A Collaborative Detection Method of Small Unmanned Aerial Vehicle Target via Multi-modal Feature Fusion in Complex Background

Wen Jiang, Keyi Zhang, Yanping Wang, Yun Lin, Fukun Bi

School of Artificial Intelligence and Computer, North China University of Technology, Beijing, China
{jiangwen, zhangkeyi_ncut, wangyp, ylin, bifukun}@ncut.edu.cn

Keywords: Small UAV Target, Low-altitude Safety, Multi-modal Feature Fusion, Collaborative Detection, Fusion Detection.

Abstract
Currently, the state-of-the-art methods for detecting small unmanned aerial vehicles (UAVs) continue to struggle in complex urban settings due to several persistent challenges, namely, frequent target occlusion, high similarity in thermal radiation signatures between UAVs and their surroundings, and the inherently low visual saliency of small UAV targets, all of which contribute to degraded detection performance. To tackle these issues, this paper introduces a novel multi-modal feature fusion collaborative detection (MFFCD) framework grounded in learnable spatial mapping. The architecture consists of three key components: firstly, a multi-branch parallel feature extraction module (MBPFE) that simultaneously processes infrared, visible, and radar range-azimuth images, complemented by a feature fusion module (FFM) designed to enhance both intra-modal and inter-modal feature interactions; then, an adaptive spatially-aware dynamic detection head module (DDH) that dynamically recalibrates feature weights to strengthen target representation and boost detection accuracy; and a feature collaborative enhancement module (FCE) that employs a learnable affine transformation to align and fuse multi-modal features, thereby producing more robust and reliable detection outcomes. Extensive experiments show that the proposed MFFCD framework substantially outperforms existing methods under challenging urban conditions, achieving a 56.89% gain in Mean Average Precision (mAP) for small UAV detection.

1. Introduction

Unmanned aerial vehicles (UAVs), particularly small-scale UAVs, have seen rapidly expanding adoption across a wide range of applications, including environmental monitoring, infrastructure inspection, and public security operations, etc. However, this growing prevalence also introduces serious safety and privacy risks, as UAVs can be misused for illicit activities such as unauthorized surveillance, smuggling, or even acts of terrorism, thereby threatening public safety and social stability. As a result, the development of accurate, reliable, and robust UAV detection technologies has emerged as a critical and urgent research imperative (Chamola et al., 2021).

In recent years, driven by the extensive use of UAVs in both civilian and military domains, substantial progress has been made in UAV detection and recognition. Researchers have developed a variety of techniques to meet the demanding requirements of complex operational environments. Current approaches leverage multiple sensing modalities, including radar-based systems (Jiang et al., 2024), radio-frequency (RF) signal analysis (Martian et al., 2021), optical imaging (Yang et al., 2024), and acoustic sensing (Al-Emadi et al., 2021). Each modality offers distinct advantages: radar enables long-range detection; RF-based methods can directly intercept UAV communication signals; infrared imaging ensures reliable performance under low-light or nighttime conditions; acoustic sensors capture the unique sound signatures generated by UAV propellers; and vision-based techniques, particularly those powered by deep learning, have shown exceptional capabilities in processing visual data from images and video streams.

However, in complex low-altitude urban environments, conventional single-modality detection approaches struggle significantly to meet the demands of detecting small UAV targets. Specifically, visible-light sensors are hampered by poor visibility, frequent occlusions, and atmospheric scattering effects, such as attenuation caused by rain or fog, exacerbated by the low-altitude flight profiles of UAVs. Moreover, the inherently small physical size of these UAVs results in limited pixel occupancy, leading to indistinct edges and substantial degradation of discriminative visual features. Infrared sensors, despite their ability to capture thermal signatures, encounter reduced thermal contrast in low-altitude urban settings due to interference from numerous ambient heat sources (Fu et al., 2024). Meanwhile, radar-based systems suffer from strong ground clutter, which causes the Doppler shifts of slow-moving UAVs to blend with those of stationary or slowly varying background objects, making it difficult to isolate true target returns from noise. In light of these limitations, multi-modal fusion strategies have attracted growing interest in recent years. By synergistically combining complementary information from heterogeneous sensing modalities, these approaches offer enhanced robustness, reliability, and adaptability, particularly for the challenging task of detecting small, low-altitude UAV targets in cluttered urban scenes (Ma et al., 2024).

