

Article

Target Detection for UAV Inspection in Complex Illumination Environments: A Lightweight Dual-Modal Fusion Method with Multi-Directional Mamba Cross-Attention

Ziyue Luo ^{1,†}, Ziyi Zhou ^{1,†}, Haishan Tian ^{1,*}, Ziyi Zhong ¹, Kewei Li ¹, Yujie Ma ¹, Weiwei Tang ¹ and Kui Wang ^{2,3}

¹ School of Physics and Electronics, Hunan Normal University, Changsha 410081, China; 202370112335@hunnu.edu.cn (Z.L.); zhouziyi@hunnu.edu.cn (Z.Z.); 202370112346@hunnu.edu.cn (Z.Z.); 202270112395@hunnu.edu.cn (K.L.); mayujie@hunnu.edu.cn (Y.M.); vivitang@hunnu.edu.cn (W.T.)

² National Engineering and Technology Research Center for Emergency and Disaster Relief Equipment, PLA Joint Logistics Support Force University of Engineering, Chongqing 401331, China; kuiwang.nudt@hotmail.com

³ Chongqing Gangli Environmental Protection Co., Ltd., Chongqing 400074, China

* Correspondence: tianhaishan@hunnu.edu.cn

† These authors contributed equally to this work.

Highlights

What are the main findings?

- A visible-light and infrared target detection method based on multi-directional Mamba cross-attention feature fusion is proposed. On the self-built OPVM-VIRD dataset, the model yields an mAP50 of 95.7% while notably reducing the parameter count.
- Three core modules are introduced: a Mixed Scanning Fusion Module (MSFM), a Separate Spectrum Enhancement Module (SSEM), and a Multi-Scale Enhanced Convolutional Module (MSEC). Their combination exhibits strong synergistic effects, effectively balancing parameter efficiency and detection accuracy.

What are the implications of the main findings?

- The proposed method demonstrates robust detection performance under various challenging conditions, including low illumination, severe occlusion, small targets, and dense scenes. In UAV-based security patrols over urban and industrial areas, it effectively addresses detection difficulties caused by complex lighting, such as low light, fog, and strong backlight.
- This algorithm not only provides a high-precision solution for visible-light and infrared dual-modal object detection but also introduces a new research direction for the application of the Mamba architecture to cross-modal fusion tasks.

Academic Editor: Haipeng Wang

Received: 8 May 2026

Revised: 18 June 2026

Accepted: 10 July 2026

Published: 12 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Abstract

Unmanned aerial vehicles (UAVs) frequently operate under complex illumination conditions during urban and industrial security patrols, including low light, fog, and strong backlighting. These conditions severely degrade the performance of existing dual-modal object detection methods. To overcome this limitation, we introduce a lightweight dual-modal object detection algorithm called M²CA-FFM, built upon multi-directional Mamba cross-attention feature fusion. First, a multi-directional Mamba enhancement module is developed to model spatial feature sequences via four-directional scanning. Second, a dedicated spectral enhancement module is designed to process spectral components in a

differentiated manner. Together, these two modules enhance single-modal features from both the spatial and frequency domains, leading to improved feature robustness. For dual-modal object detection, we propose a Mixed Scanning Fusion Module that dynamically generates scanning paths focused on the target regions. This dynamic mechanism complements the fixed four-direction scanning and strengthens global modeling, enabling efficient fusion and deep complementarity between visible light and infrared features. Extensive experiments are conducted on a self-built dataset (OPVM-VIRD) and three public benchmarks (FLIR, M3FD, and VEDAI). Our method consistently outperforms current mainstream dual-modal detection algorithms across all datasets. On the OPVM-VIRD dataset, it achieves an mAP50 of 95.7%, which is 2.2% higher than the best result obtained with other fusion methods. Preliminary verification on the edge computing platform Atlas 200I DK A2 shows that the algorithm has significant lightweight advantages, with a parameter volume of only 72.07 M and a computational complexity of 134.41 GFLOPs, demonstrating the reasoning potential to meet real-time processing requirements. Overall, the proposed M²CA-FFM method offers a real-time detection strategy with strong generalization capabilities, holding significant practical value for around-the-clock security patrols in urban and industrial parks.

Keywords: dual-modal object detection; urban security patrol; cross-attention; feature fusion; Mamba; Fourier transform

1. Introduction

Multispectral imaging, combined with the maneuverability, wide coverage, and low cost advantages of UAV platforms, has been widely adopted in security patrols in urban areas and industrial parks. By integrating cross-spectral information, multimodal fusion overcomes the inherent deficiencies of single-spectrum approaches, leading to marked improvements in both the reliability and precision of target detection under complex conditions. It has unique advantages, especially in UAV application scenarios with stringent requirements for detection efficiency and reliability, such as high-precision inspections and rapid emergency responses.

In multispectral fusion target detection tasks, Convolutional Neural Networks (CNNs) [1] and Transformers [2] are the current mainstream feature extraction and fusion methods dominating the development direction of the technology in this field. CNNs rely on local receptive fields and hierarchical architecture to efficiently mine the local textures, edges, and other detailed features of each modality. The computational cost of the fusion process is controllable. It has strong robustness to interference such as slight blurring and local occlusion caused by UAV flight [3–5] and is suitable for the real-time processing requirements of low-altitude scenes. The self-attention mechanism in a Transformer enables global modeling [6–8], overcoming the local perception constraint of CNNs to effectively capture global correlations across cross-modal features. This capability significantly enhances the semantic consistency of fused features, making Transformers a growing research focus in cross-modal fusion [9–11]. Meanwhile, effectively exploiting cross-modal spectral complementary information remains a critical challenge. The Cross-Modal Spectral Complementary and Memory Prompt Network proposed by Jiang et al. [12] offers valuable insights for dynamic feature modeling through its memory prompting mechanism.

Despite these advances, fusion methods based on CNNs and Transformers still face several unresolved challenges: (1) CNNs remain constrained by their inherently local re-

ceptive fields, which limits their capacity for global context extraction. (2) The computational complexity of Transformer self-attention is quadratic with regard to feature map resolution, making it difficult to meet the real-time detection demands of high-resolution visible–infrared images. (3) The processing and fusion design capabilities of frequency domain features are insufficient, making it difficult to fully utilize the insensitivity of frequency domain features to sudden changes in illumination and background interference, thus failing to effectively supplement spatial domain features.

With its global perception capability of linear complexity, Mamba significantly reduces computational resource consumption while maintaining excellent performance [13,14], making it a potential alternative to a Transformer. Therefore, in this paper, we propose a visible light and infrared fusion target detection method based on a multi-directional Mamba cross-attention feature fusion module (M²CA-FFM). The attention mechanism based on multi-directional Mamba is used to enable the long-distance feature interaction of multispectral images. First, we extend the single-input Mamba module to support dual inputs and create a Mixed Scanning Fusion Module (MSFM). Within the single-modal feature enhancement stage, we designed a Separable Spectral Enhancement Module (SSEM) based on Fast Fourier Transform to improve the information processing capability of the algorithm in the frequency domain. Inspired by advancements in hyperspectral image processing, DFBSNet [15] was used to construct a dual-branch fusion and selection mechanism in the frequency domain. This approach has demonstrated strong robustness against abrupt illumination changes and background interference, further validating the rationale for introducing a frequency-domain enhancement module in our proposed method. In addition, a dual-stream feature fusion network framework was designed based on YOLOv8 and incorporated into the MSFM and SSEM. This method not only integrates global contextual information efficiently but also captures complementary cross-modal features more effectively. As a result, it enhances both the accuracy and robustness of multispectral target detection while keeping the parameter count low.

The main contributions of this work are as follows:

1. A multi-directional Mamba cross-attention feature fusion framework is introduced for the first time, which comprises the following:
 - (a) A Mixed Scanning Fusion Module (MSFM) that enables efficient global context learning and feature fusion through a complementary scheme combining fixed four-directional scanning with dynamic focus scanning.
 - (b) A method is proposed for a Separable Spectral Enhancement Module (SSEM) based on Fast Fourier Transform, which differentiates the amplitude spectrum and phase spectrum, effectively suppressing frequency-domain redundancy and improving frequency-domain processing capabilities.
2. The system was validated on OPVM-VIRD (a self-built urban/industrial area UAV inspection dataset) and general datasets such as FLIR, M3FD and VEDAI. The proposed method achieves an mAP50 value of 95.7% on OPVM-VIRD, which is 2.2% better than the best value obtained using other fusion methods. The results on general datasets are also the best by comparison.
3. The edge computing performance was evaluated. Tests were conducted on the Huawei Atlas 200I DK A2 development kit, revealing that M²CA-FFM achieves low computational complexity (72.07 M parameters/134.41 GFLOPs) and a preliminary inference speed of 48.3 FPS, laying the groundwork for future deployment on UAV platforms.

2. Related Work

In CNN fusion research, the main issue depends on the choice of fusion level and strategy [16]: early fusion is easily affected by modal differences; late fusion may lose the

complementary information of the bottom layer, and mid-level fusion achieves local semantic information synergy through the interaction of intermediate layer features and improves fusion effectiveness, for example, when using the Fusion R-CNN method [17]. Chen et al. proposed DEYOLO [18], which employs a bi-directional decoupling attention module to enlarge the backbone network's receptive field and enhance detection performance under low-light conditions, thus facilitating mutual enhancement and fusion of infrared and visible modalities. Li et al. introduced FD2-Net [19], a CNN-based method that innovatively applies a discrete cosine Transform to decompose visible and infrared images into high- and low-frequency components. Feature enhancement and fusion are performed prior to detection to improve input image quality. Li et al. also presented MFY-OLO [20], built on the YOLOv8 framework, which directly boosts the detection capability for small and multi-scale targets in UAV aerial imagery through multi-scale feature fusion and an improved loss function. However, despite the strong performance of CNNs in local feature extraction and fusion, they are inherently constrained by their local receptive fields. This limitation leads to a notable deficiency in capturing global context and long-range feature dependencies, which becomes especially critical in large-scene and multi-object overlap scenarios [21,22]. To mitigate the inherent limitations of local perception in CNNs, OCSCNet-Tracker [23] integrates octave convolutions with spatial-spectral capsule networks. This approach not only reduces computational complexity but also enhances the modeling capability for target deformation, providing valuable insights for the optimization of lightweight CNN-based fusion frameworks.

Transformers can solve the problem of long-distance dependency in visible-infrared fusion detection. The ICAFusion proposed by Shen et al. is based on the dual cross-attention Transformer framework [24], which models global feature interaction through an iterative interaction mechanism to capture cross-modal complementary information. During feature extraction, the CFT algorithm [25] employs the Transformer's self-attention mechanism to simultaneously conduct intra- and inter-modal fusion, thereby capturing long-distance dependencies and potential interactions between the visible and infrared domains. C2F-Trans further enhances target detection performance in complex scenes by integrating the CNN's strength in local feature extraction with the Transformer's global modeling capability [26]. Nevertheless, when processing high-resolution images, Transformers suffer from high computational complexity and substantial memory consumption. Moreover, conventional downsampling strategies may also result in the loss of spatial information. To address the challenges of template updating and state awareness, SiamSTU [27] employs a state-aware template update mechanism. This approach effectively suppresses feature drift while maintaining tracking stability, offering valuable insights for dynamic feature maintenance in multispectral fusion.

The latest breakthroughs in state-space models (SSMs) have provided a new direction for resolving the contradiction between global modeling and computational complexity [28]. Vision Mamba, proposed by Zhu et al., improves the model's sensitivity to visual details by optimizing the visual feature extraction mechanism, and the inference latency is significantly lower than that of similar Transformer models [29]. FusionMamba, proposed by Dong et al., is a dynamic feature enhancement framework that enhances local feature expression while retaining global modeling capabilities, and a cross-modal fusion Mamba module was designed to strengthen intermodal correlation and suppress redundant information [30]. WaveMamba, proposed by Zou et al., decomposes visible light and infrared images into different frequency components using a discrete wavelet Transform and then employs the Mamba architecture for feature integration, thus incorporating complementary data from both modalities [31]. In addition, Ma et al. first applied the visual Mamba auxiliary branch to the semantic segmentation of remote sensing images. He et al. used Mamba for panchromatic sharpening tasks, verifying its potential in cross-modal

spatial information learning. In addition to the outstanding aforementioned Mamba variants, excellent spatial-domain optimization algorithms, such as Succulent-YOLO (2025) [32] and EPRepSADet (2025) [33], have also emerged in the field of UAV detection. These methods focus on optimizing spatial-domain features and network architectures, demonstrating superior performance in scenarios that require rapid adaptation and morphological feature extraction under resource-constrained conditions. In contrast, our proposed method emphasizes the exploitation of cross-modal complementarity and frequency-domain robustness. With the MSFM, we achieve deep fusion between visible and infrared modalities. Furthermore, combined with the SSEM and Mamba architecture, our algorithm effectively suppresses noise and captures long-range dependencies in environments with severe illumination interference. In summary, the above studies collectively demonstrate the effectiveness of employing Mamba for feature enhancement and cross-modal fusion. Additionally, from the perspective of feature matching efficiency, methods based on spectral difference matching reduction [34] achieve more efficient target perception by decreasing redundant matching computations, which also provides valuable insights for designing lightweight Mamba fusion modules. Due to its global perception capability with linear computational complexity, Mamba serves as a potential alternative to Transformers, significantly reducing computational resource consumption while maintaining superior performance [35,36].

3. Method

3.1. Architecture of a Dual-Mode Object Detection Algorithm Based on Multi-Directional Mamba Cross-Attention Feature Fusion

For visible–infrared dual-modal target detection scenarios, we propose a cross-attention feature fusion approach leveraging multi-directional Mamba. This approach not only strengthens the model’s capacity to deeply extract single-modal features and model long-range global dependencies but also keeps the parameter count under control while preserving detection accuracy. As a result, it strikes a balance between detection performance and inference efficiency. Figure 1 illustrates the overall architecture of the proposed dual-stream feature extraction network.

