrared contour advantages). Under complex UAV-view scenes, such undifferentiated fusion not only easily causes receptive field mismatches of heterogeneous features and triggers modal conflicts, but also leads to degradation as the semantics of tiny objects are

submerged again by stacked background clutter during propagation to deeper layers of the network [31–36]. In summary, existing research still has substantial room for improvement in constructing fusion architectures that jointly address multi-modal feature enhancement and extraction backbone, mitigate deep semantic degradation, and design cross-modal feature interaction mechanisms [37–40].

Based on the above analysis of the limitations of existing algorithms in direction perception and heterogeneous fusion strategies, this paper constructs the DMAC-Net detection network, with improvements and reconstructions targeting both the feature extraction paradigm and cross-modal feature interaction strategies. In the feature extraction stage, the Direction-Aware Granularity Enhancement (DAGE) module is introduced to capture and enhance edge textures and thermal source contours of tiny objects through direction-aware mechanisms, thereby reinforcing target features. Additionally, through multi-granularity cross-stage connections, shallow spatial details are adaptively reconstructed and injected into deep high-level semantics. In the feature fusion stage, the traditional undifferentiated concatenation module is replaced with the Asymmetric Context Guided Fusion (ACGF) module, which constructs distinct feature extraction patterns based on the different advantages of each modality. By leveraging asymmetric dynamic decoupling, alignment, and fusion of multi-modal features, a logical cross-modal feature interaction mechanism is formed, effectively avoiding modal conflicts and information redundancy between visible and infrared modalities.

3. Methods

The proposed model is built upon the mainstream YOLOv11 single-modal object detection architecture [41], which consists of three components: the backbone, the neck, and the head. The backbone is responsible for feature extraction and is primarily composed of modules such as SPPF (Spatial Pyramid Pooling Fast), C2PSA, and C3K2. The neck network mainly comprises C3K2 modules and convolutional layers. The head consists of three Detect detection heads, which produce the final detection outputs.

3.1. Overall Model Architecture

To meet the practical requirements of visible and infrared dual-modal small object detection tasks in UAV aerial scenarios, the YOLOv11 detection architecture, with its multi-scale feature extraction and complex spatial context capture capabilities, demonstrates significant potential for small object processing and multi-scale feature representation. Built upon the YOLOv11 single-branch detection framework, DMAC-Net is tailored to the task characteristics and further performs systematic architectural reform of the original architecture from three perspectives: structural expansion of the backbone network, cross-modal information interaction, and multi-scale feature enhancement in the neck. This includes the redesign and reconstruction of core extraction and fusion modules, such as C3K2, Concat, and downsampling convolutions. The overall architecture is illustrated in Figure 2.

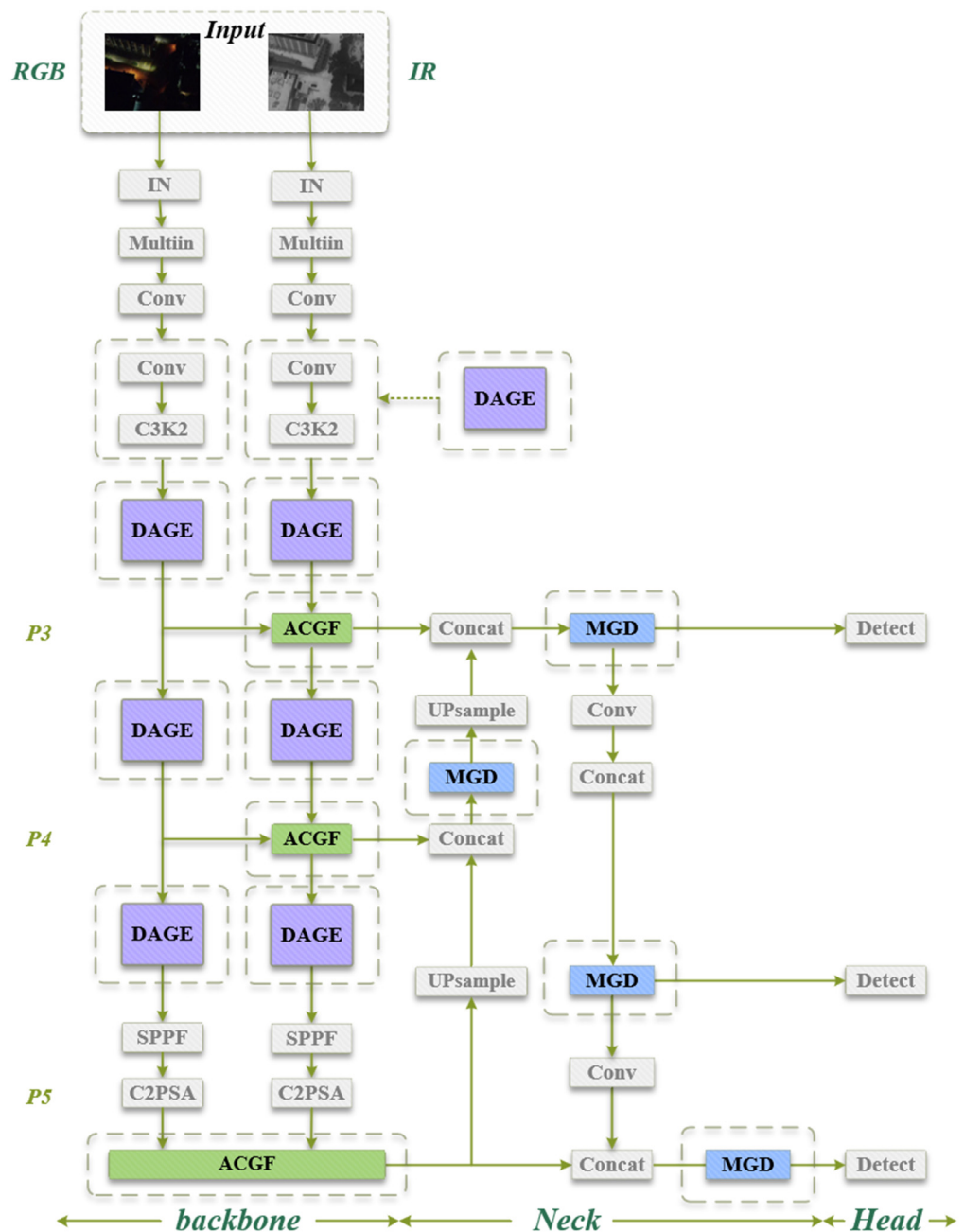


Figure 2. Architecture of the proposed DMAC-Net for multimodal remote sensing small object detection. The architecture mainly consists of three parts: a dual-branch backbone, a neck, and a detection head. The dark modules represent the original components, while the combination of bright-colored modules and dashed boxes indicates the improved modules designed in this study.

In terms of backbone architecture design, the original single-branch backbone is transformed into two structurally symmetric, parameter-independent feature extraction pathways, which accept visible and infrared image inputs, respectively. The feature streams of both branches first pass through **IN** and **Multiin** modules for feature map reception and downsampling. Then, prior to cross-modal feature fusion, the dual-branch feature maps undergo progressive multi-scale feature enhancement and extraction through **DAGE** modules at each layer. Regarding the cross-modal feature fusion mechanism, the **ACGF** modules are embedded at three semantic levels (**P3**, **P4**, **P5**) of the backbone to promote deep interaction of cross-modal features in a progressive manner. After joint concatenation and global channel attention refinement, the fused results are written back via asymmetric

residual connections, forming a cross-modal enhanced representation that uses target contours as anchors and environmental context as supplementation. Finally, the dual-branch feature maps, after feature extraction and fusion, pass through SPPF and C2PSA modules to complete spatial multi-scale aggregation and channel attention enhancement, thus accomplishing the multi-scale feature extraction and fusion operations in the backbone stage. In particular, the above systematic architectural reform distinguishes itself from simple module stacking, representing an adaptive deep architectural reconfiguration tailored to the inherent physical properties of multi-source imagery captured from UAV sensor perspectives. On one hand, the DAGE unit in the backbone deeply couples direction-aware and multi-granularity mechanisms, breaking the conventional logic of general-purpose feature extraction and establishing domain-specific priors customized for aerial small-object detection. On the other hand, the ACGF module abandons the conventional mindset of equal-weight feature concatenation and dynamically filters visible semantics using infrared anchors. This asymmetric guidance mechanism reconstructs the underlying logic and fusion paradigm of cross-modal information interaction at the fusion level.

In the neck structure, to further perform secondary mining of the multi-level features transmitted from the backbone, the proposed architecture replaces the original C3K2 modules in the neck with the MGD modules from the DAGE unit. While maintaining the spatial resolution of the feature maps, this replacement expands the effective receptive field coverage of individual modules and conducts multi-dimensional secondary context mining on the fused features, thereby further enhancing the target features delivered to the detection heads.

In the head detection output stage, the three-scale feature maps produced by the neck network are respectively fed into the corresponding Detect detection heads to perform bounding box regression and target category prediction, enabling joint detection of aerial targets at different scales.

3.2. Direction-Aware Granularity Enhancement Block (DAGE)

In multi-small-object detection tasks from UAV perspectives, the feature enhancement and extraction capability of the network directly determines the model's perception accuracy for multi-scale targets. Although traditional feature enhancement modules effectively alleviate the gradient vanishing problem in deep networks through cross-stage partitioning and merging mechanisms, their internal bottleneck layers have fixed receptive fields. In the absence of directional sensitivity, it is also difficult for them to simultaneously capture local texture details and wide-area contextual semantics.

To address the challenge that a single module cannot balance fine-grained geometric direction perception and global context aggregation, this study uniformly models the directional geometric information of aerial small objects and multi-granularity aggregation, and designs a Direction-Aware Granularity Enhancement (DAGE) block, whose computational mechanism is illustrated in Figure 3.

Which employs a four-directional receptive field to perform rotational encoding on the input features, enhancing responses to small-object edges and fine-grained directional textures. Subsequently, within a cross-stage partial connection framework, the module integrates multi-scale dilated convolution with the concept of cross-stage local connection, reconstructing the C3K2 module into a Multi-Granularity Dilation Aggregator (MGD). Through multi-branch dilated convolutions, MGD performs expansive modeling of features, capturing local details, medium-range structures, and wide-area semantic information in parallel. Finally, cross-stage aggregation is employed to achieve collaborative expression of directional priors and multi-scale context, realizing the unification of direction awareness and multi-granularity aggregation.

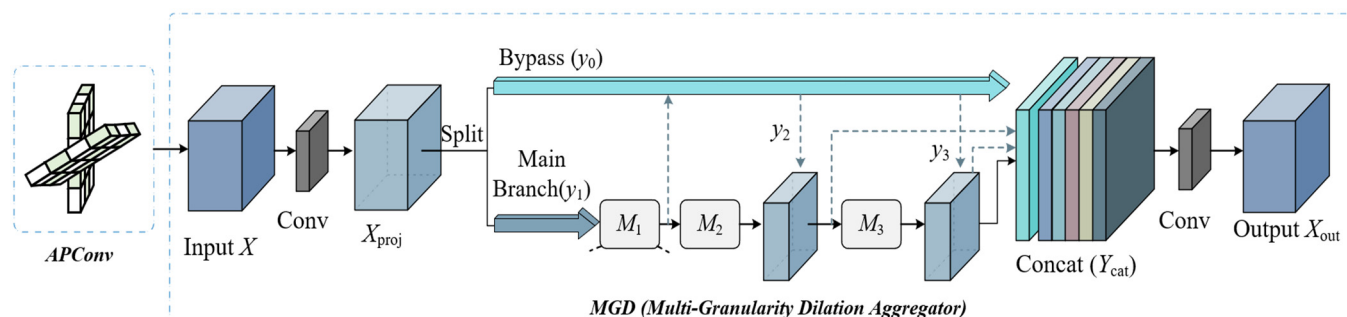


Figure 3. Architecture of the Direction-Aware Multi-Granularity Enhancement module (DAGE). It primarily comprises APConv and MGD. Input features are first encoded by APConv for directional edge enhancement, then channel-split and processed through MGD via cascaded bypass and main branches for cross-stage interaction before being fused by concatenation.