Existing bimodal target detection methods fall into two main categories: traditional and deep learning-based approaches. Traditional methods rely on handcrafted features (e.g., HOG, SIFT, LBP) and multi-modal fusion strategies, typically implemented via sliding-window or region-proposal techniques (Girshick et al., 2014). The former scans images with fixed windows to extract features, while the latter uses algorithms like Selective Search to generate candidate regions and fuses modality-specific classification scores using Bayesian rules (Uijlings et al., 2013). However, these handcrafted features often lack representational capability, limiting detection performance. In contrast, deep learning-based methods automatically learn and adaptively fuse cross-modal features through neural networks, leveraging CNNs, attention mechanisms, and end-to-end optimization to exploit modality complementarity. For instance, König et al. (2017) introduced a fully convolutional RPN for visible-infrared pedestrian

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-143-2026 | © Author(s) 2026. CC BY 4.0 License.

detection firstly to integrate multi-spectral fusion into detection. Yang et al. (2024) employed multi-scale CNNs with saliency-aware attention for dynamic infrared-visible fusion. Yang et al. (2023) combined latent low-rank representation with CNNs, low-rank separating background from sparse foreground before enhancing targets via joint optimization. Wu et al. (2024) used adversarial training in a cross-domain contrastive framework to align infrared and visible feature distributions.

Recent progress in bimodal fusion has motivated researchers to expand the approach to multi-modal fusion, aiming to harness richer complementary information from heterogeneous sensors. For example, Wen et al. (2023) combined RGB, depth, and thermal infrared images for salient object detection in robotic grasping, highlighting the benefits of integrating structural and thermal cues though their method targets static scenarios and overlooks challenges like low target saliency and dynamic background clutter. Sakellariou et al. (2024) fused visible, thermal, and radar data for UAV classification using decision-level (late) fusion; however, their system focuses on coarse UAV vs. non-UAV discrimination rather than precise small-target detection, lacking fine-grained cross-modal feature interaction and spatial alignment. Similarly, other work has merged radar, LiDAR, and thermal imaging for hazard detection in low-visibility conditions (Fritsche et al., 2017), enhancing robustness in adverse weather but neglecting visible-spectrum data essential for urban environments and not addressing small-target detection. Consequently, despite significant advances, existing multi-modal methods remain largely limited to specific applications (e.g., salient object detection) or rely on late fusion, rendering them inadequate for the demanding task of detecting small UAVs in complex, low-altitude urban settings.

To enable robust detection of small UAVs in complex urban environments and overcome the limitations of current multi-modal fusion approaches for small UAV target detection, this paper proposes a novel multi-modal feature fusion collaborative detection method (MFFCD) via learnable spatial mapping mechanism. First, multi-modal features from the input infrared, visible, and radar range-azimuth images are extracted through the multi-branch parallel feature extraction module (MBPFE), during the feature extraction process, the feature fusion module (FFM) is used to enhance both intra-modal and inter-modal feature interactions. Next, the feature weights are adaptively adjusted by the dynamic detection head module (DDH) to strengthen the representation of the target region. Finally, the visible and radar feature maps are aligned through affine transformation by the feature collaborative enhancement module (FCE), followed by feature fusion to achieve collaborative enhancement of multi-modal features, further improving the robustness and accuracy of the detection.

The remainder of this paper is organized as follows. In Section 2, related work on small target detection and multi-modal fusion detection are reviewed. Section 3 provides a detailed description of the proposed MFFCD framework. Then, the experimental results of the proposed method and performance analysis on the Anti-UAV dataset are presented in Section 4. Finally, the conclusion is drawn in Section 5.

2. Related Work

2.1 Small Target Detection

Small target detection represents a key research area in computer vision and pattern recognition, centering on the fundamental challenge of accurately identifying and localizing targets that are small, low-resolution, and exhibit weak discriminative features against complex backgrounds (Chen et al., 2022). Advances in this field are accelerating the intelligent transformation of critical applications such as security surveillance, medical diagnostics, autonomous driving, and remote sensing (Chu et al., 2025).

Recent approaches to small target detection fall broadly into traditional methods and deep learning-based techniques. Traditional strategies include filter-based, contrast-based, and low-rank/sparse decomposition methods. Filter-based approaches employ operators like high-pass or edge-detection filters to enhance target-related features in images (Hadhoud et al., 1988). Contrast-based methods aim to amplify regional intensity differences, particularly in low-contrast scenes, to improve target visibility (Han et al., 2014). Low-rank and sparse-based techniques exploit the inherent structure of natural images, modelling backgrounds as low-rank components and targets as sparse outliers; this separation facilitates target extraction while suppressing noise (Gao et al., 2013). Despite their theoretical foundations, traditional methods are often limited by hand-crafted feature design, poor generalization in cluttered environments, and an inability to capture the subtle, context-dependent characteristics of small targets.

In contrast, deep learning–based methods have become the dominant paradigm in small target detection, owing to their ability to adaptively learn multi-level representations and extract high-level semantic cues from images, offering significant advantages in complex and cluttered scenes. The prevailing strategy in current deep learning approaches focus on multi-scale feature fusion, which aims to capture target information across varying spatial resolutions. This is commonly achieved using convolutional kernels of diverse sizes or dilated convolutions with different rates to modulate receptive fields and generate rich multi-scale features (Ma et al., 2024). Representative works illustrate this trend: Jiang et al. (2024) introduced a multi-scale feature fusion network that employs a parallel multi-branch architecture to extract scale-diverse features and fuses them via channel attention for adaptive weighting. Yu et al. (2023) developed the stepwise localization-based bidirectional pyramid network, which leverages parallel branches at each backbone stage to jointly encode local details and global context, followed by cross-resolution feature concatenation.