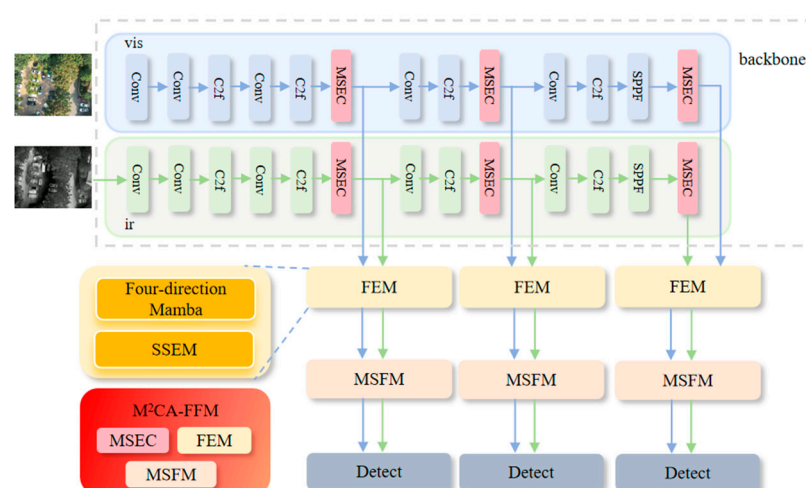


Figure 1. Overall architecture of the proposed dual-stream feature extraction network.

This two-stream network operates in three core stages: extracting and enhancing single-modal features, fusing cross-attention features via multi-directional Mamba, and finally aggregating multi-scale features at the neck network and detection head. The core component of the entire algorithm network involves feature interaction and cross-modal

fusion, which is achieved through the multi-directional Mamba cross-attention feature fusion module designed in this study. The overall process is as follows:

First, single-modal feature extraction and enhancement are performed on input I . The extracted features are

$$F^i = \varphi_{backbone}(I) \quad (1)$$

Here, $F^i \in \mathbb{R}^{W \times H \times C}$ denotes the feature map of the i -th layer ($i = 1, 2, 3$) from either the visible light or infrared branch, referred to as F_R^i and F_I^i , respectively. Symbols H , W , and C represent the height, width, and channel number of the feature map. The backbone network responsible for single-modal feature extraction is denoted as $\varphi_{backbone}$.

Then, the enhanced feature map is aggregated with visible and infrared single-modal deep features through the multi-directional Mamba cross-attention feature fusion module M²CA-FFM. The detailed computation is as follows:

$$F_{fus}^i = \lambda_{fus}(F_R^i, F_I^i) \quad (2)$$

Here, $F_{fus}^i \in \mathbb{R}^{W \times H \times C}$ denotes the fused feature of the layer, and λ_{fus} represents the proposed multi-directional Mamba cross-attention feature fusion module, called M²CA-FFM.

Finally, the resulting aggregated feature F_{fus}^i is passed to the neck network for multi-scale feature fusion, after which the downstream detection head performs dual-modal target detection. The detailed computation is as follows:

$$\{(b_k, c_k, s_k)\}_{k=1}^N = \mathcal{F}_\theta(F_{fus}^{i=1,2,3}) \quad (3)$$

Here, $\mathcal{F}_\theta$ represents the YOLOv8 detection head with parameter θ , which integrates the regression, classification, and objectness branch outputs to generate the detection result; b_k , c_k , and s_k are the coordinates, class probabilities, and detection box confidence information after decoding the predicted bounding boxes, respectively; N represents the total number of valid detected targets output by the detection head; and $\{\}_{k=1}^N$ indicates traversing from $k = 1$ to N .

3.2. Multi-Scale Enhanced Convolutional Module Design

Before feature fusion, we inserted a Multi-Scale Enhanced Convolutional Module (MSEC) into the feature extraction network to separately enhance the visible light and infrared features. As illustrated in Figure 2, MSEC comprises three main components: a depth-wise convolution that aggregates local information, a multi-branch depth-wise strip convolution that captures the multi-scale context, and a 1×1 convolution that models cross-channel relationships.

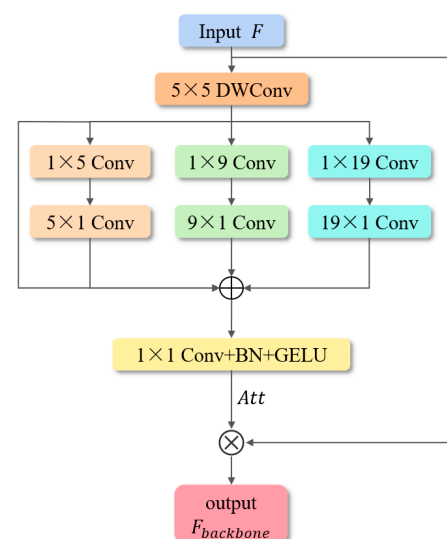


Figure 2. The architecture of the Multi-Scale Enhanced Convolutional Module (MSEC).

MSEC first extracts basic features from the input feature map using a 5×5 depth-wise separable convolution, thus obtaining intermediate features. These intermediate features are then input into three convolutional branches of different scales to capture local features at multiple spatial scales. Next, the resulting multi-scale branch features are residually connected to the basic features. The connected features are then integrated into channel features through a 1×1 convolution, while simultaneously compressing the channel dimensions to obtain the final attention weights. Finally, the attention weights interact with the original input features through residual branches, multiplying them channel-by-channel and pixel-by-pixel to enhance the expression of key semantic features while mitigating interference from invalid background features. The specific calculation is shown below:

$$Att = Conv_{1 \times 1}(\sum_{i=0}^3 Scale_i(DWConv(F))) \quad (4)$$

$$F_{backbone} = Att \odot F \quad (5)$$

where F represents the visible light or infrared single-modal input feature; $Scale_i (i = 0, 1, 2, 3)$ represents branches with different convolution scales; and $DWConv$ represents depth-wise separable convolution.

3.3. Multi-Directional Mamba Cross-Attention Feature Fusion Module

Compared with previous work, M²CA-FFM not only enables long-range cross-modal modeling and captures complementary features between visible and infrared modalities but also reduces computational complexity—striking a balance between computational efficiency and cost. The operation of M²CA-FFM consists of three steps: feature preprocessing, single-modal feature enhancement, and fused feature output. The detailed workflow is as follows.

(1) Feature Preprocessing

Although Mamba's cross-attention mechanism already has the advantage of linear computation compared to a Transformer, the spatial dimension compression is performed on the feature maps to be fused to further reduce computational cost. As illustrated in Figure 3, hybrid pooling is employed to compress the feature maps. This approach preserves strong local responses and global context information to the greatest extent without introducing significant additional parameters, thereby avoiding local feature loss. Hybrid pooling computes a linear weighted combination of max and average pooling outputs. The detailed computation is as follows:

$$F_{Mix} = \alpha \times Avg(F_{backbone}, S) + (1 - \alpha) \times Max(F_{backbone}, S) \quad (6)$$

where Avg represents the average pooling layer operation; Max represents the max pooling layer operation; α represents the learnable weight parameter between 0 and 1; and $F_{backbone}$ and S represent the input feature map and scaling factor, respectively. The original input feature dimensions are $H \times W \times C$, and the compressed feature map dimensions are reduced to $(H \times W)/S \times C$.

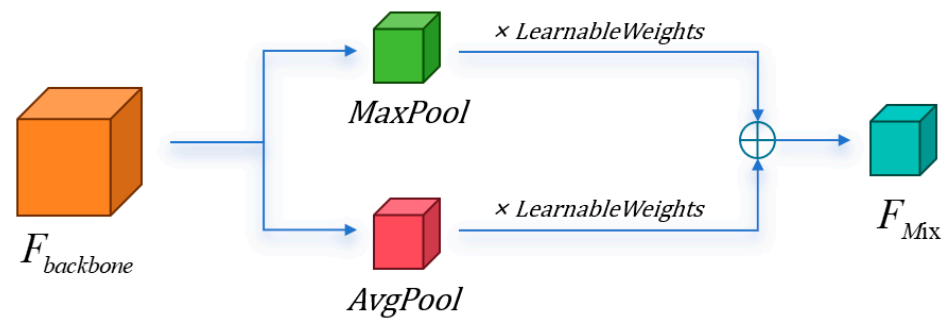


Figure 3. Feature preprocessing module.

(2) Single-modal feature enhancement

The Feature Enhancement Module (FEM) consists of a four-way Mamba module and a Separable Spectral Enhancement Module (SSEM), which is responsible for realizing the feature enhancement of single-modal features in the spatial domain and frequency domain. First, the features of the single mode are enhanced through the four-way Mamba module to provide a better feature benchmark for subsequent operations. The specific process is shown in Figure 4.

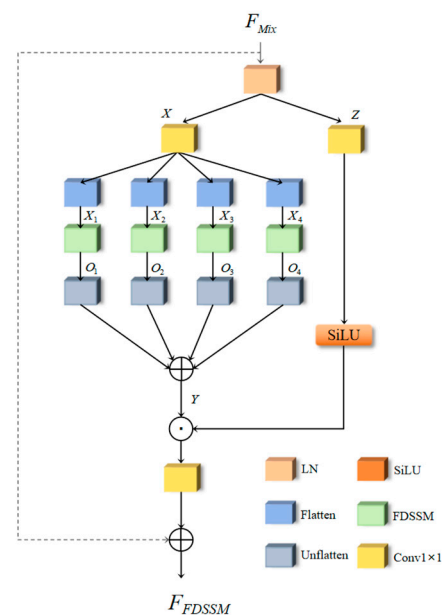


Figure 4. Four-way Mamba module.

For the input feature map $F_{Mix} \in \mathbb{R}^{H \times W \times C}$, the original input features are first mapped to the internal high-dimensional space through the linear embedding layer, allowing the model to process higher-dimensional feature spaces to capture more detailed and complex feature information, and the number of channels is expanded to four times the original. Then, F_{Mix} is segmented along the channel level to generate two different feature maps: $X \in \mathbb{R}^{H \times W \times 2C}$ and $Z \in \mathbb{R}^{H \times W \times 2C}$. The process is expressed as follows:

$$X, \quad Z = \text{Split}(\text{Norm}(F_{Mix})) \quad (7)$$

After the linear embedding layer completes the feature space mapping and channel expansion, the feature maps X and Z are dimensionally replaced, and the original dimension order (B, H, W, 2C) is converted into (B, 2C, H, W). Then, the feature map X is

flattened into 1D sequences $X_a \in \mathbb{R}^{B \times 2C \times L}$, $X_b \in \mathbb{R}^{B \times 2C \times L}$, and $L = H \times W$. The specific calculation formula is as follows:

$$X_a = \text{flatten}(X, \text{dim} = -2) \quad (8)$$

$$X_b = \text{flatten}(\text{transpose}(X; -2, -1), \text{dim} = -2) \quad (9)$$

As shown in Figure 5, the expanded 1D sequence is along four directions: upper left to lower right, lower right to upper left, upper right to lower left, and lower left to upper right. Each sequence is processed through the SSM in Mamba to capture the long-distance dependency information of the direction-containing sequence so that each pixel in the image can integrate information from other surrounding elements in different directions to model the global receptive field.

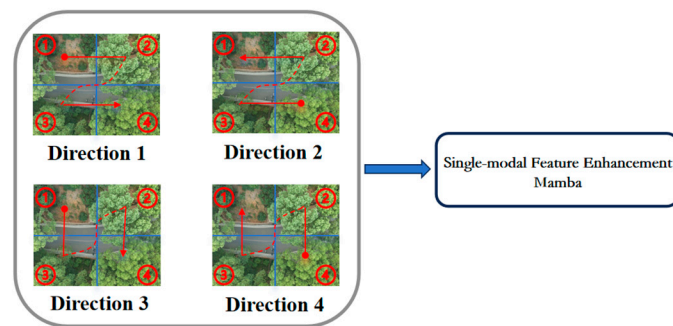


Figure 5. Mamba scan direction.

Row-forward: The flattened 1D sequence $X_a \in \mathbb{R}^{B \times 2C \times L}$ is directly fed into the row-forward FDSSM, which models long-range dependencies along the forward row-wise direction. The computation is formulated as

$$O_1 = \text{FDSSM}(X_a; \theta_1) \quad (10)$$

Row-backward: To capture reverse dependencies along the row direction, $X_a \in \mathbb{R}^{B \times 2C \times L}$ is first reversed along the sequence dimension before being fed into the row-backward FDSSM. Since the reversed output O_2' suffers from temporal misalignment, it is flipped back to preserve spatial semantic consistency:

$$O_2' = \text{FDSSM}(\text{flip}(X_a; \text{dim} = -1); \theta_2) \quad (11)$$

$$O_2 = \text{flip}(O_2'; \text{dim} = -1) \quad (12)$$

Column-forward: The pre-constructed column-wise 1D sequence $X_b \in \mathbb{R}^{B \times 2C \times L}$ is directly processed by the column-forward FDSSM to model forward column-wise dependencies. Meanwhile, the output sequence O_3' needs to be restored to the 2D feature map before it is transposed and flattened into a 1D sequence for fusion with the row sequence. The calculation formula is as follows:

$$O_3' = \text{FDSSM}(X_b; \theta_3) \quad (13)$$

$$\tilde{O}_3' = \text{unflatten}(O_3', \text{dim} = -1) \quad (14)$$

$$O_3 = \text{flatten}(\text{transpose}(\tilde{O}_3'; -2, -1), \text{dim} = -2) \quad (15)$$

Column-backward: To model backward column-wise dependencies, $X_b \in \mathbb{R}^{B \times 2C \times L}$ is first reversed along the sequence dimension before being fed into the column-backward FDSSM. The output O_4' is then aligned through reversal and transposition:

$$O_4' = \text{FDSSM}(\text{flip}(X_b; \text{dim} = -1); \theta_4) \quad (16)$$

$$O_4'' = \text{flip}(O_4'; \text{dim} = -1) \quad (17)$$

$$\tilde{O}_4'' = \text{unflatten}(O_4'', \text{dim} = -1) \quad (18)$$

$$O_4 = \text{flatten}(\text{transpose}(\tilde{O}_4''; -2, -1), \text{dim} = -2) \quad (19)$$

where θ represents parameters in, for example, the convolution, projection, and attenuation matrices in the FDSSM calculation. Subsequently, an inverse flattening operation is performed on the output of the FDSSM, and the outputs are added element-wise to obtain the four-direction Mamba enhanced feature map Y , denoted as $Y \in \mathbb{R}^{H \times W \times 2C}$.

$$Y = \sum_{i=1}^4 \text{unflatten}_i(O_i, \text{dim} = -1) \quad (20)$$

After gate processing, another feature map sequence Z is fused through 1×1 convolution to integrate the channel dimension information and then connected residually with the input feature map F_{Mix} , resulting in the output $F_{\text{FDSSM}} \in \mathbb{R}^{H \times W \times C}$:

$$F_{\text{FDSSM}} = \text{Conv}_{1 \times 1}(Y \cdot \text{SiLU}(Z)) + F_{\text{Mix}} \quad (21)$$

Building on this, given that spatial-domain information is mainly embedded in low-frequency components, and that both global (low-frequency) and local detail (high-frequency) information are valuable for representation, we introduce an FFT-based Spatial-Spectral Enhancement Module (SSEM) to enhance frequency-domain feature processing across diverse spectral bands. Figure 6 illustrates the module structure.

