

Article

Not peer-reviewed version

CMAFNet: Efficient Cross-Modal Alignment and Fusion for Real-Time RGB-Infrared Object Detection in Autonomous Driving

[Zi-Han Huang](#), [Chen-Wei Liang](#), [Mu-Jiang-Shan Wang](#) *

Posted Date: 5 March 2026

doi: 10.20944/preprints202603.0404.v1

Keywords: RGB-infrared fusion; multi-modal object detection; cross-modal alignment; state-space model; real-time perception; automotive embedded vision



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

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.

Article

CMAFNet: Efficient Cross-Modal Alignment and Fusion for Real-Time RGB–Infrared Object Detection in Autonomous Driving

Zi-Han Huang ¹, Chen-Wei Liang ¹ and Mu-Jiang-Shan Wang ^{2,*}

¹ The University of New South Wales, Sydney, NSW 2052, Australia
² Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
* Correspondence: mjs.wang@siat.ac.cn

Abstract

RGB–infrared (IR) fusion is an effective way to improve object detection robustness for automotive perception under low-light and adverse-weather conditions. Yet, practical multi-modal detectors still face three issues: imperfect cross-modal alignment, inefficient long-range interaction, and unstable query initialization when modalities exhibit inconsistent evidence. This paper presents CMAFNet, a deployment-oriented cross-modal alignment and fusion network with three key designs. (1) A Dynamic Receptive Backbone (DRB) extracts multi-scale features with adaptive receptive fields for both modalities. (2) A Channel-Split Mamba Block (CSM-Block) models long-range cross-modal dependencies using selective state-space modeling with linear complexity in token length, enabling an efficient accuracy–latency trade-off. (3) A Global Multi-modal Interaction Network (GMIN) performs fine-grained alignment and adaptive fusion via dual-branch cross-attention guided by global average/max pooling. In addition, an uncertainty-minimal query selection strategy and a separable dynamic decoder further enhance detection stability and efficiency. Experiments on M3FD and FLIR-Aligned show that CMAFNet achieves 83.9% mAP₅₀ and 84.2% mAP₅₀, respectively, while maintaining competitive inference efficiency, supporting real-time automotive deployment on compute-constrained platforms.

Keywords: RGB–infrared fusion; multi-modal object detection; cross-modal alignment; state-space model; real-time perception; automotive embedded vision

1. Introduction

Autonomous driving systems, as representative cyber-physical platforms in modern automotive electronics, demand robust and reliable perception capabilities to ensure safe navigation across diverse and challenging environmental conditions [1]. Object detection, as a fundamental perception task in the on-board sensing and perception stack, must maintain high accuracy under strict latency and computational constraints regardless of illumination variations, adverse weather, and complex traffic scenarios. While RGB cameras provide rich texture and color information under well-lit conditions, their performance degrades significantly in low-light environments, fog, rain, and glare [2]. Infrared (IR) sensors, which capture thermal radiation emitted by objects, offer complementary advantages by providing reliable detection capabilities in darkness and adverse weather conditions [3]. Consequently, the fusion of RGB and IR modalities has attracted increasing research attention as a promising approach to achieve robust all-weather object detection for autonomous driving and real-time automotive perception systems [4].

Despite significant progress in multi-modal object detection, several fundamental challenges remain unresolved. **First**, existing fusion methods often employ simple concatenation or element-wise addition to combine features from different modalities [5], which fails to capture the complex complementary relationships between RGB and IR features. The heterogeneous nature of these two

modalities—RGB images encode appearance, texture, and color information while IR images capture thermal signatures—necessitates more sophisticated alignment and interaction mechanisms. **Second**, most current approaches rely on convolutional neural networks (CNNs) or standard Transformer architectures for feature extraction and fusion [6]. CNNs are inherently limited by their local receptive fields, making it difficult to model long-range dependencies across modalities. While Transformers can capture global relationships, their quadratic computational complexity with respect to sequence length poses significant challenges for real-time and compute-constrained automotive deployment [7]. **Third**, the quality of object queries in detection heads significantly impacts final detection performance, yet most methods do not explicitly address the uncertainty in query selection when dealing with multi-modal inputs [8]. Beyond perception accuracy, robustness and reliability under adverse conditions have also been conceptually discussed from connectivity and diagnosability perspectives in complex networks [9–12], highlighting the importance of redundancy and complementary information sources.