In a subsequent study, Yu et al. (2023) proposed a task-oriented heterogeneous framework integrating ResNet and Transformer modules to capture both local textures and long-range dependencies, augmented with an adaptive channel weighting mechanism that dynamically adjusts fusion weights according to target scale. Further advances include Liu et al. (2024) introduced a DNT framework, which combines a denoising feature pyramid network with transformer-enhanced R-CNN to tackle micro-target detection under extreme noise and minimal pixel footprint. Wu et al. (2022) presented a collaborative optimization approach that unifies feature and spatial alignment through multi-scale fusion guided by attention mechanisms.

In summary, while traditional methods provided the groundwork for small target detection, their dependence on handcrafted features limits adaptability in complex scenes. By contrast, deep learning-based approaches, especially those leveraging multi-scale feature fusion and attention mechanisms, have become the dominant paradigm, delivering significantly improved robustness and accuracy for detecting small targets.

2.2 Multi-modal Fusion Detection

Single-modal detection methods are inherently constrained by the physical limitations of their sensors. In contrast, multi-modal sensing leverages heterogeneous perception mechanisms to capture complementary and richer target information, leading to more stable and reliable detection performance (Zhang et al., 2021). This advantage has spurred growing research interest in multi-modal fusion strategies for target detection.

In dual-modal fusion, Deng et al. (2021) proposed a dynamic fusion strategy that strengthens cross-modal interaction and feature complementarity between visible and infrared data by adaptively combining texture-rich visible details with thermal signatures from infrared images, significantly boosting weak target discrimination under varying lighting conditions. Qing et al. (2021) introduced a Transformer-based cross-modality fusion framework for multispectral detection, using self-attention to model long-range dependencies and complex inter-modal relationships, thereby enhancing global semantic understanding beyond local fusion. Shen et al. (2024) developed an image fusion method based on independent component analysis, decomposing multi-modal inputs into independent components and recombining them in both spatial and frequency domains to preserve visible structural clarity while enhancing infrared thermal contrast. Similarly, Zhang et al. (2023) designed an enhanced YOLO architecture for infrared-visible detection, incorporating a super-resolution feature enhancement module to mitigate low pixel occupancy of small targets and employing attention mechanisms to facilitate effective cross-modal feature interaction. In radar-image fusion-based detection, researchers have expanded to more diverse sensor pairings, such as camera-LiDAR integration. Liu et al. (2024) introduced a cross-modal attention guidance mechanism that aligns visual and LiDAR features, allowing each modality to emphasize complementary cues while reducing redundancy.

More recently, advanced fusion paradigms have emerged. For instance, Zhang et al. (2021) presented an attention-guided, multi-level dynamic fusion framework that adaptively assigns weights across modalities and scales, achieving robust detection in low-light and cluttered environments. This approach not only aligns cross-modal features but also enables multi-scale collaborative optimization, outperforming static fusion methods. Similarly, Qing et al. (2022) proposed a dual-path attention mechanism that jointly enhances shared and modality-specific features. By dynamically balancing common and differential information, their method significantly boosts detection robustness under challenging lighting conditions where conventional visible-infrared fusion often falters.

These methods clearly benefit from fusing complementary sensing cues, especially for improving detection in low-light, cluttered, or adverse-weather conditions. However, they remain inadequate for detecting small UAVs in complex low-altitude urban environments. The primary reason is that most existing approaches are tailored for scenarios with minimal background interference or highly salient targets. Consequently, their performance sharply declines when faced with urban-specific challenges, such as occlusions, thermal contrast degradation due to ambient heat sources, and dynamic clutter from elements like swaying trees. Moreover, intrinsic characteristics of small UAVs, including limited pixel footprint, low visual or thermal saliency, and slow motion, further compound these limitations, creating significant bottlenecks for current detection systems in real-world urban settings.

3. Methodology

The proposed MFFCD method is described in detail in this section, and illustrated in Figure.1. Firstly, multi-modal features from the input infrared, visible, and radar range-azimuth images are extracted through the MBPFE module. During the feature extraction process, the FFM module is used to enhance both intra-modal and inter-modal feature interactions. Next, the feature weights are adaptively adjusted by the DDH module to strengthen the representation of the target region. Finally, the visible and radar feature maps are aligned through affine transformation by the FCE module, followed by feature fusion to achieve collaborative enhancement of multi-modal features.