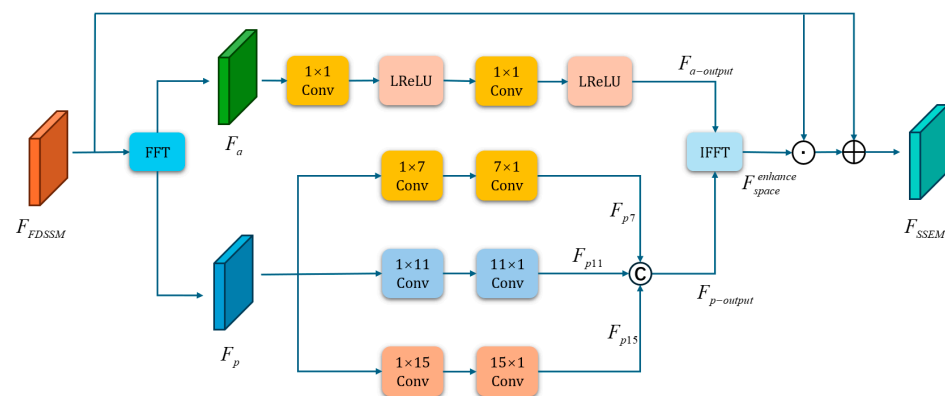


Figure 6. The architecture of the Separable Spectrum Enhancement Module (SSEM).

First, given an input feature map $F_{\text{FDSSM}} \in \mathbb{R}^{C \times H \times W}$, we first apply the Fast Fourier Transform (FFT) to obtain its spectral representation:

$$Y = \mathcal{F}(F_{\text{FDSSM}}) \quad (22)$$

where $Y \in \mathbb{C}^{C \times H \times (W/2+1)}$ is a complex-valued frequency-domain tensor. Each frequency component can be expressed in terms of its real and imaginary parts:

$$Y(u, v) = R(u, v) + jI(u, v) \quad (23)$$

where $R(u, v)$ and $I(u, v)$ denote the real and imaginary parts, respectively, and j is the imaginary unit. Subsequently, the frequency-domain features are decomposed into amplitude and phase spectra using the polar coordinate representation:

$$F_a = |Y| = \sqrt{R^2 + I^2} \quad (24)$$

$$F_p = \angle Y = \arctan\left(\frac{I}{R}\right) \quad (25)$$

Here, F_a denotes the amplitude spectrum, which reflects the energy distribution across different frequency components and encapsulates high-frequency detail information; and F_p denotes the phase spectrum, which characterizes the spatial arrangement among frequency components and carries local spatial structural information. Given the distinct information properties of the amplitude and phase spectra, the SSEM adopts a parallel architecture to model these two spectral components separately.

In the amplitude spectrum processing branch, the amplitude spectrum primarily reflects the energy distribution of image features across different frequency bands. Its low-frequency components are typically associated with global luminance, contrast, holistic contours, and background distributions, while its high-frequency components correspond to edge variations, texture intensity, and local detail responses. Considering the global distribution characteristics of the amplitude spectrum, the module adopts two cascaded 1×1 convolutions to achieve global information interaction and adaptive feature enhancement between channels without destroying the frequency domain spatial structure. Each convolution is followed by an *LReLU* activation function to introduce nonlinear mapping. The calculation is as follows:

$$F_{a-output} = LReLU\left(Conv_{1 \times 1}^2\left(LReLU\left(Conv_{1 \times 1}^1(F_a)\right)\right)\right) \quad (26)$$

In the formula, $Conv_{1 \times 1}^1$ and $Conv_{1 \times 1}^2$ are two-stage 1×1 convolutions, and $F_{a-output}$ is the enhanced amplitude spectrum feature.

The phase spectrum primarily describes the relative positional relationships among different frequency components and exerts significant influence on the spatial structure, edge localization, contour shape, and geometric relations between targets and backgrounds in an image. Unlike the amplitude spectrum, which emphasizes the energy intensity of frequency components, the phase spectrum places greater emphasis on the organizational manner and structural consistency of these components within the spatial domain. Therefore, it is insufficient to regard the phase spectrum as merely carrying local spatial information; rather, it not only contains the positional information of local edges and details but also plays a critical role in preserving global structural layout and semantic contours. The phase spectrum encapsulates core information including spatial position, edges, texture, and fine structural details, which are essential determinants of image content integrity. To completely preserve and enhance the structural information of an image, it is necessary to capture multi-scale spatial contextual associations. To address the multi-scale spatial information embedded in the phase spectrum, the module employs a multi-scale depth-wise separable convolutional architecture, utilizing three decomposed convolutional kernels with distinct receptive fields to extract local features at different scales. This design substantially reduces the computational cost while effectively capturing long-range spatial dependencies. The computational procedure of this branch is presented as follows:

$$\begin{cases} F_{p7} = Conv_{1 \times 7}\left(Conv_{7 \times 1}(F_p)\right) \\ F_{p11} = Conv_{1 \times 11}\left(Conv_{11 \times 1}(F_p)\right) \\ F_{p15} = Conv_{1 \times 15}\left(Conv_{15 \times 1}(F_p)\right) \end{cases} \quad (27)$$

$$F_{p-output} = Concat(F_{p7}, F_{p11}, F_{p15}) \quad (28)$$

In the equation, *Concat* denotes channel-wise feature concatenation, while $F_{p-output}$ represents the enhanced phase-spectrum feature. Through a cross-fusion mechanism, this structure thoroughly extracts multi-scale feature information from the phase spectrum.

After enhancing the spectral components, to recover the complete frequency-domain representation, we first reconstruct the real and imaginary parts of the enhanced frequency-domain tensor based on Euler's formula, thereby obtaining the reconstructed complex-valued frequency-domain tensor:

$$R' = F_{a-output} \cos(F_{p-output}) \quad (29)$$

$$I' = F_{a-output} \sin(F_{p-output}) \quad (30)$$

$$Y' = R' + jI' \quad (31)$$

Finally, the reconstructed frequency-domain features are transformed back to the spatial domain using the two-dimensional Inverse Fast Fourier Transform (IFFT):

$$F_{ifft} = \mathcal{F}^{-1}(Y') \quad (32)$$

Considering that negligible imaginary components may arise from numerical computation, we retain only the real-valued part after IFFT as the recovered result. To further enhance numerical stability, we take the absolute value of the recovered features as follows, where $F_{space}^{enhance}$ denotes the spatial-domain features recovered after frequency-domain enhancement:

$$F_{space}^{enhance} = |F_{ifft}| \quad (33)$$

On this basis, the module incorporates an adaptive gating mechanism and a residual connection structure. Specifically, the adaptive weighted selection of the original features is achieved through element-wise multiplication, followed by residual fusion via element-wise addition, yielding the final enhanced output F_{SSEM} . The introduction of the residual connection not only reinforces the feature fusion efficacy but also effectively alleviates the vanishing gradient problem during deep network training, thereby ensuring training stability. The computation process is formulated as follows:

$$F_{SSEM} = (F_{space}^{enhance} \odot F_{FDSSM}) \oplus F_{FDSSM} \quad (34)$$

In this equation, $\odot$ denotes element-wise multiplication, and $\oplus$ represents element-wise summation.

In summary, the SSEM achieves deep exploration of global and local feature information through frequency-domain differential modeling, effectively enhancing the model's expressive power and providing a supplementary frequency-domain dimension modeling capability to the Mamba fusion architecture.

(3) Fusion feature output

In this study, the architecture of the original SSM is extended by designing a symmetrical dual-input structure to ensure the equality and interactivity of the two types of modal information during the modeling process. Simultaneously, to overcome the limitations of fixed-direction scanning in modeling irregular targets and complex scene features, a combination of fixed four-directional sequence scanning and dynamic adaptive scanning mechanisms is integrated into the module, ultimately constructing the MSFM. This module's core task is to mine information from visible light and infrared modal features separately through a symmetrical parallel dual-branch architecture, combined with multi-branch sequence scanning modeling, dynamic channel adaptive filtering, and multi-level residual connection mechanisms, ultimately achieving dual-modal feature fusion. This

enhances both the feature representation capability and robustness of the target detection network under complex conditions. Figure 7 illustrates the overall module structure.

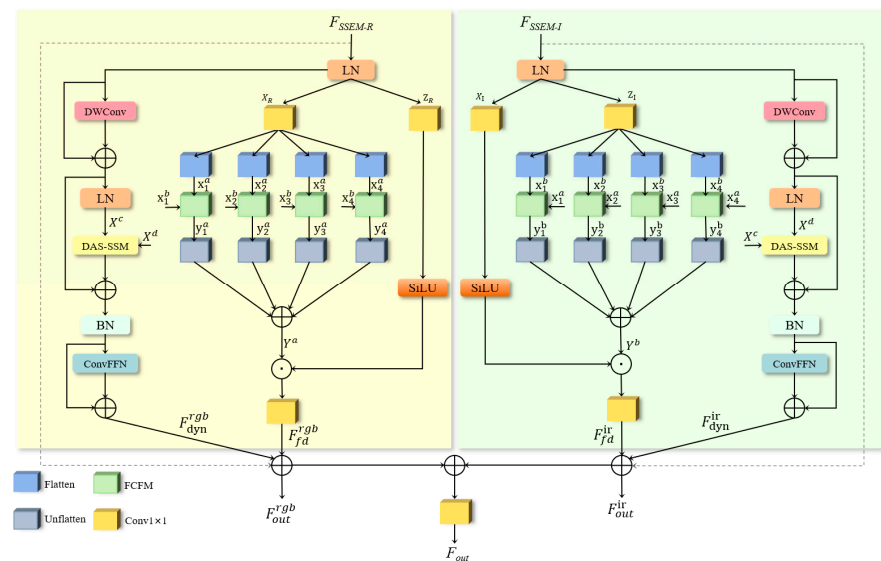


Figure 7. The architecture of the Mixed Scanning Fusion Module (MSFM).

This module employs a fully symmetrical dual-branch parallel design, with both branches using consistent computational logic and module composition to achieve independent enhancement and parallel modeling of dual-modal features. The visible and infrared branches adopt entirely independent and non-shared network parameters. Such a parameter-decoupled design facilitates the learning of modality-specific deep representations while preventing cross-modal parameter interference. Moreover, the symmetric computational paradigm guarantees equitable bi-modal information exchange and consistent computational overhead. The visible light and infrared inputs originate from the F_{SSEM} outputs of the SSEM, denoted as F_{SSEM-R} and F_{SSEM-I} , respectively. The computation process is detailed below.

For the fixed four-directional sequence-scanning portion, a method similar to the four-directional Mamba block in single-modal feature enhancement is used to generate two sets of feature maps: X_R and Z_R derived from the visible feature map and X_I and Z_I derived from the infrared feature map.

$$X_R, Z_R = \text{Split}(\text{Norm}(F_{SSEM-R})) \quad (35)$$

$$X_I, Z_I = \text{Split}(\text{Norm}(F_{SSEM-I})) \quad (36)$$

Subsequently, X_R and X_I are flattened along the four spatial directions, and the resulting one-dimensional sequence is input into the Four-direction Cross-modal Fusion Module (FCFM) for cross-modal fusion:

$$x_i^a, x_i^b = \text{Flatten}_i(X_R), \quad \text{Flatten}_i(X_I) \dots \dots i = 1, 2, 3, 4 \quad (37)$$

$$y_i^a, y_i^b = \text{FCFM}_i^a(x_i^a, x_i^b), \quad \text{FCFM}_i^b(x_i^b, x_i^a) \dots \dots i = 1, 2, 3, 4 \quad (38)$$

In the FCFM, the preceding input serves as the core subject to be processed, which will be used to transfer core modal features; the following input is specifically used to dynamically generate the projection and time scale parameters required for the SSM, providing a control basis for modeling the subject sequence. The visible light and infrared modes serve as each other's preceding and following inputs, resulting in a symmetrical

fusion structure that ensures the equal interactive status of the two modes while fully exploring bi-directional cross-modal correlation information.

The outputs in the four spatial directions are expanded to restore them as 2D spatial feature maps, and $Y^a \in \mathbb{R}^{H \times W \times C}$ and $Y^b \in \mathbb{R}^{H \times W \times C}$ are obtained via element-wise weighting:

$$Y^a, Y^b = \sum_{i=1}^4 \text{unflatten}_i(y_i^a), \sum_{i=1}^4 \text{unflatten}_i(y_i^b) \quad (39)$$

To preserve key information from the original input and enhanced feature correlation, the two calibration feature sets, Z_R and Z_I , are then element-wise multiplied with their respective preceding input feature maps to produce F_{fd}^{rgb} and F_{fd}^{ir} .

$$F_{fd}^{rgb} = \text{Conv}_{1 \times 1}^a(Y^a \cdot \text{SiLU}(Z^a)) + F_{SSEM-R} \quad (40)$$

$$F_{fd}^{ir} = \text{Conv}_{1 \times 1}^b(Y^b \cdot \text{SiLU}(Z^b)) + F_{SSEM-I} \quad (41)$$

where $\text{Conv}_{1 \times 1}^a$ and $\text{Conv}_{1 \times 1}^b$ represent the 1×1 convolutional layers that generate F_{out}^a and F_{out}^b .

To address the shortcomings of fixed four-way scanning paths and insufficient focus on salient local features, a new Dynamic Adaptive Scan (DAS) module is introduced, forming a hybrid mechanism with fixed scanning. The core of this mechanism is to dynamically predict the scanning path, guided by important regions of the feature map, to concentrate on modeling key target regions. Details of the Dynamic Adaptive Scan module are shown in Figure 8.