3.2.1. Angular Pinwheel Convolution Mechanism (APConv)

First, the APConv constructs complementary receptive field lobes in the four horizontal and vertical directions through asymmetric padding, enabling the DAGE unit to provide high-quality features with richer directional responses for subsequent cross-modal fusion. The computational flow is illustrated in Figure 4.

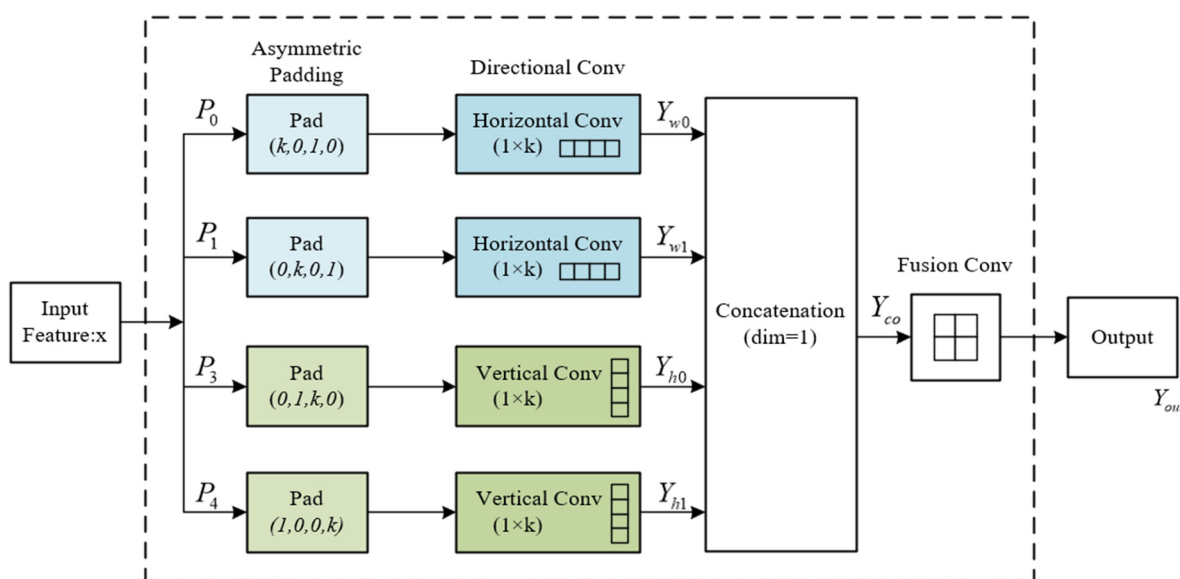


Figure 4. Schematic of the APConv mechanism. Input features are first processed along four parallel branches, each undergoing specific asymmetric spatial padding followed by 1D directional convolution, thereby constructing four directionally complementary receptive field lobes in the spatial domain. Subsequently, the four direction-sensitive feature maps are concatenated along the channel dimension and aggregated via a fusion convolution, finally producing high-quality features rich in local geometry and directional priors. This mechanism enables the module to capture fine-grained orientation cues essential for distinguishing small targets in complex remote sensing scenes.

The input feature maps of the two modalities are denoted as $X^{(v)}$ and $X^{(r)}$. Four asymmetric zero-padding schemes are defined to extend the receptive field in four directions, where each padding tuple is given as $(left, right, top, bottom)$, and k denotes the kernel size. The resulting spatial dimensions of the padded tensors are

$$P_0, P_1 \in \mathbb{R}^{B \times C \times (H+1) \times (W+k)} \quad (1)$$

$$P_2, P_3 \in \mathbb{R}^{B \times C \times (H+k) \times (W+1)} \quad (2)$$

The horizontal and vertical convolution kernels are defined as W_{cw} and W_{ch} , respectively, for horizontal and vertical perception. The branch features for the four directions are extracted as

$$Y_{w0} = \sigma(BN(W_{cw} * P_0)) \in \mathbb{R}^{\frac{B \times C \times H' \times W'}{4}} \quad (3)$$

$$Y_{w1} = \sigma(BN(W_{cw} * P_1)) \in \mathbb{R}^{\frac{B \times C \times H' \times W'}{4}} \quad (4)$$

$$Y_{h0} = \sigma(BN(W_{ch} * P_2)) \in \mathbb{R}^{\frac{B \times C \times H' \times W'}{4}} \quad (5)$$

$$Y_{h1} = \sigma(BN(W_{ch} * P_3)) \in \mathbb{R}^{\frac{B \times C \times H' \times W'}{4}} \quad (6)$$

The directional semantics of the four branches are different from each other. Y_{w0} and Y_{w1} represent the left-extending and right-extending receptive fields in the horizontal direction, while Y_{h0} and Y_{h1} represent the bottom-extending and top-extending receptive fields in the vertical direction. Finally, the outputs of the four directional branches are concatenated along the channel dimension to form a complete pinwheel feature map, and a 2×2 convolution is applied to perform cross-channel integration of the four directional features, expressed respectively as

$$Y_{co} = \text{Concat}(Y_{w0}, Y_{w1}, Y_{h0}, Y_{h1}; \text{dim} = 1) \in \mathbb{R}^{B \times C \times H' \times W'} \quad (7)$$

$$Y_{out} = \sigma(BN(W_{cat} * Y_{co})) \in \mathbb{R}^{B \times C \times H' \times W'} \quad (8)$$

To preserve the features of both modalities and perform independent feature enhancement, the two branches each perform convolution operations without sharing parameters, allowing each modality to retain its unique directional texture responses.

3.2.2. Multi-Granularity Dilation Aggregator (MGD)

After the front-end enhancement by the Pinwheel Convolution, the two feature streams separately enter the MGD structure for further multi-scale feature extraction. While retaining the advantage of multi-path gradient propagation inherent to the CSP framework, this module improves the traditional standard bottleneck layer into a dilation-aware residual. By deploying multi-branch convolutions with multiple dilation rates in parallel within the bottleneck, a single module simultaneously possesses the triple capability of fine-grained local perception, medium-range context modeling, and wide-area global semantic capture, thereby significantly enhancing the model's feature representation in multi-scale object detection scenarios.

Let the input feature map after APConv enhancement be denoted as X . It first passes through a convolutional layer for channel projection and feature enhancement, and the output is then split into two branches along the channel dimension. The input feature map is expressed as

$$X_{proj} = \text{Conv}_{1 \times 1}(X) \in \mathbb{R}^{2C_h \times H \times W} \quad (9)$$

The branches are expressed as

$$[y_0, y_1] = \text{Split}(X_{proj}), y_0, y_1 \in \mathbb{R}^{C_h \times H \times W} \quad (10)$$

where $C_h = C_2 \cdot e$ (with e being the expansion factor), y_0 serves as the cross-stage residual connection, and y_1 enters the subsequent bottleneck layer sequence for further processing.

In the recursive processing of the bottleneck layer, there are n repeated computational units M_i . The output of each layer serves as the input to the next layer, and intermediate states are preserved. The recursive procedure and computational logic are expressed as

$$y_{i+1} = M_i(y_i), i = 1, \dots, n \quad (11)$$

$$M(y) = \text{Add}(y, \phi_{\text{cor}}(\text{Conv}_{1 \times 1}(y))) \quad (12)$$

ϕ_{cor} is the core computational factor of the dilation-aware residual, and its computational flow is illustrated in Figure 5.

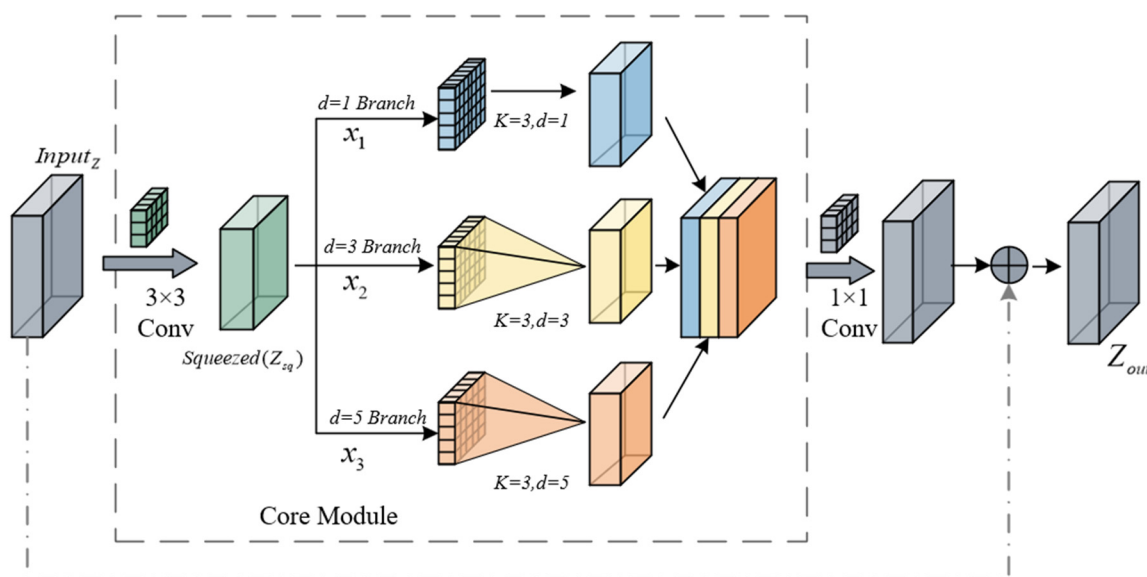


Figure 5. Illustration of the computational process of the dilation-aware residual core factor. The input features are first compressed via a convolutional layer, followed by three parallel dilated convolutions with different dilation rates, which simultaneously capture fine-grained local details and multi-scale wide-area contextual information. The features from each branch are then concatenated and fused at the channel level. Finally, the result is added to the original input features through a residual connection, producing multi-granularity aggregated features.

Internally, the module performs dilation-related residual learning through multi-branch dilated convolutions. First, a 3×3 convolution is applied to the local input Z for dimensionality reduction, expressed as

$$Z_{sq} = \delta(\text{BN}(\text{Conv}_{3 \times 3}(Z))) \in \mathbb{R}^{\frac{C_h}{2} \times H \times W} \quad (13)$$

Subsequently, multi-scale dilated convolution extraction is performed, with three branches of different receptive fields executed in parallel:

$$x_{1,2,3} = \text{Conv}_{3 \times 3, d=1,3,5}(Z_{sq}) \quad (14)$$

The multi-scale information is activated through channel concatenation, expressed as

$$Z_{out} = \text{GELU}(\text{BN}(\text{Conv}_{1 \times 1}[x_1, x_2, x_3]) + Z) \quad (15)$$

Returning to the DAGE structure, the initial cross-stage feature y_0 is concatenated with all intermediate feature maps after n transformations, which greatly enhances the

propagation efficiency of feature gradients and preserves semantic information at different depths. This is expressed as

$$Y_{cat} = \text{Concat}([y_0, y_1, \dots, y_n]) \in \mathbb{R}^{(n+2)C_h \times H \times W} \quad (16)$$

Finally, a convolutional layer maps the channel dimension back to the target dimension:

$$X_{out} = \text{Conv}_{1 \times 1}(Y_{cat}) \in \mathbb{R}^{C \times H \times W} \quad (17)$$

Overall, the DAGE module constructs an efficient feature collaboration unit that transitions from directional geometric perception to multi-scale semantic aggregation. Within this mechanism, the deep coupling of the APConv and the MGD module accomplishes precise encoding of edge and directional texture priors for small objects and successive multi-scale dilated convolution, seamlessly extending the receptive field from fine-grained local regions to wide-area global context without introducing additional spatial downsampling. Under this internal synergy, the multi-path gradient flow of cross-stage connections plays a critical bridging role: it not only ensures stable propagation of deep features but also enables early directional-sensitive features and deep multi-scale global semantics to be preserved intact and efficiently fused across levels. Finally, after projecting to a unified output dimension via a convolutional layer and strengthening the nonlinear expression capacity with the activation function at the end of the module, this deep complementarity between directional details and global multi-granularity provides more accurate artifact suppression and target boundary anchoring for subsequent neck processing under complex aerial background clutter, thereby improving the discriminability and robustness of high-altitude small object detection.