3.1 Multi-branch Parallel Feature Extraction Module

The input infrared, visible, and RA images, denoted as I_I , I_V , and I_R , are first processed through the feature extraction network to generate multi-scale feature maps at five levels. For the infrared and visible modalities, intermediate feature maps, e.g., the second-layer maps F_{I2} and F_{V2} , are input into the FFM, where attention-based fusion adaptively models intra-modal dependencies and inter-modal correlations, yielding enhanced features $F_{I2}^{'}$ and $F_{V2}^{'}$, which is depicted as follows.

$$\begin{aligned} F_{I2}^{'} &= FFM(F_{I2}) \\ F_{V2}^{'} &= FFM(F_{V2}) \end{aligned} \tag{1}$$

The fused features $F_{I2}^{'}$ and $F_{V2}^{'}$ are then element-wise added with F_{I2} and F_{V2} , respectively, to obtain the third layer feature map as follows.

$$\begin{aligned} F_{I3} &= F_{I2} \oplus F_{I2}^{'} \\ F_{V3} &= F_{V2} \oplus F_{V2}^{'} \end{aligned} \tag{2}$$

Similarly, the third and fourth layers undergo the same feature fusion process. Finally, the feature extraction networks of the first and second streams output fused feature maps at three different scales. These three scale feature maps are then element-wise added to obtain the final infrared-visible fused feature maps, F_{IV3} , F_{IV4} , and F_{IV5} .Then, by down-sampling F_{IV3} and up-sampling F_{IV5} , the scales are adjusted to match that of F_{IV4} , after which they are concatenated along the channel dimension. This result is the hierarchical feature tensor F_{IV} , which comprises three dimensions.

$$F_{IV} = down(F_{IV3}) © F_{IV4} © up(F_{IV5}) \tag{5}$$

In the same way, the radar also generates a hierarchical feature tensor F_R , as shown below:

$$F_R = down(F_{R3}) © F_{R4} © up(F_{R5}) \tag{6}$$

In summary, after the processing of the MBPFE, two hierarchical feature tensors and are produced: one derived from the fused infrared-visible streams and the other from the radar stream. The former enhances texture and thermal features, while the latter provides radar spatial information, jointly forming a multi-modal feature space that underpins adaptive detection.

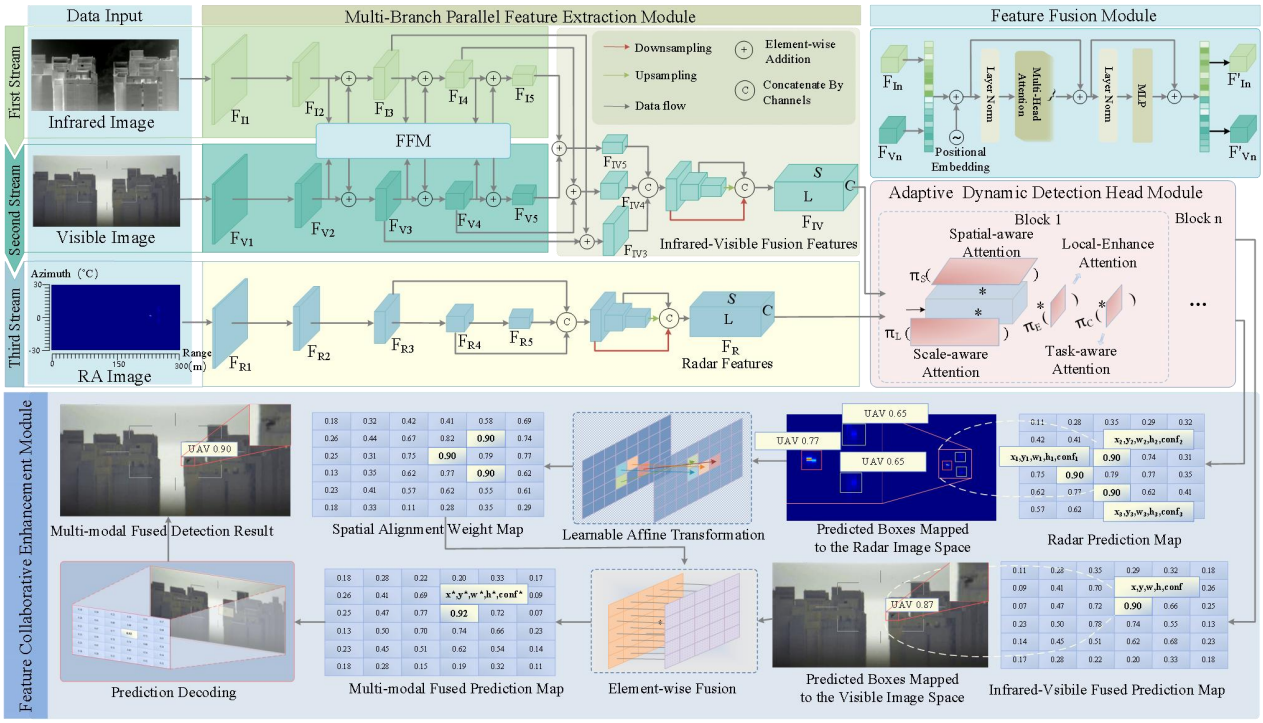

Figure 1. The framework of multi-modal feature fusion collaborative detection for small UAV targets

3.2 Dynamic Detection Head Module

A detailed description of the proposed DDH is provided, which adaptively adjusts feature importance to emphasize high-resolution details of small targets, enhancing sensitivity to UAVs while suppressing background interference.