To address these challenges under practical real-time constraints, we propose CMAFNet (Cross-Modal Alignment and Fusion Network), a comprehensive yet deployment-aware framework for robust multi-modal object detection in autonomous driving. Our approach introduces several novel components that work synergistically to achieve effective cross-modal feature alignment, interaction, and fusion while maintaining a favorable accuracy–efficiency trade-off. The overall architecture of CMAFNet is illustrated in Figure 1.

Specifically, our contributions are summarized as follows:

- We propose CMAFNet, a deployment-aware end-to-end multi-modal object detection framework that achieves robust cross-modal alignment and fusion for autonomous driving under real-time constraints. The framework integrates a Dynamic Receptive Backbone, Channel-Split Mamba Block, Global Multi-modal Interaction Network, and Separable Dynamic Decoder into a unified and efficiency-oriented architecture.
- We introduce the Channel-Split Mamba Block (CSM-Block), which leverages selective state space models [13] to efficiently capture long-range cross-modal dependencies. By splitting feature channels into multiple groups and processing them through parallel CS-Mamba branches with subsequent attention-based recalibration, the CSM-Block enables effective global context modeling with linear computational complexity, making it suitable for compute-constrained automotive perception systems.
- We design the Global Multi-modal Interaction Network (GMIN), a dual-branch cross-modal attention module that performs fine-grained feature alignment through global average pooling and global max pooling guided cross-attention. GMIN generates modality-specific gating signals to adaptively weight and fuse complementary information from RGB and IR features, improving robustness under adverse sensing conditions.
- Extensive experiments on two widely-used benchmarks (M3FD and FLIR-Aligned) demonstrate that CMAFNet achieves state-of-the-art performance while maintaining a favorable accuracy–latency trade-off, supporting practical real-time deployment in automotive embedded vision scenarios.

The remainder of this paper is organized as follows. Section 2 reviews related work on multi-modal object detection, state space models, and attention mechanisms. Section 3 presents the proposed CMAFNet framework in detail. Section 4 describes the experimental setup and presents comprehensive results, including efficiency analysis. Section 6 concludes the paper.

2. Related Work

2.1. Multi-Modal Object Detection

Multi-modal object detection aims to leverage complementary information from multiple sensor modalities to improve detection robustness in real-world perception systems [1]. In the context of autonomous driving and automotive perception stacks, RGB-infrared fusion has been extensively studied. Early approaches adopted simple fusion strategies such as early fusion (input-level concatenation),

late fusion (decision-level combination), and halfway fusion (feature-level merging) [5]. Liu et al. [5] systematically compared these strategies for multispectral pedestrian detection and demonstrated that halfway fusion generally achieves a reasonable trade-off between performance and computational complexity.

More recent methods have explored attention-based fusion mechanisms to improve cross-modal interaction. Zhang et al. [14] proposed Guided Attentive Feature Fusion (GAFF) that uses illumination-aware gating to adaptively weight modality contributions. Zhou et al. [15] introduced MBNet to address modality imbalance through differential modality-aware feature learning. The Cross-modality Interactive Attention Network (CIAN) [16] employs cross-attention to model inter-modal relationships. CFT [6] introduced a cross-modality fusion transformer that uses self-attention and cross-attention to capture both intra-modal and inter-modal dependencies. ICAFusion [17] proposed iterative cross-attention for progressive feature alignment. BAANet [18] designed bi-directional adaptive attention gates for multispectral pedestrian detection. EAEFNet [19] introduced explicit attention-enhanced fusion for RGB-thermal perception. TFDet [20] proposed target-aware fusion mechanisms for pedestrian detection.