Notably, in the construction stage of the neck feature pyramid within the overall detection network architecture, shallow high-resolution spatial details and deep low-resolution abstract semantics undergo frequent cross-scale concatenation through upsampling and downsampling. However, due to the large scale jumps of targets in the aerial receptive field and the complexity of background receptive fields, the original C3K2 module, with its fixed single receptive field, struggles to effectively and adaptively aggregate and reconstruct cross-layer features after concatenation. Therefore, to break through the performance bottleneck of the cross-scale fusion stage in the neck network, this study further replaces the C3K2 modules in the neck with the MGD module, a component of the DAGE unit. The purpose is to immediately leverage the multi-level parallel dilated convolution mechanism of the MGD module after cross-layer features are concatenated, performing multi-dimensional secondary context mining on the heterogeneously fused features. With the built-in cross-stage local connections, this design ensures the stable flow of key feature gradients during fusion, achieving deep complementarity between local focus and global perception, and ultimately delivering multi-scale features with stronger robustness and richer semantic expression to the final detection heads.

3.3. Asymmetric Context Guided Fusion Module (ACGF)

In multimodal object detection tasks, effectively fusing complementary information from different modalities and establishing a robust feature interaction mechanism are key to improving detection performance. Inspired by context-guided networks, we design an Asymmetric Context Guided Fusion (ACGF) module tailored for multimodal feature fusion.

Unlike homogeneous modeling that only performs local feature and surrounding context extraction on a single modality input, the core of this module lies in departing from symmetric equal-weight extraction. By constructing asymmetric receptive fields to

capture differentiated contextual information, it decouples local feature extraction and surrounding context extraction into two different modality branches, anchoring on fine-grained local features and adaptively interacting with contextual features from the other side, thereby achieving cross-modal feature fusion within a richer feature interaction space. Specifically, the infrared modality, which provides stable target contours under all illumination conditions, is set as the primary backbone modality and uses a small receptive field convolution to extract precise boundaries. The visible modality, which provides fine environmental semantics but exhibits significant performance variations under different illumination conditions, is set as the secondary modality and employs dilated convolution to expand its spatial search range for extracting contextual information as an enhancement.

The module consists of modality-specific channel mapping and four core components: an infrared modality feature extractor f_{ir} , a visible modality feature extractor f_{rgb} , a multimodal feature fusion layer f_{agg} , and a global cross-modal attention module f_{att} . The architecture is illustrated in Figure 6.

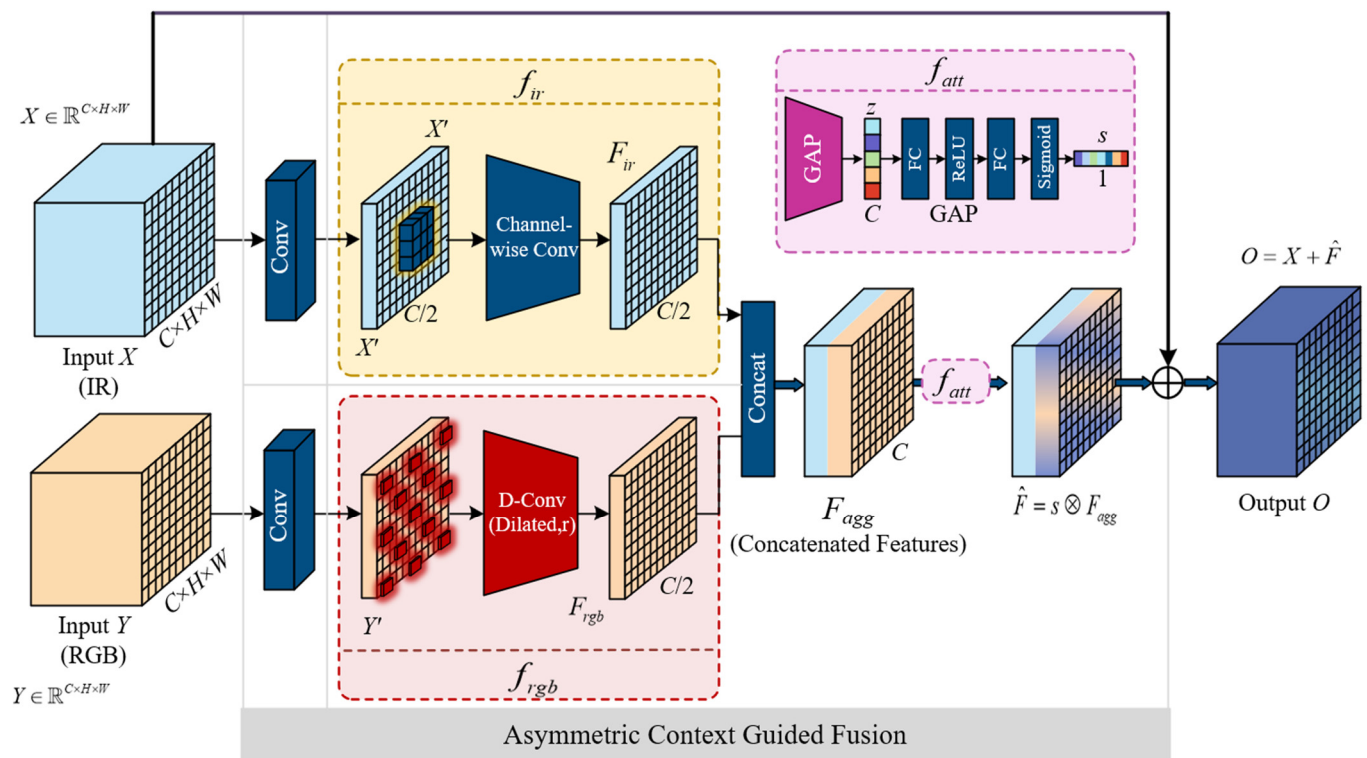


Figure 6. Overview of the Asymmetric Context Guided Fusion module (ACGF). Infrared and visible features are separately processed by asymmetric extractors (small-receptive-field depthwise convolution for infrared, large-receptive-field dilated convolution for visible), then fused with channel-wise attention. The recalibrated features are added to the original infrared input via a residual connection, using infrared as an asymmetric anchor. This design enables mutual complementarity: infrared provides stable contours, while visible supplies contextual details to suppress cross-modal artifacts.

Let the two modality outputs from the dual-branch backbone network at the same semantic level be denoted as X and Y . To perform independent nonlinear transformation and channel alignment for features of different modalities while reducing computational cost, a convolution is first applied to each branch for channel dimensionality reduction, expressed as

$$X' = \sigma(BN(W_x * X)) \in \mathbb{R}^{\frac{C \times H \times W}{2}} \quad (18)$$

$$Y' = \sigma(BN(W_y * Y)) \in \mathbb{R}^{\frac{C \times H \times W}{2}} \quad (19)$$

This step compresses the modality channel dimension by half, where W_x and W_y are the independent convolution kernel parameters for the two modalities, BN denotes batch normalization, σ is the PreLU activation function, and $*$ denotes the convolution operation. The reduced-dimensional modality features then undergo cross-modal feature extraction, with adaptive extraction operations performed according to the characteristics of each modality.

The infrared modality branch performs spatial convolution on each channel, mainly using the f_{ir} component to extract edge spatial structure features of each pixel, expressed as

$$F_{ir} = f_{ir}(X') = DWConv_{3 \times 3}(X') \in \mathbb{R}^{\frac{C \times H \times W}{2}} \quad (20)$$

In the operation, $DWConv_{3 \times 3}$ denotes a depthwise separable convolution with a kernel size of 3×3 .

The visible modality branch applies a channel-wise dilated convolution to the reduced-dimensional features, aiming to extract larger-receptive-field spatial information of the targets, expressed as

$$F_{rgb} = f_{rgb}(Y') = DWDCConv_{3 \times 3, r}(Y') \in \mathbb{R}^{\frac{C \times H \times W}{2}} \quad (21)$$

where $DWDCConv_{3 \times 3, r}(\cdot)$ denotes a depthwise separable dilated convolution with a kernel size of 3×3 and a dilation rate of r , capturing fine semantic information of targets over a larger spatial fault-tolerance range in the second modality.

Subsequently, F_{rgb} and F_{ir} are concatenated along the channel dimension, followed by normalization and activation function processing, to obtain a joint feature. After concatenation, the channel dimension is restored to C . The joint feature F_{agg} simultaneously encodes the contour context semantics of the first modality and the enlarged-receptive-field spatial features of the second modality, forming a cross-modal joint representation, expressed as

$$F_{agg} = f_{agg}(F_{ir}, F_{rgb}) = \sigma(BN(Concat[F_{ir}, F_{rgb}])) \in \mathbb{R}^{C \times H \times W} \quad (22)$$

In the feature extraction operation, by applying the infrared modality feature extractor f_{ir} and the visible modality feature extractor f_{rgb} to different modalities respectively, the module makes full use of the respective modality advantages of infrared and visible images, effectively complementing and fusing rich local details such as textures and edges with the global saliency context related to target thermal radiation, thereby improving the robustness and discriminability of target representation.

To further enhance the discriminative ability of the joint features, a global cross-modal attention module f_{att} is introduced to perform channel-wise adaptive refinement on the joint features. Specifically, global spatial information is first aggregated via global average pooling, then the dependencies among channels are learned through a multi-layer perceptron, and finally channel-wise recalibration is applied to the joint features.

In the global feature compression stage, the multimodal feature fusion layer f_{agg} performs global average pooling, compressing the spatial information of each channel into a scalar to obtain a global descriptor z , which encodes channel-level statistical information of the joint feature over the entire spatial domain, expressed as:

$$Z = GAP(F_{agg}) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W F_{agg}(i, j, k) \in \mathbb{R}^C \quad (23)$$

Subsequently, Z is fed into a bottleneck structure consisting of two fully connected layers to compute the nonlinear dependencies among channels:

$$S = \delta(W_2 \cdot \text{ReLU}(W_1 \cdot z)) \in \mathbb{R}^C (W_1, W_2 \in \mathbb{R}^{\frac{C \times C}{\gamma}}) \quad (24)$$

where W_1 is the weight of the dimensionality-reduction fully connected layer, W_2 is the weight of the dimensionality-expansion fully connected layer, γ is the channel reduction ratio, and ReLU is the Rectified Linear Unit activation function. δ denotes the Sigmoid activation function. Each element of the vector S represents the importance weight of the corresponding channel, with values in the range (0, 1).

Then, the channel attention weights and the joint features are used to perform channel-wise feature recalibration, achieving adaptive channel-level feature refinement:

$$\hat{F} = f_{att}(F_{agg}) = S \otimes F_{agg} \in \mathbb{R}^{C \times H \times W} \quad (25)$$

This operation broadcasts the scalar weights to all spatial positions of their corresponding channels via element-wise multiplication. The global contextual information is encoded as channel attention weights, which are used to adaptively emphasize cross-modal complementary feature channels beneficial to the detection task while suppressing redundant and noisy channels.

Finally, a global residual learning strategy is adopted to connect the original infrared input with the refined channel features via a residual connection, yielding the final joint output:

$$O = X + \hat{F} \in \mathbb{R}^{C \times H \times W} \quad (26)$$

The residual connection is only linked to the infrared branch. This asymmetric residual structure, with the high-contrast infrared contours serving as a strong guiding anchor, preserves the original spatial features of the primary modality while incorporating cross-modal enhanced features refined by the wide-area visible context. This design effectively alleviates feature degradation caused by information conflicts, promotes stable gradient backpropagation during training, and enhances the cross-modal feature interaction for small objects.

4. Experiments

4.1. Datasets and Preprocessing

The experimental part of this study primarily employs the RGBTDronePerson, AVMS, and SODA-A datasets, all of which are adapted for UAV vision tasks and consist of aerial small-object images captured by drones.

The RGBTDronePerson dataset is a cross-modal pedestrian detection dataset collected by UAVs, containing over 6000 visible and infrared image pairs. The scenes were captured at altitudes of 50–80 m, covering multiple viewing angles under various weather and illumination conditions. This dataset focuses on extremely small object detection, with a total of 70,880 annotated instances, of which 98% of the targets are smaller than 20 pixels. There is also a significant cross-modal spatial misalignment. The dataset includes three categories: crowd, person, and rider.