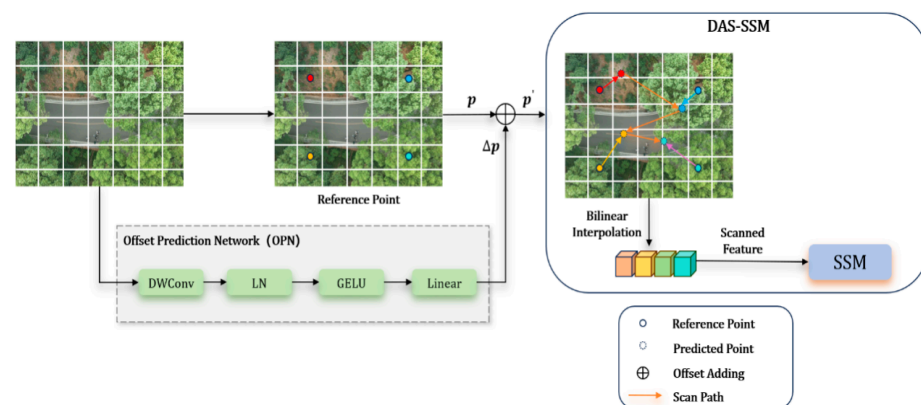


Figure 8. Schematic of the Dynamic Adaptive Scanning Architecture.

The DAS module is built around two components: an Offset Prediction Network (OPN) and a Dynamic Adaptive Scan State Space Model (DAS-SSM). The focal region is determined using learnable sampling points learned by the OPN, and the DAS position can provide a strong relative position offset, improving the SSM modeling performance.

Taking the left-side visible-light input as the primary modality, the Offset Prediction Network (OPN) first performs local feature extraction on the input X^c via depth-wise convolution, normalization, and activation functions, yielding a local feature activation map $\tilde{X}^c$. Based on this activation map, the OPN predicts the 2D coordinate offsets $\Delta p \in \mathbb{R}^{H \times W \times 2}$ of the image patches of interest relative to the original blocks:

$$\tilde{X}^c = \text{GELU}(\text{LN}(\text{DWConv}(X^c))) \quad (42)$$

$$\Delta p = \text{OPN}([x_1, x_2, \dots, x_N]) = W_{\text{offset}} \cdot \tilde{X}^c + b_{\text{offset}} \quad (43)$$

where W_{offset} and b_{offset} denote the learnable weights and bias of the linear layer within the OPN. The magnitude of the offset Δp is positively correlated with the local

feature activation strength—regions with stronger activation undergo larger coordinate shifts, steering the scanning path toward those salient areas.

Subsequently, the sampling locations are obtained by adding the offsets to the original block positions:

$$p'[h, w, :] = p[h, w, :] + \Delta p[h, w, :] \quad (44)$$

Finally, the scanning features are sampled at the dynamically adjusted coordinates p' via bilinear interpolation. The sampling process is defined as

$$X'[h, w] = \phi(p'[h, w], X^c) \quad (45)$$

where ϕ denotes the bilinear interpolation operator, which can be expanded as

$$\phi(a, b) = \sum_{(r_x, r_y)} g(a_x, r_x, a_y, r_y) \cdot b[r_y, r_x, :] \quad (46)$$

Here, $a_x = p'[h, w, 0]$ and $a_y = p'[h, w, 1]$ are the horizontal and vertical coordinates of the sampling point, respectively; (r_x, r_y) enumerate all integer grid coordinates of the feature map; and $b[r_y, r_x, :]$ denotes the original feature vector at the integer grid location.

The sampling weight function $g(\cdot)$ is activated only for the four nearest integer grids around the sampling point, and is formulated as

$$g(c, d, e, f) = \max(0, 1 - |c - d|) \times \max(0, 1 - |e - f|) \quad (47)$$

Finally, the scan features are obtained through bilinear interpolation sampling and modeled using DAS-SSM, and their dimensions are restored to complete dynamic enhancement. The module employs dual-modal, bi-directional interactive modeling, similarly to the processing logic of the four-way fixed scan method. The preceding input serves as the core subject to be processed, which will be used to transfer core modal features; the following input is specifically used to dynamically generate the projection and time-scale parameters required by the SSM, providing adaptive control for the modeling of the main sequence. The visible light and infrared modes are the preceding and following inputs, respectively, resulting in a symmetrical fusion structure, ultimately outputting F_{dyn}^{rgb} and F_{dyn}^{ir} .

After obtaining the fused features output from the fixed four-way and dynamic adaptive scans, the global enhancement features output from the fixed four-way scan are fused with the local focused enhancement features output from the dynamic adaptive scan through element-wise addition, yielding the final sequence enhancement features after hybrid scan modeling:

$$F_{out}^{rgb} = F_{fd}^{rgb} + F_{dyn}^{rgb} \quad (48)$$

$$F_{out}^{ir} = F_{fd}^{ir} + F_{dyn}^{ir} \quad (49)$$

The obtained visible-light fusion output F_{out}^{rgb} and infrared fusion output F_{out}^{ir} are then subjected to a 1×1 convolution operation to obtain the final fusion output F_{out} :

$$F_{out} = \text{Conv}_{1 \times 1}(F_{out}^{rgb} + F_{out}^{ir}) \quad (50)$$

4. Experiment

To evaluate the proposed method, we trained and tested our model on three public benchmarks—FLIR, M3FD, and VEDAI—along with our self-built dataset OPVM-VIRD, and compared its detection performance against other algorithms. Furthermore, to investigate the actual roles and contributions of each component module in the M²CA-FFM algorithm, ablation experiments were designed and implemented to verify the independent role and synergistic effect of each module. Finally, targeted experimental tests were conducted in real-world application scenarios to comprehensively evaluate the practical

application value and performance of the proposed algorithm. The publicly available datasets FLIR, M3FD, and VEDAI employed in this study were used in accordance with their original licenses. As for the self-constructed OPVM-VIRD dataset, its collection area was situated within the Yuelu Mountain National Scenic Area, which encompasses geographically sensitive regions. Due to applicable security regulations, the original imagery cannot be publicly released. A desensitized test subset can be made available to qualified researchers upon request, subject to a data usage agreement.

4.1. Datasets

4.1.1. Public Datasets

FLIR Dataset: This dataset was open-sourced by Teledyne FLIR, and subsequently, image alignment and optimization were completed by Zhang et al. It is a dataset specifically designed for dual-modal target detection tasks [37]. The dataset includes 4129 training image pairs and 1013 test image pairs. Each pair contains one infrared and one visible light image, strictly aligned in both spatial and temporal dimensions at a uniform resolution of 640×512 pixels. These images were captured with vehicle-mounted cameras, covering diverse scenes such as road traffic, vehicle flow, and pedestrian movement, exhibiting significant scene diversity.

M3FD Dataset: This dataset was independently built by the School of Software at Dalian University of Technology [38]. It is specifically designed for visible–infrared dual-modal target detection, aiming to enhance detection accuracy through the fusion of infrared and visible images. The dataset comprises 4200 calibrated infrared–visible image pairs, covering a wide range of real-world scenes and varying pixel scales. In total, it includes 8400 images, supporting tasks such as image fusion, object detection, and detection based on fusion techniques.

VEDAI: This dataset was developed by the GREYC laboratory in Grenoble, France, and released in 2015. It aims to provide benchmark data for vehicle detection tasks in aerial imagery [39]. The dataset contains 1210 aerial images covering nine vehicle categories (such as cars, trucks, boats, campervans, tractors, etc.) and provides image versions in two resolutions: 512×512 and 1024×1024 . The dataset covers a variety of scenes, including urban, rural, and industrial areas, and the images were collected from different heights and perspectives, exhibiting rich diversity and challenges.

4.1.2. Self-Built Dataset

The OPVM-VIRD dataset [40] constructed in this work aims to address the limitations of existing mainstream datasets (e.g., M3FD and FLIR), which suffer from insufficient scene coverage under unmanned aerial vehicle (UAV) perspectives and limited target scale variation due to fixed flight altitudes. The data acquisition system is mounted on a Z-3N UAV (compatible with industrial-grade flight platforms such as the DJI M350), with the core sensing unit employing a dual-modal fusion sensor pod developed by our project team, centered around the GDU PDL-1K dual-optical gimbal camera. This camera integrates a high-sensitivity visible-light sensor with a resolution of 1920×1080 and an infrared sensor with a resolution of 640×512 , both operating at 30 fps. The infrared thermal imaging covers the 8–14 μm spectral band, enabling the synchronous acquisition of spatially aligned dual-spectral data.

To simulate complex low-altitude inspection scenarios, data were collected under all-day conditions (daytime and nighttime) and all-weather conditions (sunny and rainy). In accordance with the operational characteristics of high-altitude UAVs, a hierarchical variable-altitude flight path was designed, with flight altitudes strictly ranging from 30 to 200 m, to capture target features at varying scales. Additionally, fixed-point hovering strategies were implemented in key areas to acquire multi-angle and multi-temporal data for

the same targets. During the data preprocessing stage, to eliminate lens distortion artifacts, the infrared images underwent checkerboard calibration and geometric correction. Image registration based on feature point matching was employed to achieve pixel-level alignment between visible-light and infrared images, with the registration accuracy controlled within two pixels, thereby ensuring spatial consistency across the dual-modal data. To address inter-frame redundancy caused by hovering acquisitions, frame deduplication and similar sample filtering strategies were adopted, with three annotators performing cross-validation in rotation to remove duplicate and low-quality samples. As shown in Figure 9, the registration results for two representative scenes are presented. The visible-light and infrared images exhibit highly consistent edge contours after overlay, with target features such as vehicles and trees clearly aligned, qualitatively validating that our data preprocessing pipeline effectively eliminates geometric distortions and ensures spatial consistency of the dual-modal data.



Figure 9. Registration examples of the OPVM-VIRD dataset.

The dataset focuses on “pedestrian” and “vehicle” as the primary object categories, representing the most typical targets in low-altitude UAV surveillance. All images were resized to 1241×993 pixels and manually annotated using Labellmg following the YOLO format, with minimum bounding rectangles tightly fitted to object contours. A rigorous multi-round visual inspection and cross-annotation review pipeline was enforced to ensure labeling quality. The dataset was partitioned into training, validation, and test sets at 7:2:1 via stratified sampling to preserve class distribution balance. After meticulous screening, the final dataset contains 2023 registered visible–infrared image pairs with 20,025 valid labels (8584 for pedestrians and 11,441 for vehicles). As shown in Figure 10, it exhibits superior scene diversity, scale variability, and environmental adaptability compared to existing public datasets, offering a solid benchmark for evaluating algorithmic robustness in challenging scenarios.

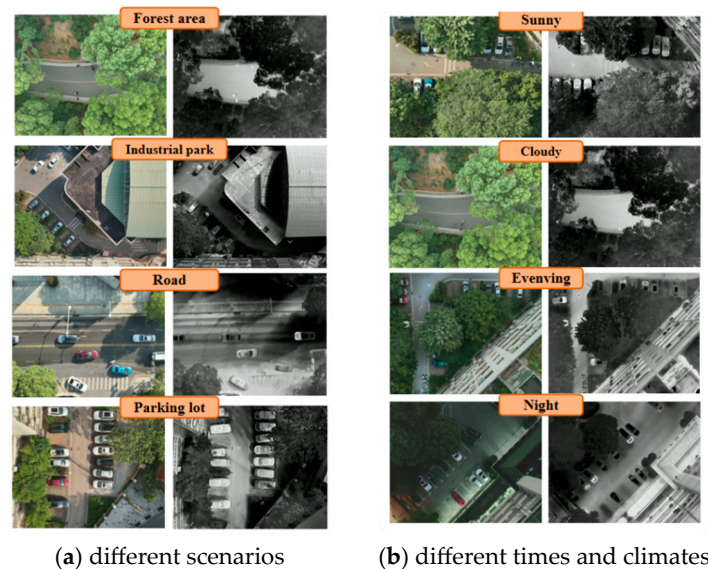


Figure 10. Visualization of the OPVM-VIRD dataset.

4.2. Experimental Setup and Configuration

To systematically and comprehensively evaluate model performance, commonly used and representative quantitative evaluation metrics within the computer vision community were used in this study, including the following:

Intersection over Union (IoU) is defined as the ratio between the overlapping area and the combined area of the predicted box B_{prea} and the ground-truth box B_{gt} . Precision (P) measures the proportion of correctly predicted positive samples among all samples the model labels as positive.

Recall and precision complement each other. A high recall value reflects a model's strong capability to capture positive samples, minimizing the risk of false negatives. Recall is defined as the fraction of actual positive samples that are correctly identified as positive by the model. The corresponding formula is

$$R = \frac{TP}{TP + FN} \quad (51)$$

In this context, True Positive (TP) denotes positive targets that are correctly detected; False Positive (FP) denotes negative samples incorrectly classified as positive; False Negative (FN) denotes positive targets that are missed; and True Negative (TN) denotes negative samples correctly identified, though TN is typically excluded from target detection evaluation metrics.

Average Precision (AP) serves as a key evaluation metric tailored for object detection tasks. Its core function is to measure the model's precision at different recall thresholds, comprehensively reflecting the model's overall detection capability for a single class of objects. The AP value corresponds to the area under the PR curve, where recall (R) is on the horizontal axis and precision (P) is on the vertical axis, and its integral expression is

$$AP = \int P(R) dR \quad (52)$$

mAP is computed by averaging the AP scores across all target classes, reflecting the model's multi-category detection performance as a whole. Its calculation formula is

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \quad (53)$$

Here, N denotes the total number of target categories to be detected. In evaluating target detection performance, two widely used metrics are mAP50 and mAP. Specifically, mAP50 refers to the average AP across all categories with a fixed IoU threshold of 0.5. Meanwhile,

mAP is calculated by averaging the AP values over all categories under multiple IoU thresholds ranging from 0.5 to 0.95 at a step size of 0.05.