For the input hierarchical feature tensor F, it is processed through scale-aware attention $\pi_L(\cdot)$, spatial-aware attention $\pi_S(\cdot)$, and task-aware attention $\pi_C(\cdot)$, resulting in a multi-dimensional attention-enhanced feature representation. The formula is as follows:

$$W(F) = \pi_C(\pi_S(\pi_L(F)\cdot F)\cdot F)\cdot F \tag{7}$$

The $\pi_L(\cdot)$ mechanism adaptively fuses multi-scale features to balance fine details and semantic information for targets of different sizes.

$$\pi_L(F)\cdot F = \sigma(f(\frac{1}{SC}\sum_{S,C}F))\cdot F \tag{8}$$

where $f(\cdot)$ is a linear function approximated by a 1×1 convolutional layer, and the activation function $\sigma(\cdot)$ used is the hard-sigmoid function.

The $\pi_S(\cdot)$ mechanism uses deformable convolution and cross-level fusion to focus on critical spatial regions, capturing both fine details and high-level semantic context.

$$\pi_S(F)\cdot F = \frac{1}{L}\sum_{l=1}^{L}\sum_{k=1}^{K} w_{l,k}\cdot F(l; p_k + \Delta p_k; c)\cdot \Delta m_k \tag{9}$$

where k denotes the number of sparse sampling locations, $p_k + \Delta p_k$ represents a shifted location and Δm_k denotes a self-learned importance scalar at location p_k.

The $\pi_C(\cdot)$ mechanism dynamically adjusts feature importance for different detection tasks, emphasizing relevant features and suppressing irrelevant ones to improve performance and robustness.

$$\pi_c(F) = \max(\alpha_1(F)\cdot F_c + \beta_1(F), \alpha_2(F)\cdot F_c + \beta_2(F)) \tag{10}$$

where F_c represents the feature slice at the c-th channel, and α_1, α_2, β_1, β_2 are hyperfunctions that learn to control the activation thresholds.

Loss Function: The loss function includes classification loss L_{cls}, confidence loss L_{obj}, bounding box regression loss L_{box}, and alignment loss L_{align}. The L_{align} aligns radar features with fused infrared-visible features by minimizing their spatial differences, which can be computed as follows.

$$L_{align} = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\left\| F_{radar}^{trans}(i,j) - F_{IVf}(i,j)\right\|_2^2 \tag{11}$$

where F_{radar}^{trans} represents the transformed radar prediction map, F_{IVf} represents the fused infrared-visible prediction map.

Therefore, the overall loss function L_{total} can be defined as follows.

$$L_{total} = \lambda_{cls}\cdot L_{cls} + \lambda_{obj}\cdot L_{obj} + \lambda_{box}\cdot L_{box} + \lambda_{align}\cdot L_{align} \tag{12}$$

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-143-2026 | © Author(s) 2026. CC BY 4.0 License.

where λ_{cls} , λ_{obj} , λ_{box} , λ_{align} are the weight coefficients for each component.

3.3 Feature Collaborative Enhancement Module

This proposed FCE module aligns radar and infrared-visible features through affine transformation and bilinear interpolation, enabling multi-modal feature fusion and precise target localization.

The radar prediction map is spatially aligned with the infrared-visible map via a 2D affine transformation, represented by a homogeneous matrix H that encodes scaling, rotation, and translation.

$$\begin{aligned} H(\Theta) &= T(t_x,t_y)R(\theta)S(s_x,s_y) \\ &= \begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{aligned} \quad (13)$$

where $\Theta = \{s_x, s_y, \theta, t_x, t_y\}$ denotes the learnable parameters.

Then, the matrix H is applied to the radar prediction map to generate a spatial alignment weight map, synchronizing radar and infrared-visible features by adjusting position and orientation.

$$\begin{bmatrix} x_{align} \\ y_{align} \\ 1 \end{bmatrix} = H(\Theta) \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (14)$$

where (x, y) are the original coordinates of the radar prediction map, and (x_{align}, y_{align}) are the transformed and aligned coordinates after the transformation.

Next, the aligned coordinates are bilinearly interpolated to generate a weight map matching the infrared-visible prediction map, which is then element-wise multiplied to fuse radar and infrared-visible features.

$$\begin{aligned} F_{fusion} &= F_{IVf} \odot F_{radar}^{align} \\ F_{radar}^{align}(x',y') &= \sum_{i=0}^{1}\sum_{j=0}^{1} \omega_{ij} F_{radar}(x_0+i, y_0+j) \\ \omega_{ij} &= (1-|i-\alpha|)(1-|j-\beta|), \alpha = x-x_0, \beta = y-y_0 \end{aligned} \quad (15)$$

where (x', y') denotes the interpolated coordinates, (x_0, y_0) represents the integer grid points, α and β are the offset values, while i and j are indices used in the bilinear interpolation process to calculate the weights ω_{ij} . Finally, the fused multi-modal prediction map is decoded by the detection head to regress bounding box parameters and confidence scores, which are then mapped onto the visible image for accurate localization and visualization of targets.