The AVMS (Aerial View Multi-Scene) dataset is a self-constructed UAV multi-modal image fusion dataset proposed in the FS-Diff paper, aimed at addressing the shortcomings of existing datasets in terms of long-distance shooting, diverse scenes, and modality clarity inconsistency. This dataset provides fine-grained annotations for a total of 8991 targets, mostly under daytime illumination conditions, and covers more than seven typical scenes

including residential areas, roads, campuses, parks, streets, stadiums, parking lots, and farmlands. Considering practical application scenarios and the instance characteristics of this dataset, we cleaned the dataset into four categories: car, person, truck, and tent.

To further explore the scene adaptability of the improved modules in our model for small object detection from UAV perspectives, this study introduces the SODA-A (Small Object Detection in Aerial Images) large-scale high-resolution aerial single-modal dataset for auxiliary validation. In this dataset, all categories are small-scale targets, with pixel occupancy smaller than 32×32 pixels. During experimental processing, the original dataset was first tiled, and oriented bounding boxes were converted to horizontal bounding boxes, then further cropped to a size of 640×640 pixels. Through data cleaning, a total of 14,000 images were used, covering ten categories: Airplane, Helicopter, Small-vehicle, Large-vehicle, Ship, Container, Storage-tank, Swimming-pool, Windmill, and ignore. The dataset was split into training and test sets at a ratio of 8:2.

Furthermore, to evaluate the cross-scene generalization capability of the proposed model under extreme low-illumination environments and near-ground medium-scale overhead targets, this study introduces the LLVIP dataset for supplementary assessment in the comparative experiments. LLVIP is a visible-infrared paired pedestrian dataset specifically designed for low-light vision tasks, containing 16,836 pairs of images captured during the low-light hours from 18:00 to 22:00. The dataset uses infrared images to back-map and annotate pedestrians in visible images, covering rich pedestrian targets across 26 scenes, making it suitable for validating algorithm robustness under extreme overhead dark-light conditions.

4.2. Experimental Settings

To ensure the unbiasedness of experimental data, prevent cross-modal interference between multimodal samples, and maintain the stability of modality-specific branch data, this study maintains the spatial correspondence between dual-modal image pairs and does not use pre-trained weights during training, validation, testing, and all ablation and comparative experiments. In the data preprocessing stage, all input images are resized to a consistent resolution, and mixup augmentation is disabled. The maximum number of training epochs is set to 200, and the stochastic gradient descent optimizer is adopted, with the learning rate dynamically adjusted via the cosine annealing algorithm. The remaining experimental settings and parameter configurations are summarized in Table 1.

Table 1. Experimental environment and related parameter configurations.

Parameter	Value
GPU	NVIDIA GeForce RTX 5070 Ti (NVIDIA Corp., Santa Clara, USA)
Python version	3.11 (Python Software Foundation, Beaverton, OR, USA)
CUDA version	12.8 (NVIDIA Corp., Santa Clara, CA, USA).
PyTorch version	2.9.0 (PyTorch Foundation, San Francisco, CA, USA)
Momentum	0.937
Weight Decay	0.0005

4.3. Ablation Studies

The ablation studies are conducted on the RGBTDronePerson and SODA-A datasets. During the experiments, multiple evaluation metrics are emphasized, including Precision (P), Recall (R), mean Average Precision (mAP) under different IoU thresholds, model parameters (Params), and floating point operations (GFLOPs).

To systematically evaluate the performance advantages of the proposed modules, this study follows a progressive optimization and merging criterion, introducing each

optimized module sequentially and conducting ablation experiments to perform qualitative performance analysis on the enhanced detection network.

4.3.1. Synergistic Ablation Analysis of Core Components Within the DAGE Module

To preliminarily validate the adaptability of the DAGE module in UAV aerial scenes and the soundness of its fusion design, this study first conducts ablation experiments on the single-modal aerial small-object scenario using the SODA-A dataset, focusing on the direction-aware guided multi-granularity cross-stage unit. The quantitative results are summarized in Table 2, where the symbol “×” indicates that the module is not added to the model, and “√” indicates that the module is added. Bold indicates the best value among the indicators, this applies to all tables below.

Table 2. Ablation analysis of core components of DAGE on the SODA-A dataset.

Methods	APConv	MGD	Precision (%)	Recall (%)	F1	Map50 (%)	Map50-95 (%)	Map75 (%)	GFLOPs
Base	×	×	58.0	49.9	53.6	50.7	27.3	26.0	6.3
Only APConv	√	×	63.7	50.6	56.4	53.9	29.6	28.6	6.8
Only MGD	×	√	63.3	53.0	57.6	55.6	30.3	29.6	7.3
DAGE	√	√	63.5	53.4	58.0	56.1	30.6	29.7	7.8

Table 2 details the specific impact of the internal core components of the DAGE module on the detection performance on the SODA-A dataset. In the decoupling validation of internal components, using only the APConv as the front-end convolutional processing yielded a significant precision gain, demonstrating that its asymmetric four-directional receptive field injects keen directional perception into the feature stream and enhances the model’s ability to accurately encode geometric edges of small targets. When the MGD was introduced alone, the recall improvement became more prominent, reaching 53.0%, verifying that the multi-scale dilated convolution mechanism effectively captures richer contextual semantics by enlarging the local receptive field, thereby enabling the network to re-localize tiny targets that are otherwise easily missed.

Finally, when the complete DAGE module was deployed as a whole in the network, the model achieved the best balance between precision and recall, attaining optimal comprehensive performance metrics with mAP50 of 56.1% and mAP50-95 of 30.6%. From a module synergy perspective, this performance improvement is not a simple linear superposition. The front-end APConv provides fine-grained directional features for the feature stream, laying accurate geometric prior guidance for subsequent multi-scale semantic extraction, while the MGD compensates for multi-scale limitations with its broadly extracted contextual information. The two components form a complementary relationship in the feature stream, combining local directional refinement with global scale coverage.

This synergistic mechanism ensures that when the model faces densely distributed, extremely small-scale targets in aerial scenes, it can accurately identify target boundaries through directional sensitivity while recovering easily missed targets using multi-scale features, ultimately achieving a steady improvement in overall detection performance.

4.3.2. Ablation Study on the Overall Architecture of the Multimodal Detection Network

To address the inherent perceptual limitations of single-modality scenes in complex aerial scenarios, this study conducts systematic ablation experiments on the proposed DMAC-Net multi-modal architecture using the multi-modal RGBTDronePerson dataset. The experiments focus on examining the necessity of fusing infrared and visible information, as well as the specific contribution of each core module to cross-modal interaction. The experimental results are summarized in Table 3, where the symbol “×” indicates that the

module is not added to the model, “√” indicates that the module is added, and bold font denotes the best performance.

Table 3. Ablation experiments of the DMAC-Net architecture on the RGBTDronePerson dataset.

Methods	DAGE	ACGF	MGD	Modality	Precision (%)	Recall (%)	Map50 (%)	Map50-95 (%)	Map75 (%)
Base	×	×	×	single	48.3	45.6	40.2	14.9	6.6
Dual-Base	×	×	×	dual	52.3	48.8	44.9	16.6	7.3
Only DAGE	√	×	×	dual	54.3	50.8	49.0	18.4	8.0
Only ACGF	×	√	×	dual	53.8	50.1	46.5	17.1	7.5
ACGF+DAGE	√	√	×	dual	55.9	53.4	50.0	18.8	8.3
Only MGD	×	×	√	dual	55.6	49.8	48.0	17.7	7.3
DMAC-Net	√	√	√	dual	56.4	50.9	50.2	19.1	8.6

As shown in Table 3, compared with the single-modal baseline model, the introduction of dual-modal branches substantially raises the mAP50 of the basic fusion model from 40.2% to 44.9%, further demonstrating that leveraging feature complementarity between multi-modal data is a key prerequisite for breaking through the bottleneck of small object detection in complex environments.

First, based on this multi-modal fusion baseline, introducing the unified backbone extraction module DAGE alone increases precision and recall to 54.3% and 50.8%, respectively, with mAP50 reaching 49.0%. This indicates that the direction-aware component inside DAGE can keenly extract thermal radiation edges from infrared targets and geometric details from visible targets, effectively filtering out cross-modal background clutter. Meanwhile, its multi-granularity component effectively enlarges the receptive field, further capturing densely distributed small targets with drastic scale variations.

After introducing the asymmetric context-guided fusion module designed specifically for multi-modal networks, all core metrics show balanced and steady improvements over the basic fusion: precision increases from 52.3% to 53.8%, recall rises from 48.8% to 50.1%, and the comprehensive mAP50 metric improves by 1.6 percentage points. This further demonstrates the asymmetric decoupling philosophy of the ACGF module, which leverages the stable imaging advantage of the infrared modality in complex environments by using a small receptive field to precisely anchor local spatial boundaries of targets, while employing the visible modality as a supplement to extract broader environmental context through large-receptive-field dilated convolution. This design fully exploits the physical imaging characteristics of both modalities, effectively alleviating modal conflicts caused by direct concatenation of dual-modal features.

Building upon this multi-modal feature interaction, the network replaces the neck’s C3K2 module with the MGD module from the DAGE unit, ensuring stable flow of key feature gradients during the fusion process and mitigating feature degradation of weak targets. Compared with the original neck network, the introduction of the MGD module brings notable performance gains, with mAP50 improving from 44.9% to 48.0%, accompanied by simultaneous increases in both precision and recall. This further demonstrates that the proposed modification delivers multi-scale features with stronger robustness and richer semantic expression to the final detection heads.

In the final integrated DMAC-Net model, the cross-modal synergy effect is maximized. In the feature encoding stage, the DAGE module in the backbone, through the coupling of direction-aware and multi-granularity mechanisms, supplies low-level feature representations that possess high-order geometric priors and rich local context. In the cross-modal interaction stage, the ACGF mechanism effectively leverages asymmetric guidance to achieve deep decoupling of infrared spatial contours and visible global semantics. Finally,

in the feature decoding and pyramid construction stage, the MGD module introduced into the neck takes over these high-quality fused features and effectively suppresses the dilution and degradation of weak small-target features during the top-down propagation process. Benefiting from this synergistic mechanism, DMAC-Net alleviates the inherent background interference and scale sensitivity issues of multi-modal small object detection in complex aerial scenes, ultimately achieving a breakthrough in overall detection accuracy.

4.3.3. False Positive Error Evolution Analysis

To more clearly quantify the specific contribution of each module in complex aerial scenes, this paper conducts an in-depth error decomposition of the prediction results from the single-modal baseline model (Base), the multi-modal basic fusion model (Dual-Base), and the final proposed DMAC-Net model, focusing on the Crowd and Person categories based on experimental data from the RGBTDronePerson dataset.

In multi-class object detection, false positives (FP) are further divided into background false positives (Bg-FP), where background regions are predicted as targets, and inter-class false positives (Cls-FP), where objects of other categories are misclassified as the target class. The distribution statistics of these two error types for the three models are summarized in Table 4.

Table 4. Comparison of false positive and fine-grained error distribution across different models.

Class	Model	Bg-FP	Cls-FP
Crowd	Base	755	260
	Dual-Base	894	397
	DMAC-Net	663	252
Person	Base	2164	322
	Dual-Base	2203	287
	DMAC-Net	1275	315

Observing the evolutionary trend of the data in Table 4, it can be seen that the model's ability to suppress false positives exhibits a phased variation during cross-modal feature interaction.

Comparing the single-modal baseline with the multi-modal basic fusion model, after introducing multi-modal features, background false positives for both categories increase to some extent. This indicates that naive concatenation of heterogeneous features without deep alignment tends to introduce additional redundant information, causing some background regions with thermal radiation or visual interference to be misjudged as targets, thereby increasing background noise to a certain degree.

In terms of the background false positive metric, the improved model achieves substantial reductions for both categories. This comprehensive optimization demonstrates that extracting multi-directional edge features combined with a cross-modal asymmetric fusion mechanism greatly enhances the network's ability to strip away complex background clutter, effectively avoiding the misclassification of meaningless terrain or light spots in UAV aerial images as targets.