The parameter count of a model serves as a key indicator of its complexity and computational overhead, directly determining the model's storage requirements and training cost. Taking a standard convolutional layer as an example, the equations for computing its parameter count and computational cost ($FLOPs$) are as follows:

$$Params = (C_{in} \times C_{out} \times K_H \times K_W) + C_{out} \quad (54)$$

$$FLOPs = Params \times W_{out} \times H_{out} \quad (55)$$

Here, K_H and K_W denote the kernel dimensions (height and width), while C_{in} and C_{out} represent the input and output channel counts, respectively. The spatial size of the output feature map is given by W_{out} (width) and H_{out} (height).

Frame rate (FPS) serves as a primary metric for assessing processing speed in image and video inference tasks. It quantifies how many frames a model can handle per second, directly reflecting its real-time inference performance. Its calculation formula is

$$FPS = \frac{N}{T} \quad (56)$$

In the formula, N denotes the total number of frames processed, and T stands for the total time taken to process them.

All experiments were carried out on PyTorch 2.0.1 under Ubuntu 20.04 with an Intel Xeon Gold 6338 CPU, 1 TB of RAM, and four RTX 4090 GPUs (Table 1). Training was conducted for 250 epochs, with a batch size of 16, an input size of 640×640 , the AdamW optimizer ($lr = 1e-3$, weight decay = $5e-4$), and mosaic/random-flip augmentations. The loss setup followed that of YOLOv8. A fixed protocol (confidence = 0.001, NMS IoU = 0.7) was used for the evaluation to ensure fair recall estimation. To assess statistical significance, we performed five independent runs per method and applied a paired t-test (null hypothesis: no difference). The experimental results demonstrate that the p -values of the proposed method are all below 0.05, with standard deviations across all trials remaining under 0.1. This provides statistically sound evidence that the improvements achieved by the proposed method in both localization accuracy and robustness are genuine and significant.

Table 1. Experimental environment and training parameters.

Parameter Items	Configuration
Framework/System	PyTorch 2.7.0/Ubuntu 20.04
Hardware	Intel(R) Xeon(R) Gold 6338 + 4 × RTX 4090 24 G
Epoch/Batch Size	250/16
Optimizer/Learning Rate	AdamW/ 1.0×10^{-3}
Weight Decay	0.0005
Input Size (Training/Testing)	$640 \times 640/640 \times 640$

4.3. Ablation Experiments

To systematically validate the effectiveness of the proposed Mixed Scanning Fusion Module (MSFM), Separated Spectrum Enhancement Module (SSEM), Multi-Scale Enhancement Convolution (MSEC) module, hybrid pooling feature preprocessing mechanism, and frequency-domain amplitude-phase branches, ablation experiments were conducted on both the M3FD and FLIR datasets. All experimental groups strictly adhered to identical training strategies, hyperparameter settings, and dataset splits. The optimal results for each evaluation metric are highlighted in bold.

To thoroughly evaluate the contribution of each individual component, a total of twelve experimental groups were designed for the M3FD and FLIR datasets:

- (1) Experimental Groups 1 and 2: Single-modality YOLOv8 baseline models using visible and infrared inputs, respectively.
- (2) Experimental Group 3: A dual-modality fusion model incorporating only the four-direction Mamba block, denoted as YOLOv8-FD.
- (3) Experimental Group 4: A dual-modality fusion model incorporating only the Dynamic Adaptive Scan (DAS) module, denoted as YOLOv8-DAS.
- (4) Experimental Group 5: A dual-modality fusion model incorporating only the MSFM (abbreviated as C in the table).
- (5) Experimental Group 6: A dual-modality fusion model incorporating both the MSFM and MSEC (abbreviated as M in the table).
- (6) Experimental Group 7: A dual-modality fusion model incorporating both the MSFM and SSEM (abbreviated as S in the table).
- (7) Experimental Group 8: A dual-modality fusion model incorporating the MSFM with only the amplitude-spectrum enhancement branch, denoted as YOLOv8-C-Amp.
- (8) Experimental Group 9: A dual-modality fusion model incorporating the MSFM with only the phase-spectrum enhancement branch, denoted as YOLOv8-C-Phs.
- (9) Experimental Group 10: The complete M²CA-FFM model but with average pooling only, denoted as M²CA-FFM(Avg).
- (10) Experimental Group 11: The complete M²CA-FFM model but with max pooling only, denoted as M²CA-FFM(Max).
- (11) Experimental Group 12: The full M²CA-FFM model, which serves to evaluate the synergistic gains achieved by integrating all proposed modules and mechanisms.

On the M3FD dataset, which encompasses multi-category complex urban scenarios, the single-modality YOLOv8 models in Groups 1 and 2 achieve mAP50 scores of 84.0% and 80.1% for visible-light and infrared inputs, respectively, indicating that the visible modality generally outperforms its infrared counterpart in overall detection performance. Group 3 incorporates only the four-direction Mamba block into the dual-modality fusion model (YOLOv8-FD), attaining an mAP50 of 85.4%, which represents a 1.4-percentage-point improvement over the visible-light baseline, along with an mAP of 56.8%. These results verify that the long-range dependency modeling capability of the four-direction Mamba architecture can preliminarily exploit the complementarity between modalities, yielding modest enhancements even for small targets such as Lamp; however, the overall fusion gain remains limited. Group 4, which integrates only the Dynamic Adaptive Scan (DAS) module (YOLOv8-DAS), further raises the mAP50 to 86.2%, which is a 0.8-percentage-point increase over YOLOv8-FD, and improves recall (R) to 74.1%. This demonstrates that the dynamic adaptive scanning mechanism can more precisely align cross-modal feature spaces and strengthen the discriminability of fused features, with particularly pronounced improvements on targets such as Motorcycle, as shown in Table 2.

Building upon these components, Group 5 incorporates the complete MSFM (Mixed Scanning Fusion Module), which unifies the four-direction Mamba and DAS into YOLOv8-C, achieving deep cross-modal fusion. Compared with the visible-light single-modality baseline, this model yields gains of 4.0 percentage points in mAP50 (reaching 88.0%), 4.7 percentage points in mAP (reaching 60.1%), and 5.3 percentage points in recall (reaching 81.5%). Relative to the standalone four-direction Mamba and DAS configurations, mAP50 improves by 2.6 and 1.8 percentage points, respectively. These results firmly

substantiate that the MSFM, by synergistically combining long-range dependency modeling with dynamic feature scanning, attains a “1 + 1 > 2” collaborative gain, thereby validating the efficacy of its internal design.

Group 6, which augments the MSFM with the MSEC module, further elevates mAP50 to 88.6%. Group 7, incorporating the SSEM in addition to the MSFM, raises recall by 0.7 percentage points to 82.2% relative to Group 5. In contrast, Groups 8 and 9, which employ only a single frequency-domain branch—namely, the MSFM with the amplitude-spectrum enhancement branch and the MSFM with the phase-spectrum enhancement branch, respectively—exhibit performance degradation, achieving mAP50 scores of 87.7% and 87.4%. Their mAP and recall values also fall below those of the multi-module combination schemes, indicating that the standalone use of either the amplitude or phase branch cannot fully unleash the potential of frequency-domain enhancement.

To investigate the influence of pooling strategies, Group 10 (employing average pooling only) and Group 11 (employing max pooling only) were evaluated, both exhibiting conspicuous performance deficiencies, with mAP50 scores of 87.8% and 87.6%, respectively. Ultimately, the complete M²CA-FFM model incorporating the hybrid pooling feature preprocessing mechanism (Group 12) achieves the best overall performance across all metrics, attaining an mAP50 of 89.4% and a recall of 83.8%. Under nearly identical parameter budgets, this model outperforms its average-pooling-only counterpart by 1.6 percentage points in mAP50, 0.4 percentage points in mAP, and 2.5 percentage points in recall; relative to the max-pooling-only scheme, the improvements are 1.8 and 1.3 percentage points in mAP50 and mAP, respectively. In terms of per-category detection performance, compared with Groups 5, 6, and 7, the proposed model improves the AP50 for Motorcycle by 2.3–3.1 percentage points, raising it to 91.1%. More notably, relative to the two single-pooling baselines, the AP50 for extremely small targets, such as those under the Lamp category, surges from 78.2% and 77.9% to 84.9%, achieving substantial gains of 6.7–7.0 percentage points, while the AP50 for medium-scale targets, such as those under the Motorcycle category, also increases by 3.1–5.3 percentage points. These comprehensive results collectively demonstrate that hybrid pooling can adaptively integrate the advantages of both average and max pooling; when coupled with the three core modules—the MSFM, MSEC, and SSEM—the model significantly enhances small-target recognition capability and effectively mitigates missed-detection issues.

Table 2. Performance comparison on the M3FD dataset in terms of mAP50 and mAP.

Group	Modality	Model	Person	Car	Bus	Motorcycle	Lamp	Truck	mAP50	mAP	R	Param.(M)
1	RGB	YOLOv8	74.2	93.5	91.8	86.0	73.1	85.1	84.0	55.4	76.2	43.7
2	IR	YOLOv8	70.4	87.6	85.8	80.0	70.1	79.1	80.1	49.5	70.5	43.7
3	RGB+IR	YOLOv8-FD	79.3	91.0	86.4	88.9	82.2	84.5	85.4	56.8	72.6	63.23
4	RGB+IR	YOLOv8-DAS	88.8	94.0	79.9	89.9	80.7	84.0	86.2	57.1	74.1	62.38
5	RGB+IR	YOLOv8-C	88.9	94.4	92.4	88.0	79.3	84.6	88.0	60.1	81.5	67.07
6	RGB+IR	YOLOv8-C-M	88.7	94.7	93.9	88.4	80.3	85.5	88.6	60.1	81.6	68.45
7	RGB+IR	YOLOv8-C-S	88.1	94.5	93.2	88.8	79.8	86.6	88.5	60.2	82.2	70.71
8	RGB+IR	YOLOv8-C-Amp	86.6	94.3	93.0	85.1	81.1	86.2	87.7	58.4	81.6	69.38
9	RGB+IR	YOLOv8-C-Phs	84.7	94.0	92.3	83.9	80.8	88.9	87.4	59.1	81.2	68.35
10	RGB+IR	M ² CA-FFM(Avg)	87.1	94.1	93.7	85.8	78.2	87.6	87.8	59.7	81.3	72.07
11	RGB+IR	M ² CA-FFM(Max)	88.4	94.6	92.5	88.0	77.9	84.4	87.6	58.8	83.6	72.07
12	RGB+IR	M ² CA-FFM(our)	88.9	94.7	92.4	91.1	84.9	84.4	89.4	60.1	83.8	72.07

On the FLIR dataset, the visible-light single-modality YOLOv8 baseline (Group 1) achieves an mAP50 of 75.7% and an mAP of 38.6%, while the infrared single-modality counterpart (Group 2) yields 70.5% and 35.0%, respectively. Group 3, which incorporates

only the four-direction Mamba block into the dual-modality fusion model (YOLOv8-FD), attains an mAP50 of 83.9%, corresponding to an 8.2-percentage-point improvement over the visible-light baseline. This result preliminarily demonstrates the benefits of cross-modal fusion combined with long-range dependency modeling. Group 4, integrating only the Dynamic Adaptive Scan (DAS) module (YOLOv8-DAS), further increases the mAP50 to 86.5%, which is a 2.6-percentage-point gain over YOLOv8-FD, thereby validating the reinforcing effect of the dynamic scanning mechanism on infrared–visible feature alignment, as shown in Table 3.

The introduction of the complete MSFM further amplifies the gains from dual-modality fusion. Group 5, equipped with the full MSFM, achieves a substantial increase of 12.1 percentage points in mAP50 (reaching 87.8%) and 11.9 percentage points in mAP (reaching 51.5%). Relative to the standalone four-direction Mamba and DAS configurations, mAP50 improves by 3.9 and 1.3 percentage points, respectively, once again corroborating the synergistic optimization effect achieved by integrating both components within MSFM. Group 6, which augments the MSFM with the MSEC module, raises mAP50 to 88.0% while maintaining mAP at 51.5%. Group 7, incorporating the SSEM in addition to the MSFM, increases mAP to 51.6% with mAP50 remaining at 87.8%. In contrast, Groups 8 and 9, which introduce only the amplitude- and phase-spectrum enhancement branches, respectively, exhibit performance degradation, achieving mAP50 scores of 87.8% and 87.2%, indicating that a single frequency-domain branch alone is insufficient to yield additional performance gains.

With respect to the pooling strategy ablations, Group 10 (average pooling only) attains an mAP50 of 87.3%, while Group 11 (max pooling only) achieves only 86.6%, both underperforming relative to the hybrid pooling scheme. The complete M²CA-FFM model equipped with the hybrid pooling mechanism emerges as the best-performing model on this dataset, achieving an mAP50 of 88.4% and a recall of 81.3%, with the AP50 for the Bicycle category reaching 83.2%, the highest among all experimental groups. This model consistently outperforms all single-component, single-module, dual-module, and single-frequency-domain-branch configurations, intuitively demonstrating the significant synergistic optimization effects of the three core modules—the MSFM, MSEC, and SSEM. Meanwhile, the hybrid pooling mechanism also exhibits favorable effectiveness on this dataset: compared with the average-pooling-only scheme, it yields improvements of 1.1 percentage points in mAP50 and 0.4 percentage points in mAP; relative to the max-pooling-only scheme, the gains are 1.8 and 1.5 percentage points, respectively. For the Bicycle category, which is susceptible to noise interference in infrared imaging, the AP50 increases from 81.7% and 80.5% to 83.2%, representing improvements of 1.5–2.7 percentage points, thereby reflecting the enhanced noise robustness conferred by hybrid pooling. Moreover, the AP50 for the Person category rises from 88.0% and 87.4% to 89.5%, with gains of 1.5–2.1 percentage points, further verifying that hybrid pooling can effectively reinforce target contour features and demonstrating the general applicability of this mechanism.

Meanwhile, despite achieving optimal overall performance across both datasets, M²CA-FFM does not attain the highest per-category AP50 for certain classes, specifically Bus and Truck on the M3FD dataset and Car on the FLIR dataset. Individual modules exhibit specialized enhancement advantages for large-scale rigid objects and contour-dominated categories. However, since the complete model prioritizes global equilibrium and improvements on small and challenging targets as its core optimization direction, the focal concentration on local features becomes somewhat distributed under multi-module collaboration.

Table 3. Performance comparison on the FLIR dataset in terms of mAP50 and mAP.