4. Experimental Results and Analysis

4.1 Datasets and Experimental Detail

The Anti-UAV dataset was adopted for experiment, which contains infrared and visible video data of UAVs. Videos recorded at 20 frames per second are sampled every 10 frames, and the resulting images are cropped to 640×312 pixels, yielding a diverse dataset of 4,656 image pairs, with 3,724 used for training and 932 for testing. Simultaneously, this paper performs inverse imaging simulation on the original radar echo data by leveraging the azimuth and pitch angles derived from the infrared and visible multimodal dataset. As a result, a set of radar feature images is constructed, each characterized by range-azimuth two-dimensional resolution, resulting in a total of 4,656 such images.

All experiments were conducted on a GeForce RTX 3060Ti GPU for 300 epochs, using a batch size of 16, a learning rate of 0.0001, and the Adam optimizer. Model performance was evaluated using four metrics: precision (P), recall (R), mean average precision (mAP), and F1-Score (F1).

4.2 Comparative Experimental Results and Analysis

To evaluate MFFCD, comparative experiments were conducted on infrared-visible and visible-radar fusion, with detection results visualized in Fig. 2 (blue boxes indicate targets with category and confidence). The quantitative results are summarized in Table 1. Comparative experiments show that the proposed MFFCD method outperforms existing approaches in small UAV detection. Using only infrared-visible fusion, CFT (Fritsche et al., 2017) achieves 0.647 recall, with noticeable false detections. ICAFusion (Chen et al., 2022) improves precision to 0.637 and balances recall at 0.633. Adding radar (YCANet (Chu et al., 2025)) raises precision to 0.831, highlighting radar’s complementary information. With infrared, visible, and radar fusion, MFFCD reaches map of 0.899, demonstrating significant gains in accuracy and robustness in complex urban environments.

Method	Modality	mAP	P	R	F1
CFT	V+IR	0.573	0.574	0.647	0.608
ICAFusion	V+IR	0.601	0.637	0.633	0.635
YCANet	V+RA	0.820	0.831	0.863	0.847
BEVFusion	V+RA	0.811	0.803	0.825	0.814
MFFCD	V+IR+RA	**0.899↑**	**0.913↑**	**0.935↑**	**0.924↑**

Table 1. Comparison results of different methods

4.3 Ablation experimental results and analysis

The ablation study results, shown in Table 2, demonstrate the impact of the DDH and FCE modules. Adding DDH increases precision to 0.895, demonstrating that adaptive feature weighting enhances target representation. With FCE, mAP rises to 0.899, indicating that affine-aligned multi-modal fusion significantly improves detection accuracy and robustness. These results confirm that the proposed module design effectively strengthens feature collaboration and contributes to more reliable UAV detection in complex urban environments.

Metrics	mAP	P	R	F1
Baseline	0.573	0.574	0.647	0.608
Baseline+DDH	0.820	0.831	0.863	0.847
Baseline+DDH+FCE	**0.899↑**	**0.913↑**	**0.935↑**	**0.924↑**