Notably, a deeper analysis of the inter-class false positive data reveals a moderate shift in the model's prediction tendency when handling dense scenes. Compared with the basic fusion model, DMAC-Net reduces the misclassification of "single-person targets misidentified as crowd" in the Crowd category from 397 to 252 cases, the lowest among the three models. Meanwhile, for the Person category, the misclassification of "actual crowd samples locally predicted as individual persons" remains at 315 cases, a slight increase

over the basic fusion model. This shift is not a performance degradation but rather a manifestation of the deep logical behavior of the DMAC-Net architecture. In the context of multi-class object detection under complex receptive fields, such a shift further indicates a refinement of the network's spatial resolution granularity. When facing dense targets, the baseline model, limited by its feature extraction capability, tends to coarsely predict an entire region as "crowd". In contrast, the improved model, benefiting from deep interaction through asymmetric context guidance, acquires a stronger ability to separate individual instances, decoupling originally clustered crowds and accurately locating individual persons within them. This marks a substantial improvement in feature discriminability in crowded scenes. From a practical engineering perspective, this prediction tendency to separate aggregated crowds into individual person targets essentially endows the aerial detection system with higher precision and finer-grained personnel localization capability, reflecting the stable characteristic of the proposed model when handling complex crowded scenarios.

4.3.4. Visualization Analysis

To intuitively analyze the robustness of each architecture in complex receptive field environments, this paper selects several challenging typical scenes for visual comparison in the ablation study, as shown in Figure 7, which more directly illustrates the internal detection logic and working mechanism of the model.

Due to the highly complex backgrounds in the original dataset images and the extremely small proportion of target objects, to enhance the visual contrast of target regions and improve the readability of the visualization panels—thereby better illustrating the performance variations during the ablation process—the panels in this section are appropriately cropped for emphasis based on the comparison regions, and are supplemented with annotations in various colors and shapes to facilitate analysis. The specific meanings of all markers are detailed in the respective figure captions of each visualization panel.

Comparing the detected "person" and "crowd" categories in the first and third column (yellow circles), as modules are progressively introduced, the model tends to classify dense areas as crowd, while the final DMAC-Net is more inclined to separate individual pedestrians from the crowd. This finding is consistent with the error analysis presented earlier; after deep feature interaction, the model acquires a stronger capability to decouple originally clustered crowds into individual targets.

In the first-column scene, the person targets are located in marginal areas with complex occlusion and illumination distribution. Particularly in night vision scenarios, inter-class discriminability is challenging, especially for morphologically similar targets such as riders and ordinary pedestrians. Due to the lack of deep context guidance, the basic fusion model and the model using only the DAGE module fail to recognize weakly featured pedestrians under tree shadows, resulting in missed detections. With the introduction of the ACGF module, the network begins to perceive the existence of targets through asymmetric context guidance. Ultimately, DMAC-Net, equipped with a more comprehensive global receptive field, captures pedestrian contours under occlusion and recovers the previously missed targets.

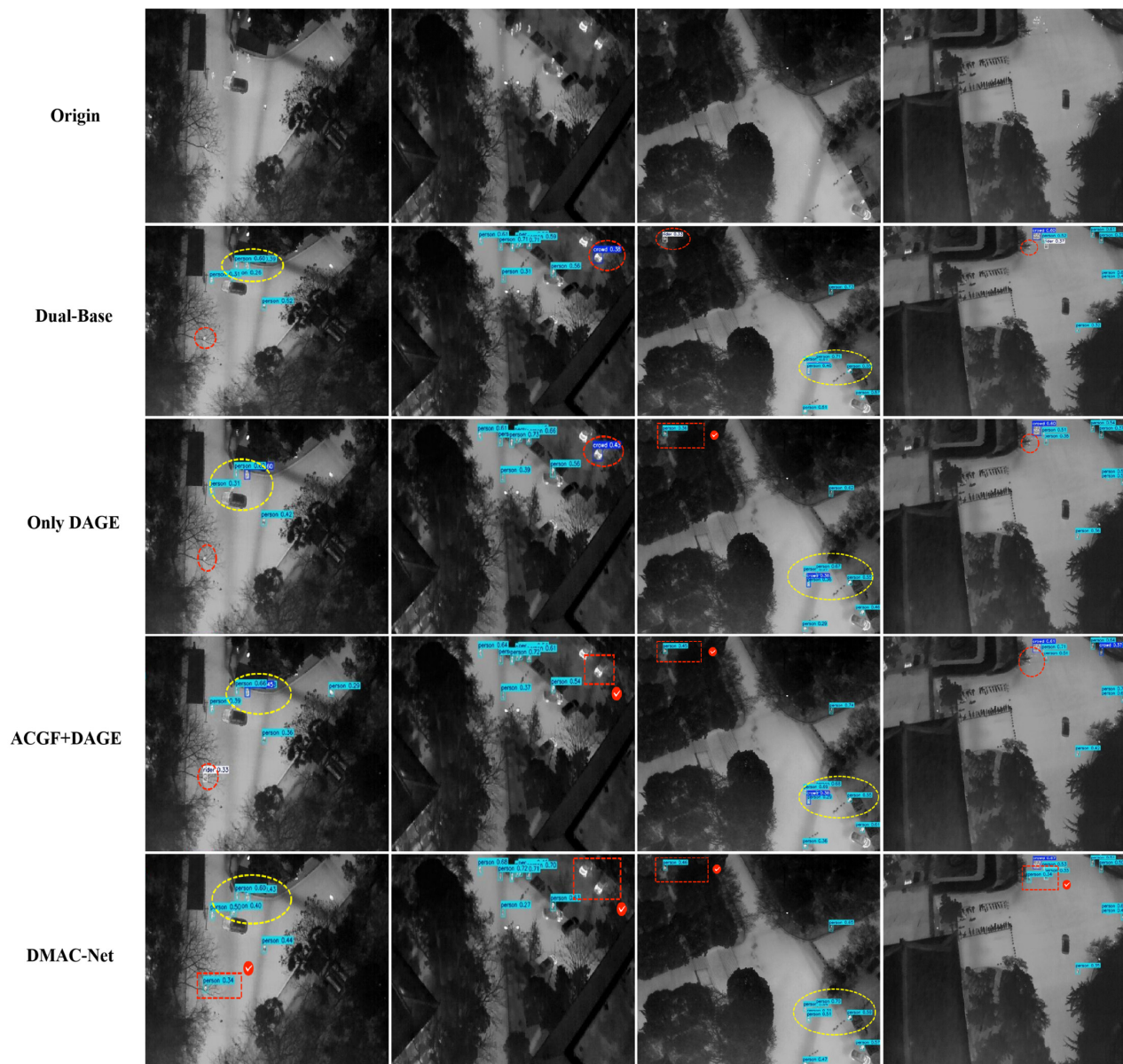


Figure 7. Visual ablation results of the DMAC-Net architecture on the RGBTDronePerson dataset. The combination of red dashed boxes and red check marks indicates correct detections. Red circular dashed boxes denote missed or false detections, while yellow circular dashed boxes highlight the different detection behaviors between the “crowd” and “person” categories across ablation stages. Column markers correspond to the ablation stage models listed in Table 3, and the module composition names follow the same convention as in Table 3. In each of the figures, the light blue labels represent the “person” class, the white ones represent the “rider” class, and the dark blue ones represent the “crowd” class.

The second column demonstrates the model’s ability to discriminate complex background noise and further reveals the complementarity of infrared and visible fusion. Under low-illumination conditions, visible images often fail to accurately locate targets, whereas the infrared modality can more precisely position targets through thermal radiation. However, for interfering objects such as vehicle hoods, the infrared modality exhibits high thermal radiation contrast, which, while enabling precise localization, also makes it prone to misclassifying the front or rear areas of vehicles as detection targets. This issue is substantially mitigated after introducing the asymmetric context guided fusion mechanism. The ACGF structure leverages information from the visible branch, which is unaffected by

thermal radiation, to compensate for interference in the infrared branch, achieving semantic alignment between modalities and effectively distinguishing non-person thermal radiation information, thereby suppressing false positives. This provides strong evidence for the unique advantage of the asymmetric fusion mechanism in filtering heterologous artifacts.

In the third and fourth columns, dense target scenes contain occluded targets at image edges and extremely small-scale person targets. The baseline model exhibits degraded detection performance near image boundaries. Through the synergy of its modules, DMAC-Net demonstrates edge robustness and noise suppression. Whether for pedestrians at the image periphery or those occluded by trees, DMAC-Net achieves precise bounding box localization, with significantly higher detection completeness compared to all ablation stages.

Overall, the visualization results are consistent with the quantitative analysis. Through the organic integration of its modules, DMAC-Net achieves optimal performance across various complex scenes. In the front-end backbone stage, the DAGE unit enhances the representation of local directional details; ACGF suppresses background false positives through asymmetric alignment; the filtered target information is then delivered to the neck network, where multi-scale dilated cross-stage feature extraction ensures detection accuracy for small and occluded aerial targets. This multi-dimensional feature enhancement mechanism further ensures the strong reliability of the model under the complex receptive fields characteristic of UAV perspectives.

4.3.5. Validation of the Direction-Aware Front-End Strategy

To verify the optimal topological deployment of convolutional reconstruction in the feature extraction stage, this study designs two different integration schemes to evaluate the logical order and computational complexity of direction-aware and feature extraction within the module. The experimental results on the RGBTDronePerson dataset are summarized in Table 4.

APConv-pre denotes a downsampling replacement design that, based on direction-sensitive downsampling, restructures the APConv at the beginning of the DAGE unit. APConv-post denotes a feature extraction block replacement scheme that retains the original downsampling layer and adds the pinwheel convolution as an attention mechanism to the output of the core feature extraction block within the original C3K2 module. The experimental data are presented in Table 5.

Table 5. Analysis of direction-aware front-end strategies on RGBTDronePerson.

Methods	Modality	Precision	Recall	Map50	Map50-95	Parameters(M)	GFLOPs(G)
Conv	dual	0.523	0.488	0.449	0.166	3.72	9.4
APConv-pre	dual	0.567	0.510	0.487	0.178	3.67	9.2
APConv-post	dual	0.548	0.514	0.482	0.181	4.80	16.8

As can be observed from the experimental results, both schemes achieve significant improvements in various accuracy metrics, yet they exhibit different characteristics in terms of efficiency and computational cost.

For the APConv-post reconstruction scheme, although this high-frequency invocation approach achieves the highest mAP50-95 under strict thresholds (0.181), it demands higher computational requirements and underperforms the pre-fusion strategy in terms of precision and mAP50. The pre-fusion design adopted in this paper demonstrates more stable parameter efficiency. Meanwhile, this scheme improves precision by 4.4% and mAP50 by 3.8%. This indicates that, during the downsampling stage where spatial resolution is reduced, the pre-positive convolution mechanism can serve as an efficient operator to

suppress background noise and preserve key directional geometric information while reducing redundant computational overhead.

From the perspective of feature evolution, reconstructing the convolution at a pre-positive position has heuristic significance for processing logic. During the downsampling process of UAV images, the reduction in spatial resolution is often accompanied by irreversible loss of edge information from tiny objects. By leveraging asymmetric directional receptive fields to extract geometric extreme-value features precisely when information loss occurs, the detection structure acts as a structural prior filter. This design ensures that the feature stream injected into subsequent extraction branches possesses a higher signal-to-noise ratio and directional discriminability, thereby enabling the network to maintain accurate capture of target contours even when dealing with low-quality images from UAV perspectives.

4.4. Comparative Experiments

To further validate the effectiveness of DMAC-Net in complex remote sensing tasks, this study selects multiple mainstream detection algorithms for comparative evaluation. Other types of models under the same task are trained and validated on the same dataset, and the experimental results and data are compared accordingly. The models and architectures selected for comparison include YOLOv12 [42], MASAG [43], ICAFusion [28], MDAF [44], and DEYOLO [27]. The experimental results on the RGBTDronePerson dataset are shown in Table 6.

Table 6. Comparison of DMAC-Net with other algorithms on the RGBTDronePerson dataset.

Methods	Precision (%)	Recall (%)	F1 (%)	Map50 (%)	Map50-95 (%)
YOLOv12 [42]	52.6	48.1	50.2	44.7	16.1
MASAG [43]	51.1	47.5	49.2	42.8	15.7
ICAFusion [28]	52.9	47.7	50.2	44.4	15.9
MDAF [44]	55.7	50.1	52.7	47.3	17.2
DEYOLO [27]	55.8	51.2	53.4	47.5	17.2
DMAC-Net	56.4	50.9	53.5	50.2	19.1

Experimental results demonstrate that the proposed DMAC-Net model is competitive across multiple core evaluation metrics. In terms of mAP50, DMAC-Net achieves 50.2%, outperforming DEYOLO (47.5%) and MDAF (47.3%) by 2.7% and 2.9%, respectively. For the mAP50-95 metric, which reflects high-precision localization capability, the proposed model attains 19.1%, surpassing all other compared networks. This further demonstrates that the model can achieve more precise fitting of target boundaries while maintaining accurate classification.