Group	Modality	Model	Person	Car	Bicycle	mAP50	mAP	R
-------	----------	-------	--------	-----	---------	-------	-----	---

1	RGB	YOLOv8	71.8	86.1	69.2	75.7	38.6	68.8
2	IR	YOLOv8	68.8	79.1	68.0	70.5	35.0	60.7
3	RGB+IR	YOLOv8-FD	85.5	89.8	76.4	83.9	46.4	75.0
4	RGB+IR	YOLOv8-DAS	88.0	91.5	79.9	86.5	48.9	77.1
5	RGB+IR	YOLOv8-C	89.0	92.6	82.0	87.8	51.5	80.6
6	RGB+IR	YOLOv8-C-M	89.1	92.5	82.6	88.0	51.5	80.8
7	RGB+IR	YOLOv8-C-S	89.9	92.2	81.4	87.8	51.6	80.7
8	RGB+IR	YOLOv8-C-Amp	88.9	92.2	82.3	87.8	51.1	79.8
9	RGB+IR	YOLOv8-C-Phs	87.8	91.0	82.9	87.2	50.5	80.9
10	RGB+IR	M ² CA-FFM(Avg)	88.0	92.1	81.7	87.3	51.1	81.0
11	RGB+IR	M ² CA-FFM(Max)	87.4	91.8	80.5	86.6	50.0	79.1
12	RGB+IR	M ² CA-FFM (ours)	89.5	92.5	83.2	88.4	51.5	81.3

The convergence process of the learnable weight parameters in the hybrid pooling mechanism is analyzed. The results indicate that parameter α undergoes rapid adjustment during the early stage of training and gradually converges to a stable interval in the later stage, demonstrating that the model can adaptively learn the optimal fusion ratio between the two pooling strategies throughout the training process. Based on the M3FD dataset, the convergence curves of the learnable weight parameters for the visible and infrared modalities are presented in Figure 11, where w_1 denotes α and w_2 denotes $1-\alpha$.

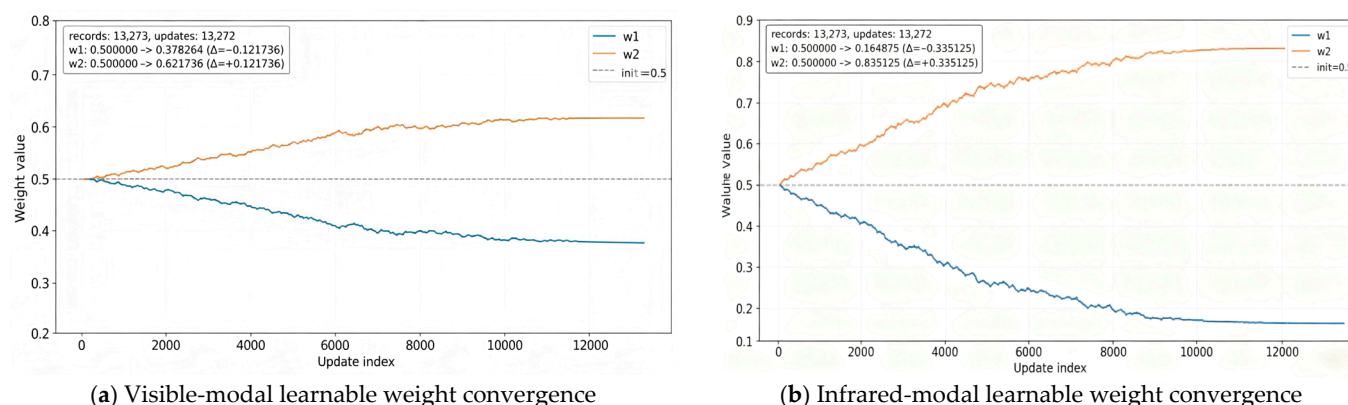


Figure 11. Convergence curves of learnable weight parameters.

4.4. Visual Interpretation

To assess the detection performance of M²CA-FFM across varying scenarios, inference tests were conducted on typical scenes extracted from the FLIR, M3FD, VEDAI, and OPVM-VIRD datasets for comparison with the ICAFusion method. The qualitative comparison results for the three datasets have the same layout; the three columns present, respectively, the ground-truth bounding boxes as the accuracy reference, the detection output from ICAFusion, and the results produced using our M²CA-FFM algorithm.

Figure 12 compares the detection performance of M²CA-FFM and ICAFusion across representative scenes from the FLIR dataset.

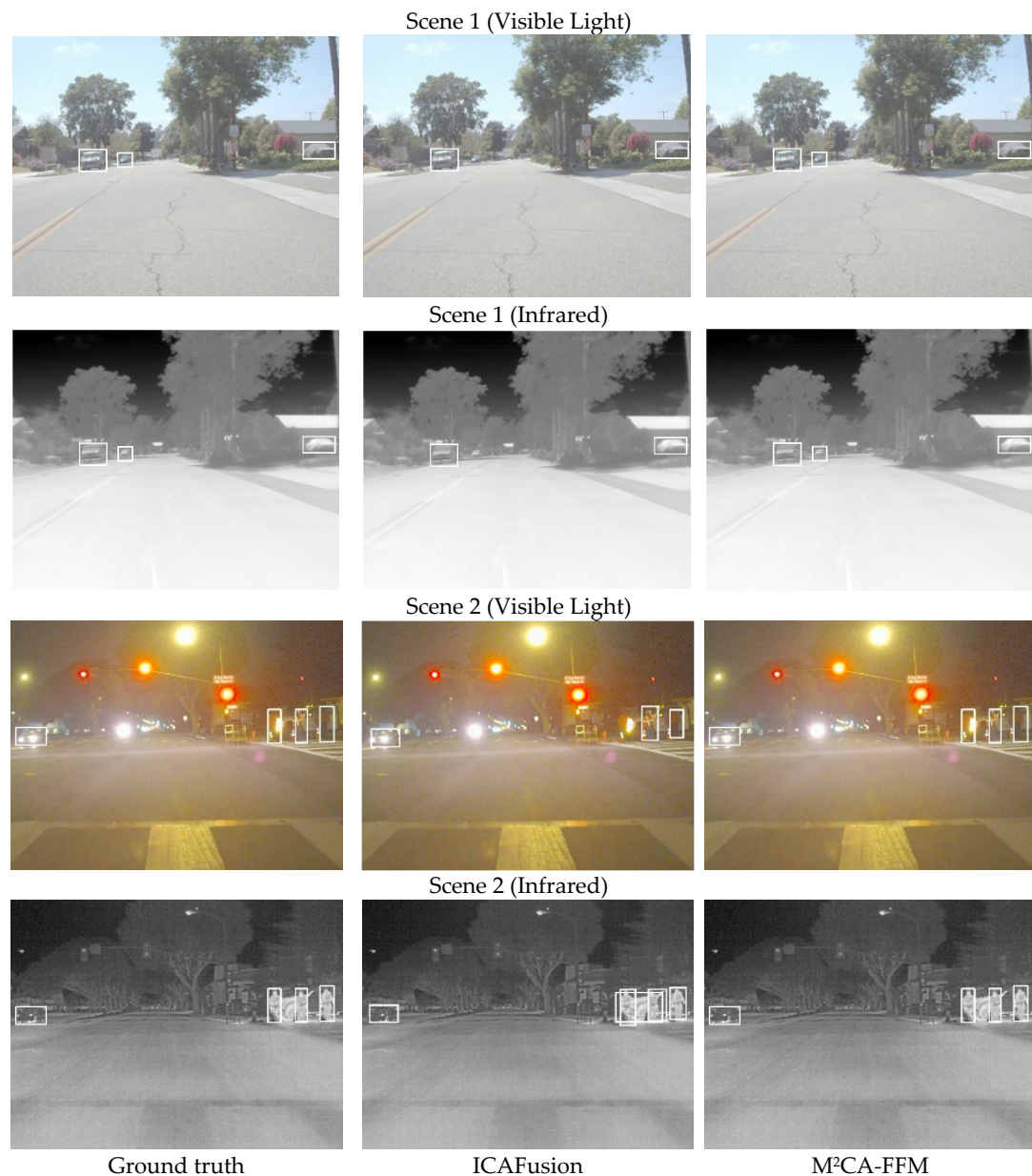


Figure 12. Visual comparison of detection results based on the FLIR dataset.

The visualization results clearly show that in Scene 1, the ICAFusion algorithm missed detecting a car target in the center of the background, while the M²CA-FFM algorithm fully utilized the thermal radiation features of the infrared mode and the edge information of the visible light mode to detect all targets in the scene and thereby demonstrate stronger target detection capabilities. In Scene 2, the ICAFusion algorithm failed to detect backlit targets in the visible light background image and mistakenly detected interfering heat sources as vehicles in the infrared background image, resulting in false detections. In contrast, the M²CA-FFM algorithm successfully detected all targets in the dense crowd without any false detections, and the detection bounding box positions were consistent with the height of the ground-truth bounding boxes.

Based on images from the M3FD dataset, Figure 13 compares the detection results of the M²CA-FFM and ICAFusion algorithms on typical scenes.

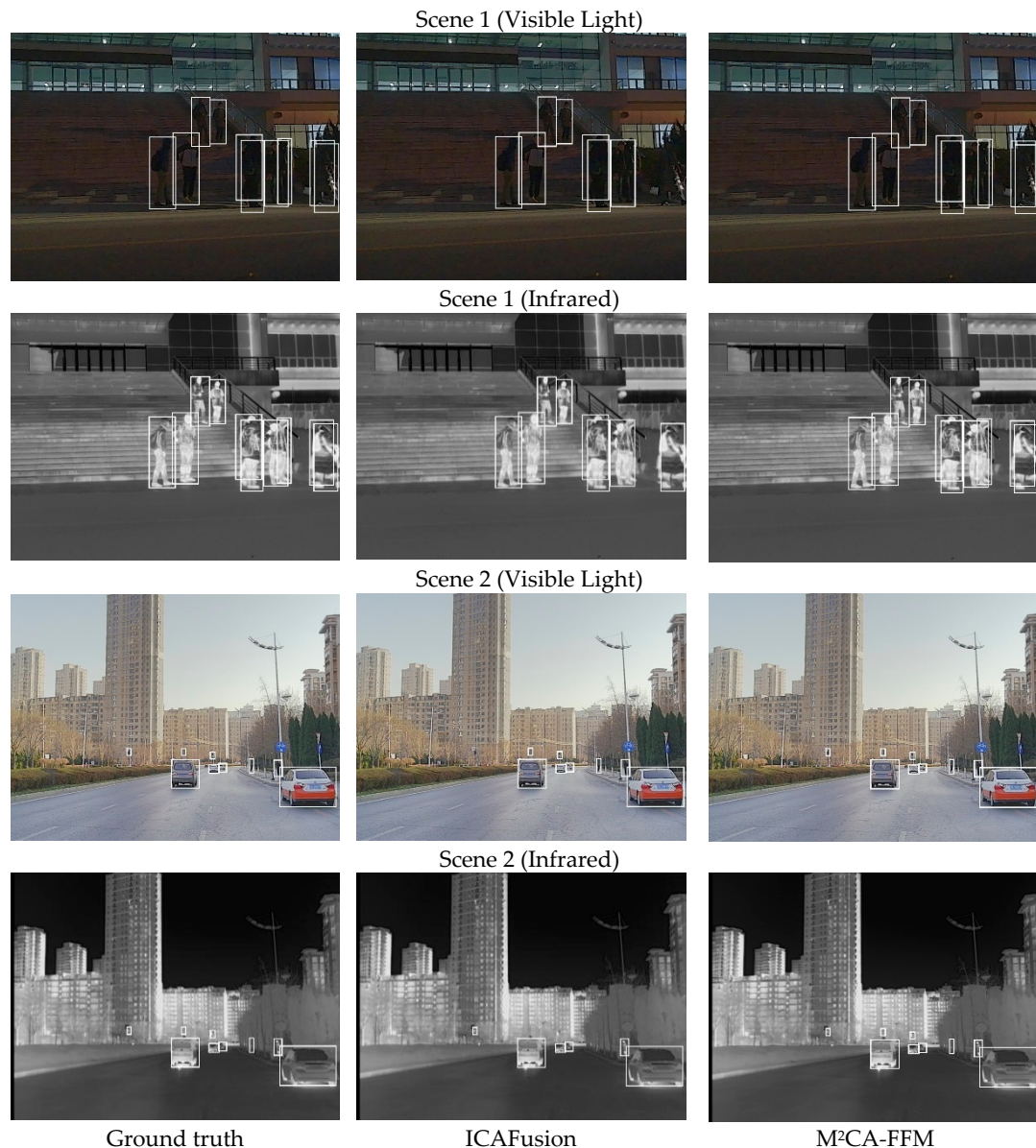


Figure 13. Visual comparison of detection results based on the M3FD dataset.

The visualization results clearly show that in Scene 1, the ICAFusion algorithm missed pedestrian targets under low-light conditions and failed to detect overlapping pedestrian targets. In contrast, the M²CA-FFM algorithm detected all targets in the scene, overcoming interference from low light by integrating infrared features and demonstrating stronger target detection capabilities. In Scene 2, the ICAFusion algorithm failed to detect smaller traffic light targets. By contrast, M²CA-FFM accurately identified small targets, with its detection box positions aligning closely with the ground-truth bounding boxes.

Figure 14 compares the detection performance of M²CA-FFM and ICAFusion on representative scenes from the VEDAI dataset.



Figure 14. Visual comparison of detection results based on the VEDAI dataset.

The visualization results clearly show that in Scene 1, the ICAFusion algorithm missed detecting adjacent vehicle targets and failed to detect the tractor target. In contrast, the M²CA-FFM algorithm detected all targets in the scene, demonstrating stronger target detection capabilities by integrating infrared features. In Scene 2, the ICAFusion algorithm failed to detect vehicle targets under low-light conditions and highly interfering vehicle targets. By contrast, M²CA-FFM accurately identified small targets, with its predicted bounding box height closely matching the ground-truth box height.

Figure 15 compares the detection performance of M²CA-FFM and ICAFusion on representative scenes from the OPVM-VIRD dataset.