Although these approaches significantly improve detection accuracy, many of them rely on Transformer-style attention mechanisms with quadratic complexity or introduce additional computational overhead, which may limit scalability under real-time and embedded deployment constraints. Our proposed CMAFNet addresses these limitations by introducing the CSM-Block with linear complexity and the GMIN module for comprehensive yet efficient cross-modal interaction. From a broader robustness perspective, classical connectivity and matching-preclusion studies in interconnection networks highlight how redundancy-preserving structures maintain reliable information flow under disruptions [21–24]. This conceptual viewpoint resonates with cross-modal fusion, where complementary sensing modalities provide redundancy to sustain detection reliability under adverse environmental conditions.

2.2. State Space Models for Vision

State Space Models (SSMs) have recently emerged as a promising alternative to Transformers for sequence modeling with improved computational efficiency [25]. The Mamba architecture [13] introduced a selective state space mechanism with hardware-aware design, achieving linear computational complexity while maintaining strong modeling capabilities for long sequences. Vision Mamba [26] adapted the Mamba architecture for visual representation learning by introducing bidirectional scanning strategies. VMamba [27] proposed a visual state space model with cross-scan mechanisms for hierarchical feature extraction.

The efficiency advantages of SSM-based architectures make them particularly attractive for perception modules operating under latency and compute constraints. In multi-modal fusion, the ability to efficiently model long-range dependencies across modalities is crucial for capturing complementary information without incurring prohibitive computational cost. In this work, we propose the Channel-Split Mamba Block that leverages the selective state space mechanism to capture cross-modal dependencies with linear complexity, making it suitable for real-time autonomous driving applications.

2.3. Attention Mechanisms in Object Detection

Attention mechanisms have been widely adopted in object detection to enhance feature representation [7]. Channel attention mechanisms such as SE-Net [28] and spatial attention mechanisms such as CBAM [29] have demonstrated effectiveness in single-modal detection with modest computational overhead. In the multi-modal setting, cross-attention mechanisms enable explicit interaction between features from different modalities [6], but may introduce additional complexity depending on implementation.

The DETR family of detectors [8,30–32] has reformulated object detection as a set prediction problem. Deformable DETR [31] introduced deformable attention to reduce computational cost. DINO [8] improved detection performance through denoising training and mixed query selection.

RT-DETR [32] achieved real-time performance by designing an efficient hybrid encoder. Co-DETR [33] proposed collaborative hybrid assignments for improved training efficiency. These advances in query-based detection provide a strong architectural foundation for our uncertainty-minimal query selection strategy under multi-modal settings.

3. Proposed Method

In this section, we present the proposed CMAFNet framework in detail. As illustrated in Figure 1, CMAFNet consists of four main components: (1) a Dynamic Receptive Backbone (DRB) for multi-scale feature extraction, (2) a Channel-Split Mamba Block (CSM-Block) for efficient cross-modal feature enhancement, (3) a Global Multi-modal Interaction Network (GMIN) for fine-grained cross-modal fusion, and (4) an Uncertainty-minimal Query Selection mechanism with a Separable Dynamic Decoder for final detection. The overall design emphasizes efficient cross-modal interaction and real-time feasibility under automotive perception constraints.