From the perspective of remote sensing UAV imagery, the ability to suppress missed detections of dense targets is intuitively reflected by recall. DMAC-Net achieves a recall of 50.9%, demonstrating stronger detection capability compared to most competing models. The F1 score, which balances precision and recall, reaches 53.5%, indicating that the model not only increases the number of detected targets but also maintains good prediction accuracy, striking a favorable balance between missed detections and false positives. Compared with DEYOLO, although DEYOLO exhibits a slightly higher recall (+0.3 percentage points), its mAP metrics are both lower than those of DMAC-Net. This suggests that while DEYOLO has some feature sensitivity for reducing false positives, there remains room for improvement in handling weak targets under complex backgrounds. In contrast, DMAC-Net effectively expands the detection range of target receptive fields while preserving high precision, achieving more robust detection performance.

The comparative results further validate the feasibility of the DMAC-Net architecture for multi-modal feature fusion. Through asymmetric guidance and multi-scale feature extraction mechanisms, the model effectively bridges the limitations of single-modal information and demonstrates superior comprehensive adaptability on the RGBT dataset compared to existing representative algorithms.

To visually compare the detection performance of different methods, a visual comparison of this set of comparative experiments is presented in Figure 8.

As shown in Figure 8, in the multi-target aerial scene of row (a), the performance of all models is generally stable. However, the overlapping “rider” targets located at the edges remain challenging for detection. Although YOLOv12, DEYOLO, and MASAG accurately identify individual categories, they are insensitive to overlapping targets of the same class, which further highlights the importance of multi-scale cross-stage feature extraction for ensuring robust detection of occluded or overlapping targets. In the occluded target scenario of row (b), only DMAC-Net and MDAF correctly detect the occluded “rider” category. The other models exhibit varying degrees of missed detections, while ICAFusion and YOLOv12 produce redundant and imprecise detection boxes, largely due to noise interference degrading the model’s localization discrimination. In the tiny “person” detection of row (c), the model’s ability to capture extremely weak features of aerial small targets is demonstrated. Only DMAC-Net successfully identifies the “person” target located within a vehicle. MDAF and MASAG misclassify the vehicle’s infrared signature as “crowd,” which is consistent with the observations in the previous ablation experiments, further proving the importance and necessity of the asymmetric fusion mechanism in filtering heterologous artifacts.

To further validate the generalization capability and robustness of the proposed model under the same scenario, we introduce the AVMS UAV aerial dataset for auxiliary comparative validation. The experimental results are compared in Table 7.

Table 7. Comparison of DMAC-Net with other algorithms on the AVMS dataset.

Methods	Precision (%)	Recall (%)	F1 (%)	Map50 (%)	Map50-95 (%)
YOLOv12 [42]	55.0	51.7	53.3	51.2	29.7
MASAG [43]	45.5	50.1	47.7	45.1	25.2
ICAFusion [28]	68.1	45.8	54.8	53.5	29.3
MDAF [44]	60.1	49.4	54.2	52.8	29.3
DEYOLO [27]	56.8	47.1	51.5	56.5	24.7
DMAC-Net	60.6	51.7	55.8	57.6	33.2

In terms of comprehensive evaluation metrics, the proposed model achieves the highest mAP50 and mAP50-95 scores. Compared with models of similar performance, the proposed model maintains relatively stable classification accuracy and bounding box regression precision under new data distributions.

When analyzing the trade-off between missed detections and false positives, different metrics exhibit distinct distribution tendencies. ICAFusion achieves the highest precision (68.1%) but suffers from low recall (45.8%), indicating that this algorithm tends to make conservative predictions, resulting in some targets being missed. The recall of the proposed model reaches 51.7%, which is on par with the YOLOv12 model, while its precision shows a certain improvement over YOLOv12. Based on a comprehensive consideration of these two metrics, the proposed model achieves an F1 score of 55.8, demonstrating the best detection balance among all compared algorithms.

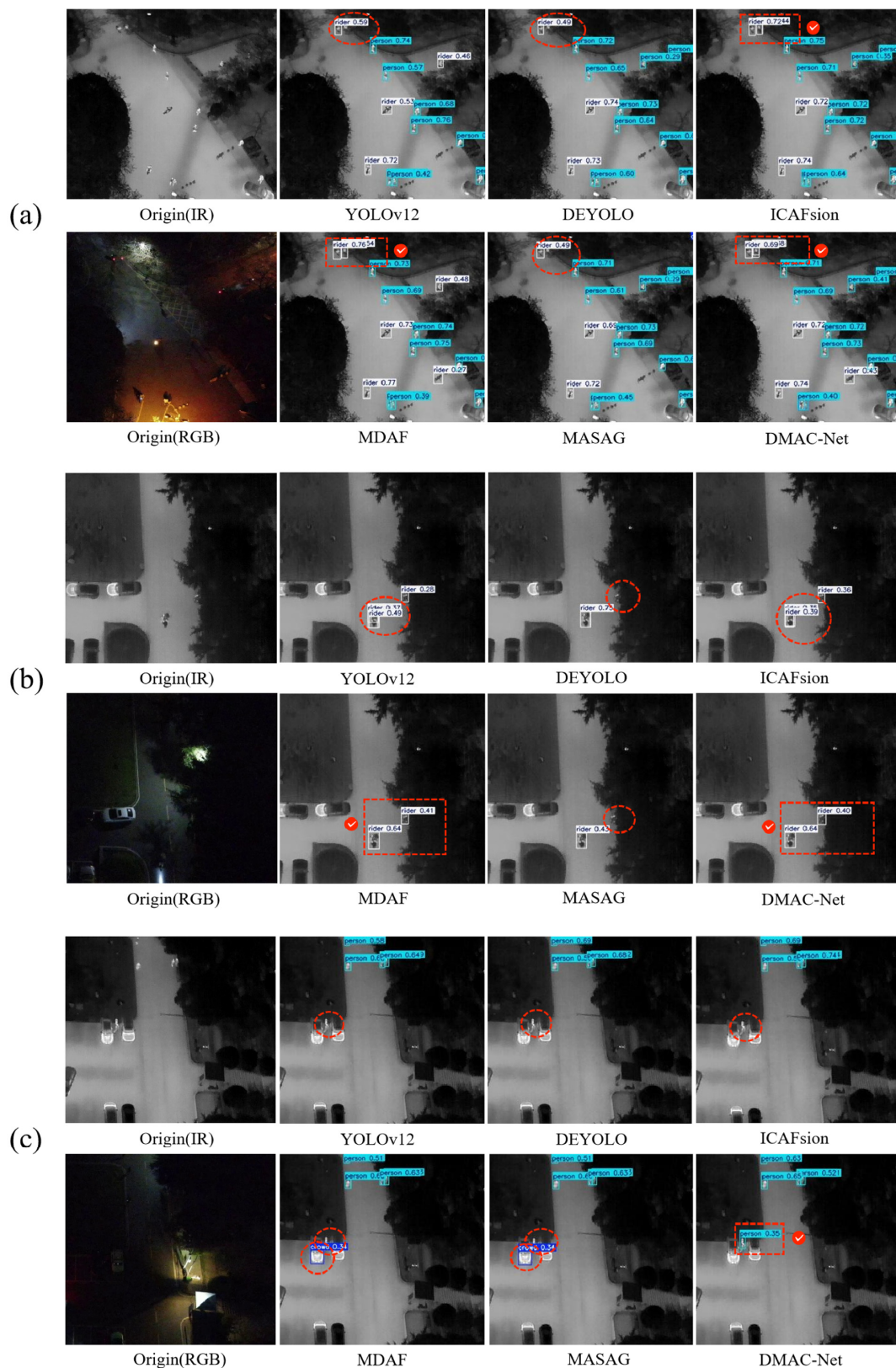


Figure 8. Detection comparison on RGBTDronePerson using DMAC-Net. Each scene is presented in two rows. Scenes (a–c) correspond to a multi-object remote sensing scene, an occluded target scene, and a scene with extremely weak features for tiny objects, respectively. Red dashed boxes with red check marks indicate correct detections, while red circular dashed boxes denote missed or false detections. The comparison illustrates how DMAC-Net outperforms other methods in handling overlapping targets, occlusion, and feature-sparse small objects.

Therefore, based on the test performance on the AVMS dataset, it can be concluded that the multi-modal feature processing mechanism of DMAC-Net possesses a certain degree of generalizability. Facing scene variations across different datasets, the model maintains target detection rates while controlling false positives, reflecting its reliable adaptability. The visual comparison on this dataset is illustrated in Figure 9.

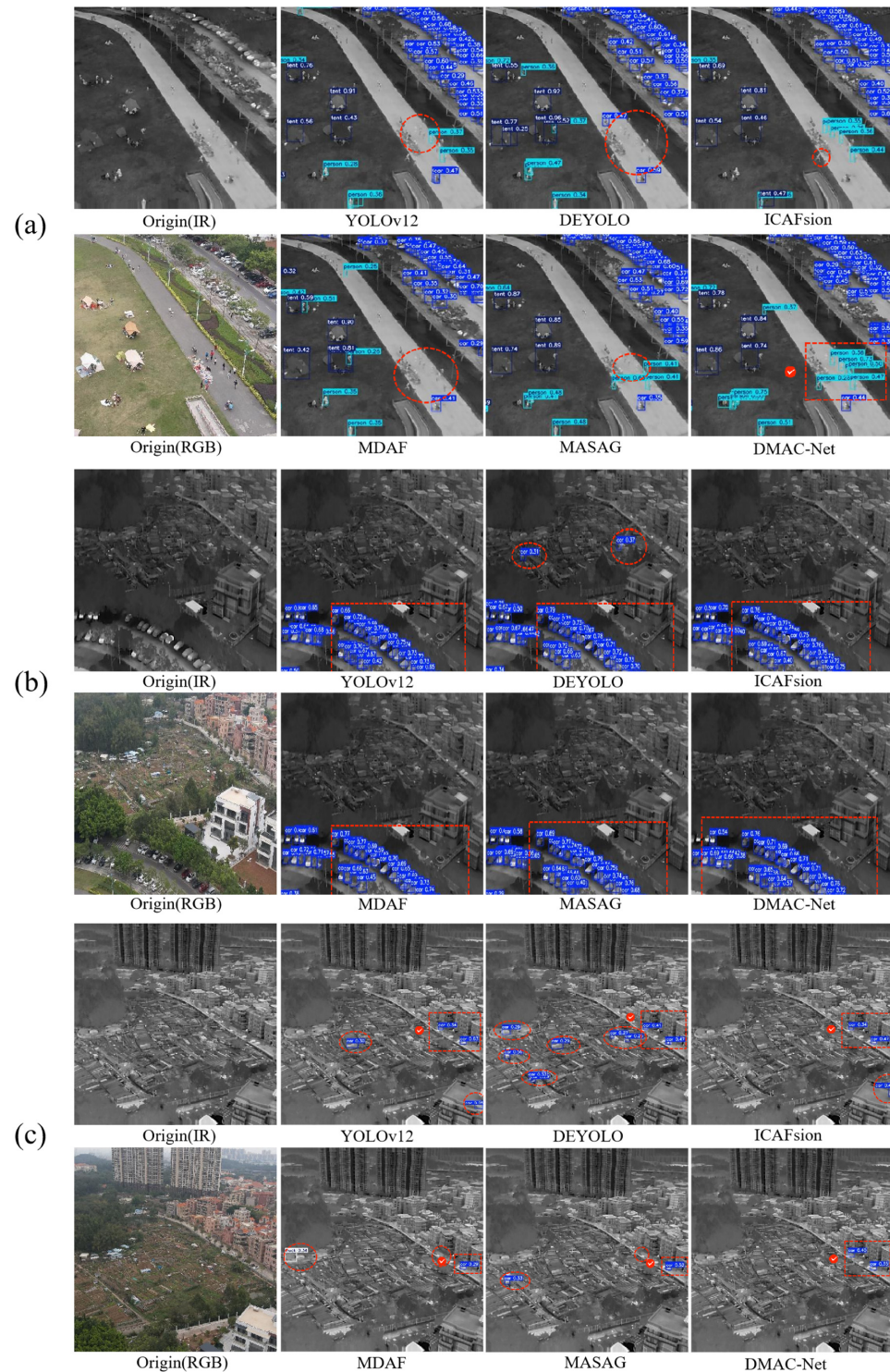


Figure 9. Visual comparison of detection results on the AVMS dataset using the DMAC-Net architecture. Each scene (a–c) is displayed in two rows. Red dashed boxes with red check marks indicate correct detections, while red circular dashed boxes denote missed or false detections. The comparison illustrates the robustness of DMAC-Net under dense target arrangements and oblique viewing angles.