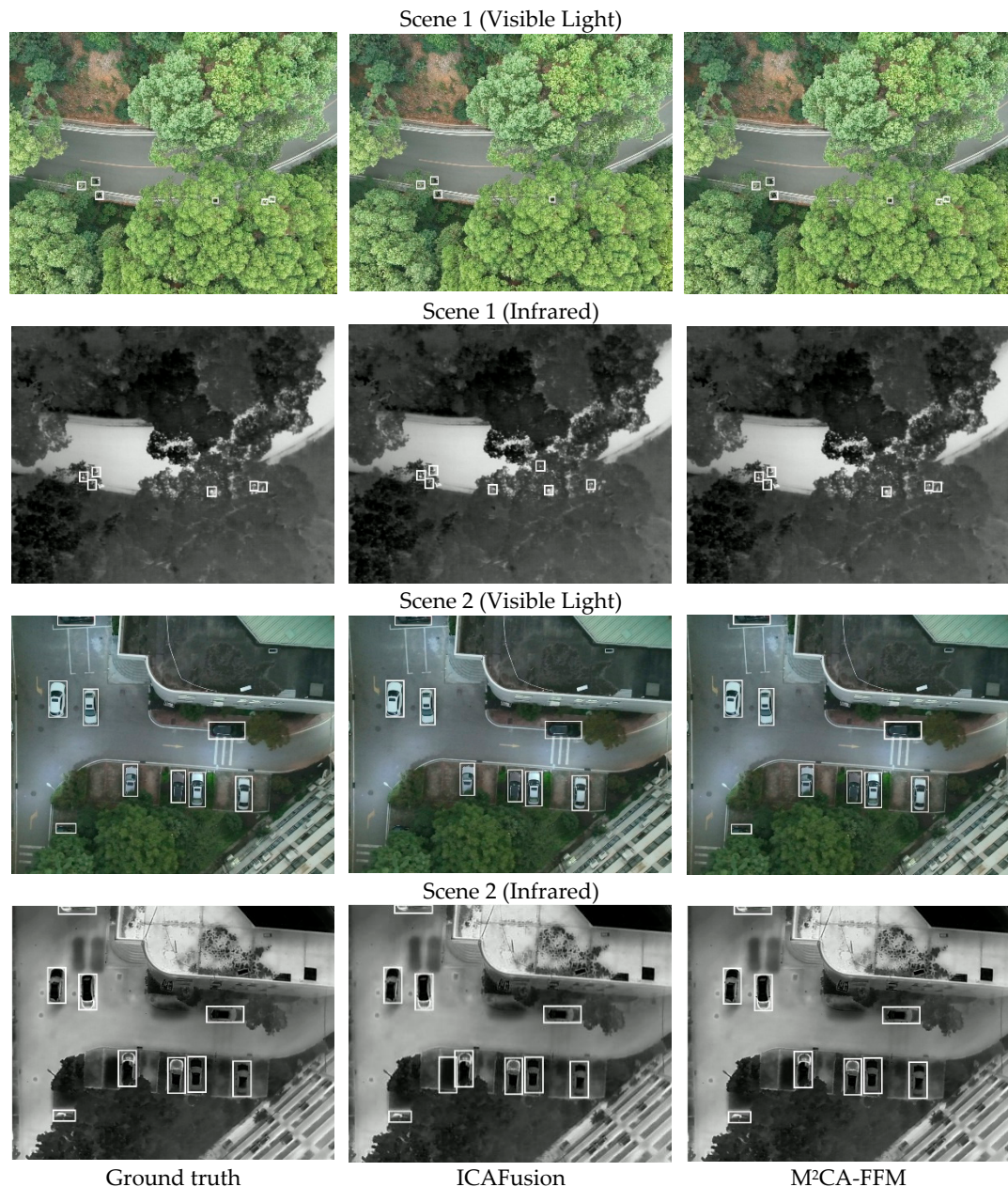


Figure 15. Visual comparison of detection results based on the OPVM-VIRD dataset.

The visualization results clearly show that in Scene 1, tree branches and leaves partially obscure the human target. Simultaneously, the thermal accumulation effect of the leaves degrades the infrared thermal features of the target. The ICAFusion algorithm exhibits significant false negatives and missed detections of the person target obscured by branches and leaves. In contrast, M²CA-FFM successfully captures all targets in the scene, effectively offsetting information loss from degraded single-modal features and exhibiting superior detection performance. In Scene 2, the ICAFusion algorithm mistakenly detects shadows as vehicle targets, with infrared features acting as interference. In comparison, the M²CA-FFM algorithm, based on texture features rich in visible light information, successfully detects all targets, and its detection bounding box positions are consistent with the height of the ground-truth annotations.

The visualization results and qualitative analysis across the three datasets corroborate the experimental findings of the comparative experiments, leading to the following main conclusions: The M²CA-FFM algorithm proposed in this paper effectively addresses the limitations of traditional dual-modal fusion algorithms, such as sensitivity to single-

modal imaging quality and susceptibility to false negatives and missed detections in complex scenes.

Compared to the mainstream ICAFusion algorithm, the M²CA-FFM algorithm demonstrates superior target detection capabilities, higher accuracy, and stronger environmental robustness in various challenging scenarios, including low-light, long-distance scenarios; dense targets; partial occlusion; and strong background interference. This method fully exploits the complementary information between visible and infrared modalities, effectively offsetting the inherent limitations of single-modal imaging. It sustains high-precision detection across diverse real-world complex environments, further confirming the effectiveness and generalization capability of the proposed approach.

4.5. Comparative Experiments

4.5.1. Performance Comparison

This section compares the performance of the M²CA-FFM model on the general datasets FLIR, M3FD, and VEDAI, as well as the self-built dataset OPVM-VIRD, verifying its generalization ability on different datasets and its strong robustness in different environments.

On the FLIR dataset, the proposed M²CA-FFM achieves the best performance across both mAP50 and mAP metrics, attaining 88.4% and 51.5%, respectively. It outperforms the state-of-the-art dual-modal algorithm DEYOLO by 1.9% and 1.1%, FusionMamba by 3.5% and 4.5%, and ICAFusion by 5.0% and 10.1%, respectively. These results indicate that our algorithm not only achieves higher detection accuracy under conventional thresholds (as evidenced by the mAP50 advantage) but also maintains stable localization precision and classification confidence under stringent IoU thresholds (as reflected by the mAP advantage). Notably, the mAP improvement of 10.1% over ICAFusion sufficiently demonstrates the effectiveness of M²CA-FFM's cross-modal attention mechanism and frequency-domain feature modulation in fine-grained detection. Furthermore, compared with the single-modality RGB YOLOv8 baseline (75.7%/38.6%), M²CA-FFM achieves substantial gains of 12.7 percentage points in mAP50 and 12.9 percentage points in mAP, corroborating the robustness advantages of dual-modality fusion under extreme illumination conditions and adverse weather scenarios, as shown in Table 4.

Table 4. Comparative experimental results on the FLIR dataset.

Dataset	Modality	Detector	mAP50	mAP
FLIR	RGB	YOLOv8	75.7	38.6
	IR	YOLOv8	70.5	35.0
	RGB+IR	FusionMamba	84.9	47.0
	RGB+IR	FD2-Net	82.9	-
	RGB+IR	MFYOLO	83.6	-
	RGB+IR	DEYOLO	86.5	50.4
	RGB+IR	ICAFusion	83.4	41.4
	RGB+IR	M ² CA-FFM(our)	88.4	51.5

On the M3FD dataset, the proposed M²CA-FFM also achieves the best detection performance, attaining an mAP50 of 89.4% and an mAP of 60.1%. Compared with the mainstream dual-modality algorithms DEYOLO and MFYOLO, it yields improvements of 2.9% and 1.5% over DEYOLO and 5.2% and 3.5% over MFYOLO, respectively. Relative to ICAFusion, M²CA-FFM achieves a substantial mAP improvement of 12.9%, alongside a 7.2% gain in mAP50. The concurrent superiority in both mAP50 and mAP indicates that M²CA-

FFM not only attains higher detection rates under conventional thresholds but also exhibits remarkable advantages in fine-grained localization and confidence-calibrated classification. Furthermore, in comparison with the single-modality RGB YOLOv8 baseline (84.0%/55.4%), the proposed method outperforms it by 5.4 percentage points in mAP50 and 4.7 percentage points in mAP. These results further corroborate that the visible and infrared modalities achieve effective complementary enhancement through cross-modal attention and frequency-domain feature modulation and that this advantage remains consistently valid under the complex multimodal scenarios of the M3FD dataset, as shown in Table 5.

Table 5. Comparative experimental results on the M3FD dataset.

Dataset	Modality	Detector	mAP50	mAP
M3FD	RGB	YOLOv8	84.0	55.4
	IR	YOLOv8	80.1	49.5
	RGB+IR	FD2-Net	83.5	-
	RGB+IR	MFYOLO	84.2	56.6
	RGB+IR	DEYOLO	86.5	58.6
	RGB+IR	ICAFusion	82.2	47.2
	RGB+IR	M ² CA-FFM(our)	89.4	60.1

On the VEDAI dataset, the proposed M²CA-FFM also achieves the best performance, attaining 73.5% in mAP50 and 46.0% in mAP. It surpasses the mainstream dual-modal algorithm DEYOLO by 0.9% and 0.7%, respectively. Although the absolute improvements appear marginal, it should be noted that in aerial scenarios such as VEDAI—characterized by dense small targets and complex backgrounds—the detection difficulty is substantially higher than that of conventional road scenes, where even minor gains signify meaningful enhancements in feature discriminability. More notably, compared with ICAFusion, M²CA-FFM achieves a 5.7% improvement in mAP50 and a substantial 9.2% improvement in mAP, demonstrating concurrent superiority on both metrics with a narrower performance gap, which indicates that the cross-modal attention and frequency-domain modulation mechanisms of M²CA-FFM effectively enhance edge details and semantic representations of small targets. Furthermore, compared with the unimodal RGB YOLOv8 baseline (68.9%/43.1%), our method yields improvements of 4.6% in mAP50 and 2.9% in mAP, validating that visible–infrared fusion confers significant gains for small-target detection in aerial perspectives and that such gains are stably propagated to high IoU thresholds, underscoring the model’s robust advantage in fine-grained localization, as shown in Table 6.

Table 6. Comparative experimental results on the VEDAI dataset.

Dataset	Modality	Detector	mAP50	mAP
VEDAI	RGB	YOLOv8	68.9	43.1
	IR	YOLOv8	65.0	42.1
	RGB+IR	DEYOLO	72.6	45.3
	RGB+IR	ICAFusion	67.8	36.8
	RGB+IR	M ² CA-FFM(our)	73.5	46.0

On the self-constructed UAV-borne OPVM-VIRD dataset, the proposed M²CA-FFM achieves outstanding performance, attaining 95.7% in mAP50 and 70.3% in mAP. Compared with DEYOLO—which yields the same mAP of 70.3%—M²CA-FFM outperforms it by 2.2% in mAP50, indicating more comprehensive detection under conventional thresholds. Relative to ICAFusion, M²CA-FFM achieves a 10.3% improvement in mAP50 and a

substantial 15.5% improvement in mAP. More notably, CFT attains only 82.7% in mAP50 and 47.6% in mAP, both of which are significantly surpassed by our method by over 13% and 22.7%, respectively. These results strongly validate that, in complex low-altitude UAV scenarios characterized by mixed pedestrian–vehicle traffic and partial occlusions, M²CA-FFM not only exhibits superior target recall capability but also maintains excellent localization accuracy and detection confidence under stringent IoU thresholds, thereby achieving concurrent superiority in both mAP and mAP50. Furthermore, compared with the unimodal RGB YOLOv8 baseline (93.5%/68.0%), our method yields improvements of 2.2% in mAP50 and 0.3% in mAP, further demonstrating that visible–infrared fusion consistently delivers stable gains in complex dynamic UAV-borne scenarios, as shown in Table 7.

Table 7. Comparative experimental results on the OPVM-VIRD dataset.

Dataset	Modality	Detector	mAP50	mAP
OPVM-VIRD	RGB	YOLOv8	93.5	68.0
	IR	YOLOv8	85.5	55.2
	RGB+IR	CFT	82.7	47.6
	RGB+IR	DEYOLO	93.5	70.3
	RGB+IR	ICAFusion	85.4	54.8
	RGB+IR	M ² CA-FFM(our)	95.7	70.3

4.5.2. Comparison of Computational Complexity and Speed

To evaluate the deployment potential of the proposed M²CA-FFM method in resource-constrained environments, we established an inference validation system based on an edge computing platform in a laboratory setting (as illustrated in Figure 15). The hardware core of this system employed the Huawei Atlas 200I DK A2 development kit, equipped with the Ascend 310B AI processor (providing 20 TOPS INT8/10 TFLOPS FP16 computational power) and 4 GB LPDDR4X memory, designed to simulate the computational constraints of industrial-grade UAV payloads. The perception front-end was equipped with a dual-optical camera (PDL-1K) integrating visible-light and infrared modules for the synchronous acquisition of dual-spectral data (as illustrated in Figure 16).

At the software level, we trained the model using the PyTorch 2.7.0 framework and subsequently performed model conversion and quantization through the Huawei CANN toolchain to accommodate the inference acceleration requirements of the on-device NPU. Note that this stage of experimentation primarily focused on benchmarking the inference performance of the algorithmic model itself, rather than conducting end-to-end full-link testing, encompassing data transmission and flight control. During the experiments, the input image resolution was uniformly fixed at 640 × 640.



Figure 16. M²CA-FFM algorithm, UAV-based inference, and verification system.

This experiment compared the computational complexity between the proposed M²CA-FFM and mainstream multispectral object detection methods, including CFT and

ICAFusion. As shown in Table 8, the results demonstrate that, owing to the efficient feature fusion design, our method achieves a substantial reduction in resource consumption while maintaining high detection performance: it comprises only 72.07 M parameters and incurs a computational cost of 134.41 G FLOPs. Compared with CFT and ICAFusion, the proposed method exhibits substantial advantages in both the model parameter count and computational overhead, demonstrating potential for real-time processing on embedded devices. These favorable properties lay a solid foundation for its future deployment and practical application in unmanned aerial vehicle (UAV) platforms.

Table 8. Complexity and speed comparison.