Table 2. Ablation results for module stacking

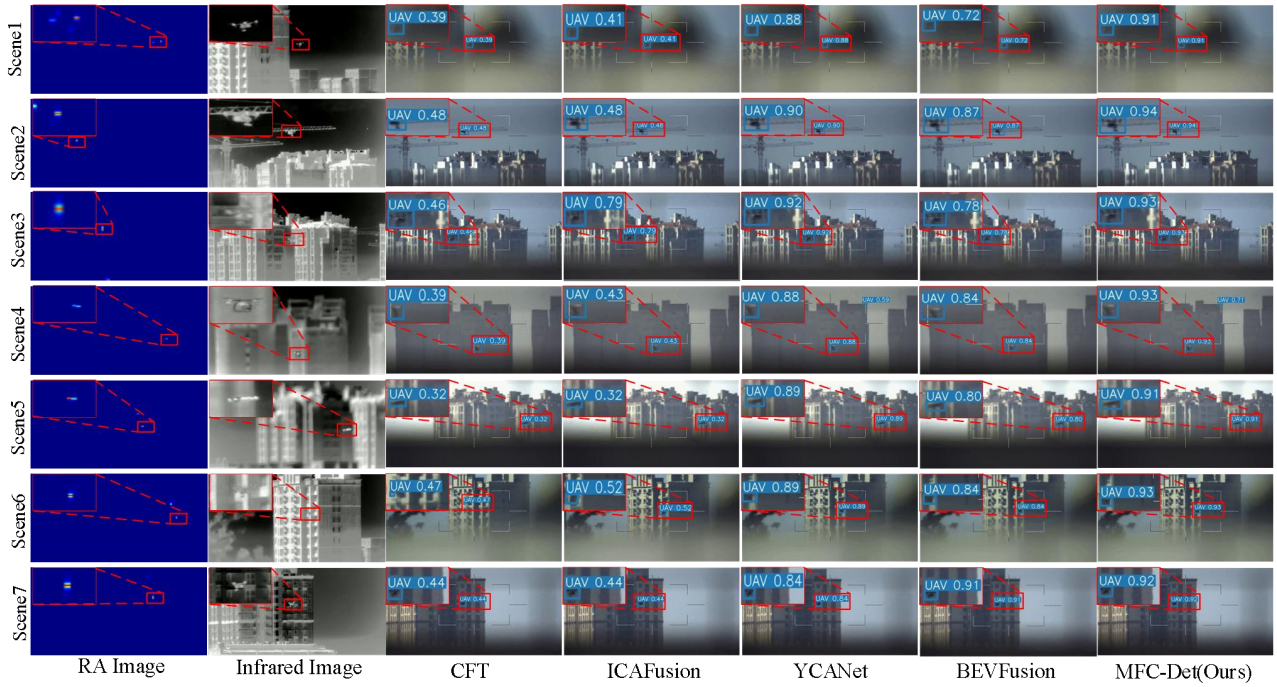

Figure 2. Comparative experimental results under different scenarios

5. Conclusion

This paper proposed the MFFCD framework for small UAV target detection in complex urban environments. By fusing infrared, visible, and radar data, it leverages the complementary strengths of information of heterogeneous sensors, in which the MBPFE module extracts infrared, visible, and radar range-azimuth features with a FFM for enhanced interactions, then, the DDH module adjusts feature weights for improved accuracy, and the FCE module aligns and fuses features via affine transformation. Experiments results show that the proposed MFFCD framework effectively addresses occlusion, low thermal contrast, and small-target challenges, achieving superior accuracy and robustness in small UAV target detection in complex background.

Acknowledgements

This work was supported by the Natural Science Foundation of China (Key Program) under Grants 62131001 and (General Program) under Grants 62371005, the Beijing Natural Science Foundation under Grant 4234082, the R&D Program of Beijing Municipal Education Commission (No. 110052972508-20), and the Youth Research Special Project of NCUT (Project No. 2025NCUTYRSP010).

References

Chamola, V., Kotesh, Agarwal, P.A., Gupta, N., Guizani, M., 2021. A comprehensive review of unmanned aerial vehicle attacks and neutralization techniques, *Ad Hoc Networks*, vol. 111, pp. 102324.

Jiang, W., Liu, Z., Wang, Y., Lin, Y., Li, Y., and Bi, F., 2024. Realizing small UAV targets recognition via multi-dimensional feature fusion of high-resolution radar, *Remote Sens.*, vol. 16, pp. 2710.

Martian, A., Chiper, F.L., Craciunescu, R., Vlădeanu, C., Fratu, O., and Marghescu, I., 2021. RF-based UAV detection and defense systems: Survey and a novel solution, in Proc. 2021 *IEEE Int. Black Sea Conf. Commun.* Netw. (BlackSeaCom), Bucharest, Romania, pp. 1-4.

Yang, B., Zhang, X., Zhang, J., Luo, J., Zhou, M., and Pi, Y., 2024. EFLNet: Enhancing feature learning network for infrared small target detection, *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1-11, Art no. 5906511.

Al-Emadi, S., Al-Ali, A., 2021. Audio-based drone detection and identification using deep learning techniques with dataset enhancement through generative adversarial networks, *Sensors*, vol. 21, pp. 4953.

Fu, H. et al., 2024. LRAF-Net: Long-range attention fusion network for visible-infrared object detection, *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 10, pp. 13232-13245.

Ma,X., and Tong, J., 2024. Enhanced small target detection via multi-modal fusion and attention mechanisms: A YOLOv5 approach. *IEEE Smart World Congr.*, Nadi, Fiji, pp. 1886-1891.

Girshick,R., Donahue, J., Darrell, T., and Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proc. *2014 IEEE Conf. Comput. Vis. Pattern Recognit*, Columbus, OH, USA, 2014, pp. 580-587.

Uijlings, J.R.R., Sande, K., Gevers, T., and Smeulders, A., 2013. Selective search for object recognition, *Int. J. Comput. Vis.*, vol. 104, pp. 154-171.

König, D., Adam, M., Jarvers, C., Layher, G., Neumann, H., and Teutsch, M., 2017. Fully convolutional region proposal networks for multi-spectral person detection,” in Proc. *IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul., pp. 49-56.

Yang, C., He, Y., Sun, C., Chen, B., Cao, J., Wang, Y., and Hao, Q., 2024. Multi-scale convolutional neural networks and saliency weight maps for infrared and visible image fusion, *J. Vis. Commun. Image Represent.*, vol. 98, pp. 104015.