Unlike the RGBTDronePerson dataset, the AVMS dataset features denser targets. Most importantly, the images in this dataset are mostly captured under relatively stable illumination conditions, making the visibility of the visible light branch significantly higher than that of RGBTDronePerson, with much richer features. Therefore, comparative experiments on this dataset can provide auxiliary insights into the model's robustness under different illumination conditions.

As shown in Figure 9, rows (a) and (b) both present dense vehicle arrays. All models in the comparative experiments exhibit stable classification and precise detection bounding boxes when handling such dense scenes. The multi-target scene in row (a) covers almost all categories in this dataset. In this scenario, only the DMAC-Net model successfully classifies multiple “person” targets on the main road, while the other models exhibit varying degrees of missed detections. In the typical aerial scene of row (b), the DEYOLO model misclassifies other targets in the field as “car”, which further reflects the difficulty of learning category features from an overhead perspective due to the challenge of capturing target characteristics in UAV small object detection. In the oblique-view scene of row (c), the receptive field is more complex and target interference is stronger. YOLOv12, DEYOLO, ICAFusion, MDAF, MASAG, and DMAC-Net all successfully identify the “car” category among buildings. However, except for DMAC-Net, all other models exhibit varying degrees of false positives. This provides strong evidence for the capability of DMAC-Net in mitigating modality degradation and enhancing multi-scale feature extraction.

Overall, under relatively stable illumination conditions, the proposed model still demonstrates stable robustness and detection capability.

4.5. Cross-Scale Scenario Comparison

Notably, to further validate the boundaries of DMAC-Net under cross-scene and multi-scale feature conditions, this section introduces the LLVIP dataset for cross-scale scenario evaluation. Unlike the high-altitude long-distance, extremely small object characteristics exhibited by the RGBTDronePerson and AVMS datasets, the LLVIP dataset focuses on near-ground medium-distance imaging, where pedestrian instances exhibit a clear medium-scale characteristic in terms of spatial occupancy. Moreover, the acquisition environment is under extreme low-illumination conditions with severe light constraints.

By extending the validation scope from UAV overhead remote sensing scenes to mid-distance overhead contour representation, drawing reference from low-altitude scenarios, the stability of feature extraction under different spatial resolutions and viewing distance spans can be examined, thereby evaluating the generalizability of the architecture under complex lighting transitions and scale variations. The experimental results are presented in Table 8.

Table 8. Cross-dataset generalization evaluation of DMAC-Net on LLVIP.

Methods	Precision	Recall	F1	Map50
YOLOv12 [42]	0.955	0.861	0.906	0.942
CFT [45]	0.934	0.870	0.901	0.939
MASAG [43]	0.946	0.880	0.912	0.948
ICAFusion [28]	0.952	0.891	0.920	0.954
MDAF [44]	0.951	0.883	0.916	0.956
DEYOLO [27]	0.932	0.890	0.911	0.948
YOLOv10 [46]	0.946	0.880	0.912	0.940
YOLOv13 [47]	0.950	0.884	0.916	0.948
DMAC-Net	0.953	0.892	0.922	0.952

As shown in Table 8, due to larger target scales and richer data, all algorithms achieve significantly improved metrics compared to the previous remote sensing datasets.

The experimental results show that the proposed model achieves outstanding recall (0.892) and F1 score (0.922), reflecting its ability to expand detection boundaries while maintaining high confidence, thereby mitigating background false alarms and target loss caused by imbalanced modality weights. This further demonstrates that under low-illumination conditions, DMAC-Net effectively extracts infrared thermal radiation features, enabling more robust capture of visible-light targets and reducing missed detections. In terms of mAP50, although the proposed model is slightly lower than ICAFusion by 0.2 percentage points, it still significantly outperforms all other models, including MASAG and DEYOLO. Moreover, the marginal fluctuations in the high-score range and the overall stable performance further validate that DMAC-Net, originally designed for high-altitude small targets, maintains good detection accuracy and stable scale convergence when handling near-ground medium-scale targets with varying spatial resolutions.

Therefore, the performance on the LLVIP dataset confirms the robust monitoring capability of DMAC-Net under extreme low-light scenarios and medium-scale targets. The architecture is not only suitable for fragmented target recognition in high-altitude aerial views but also demonstrates reliable detection efficiency in near-ground wide-area surveillance tasks, reflecting its technical adaptability in cross-scale and cross-scene applications. The visual detection results of the model on this dataset under high-noise, low-illumination, and target-dense scenarios are illustrated in Figure 10.



Figure 10. Visual detection results of the DMAC-Net architecture on the LLVIP dataset. Each column represents a different scene type: from left to right, high-noise, low-light, and dense target scenarios. Only containing one type of "person" category target, in the group of images, it is represented by blue detection boxes and labels.

In dense target scenarios, the proposed model can finely segment crowded crowds. In the second column, under extremely low illumination where visible images fail to capture human targets, the model accurately recalls targets in near-dark environments,

demonstrating its stable robustness in low-light multi-scale scenarios. In the third column, under challenging lighting and complex noise interference, the model successfully identifies a person on an electric vehicle, further reflecting its stability in multi-representation target contexts.

This comparison also provides insights into target classification. In the RGBTDronePerson dataset, although all three target categories are pedestrians in physical form, the dataset further subdivides pedestrians into “person”, “crowd”, and “rider” classes based on morphology, focusing on deeper feature analysis. This poses greater challenges for the model’s feature extraction and classification granularity. In contrast, the LLVIP dataset, while maintaining larger target scales, uniformly annotates various human forms as the “person” class, which is also well-suited for large-scale aerial detection scenarios in industrial applications. Despite different target scales and classification priorities, DMAC-Net consistently exhibits excellent detection performance, further validating its stable transferability and strong generalization capability.

5. Discussion

The DMAC-Net architecture proposed in this paper integrates Direction-Aware Granularity Enhancement (DAGE), Asymmetric Context Guided Fusion (ACGF), and multi-Granularity Dilation Aggregator (MGD). These modules collectively address the key challenges of multimodal small-object detection from UAV perspectives, including feature sparsity, strong background interference, and cross-modal heterogeneous artifacts. Although DMAC-Net demonstrates robust performance and high accuracy on the RGBTDronePerson, AVMS, and LLVIP datasets, several limitations remain. These include issues related to the underlying model mechanisms, computational cost, practical deployment, and comparison with state-of-the-art methods. Further investigation and optimization are needed in future work.

Based on the visualization results in Sections 4.3 and 4.4, DMAC-Net still shows weaknesses in some extreme scenarios. First, the ACGF module is built on an asymmetric prior assumption: the infrared modality can provide stable contour anchors. However, this assumption may fail when the infrared radiation temperature of the target is extremely close to that of the background. In such cases, a thermal crossover effect occurs. As a result, the infrared branch can no longer serve as an effective local guiding anchor, and the quality of cross-modal feature alignment decreases. In addition, the model still has room for improvement in fine-grained classification. This issue is particularly evident in the RGBTDronePerson dataset. When targets such as “rider,” “person,” and “crowd” have highly similar shapes, appear near the image boundary, and overlap heavily, the model finds it difficult to distinguish subtle inter-class differences at extremely small scales. For example, it may still struggle to determine whether a pedestrian is riding a bicycle.

Improved detection accuracy inevitably brings higher computational complexity. According to the model complexity results in Table 2, both the number of parameters (Params) and the floating-point operations (GFLOPs) increase as more modules are introduced and refined, compared with the baseline model. In practical deployment, this creates additional challenges. UAV onboard computing platforms usually operate under strict constraints in power consumption and memory bandwidth. Therefore, real-time onboard deployment will require a better balance between accuracy and efficiency. Future work may address this issue by introducing model compression techniques, such as structured pruning and knowledge distillation. In the comparative experiments in Section 4.4, DMAC-Net was evaluated against several advanced detection algorithms. The results show that DMAC-Net outperforms other fusion-based detection models in terms of F1 score and mAP50. At the same time, the objective comparison across cross-scale scenarios also reveals an-

other limitation. DMAC-Net adopts a perception-oriented design. To handle complex cross-modal semantic heterogeneity, it relies on relatively heavy asymmetric guidance and multi-granularity interactive feature flows. This design is effective in challenging scenarios, but it also makes the model relatively heavy. As a result, in simple and conventional scenes with favorable illumination, where multimodal complementarity is less necessary, the large dual-stream architecture of DMAC-Net still has room for improvement in adaptability.

The above analysis of the model's weaknesses, computational cost, and scenario-dependent limitations helps clarify the current application boundaries of DMAC-Net. It also provides useful guidance for future research on lightweight multimodal perception systems for real-time UAV edge deployment.

6. Conclusions

This study presents DMAC-Net, a direction-aware asymmetric context-guided multimodal detection network for small object detection from UAV perspectives. The model is designed to address several persistent challenges in this task, including sparse target features, noisy cross-modal fusion, and complex illumination interference. Within the overall architecture, the DAGE backbone unit and the ACGF fusion module work together as two key components. The DAGE unit uses its direction-aware design to capture thermal edge cues from infrared images and geometric details from visible images. Its multi-granularity design also enlarges the receptive field, which helps relieve the feature extraction difficulty caused by dense small targets and large scale variation. Building on this, the ACGF module avoids the weakness of conventional equal-weight fusion, which often introduces extra noise. Instead, it follows an asymmetric decoupling strategy. Infrared features provide stable local contour guidance, while visible features offer broader semantic context. This design improves cross-modal interaction and helps suppress heterogeneous artifacts. Together with the multi-scale cross-stage feature reconstruction in DAGE, the network achieves a good balance between precision and recall. Experimental results show that DMAC-Net achieves the best mAP on the RGBTDronePerson dataset. It also maintains stable detection performance on the LLVIP dataset across cross-scale scenarios, different acquisition heights, and extreme lighting conditions.

Although DMAC-Net performs competitively across multiple metrics, several issues still need further study before it can better meet more complex practical demands. First, the current asymmetric guidance mechanism in ACGF always treats the infrared modality as the primary anchor. This design may become less effective when the thermal radiation signal of the target is extremely weak. A possible next step is to introduce an adaptive modality reliability evaluation mechanism, so that the weighting of the two modalities can be adjusted according to the scene. Second, missed detections of extremely small targets under severe occlusion remain a challenge. In addition, the network's preference for extremely small target perception may slightly affect the fine localization of medium-scale targets. Future work will therefore examine the use of a super-resolution auxiliary branch or a more general multi-scale fine-grained decoupling strategy. Finally, for deployment on resource-constrained UAV platforms, lightweight optimization remains necessary. Future research will explore approaches such as knowledge distillation and channel pruning to reduce computational cost while preserving the current detection accuracy, thereby improving the practical value of the model in real deployment scenarios [48–50].