Method	Params.(M)	FLOPs (G)
CFT	206.3	224.40
ICAFusion	120.23	201.12
M ² CA-FFM	72.07	134.41

5. Conclusions

To address the industry challenges of existing visible–infrared dual-modal target detection algorithms, such as the susceptibility of spatial domain fusion to noise interference, the limitation of low-frequency redundant information on feature representation, the insufficient global modeling capabilities of CNN models, and the excessive computational complexity of Transformer architectures, we introduce a dual-modal target detection algorithm leveraging multi-direction Mamba cross-attention fusion. It integrates the global modeling capability and linear complexity of the Mamba architecture with frequency-domain feature enhancement. Our approach strengthens single-modal feature extraction and long-range dependency modeling while keeping the parameter count low, striking a balance between detection accuracy and inference efficiency. To verify the algorithm’s actual performance and generalization ability, systematic experimental tests were conducted on four datasets: FLIR, M3FD, VEDAI, and OPVM-VIRD. Our M²CA-FFM algorithm attained mAP50 scores of 88.4%, 89.4%, 73.5%, and 95.7% across the four datasets, respectively, comprehensively outperforming single-modal baseline algorithms and mainstream Transformer dual-modal fusion algorithms such as ICAFusion. Ablation experiments verified the effectiveness of the MSEC module, SSEM, and MSFM, as well as the significant synergistic effect among them. Visualization results further confirm the algorithm’s strong detection capability and robustness under challenging conditions such as low illumination, heavy occlusion, small objects, and dense scenes. Our work not only delivers a high-accuracy solution for visible–infrared dual-modal detection but also introduces a new perspective for applying the Mamba architecture to cross-modal fusion tasks.

Author Contributions: Conceptualization, methodology, investigation, Z.L.; software, validation, writing—original draft preparation, Z.Z. (Ziyi Zhou); supervision, funding acquisition, H.T.; writing—review and editing, Z.Z. (Ziyi Zhong) and K.L.; formal analysis, Y.M.; project administration, W.T.; funding acquisition, K.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Chongqing Municipal Natural Science Foundation under Grant CSTB2022NSCQ-MSX0192.

Data Availability Statement: All the data are contained in the article.

Conflicts of Interest: Author Kui Wang was employed by the company Chongqing Gangli Environmental Protection Co., Ltd. The remaining authors declare that the research was conducted in the

absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. <https://doi.org/10.48550/arXiv.1512.03385>.
2. Islam, S.; Elmekki, H.; Elsebai, A.; Bentahar, J.; Drawel, N.; Rjoub, G.; Pedrycz, W. A comprehensive survey on applications of transformers for deep learning tasks. *Expert Syst. Appl.* **2024**, *241*, 122666. <https://doi.org/10.1016/j.eswa.2023.122666>.
3. Zhu, J.; Rong, J.; Kou, W.; Zhou, Q.; Suo, P. Robust real-time recognition of drones and birds in complex scenarios: A multimodal data fusion recognize approach. *Complex Intell. Syst.* **2026**, *12*, 1–21. <https://doi.org/10.1007/s40747-025-02117-y>.
4. Jiang, J.; Ke, L.; Chen, L.; Dou, B.; Zhu, Y.; Liu, J.; Zhang, B.; Zhou, T.; Wei, G.W. Transformer technology in molecular science. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2024**, *14*, e1725. <https://doi.org/10.1002/wcms.1725>.
5. Zhao, D.; Zhang, H.; Huang, K.; Zhu, X.; Arun, P.V.; Jiang, W.; Li, S.; Pei, X.; Zhou, H. SASU-Net: Hyperspectral Video Tracker based on Spectral Adaptive Aggregation Weighting and Scale Updating. *Expert Syst. Appl.* **2025**, *272*, 126721. <https://doi.org/10.1016/j.eswa.2025.126721>.
6. Peng, B.; Alcaide, E.; Anthony, Q.; Albalak, A.; Arcadinho, S.; Biderman, S.; Cao, H.; Cheng, X.; Chung, M.; Derczynski, L.; et al. RvkV: Reinventing rnns for the transformer era. In *Findings of the Association for Computational Linguistics: EMNLP 2023*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 14048–14077. <https://doi.org/10.18653/v1/2023.findings-emnlp.936>.
7. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. <https://doi.org/10.1109/TPAMI.2022.3152247>.
8. Chen, L.; Gu, L.; Li, L.; Yan, C.; Fu, Y. Frequency dynamic convolution for dense image prediction. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025; pp. 30178–30188. <https://doi.org/10.1109/CVPR52734.2025.02809>.
9. Ibrahim, V.; Cheikh, F.A.; Asari, V.K.; Paul, J.S. Extrapolation Convolution for Data Prediction on a 2-D Grid: Bridging Spatial and Frequency Domains with Applications in Image Outpainting and Compressed Sensing. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 20315–20327. <https://doi.org/10.1109/TNNLS.2025.3598745>.
10. Aslan, M.F.; Aslan, B.; Sabanci, K. Frequency-Domain Vision Transformers: Architectures, Applications, and Open Challenges. *Appl. Sci.* **2026**, *16*, 2024. <https://doi.org/10.3390/app16042024>.
11. Gao, N.; Jiang, X.; Zhang, X.; Deng, Y. Efficient frequency-domain image deraining with contrastive regularization. In *European Conference on Computer Vision*; Springer Nature: Cham, Switzerland, 2024; pp. 240–257. https://doi.org/10.1007/978-3-031-72940-9_14.
12. Jiang, W.; Zhao, D.; Wang, C.; Yu, X.; Arun, P.V.; Asano, Y.; Xiang, P.; Zhou, H. Hyperspectral Video Object Tracking with Cross-Modal Spectral Complementary and Memory Prompt Network. *Knowl.-Based Syst.* **2025**, *330*, 114595. <https://doi.org/10.1016/j.knosys.2025.114595>.
13. Zhang, H.; Zhu, Y.; Wang, D.; Zhang, L.; Chen, T.; Wang, Z.; Ye, Z. A survey on visual mamba. *Appl. Sci.* **2024**, *14*, 5683. <https://doi.org/10.3390/app14135683>.
14. Qu, H.; Ning, L.; An, R.; Fan, W.; Derr, T.; Liu, H.; Xu, X.; Li, Q. A survey of mamba. *arXiv* **2024**, arXiv:2408.01129. <https://doi.org/10.48550/arXiv.2408.01129>.
15. Yao, Y.; Wang, Q.; Zhao, D.; You, M.; Xiang, P.; Asano, Y.; Yu, X.; Wang, C.; Zhou, H.; Ren, J. DFBSNet: Dual Frequency-domain Branch Fusion and Selection Network for Hyperspectral Anomaly Detection. *Pattern Recognit.* **2026**, *180*, 113967. <https://doi.org/10.1016/j.patcog.2026.113967>.
16. Mai, X.; Zhang, H.; Jia, X.; Meng, M.Q.H. Faster R-CNN with classifier fusion for automatic detection of small fruits. *IEEE Trans. Autom. Sci. Eng.* **2020**, *17*, 1555–1569. <https://doi.org/10.1109/TASE.2020.2964289>.
17. Zhou, G.; Zhang, W.; Chen, A.; He, M.; Ma, X. Rapid detection of rice disease based on FCM-KM and faster R-CNN fusion. *IEEE Access* **2019**, *7*, 143190–143206. <https://doi.org/10.1109/ACCESS.2019.2943454>.
18. Chen, Y.; Wang, B.; Guo, X.; Zhu, W.; He, J.; Liu, X.; Yuan, J. Deyolo: Dual-feature-enhancement yolo for cross-modality object detection. In *International Conference on Pattern Recognition*; Springer Nature: Cham, Switzerland, 2024; pp. 236–252. https://doi.org/10.1007/978-3-031-78447-7_16.

19. Li, K.; Wang, D.; Hu, Z.; Li, S.; Ni, W.; Zhao, L.; Wang, Q. Fd2-net: Frequency-driven feature decomposition network for infra-red-visible object detection. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 4797–4805. <https://doi.org/10.1609/aaai.v39i5.32507>.
20. Ning, L.; Liu, X.; Zhou, L.; Zou, X. Cross-Modal Object Detection Based on Content-Guided Feature Fusion and Self-Calibration. *Sensors* **2025**, *25*, 3392. <https://doi.org/10.3390/s25113392>.
21. Qin, Q.; Chang, K.; Huang, M.; Li, G. Denet: Detection-driven enhancement network for object detection under adverse weather conditions. In Proceedings of the Asian Conference on Computer Vision, Macao, China, 4–8 December 2022; pp. 2813–2829. https://doi.org/10.1007/978-3-031-26313-2_30.
22. Bharati, P.; Pramanik, A. Deep learning techniques—R-CNN to mask R-CNN: A survey. *Comput. Intell. Pattern Recognit. Proc. CIPR* **2019**, *2019*, 657–668. https://doi.org/10.1007/978-981-13-9042-5_56.
23. Zhao, D.; Wang, M.; Huang, K.; Zhong, W.; Arun, P.V.; Li, Y.; Asano, Y.; Wu, L.; Zhou, H. OCSCNet-Tracker: Hyperspectral Video Tracker Based on Octave Convolution and Spatial-Spectral Capsule Network. *Remote Sens.* **2025**, *17*, 693. <https://doi.org/10.3390/rs17040693>.
24. Shen, J.; Chen, Y.; Liu, Y.; Zuo, X.; Fan, H.; Yang, W. ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognit.* **2024**, *145*, 109913. <https://doi.org/10.1016/j.patcog.2023.109913>.
25. Qingyun, F.; Dapeng, H.; Zhaokui, W. Cross-modality fusion transformer for multispectral object detection. *arXiv* **2021**, arXiv:2111.00273. <https://doi.org/10.48550/arXiv.2111.00273>.
26. Zhang, P.; He, X.; Tian, B.; Li, J. Research on disease identification of highway protection engineering based on improved YOLOv8 algorithm. In *International Conference on Frontiers of Traffic and Transportation Engineering (FTTE 2025)*; SPIE: Bellingham, WA, USA, 2026; Volume 14060, pp. 682–692. <https://doi.org/10.1117/12.3101939>.
27. Zhao, D.; Zhang, H.; Arun, P.V.; Jiao, C.; Zhou, H.; Xiang, P.; Cheng, K. SiamSTU: Hyperspectral Video Tracker Based on Spectral Spatial Angle Mapping Enhancement and State Aware Template Update. *Infrared Phys. Technol.* **2025**, *150*, 105919. <https://doi.org/10.1016/j.infrared.2025.105919>.
28. Merrill, W., Petty, J., & Sabharwal, A. The illusion of state in state-space models. *arXiv* **2024**, arXiv:2404.08819. <https://doi.org/10.48550/arXiv.2404.08819>.
29. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv* **2024**, arXiv:2401.09417. <https://doi.org/10.48550/arXiv.2401.09417>.
30. Dong, W.; Zhu, H.; Lin, S.; Luo, X.; Shen, Y.; Guo, G.; Zhang, B. Fusion-mamba for cross-modality object detection. *IEEE Trans. Multimed.* **2025**, *27*, 7392–7406. <https://doi.org/10.1109/TMM.2025.3599020>.
31. Zou, W.; Gao, H.; Yang, W.; Liu, T. Wave-mamba: Wavelet state space model for ultra-high-definition low-light image enhancement. In Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne, VIC, Australia, 28 October–1 November 2024; pp. 1534–1543. <https://doi.org/10.1145/3664647.3681580>.
32. Li, H.; Zhao, F.; Xue, F.; Wang, J.; Liu, Y.; Chen, Y.; Wu, Q.; Tao, J.; Zhang, G.; Xi, D.; et al. Succulent-YOLO: Smart UAV-Assisted Succulent Farmland Monitoring with CLIP-Based YOLOv10 and Mamba Computer Vision. *Remote Sens.* **2025**, *17*, 2219. <https://doi.org/10.3390/rs17132219>.
33. Chen, Z.; Yang, J.; Li, F.; Feng, Z.; Chen, L.; Jia, L.; Li, P. Foreign object detection method for railway catenary based on a scarce image generation model and lightweight perception architecture. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *36*, 1377–1391. <https://doi.org/10.1109/TCSVT.2025.3567319>.
34. Zhao, D.; Hu, B.; Jiang, W.; Zhong, W.; Arun, P.V.; Cheng, K.; Zhao, Z.; Zhou, H. Hyperspectral Video Tracker Based on Spectral Difference Matching Reduction and Deep Spectral Target Perception Features. *Opt. Lasers Eng.* **2025**, *194*, 109124. <https://doi.org/10.1016/j.optlaseng.2025.109124>.
35. Patro, B.N.; Agneeswaran, V.S. Mamba-360: Survey of state space models as transformer alternative for long sequence modeling: Methods, applications, and challenges. *Eng. Appl. Artif. Intell.* **2025**, *159*, 111279. <https://doi.org/10.1016/j.engappai.2025.111279>.
36. Jiang, N.; Chen, J.; Zhu, J.; Liang, B.; Yang, D.; Huang, X.; Xing, M. High Frame Rate Along-Track Swarm SAR Subaperture Collaboration Imaging for Moving Target. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5211721. <https://doi.org/10.1109/TGRS.2025.3567972>.
37. FREETeledyne. FLIR Thermal Dataset for Algorithm Training. Available online: <https://oem.flir.com/solutions/automotive/adas-dataset-form/> (accessed on 1 January 2026).
38. Liu, J.Y.; Fan, X.; Huang, Z.B.; Wu, G.Y.; Liu, R.S.; Zhong, W.; Luo, Z.X. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In Proceedings of the IEEE/CVF Conference

- on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5802–5811. <https://doi.org/10.48550/arXiv.2203.16220>.
39. Kumar, S.; Jain, A.; Rani, S.; Alshazly, H.; Idris, S.A.; Bourouis, S. Deep Neural Network Based Vehicle Detection and Classification of Aerial Images. *Intell. Autom. Soft Comput.* **2022**, *34*, 119–131. <https://doi.org/10.32604/iasc.2022.024812>.
40. Li, K.; Zhong, Z.; Luo, Z.; Tian, H.; Wang, K.; Jiang, H.; Xiang, D.; Tang, W. FCAT: Frequency-Domain Cross-Attention for All-Weather Multispectral Object Detection in Low-Altitude UAV Security Inspection of Urban and Industrial Areas. *Remote Sens.* **2026**, *18*, 826. <https://doi.org/10.3390/rs18050826>.

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.