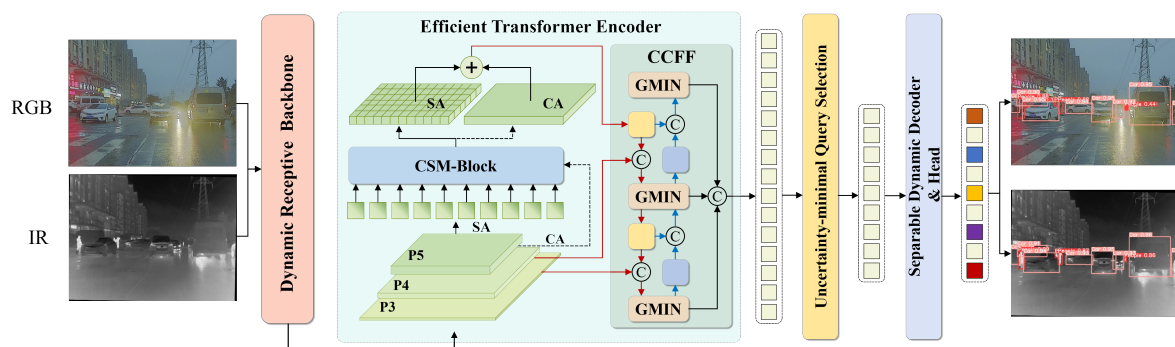


Figure 1. Overall architecture of the proposed CMAFNet. The framework consists of four main components: (1) a shared Dynamic Receptive Backbone for multi-scale feature extraction from both RGB and IR modalities, (2) an Efficient Transformer Encoder with CSM-Block for cross-modal feature enhancement, (3) GMIN modules for global multi-modal interaction and fusion at multiple scales, and (4) an Uncertainty-minimal Query Selection mechanism followed by a Separable Dynamic Decoder for final detection. The architecture is designed to balance detection accuracy and computational efficiency for real-time automotive deployment.

3.1. Dynamic Receptive Backbone

Effective multi-scale feature extraction is fundamental to object detection, particularly in autonomous driving scenarios where objects exhibit significant scale variations (e.g., distant pedestrians vs. nearby vehicles). In practical automotive perception systems, the backbone must not only capture scale diversity but also maintain computational efficiency. We adopt a Dynamic Receptive Backbone (DRB) that adaptively adjusts its receptive field to capture features at multiple scales while preserving favorable computational characteristics, as shown in Figure 2.

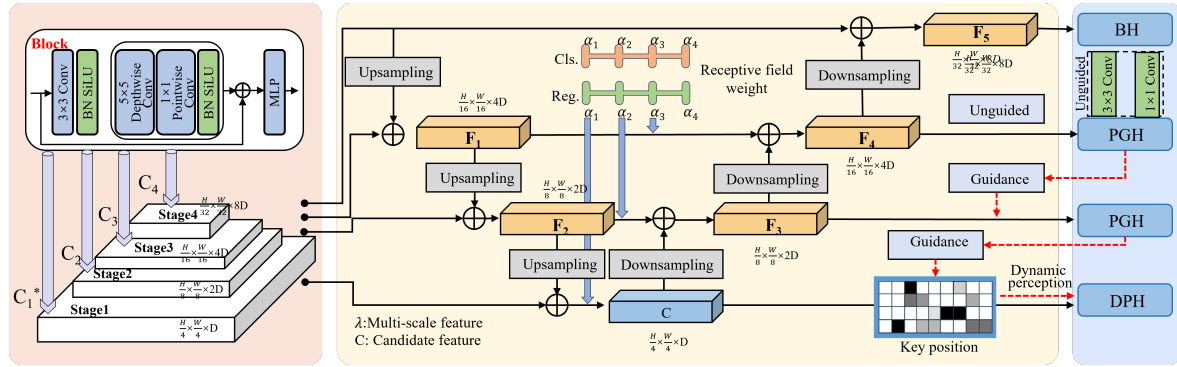


Figure 2. Architecture of the Dynamic Receptive Backbone (DRB). The backbone consists of four stages that progressively extract multi-scale features $\{C_1, C_2, C_3, C_4\}$. Each stage employs a Block module composed of 3×3 convolution, depthwise separable convolution, pointwise convolution, and MLP layers. The multi-scale features are further processed through upsampling and downsampling paths with receptive field weight adaptation to generate enhanced feature maps $\{F_1, F_2, F_3, F_4, F_5\}$. The backbone includes multiple detection heads: BH (Base Head), PGH (Perception-Guided Head), and DPH (Dynamic Perception Head). The design maintains a balance between representational capacity and computational cost.