Author Contributions: Conceptualization, Q.C. and Y.J.; methodology, Y.J. and Y.G.; software, Y.J. and Z.G.; formal analysis, Y.J.; investigation, Y.J.; resources, Q.C. and Y.G.; data curation, Y.J.; writing—original draft preparation, Y.J. and Z.G.; writing—review and editing, Y.G., S.L. and X.T.; visualization, Y.J.; supervision, Q.C.; project administration, Q.C. and Y.G.; funding acquisition, Q.C., X.T. and S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the project “Key Technologies and Demonstration Application of Manned and Unmanned Aerial Vehicle Low-Altitude Fusion Flight in Tibet” (Grant No. XZ202601ZY0091), the Fundamental Research Funds for the Central Universities (Grant No. 25CA-FUC04055), the 2026 Fundamental Research Funds for the Central Universities—Graduate Student Project (Grant No. 26CAFUC10026).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors would like to thank the College of Air Traffic Control, the College of Aviation Electronic and Electrical Engineering, and the Low-Altitude Operation Intelligent Control Research Center at the Civil Aviation Flight University of China for their support. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

- White, E.C.; Seymour, A.C.; Dale, J.; Newton, E.; Johnston, D.W. Mapping the ‘Ghost Fleet’ of Mallows Bay, Maryland Using Drone Remote Sensing. *Sci. Data* **2025**, *12*, 1547. [CrossRef] [PubMed]
- Deng, Q.; Zhang, Y.; Lin, Z.; Gao, X.; Weng, Z. The Impact of Digital Technology Application on Agricultural Low-Carbon Transformation—A Case Study of the Pesticide Reduction Effect of Plant Protection Unmanned Aerial Vehicles (UAVs). *Sustainability* **2024**, *16*, 10920. [CrossRef]
- Liu, H.; Yu, Y.; Liu, S.; Wang, W. A Military Object Detection Model of UAV Reconnaissance Image and Feature Visualization. *Appl. Sci.* **2022**, *12*, 12236. [CrossRef]
- Wang, W.; Lu, B.; Wu, C.H. Cost-effective drone monitoring and evaluating toolkits for stream habitat health: Development and application. *Environ. Monit. Assess.* **2026**, *198*, 10.
- Lyu, Z.; Gao, Y.; Chen, J.; Du, H.; Xu, J.; Huang, K.; Kim, D.I. Empowering Intelligent Low-Altitude Economy with Large AI Model Deployment. *IEEE Wirel. Commun.* **2026**, *33*, 64–72. [CrossRef]
- Matos-Carvalho, J.P.; Seman, L.O.; Stefanon, S.F.; Khreast, M.K.M.; Villarrubia González, G. A Novel YOLO26-MoE Optimized by an LLM Agent for Insulator Fault Detection Considering UAV Images. *arXiv* **2026**, arXiv:2605.19595. [CrossRef]
- Han, J.; Sun, F.; Xu, Z.; Song, L.; Fang, J. An Enhanced Algorithm Integrating YOLOv11 and ByteTrack for Small-Object Detection and Tracking in Low-Altitude Remote Sensing Imagery. *Remote Sens.* **2026**, *18*, 1547. [CrossRef]
- Lei, H.; Shang, L.; Zhao, H.; Yang, W. Polarity aware detection transformer with hierarchical cross attention for unmanned aerial vehicle small object detection. *Pattern Recognit.* **2026**, *179*, 113812. [CrossRef]
- Liu, J.; Liu, Q.; Wang, R.; Qu, M. HB-YOLOv11: A model focused on enhancing the detection of remote sensing small targets in complex backgrounds. *Signal Image Video Process.* **2026**, *20*, 249. [CrossRef]
- Zhou, M.; He, S.; Wang, C.; Wang, J. EFSL-YOLO: An Improved Model for Small Object Detection in UAV Vision. *Drones* **2026**, *10*, 243. [CrossRef]
- Yang, J.; Liu, S.; Wu, J.; Su, X.; Hai, N.; Huang, X. Pinwheel-shaped convolution and scale-based dynamic loss for infrared small target detection. *arXiv* **2024**, arXiv:2412.16986. [CrossRef]
- Wang, Y.; Jiang, Y.; Zeng, W.; Cao, S. Research on detection methods for dynamic ship targets in complex marine environment from visible light images. *Eng. Rep.* **2025**, *7*, e70000. [CrossRef]
- Li, Z.; Wang, Q.; Zhao, Z. Research on small target detection in multispectral remote sensing images based on multimodal deep learning. *J. Phys. Conf. Ser.* **2024**, *2917*, 012030. [CrossRef]
- Tang, S.; Cao, C.; Lu, J.; Fan, Q.; Hu, S.; Ding, G. Resolving reliability-generalization dilemma in computational fluid dynamics surrogate models for residential environments: A multimodal feature fusion and meta-learning approach. *J. Build. Eng.* **2026**, *127*, 116336. [CrossRef]
- Jiao, T.; Guo, C.; Feng, X.; Chen, Y.; Song, J. A Comprehensive Survey on Deep Learning Multi-Modal Fusion: Methods, Technologies and Applications. *Comput. Mater. Contin.* **2024**, *80*, 1–35. [CrossRef]
- Liu, Y.; Yang, Y.; Li, X.; Yang, F.; Xie, H.; Wang, W.; Dong, C. A Deep Learning-Based Pipeline for Detecting Rip Currents from Satellite Imagery. *Remote Sens.* **2026**, *18*, 368. [CrossRef]
- Deng, Y.; Hu, Y.; Ye, Y.; Xu, P. AD-YOLO: A Unified Method for Traffic-Dense and Small Object Detection in UAV Images. *Drones* **2026**, *10*, 338. [CrossRef]

18. Yang, Y.; Guo, F.; Niu, P. UAVDet: A CNN-Mamba hybrid network for efficient small object detection in UAV imagery. *Comput. Vis. Image Underst.* **2026**, *264*, 104637. [\[CrossRef\]](#)
19. Mahaur, B.; Mishra, K.K. Small-object detection based on YOLOv5 in autonomous driving systems. *Pattern Recognit. Lett.* **2023**, *168*, 115–122. [\[CrossRef\]](#)
20. Kisantal, M.; Wojna, Z.; Murawski, J.; Naruniec, J.; Cho, K. Augmentation for small object detection. *arXiv* **2019**, arXiv:1902.07296. [\[CrossRef\]](#)
21. Chen, Y.; Zhang, P.; Li, Z.; Li, Y.; Zhang, X.; Meng, G.; Xiang, S.; Sun, J.; Jia, J. Stitcher: Feedback-driven Data Provider for Object Detection. *arXiv* **2020**, arXiv:2004.12432. [\[CrossRef\]](#)
22. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [\[CrossRef\]](#)
23. Yang, C.; Huang, Z.; Wang, N. QueryDet: Cascaded sparse query for accelerating high-resolution small object detection. *arXiv* **2021**, arXiv:2103.09136. [\[CrossRef\]](#)
24. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788. [\[CrossRef\]](#)
25. Zhou, S.; Zhou, H.; Qian, L. A multi-scale small object detection algorithm SMA-YOLO for UAV remote sensing images. *Sci. Rep.* **2025**, *15*, 9255. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Qi, H.; Qin, H.; Xiang, X.; Yang, C.; Tan, Y. LF-DETR: A Laplacian frequency enhanced DETR for aerial RGB-infrared pedestrian detection. *Remote Sens.* **2026**, *18*, 531. [\[CrossRef\]](#)
27. Chen, Y.; Wang, B.; Guo, X.; Zhu, W.; He, J.; Liu, X.; Yuan, J. DEYOLO: Dual-feature-enhancement YOLO for cross-modality object detection. *arXiv* **2024**, arXiv:2412.04931. [\[CrossRef\]](#)
28. Shen, J.; Chen, Y.; Liu, Y.; Zuo, X.; Fan, H.; Yang, W. ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection. *arXiv* **2023**, arXiv:2308.07504. [\[CrossRef\]](#)
29. Zhang, J.; Lei, J.; Xie, W.; Fang, Z.; Li, Y.; Du, Q. SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery. *arXiv* **2022**, arXiv:2209.13351. [\[CrossRef\]](#)
30. Tang, L.; Zhang, H.; Xu, H.; Ma, J. Rethinking the necessity of image fusion in high-level vision tasks: A practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity. *Inf. Fusion* **2023**, *99*, 101870. [\[CrossRef\]](#)
31. Moesl, B.; Schaffernak, H.; Vorraber, W.; Braunstingl, R.; Koglbauer, I.V. Multimodal augmented reality applications for training of traffic procedures in aviation. *Multimodal Technol. Interact.* **2022**, *7*, 3. [\[CrossRef\]](#)
32. Roy, D.; Fragulis, G.; Cantu Campos, H.A.; Martens, W.L.; Cohen, M. Spatial navigation by seated users of multimodal augmented reality systems. *SHS Web Conf.* **2021**, *102*, 04022. [\[CrossRef\]](#)
33. Hovhannisyan, S. Robust perception in degraded visual environments: A multimodal enhancement framework. *Pattern Recognit. Image Anal.* **2026**, *35*, 1003–1014. [\[CrossRef\]](#)
34. Yu, X.; Wei, Z.; Wu, H.; Lu, C.; Wu, Z.; Zhan, T. IMENet: Infrared-guided multimodal enhancement network for low-light vision. *Multimed. Syst.* **2025**, *32*, 20. [\[CrossRef\]](#)
35. Zheng, Z.; Wu, H.; Lv, L.; Bardou, D.; Niu, S.; Yu, G. MERGE: Multimodal-enhanced representation and guided ensemble for pneumonia recognition in chest X-ray images. *J. Supercomput.* **2025**, *81*, 907. [\[CrossRef\]](#)
36. Liu, Q.; Zhang, D.; Li, S. FedVPN: A Federated Multi-Modal Perception Framework for Multi-UAV in Mountain Search and Rescue. *Electronics* **2026**, *15*, 2678. [\[CrossRef\]](#)
37. Wang, B.; Ge, F.; Xu, Y.; Li, Z. CFISRO: Cross-modal feature interaction and similarity ranking optimization for image-text retrieval. *Concurr. Comput. Pract. Exp.* **2025**, *37*, e70427. [\[CrossRef\]](#)
38. Wu, J.; Zeng, J.; Dong, W.; Shi, G.; Lin, W. Blind image quality assessment with hierarchy: Degradation from local structure to deep semantics. *J. Vis. Commun. Image Represent.* **2018**, *58*, 353–362. [\[CrossRef\]](#)
39. Wei, H.; Liu, X.; Xu, S.; Dai, Z.; Dai, Y.; Xu, X. DWRSeg: Rethinking efficient acquisition of multi-scale contextual information for real-time semantic segmentation. *arXiv* **2022**, arXiv:2212.01173. [\[CrossRef\]](#)
40. Wu, T.; Tang, S.; Zhang, R.; Zhang, Y. CGNet: A light-weight context guided network for semantic segmentation. *IEEE Trans. Image Process.* **2018**, *30*, 1169–1179.
41. Khanam, R.; Hussain, M. YOLOv11: An overview of the key architectural enhancements. *arXiv* **2024**, arXiv:2410.17725. [\[CrossRef\]](#)
42. Tian, Y.; Ye, Q.; Doermann, D. YOLOv12: Attention-centric real-time object detectors. *arXiv* **2025**, arXiv:2502.12524. [\[CrossRef\]](#)
43. Ghorbani Kolahi, S.; Chaharsooghi, S.K.; Khatibi, T.; Bozorgpour, A.; Azad, R.; Heidari, M.; Hacıhaliloglu, I.; Merhof, D. MSA²Net: Multi-scale adaptive attention-guided network for medical image segmentation. *arXiv* **2024**, arXiv:2407.21640. [\[CrossRef\]](#)
44. Yang, Y.; Yuan, G.; Li, J. SFFNet: A wavelet-based spatial and frequency domain fusion network for remote sensing segmentation. *arXiv* **2024**, arXiv:2405.01992. [\[CrossRef\]](#)

45. Fang, Q.; Han, D.; Wang, Z. Cross-modality fusion transformer for multispectral object detection. *arXiv* **2021**, arXiv:2111.00273. [[CrossRef](#)]
46. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. In Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024), Vancouver, BC, Canada, 10–15 December 2024.
47. 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. [[CrossRef](#)]
48. Chen, H.; Li, Y.; Jia, Y.; Yao, G.; Zhu, R. RKF-YOLO: A Lightweight Dual-Task Model for Illegal Parking Detection and License Plate Recognition on Edge Devices. *Electronics* **2026**, *15*, 2638. [[CrossRef](#)]
49. Zhao, F.; Lou, W.; Feng, H.; Ding, N.; Li, C. MFMG-Net: Multispectral Feature Mutual Guidance Network for Visible–Infrared Object Detection. *Drones* **2024**, *8*, 112. [[CrossRef](#)]
50. Wang, Y.; Wang, Y.; Rohra, A.; Yin, B. End-to-end model compression via pruning and knowledge distillation for lightweight image super resolution. *Pattern Anal. Appl.* **2025**, *28*, 94. [[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.