Yang, Y., Gao, C., Ming, Z., Guo, J., Leopold, E., Cheng, J., Zuo, J., and Zhu, M., 2023. LatLRR-CNN: An infrared and visible image fusion method combining latent low-rank representation and CNN, *Multimedia Tools Appl.*, vol. 82, no. 23, pp. 36303-36323.

Wu, H., Zhu, Y., and Li, S., 2024. CDYL for infrared and visible light image dense small object detection, *Sci. Rep.*, vol. 14, no. 1, pp. 3510.

Wen,H., Song, K., Huang, L., Wang, H., Wang, J., and Yan, Y., 2023. Hierarchical two-stage modal fusion for triple-modality salient object detection, *Measurement*, vol. 218, pp. 113180.

Sakellariou, N., Lalas, A., Votis, K., and Tzovaras, D., 2024. Multi-sensor fusion for UAV classification based on feature maps of image and radar data, *arXiv preprint, arXiv:2410.16089*.

Fritsche, P., Zeise, B., Hemme, P., and Wagner, B., 2017. Fusion of radar, LiDAR and thermal information for hazard detection in low visibility environments,” in Proc. *2017 IEEE Int. Symp. Saf., Secur. Rescue Robot.*, Shanghai, China, pp. 96-101.

Chen, G., et al., 2022. A survey of the four pillars for small object detection: Multiscale representation, contextual information, super-resolution, and region proposal, *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 52, no. 2, pp. 936-953.

Chu, T., Zhou,H., Ren, Z., Ye, Y., Wang, C., and Zhou, F., 2025. Intelligent detection of low-slow-small targets based on passive radar, *Remote Sens.*, vol. 17, pp. 961.

Hadhoud, M, and Thomas, D., 1988. The two-dimensional adaptive LMS (TDLMS) algorithm, *IEEE Trans. Circuits Syst.*, vol. CS-35, no. 5, pp. 485-494.

Han, J., Ma, Y., Zhou, B., Fan, F., Liang, K., and Fang, Y., 2014. A robust infrared small target detection algorithm based on human visual system, *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 12, pp. 2168-2172.

Gao, C., Meng, D., Yang, Y., Wang, Y., Zhou, X., and Hauptmann, A., 2013. Infrared patch-image model for small target detection in a single image, *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4996-5009.

Ma, T., Wang, H., Liang, J., Peng, J., Ma, Q., and Kai, Z., 2024. MSMA-Net: An infrared small target detection network by multiscale super-resolution enhancement and multilevel attention fusion, *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Art. no. 5602620.

Jiang, L., et al., 2024. MFFSODNet: Multiscale feature fusion small object detection network for UAV aerial images, *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1-14.

Yu, N., Ren, H., Deng, T., and Fan, X., 2023. Stepwise locating bidirectional pyramid network for object detection in remote sensing imagery, *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1-5, Art. no. 6001905.

Yu, Y., Yang, X., Li, J., and Gao, X., 2023. Task-specific heterogeneous network for object detection in aerial images, *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1-15, Art. no. 5620315.

Liu, H., Tseng, Y., Chang, K., Wang, P., Shuai, H., and Cheng, W., 2024. A denoising FPN with transformer R-CNN for tiny object detection, *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1-15, Art. no. 4704415.

Wu, J., Pan, Z., Lei, B., and Hu, Y., 2022. FSANet: Feature-and-spatial-aligned network for tiny object detection in remote sensing images, *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1-17, Art. no. 5630717.

Zhang, L., et al., 2021. Weakly aligned feature fusion for multimodal object detection, *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Aug. 26.

Deng, Q., Tian, W., Huang, Y., Xiong, L., and Bi, X., 2021. Pedestrian detection by fusion of RGB and infrared images in low-light environment, in Proc. *IEEE 24th Int. Conf. Inf. Fusion*, Nov., pp. 1-8.

Qing, F., Da, H., and Zhao, W., 2021. Cross-modality fusion transformer for multispectral object detection, *arXiv preprint, arXiv:2111.00273.*

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

Zhang, J., Lei, J., Xie, W., Fang, Z., Li, Y., and Du, Q., 2023. SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery, *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1-15, Art. no. 5605415.

Liu, H., Tseng, Y., Chang, K., Wang, P., Shuai, H., and Cheng, W., 2024. YCANet: Target detection for complex traffic scenes based on camera-LiDAR fusion, *IEEE Sens. J.*, vol. 24, no. 6, pp. 8379-8389, Mar. 15.

Zhang, H. Fromont,, E., Lefevre, S., and Avignon, B., 2021. Guided attentive feature fusion for multispectral pedestrian detection, in Proc. *IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2021, pp. 72-80.

Qing, F., and Zhao, W., 2022. Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery, *Pattern Recognit.*, vol. 130, p. 108786.