The DRB processes both RGB and IR inputs through a shared-weight backbone architecture, which reduces parameter redundancy and facilitates efficient multi-modal feature extraction. Given an input image $\mathbf{I} \in \mathbb{R}^{3 \times H \times W}$ (or $\mathbb{R}^{1 \times H \times W}$ for single-channel IR), the backbone extracts hierarchical features through four stages:

$$\{C_1, C_2, C_3, C_4\} = \text{DRB}(\mathbf{I}), \quad (1)$$

where $C_i \in \mathbb{R}^{c_i \times \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}}$ represents the feature map at stage i with channel dimension $c_i \in \{D, 2D, 4D, 8D\}$ and D denotes the base channel dimension. This hierarchical design enables progressive spatial reduction while maintaining representational efficiency.

Each building block in the DRB consists of a sequence of operations:

$$\mathbf{x}' = \text{MLP}(\text{BN}(\text{SiLU}(\text{PW}(\text{BN}(\text{SiLU}(\text{DW}_{5 \times 5}(\text{BN}(\text{SiLU}(\text{Conv}_{3 \times 3}(\mathbf{x}))))))))), \quad (2)$$

where $\text{Conv}_{3 \times 3}$ denotes a standard 3×3 convolution, $\text{DW}_{5 \times 5}$ represents a 5×5 depthwise separable convolution [34], PW is a 1×1 pointwise convolution, BN is batch normalization [35], and SiLU is the Sigmoid Linear Unit activation [36]. The use of depthwise separable convolution reduces computational overhead compared to standard convolution while preserving expressive capacity.

The multi-scale feature enhancement module employs bidirectional feature propagation with adaptive receptive field weights. For the top-down path, features are progressively upsampled and fused:

$$F_i^{\text{td}} = C_i \oplus \text{Up}(F_{i+1}^{\text{td}}), \quad i = 3, 2, 1, \quad (3)$$

where $\oplus$ denotes element-wise addition and $\text{Up}(\cdot)$ represents bilinear upsampling. For the bottom-up path:

$$F_i^{\text{bu}} = F_i^{\text{td}} \oplus \text{Down}(F_{i-1}^{\text{bu}}), \quad i = 2, 3, 4, \quad (4)$$

where $\text{Down}(\cdot)$ denotes strided convolution for downsampling. This bidirectional aggregation enhances multi-scale representation without introducing excessive architectural complexity.

The receptive field weight adaptation mechanism dynamically adjusts the contribution of features at different scales through learned attention weights:

$$\alpha_i = \text{Softmax}\left(\text{Reg}\left(\text{Cls}\left(\text{Up}\left(F_4^{\text{bu}}\right)\right)\right)\right), \quad (5)$$

$$F_i = \alpha_i \odot F_i^{\text{bu}}, \quad i = 1, 2, 3, 4, \quad (6)$$

where $\text{Cls}(\cdot)$ and $\text{Reg}(\cdot)$ are classification and regression branches, and $\odot$ denotes element-wise multiplication. The adaptive weighting mechanism enables dynamic emphasis on informative scales while maintaining computational tractability.

For both RGB and IR modalities, we extract multi-scale features:

$$\{P_3^{\text{rgb}}, P_4^{\text{rgb}}, P_5^{\text{rgb}}\} = \text{DRB}(\mathbf{I}_{\text{rgb}}), \quad \{P_3^{\text{ir}}, P_4^{\text{ir}}, P_5^{\text{ir}}\} = \text{DRB}(\mathbf{I}_{\text{ir}}), \quad (7)$$

where P_3 , P_4 , and P_5 correspond to feature maps at $1/8$, $1/16$, and $1/32$ of the input resolution, respectively. The shared backbone structure promotes parameter efficiency and consistent feature abstraction across modalities.

3.2. Channel-Split Mamba Block

To efficiently capture long-range dependencies across modalities while maintaining strict computational budgets, we propose the Channel-Split Mamba Block (CSM-Block). Unlike standard Transformer-based cross-attention that incurs quadratic complexity with respect to token length, the CSM-Block leverages selective state space models to achieve linear complexity, which improves scalability for high-resolution inputs in real-time automotive perception systems. The detailed architecture is shown in Figure 3.

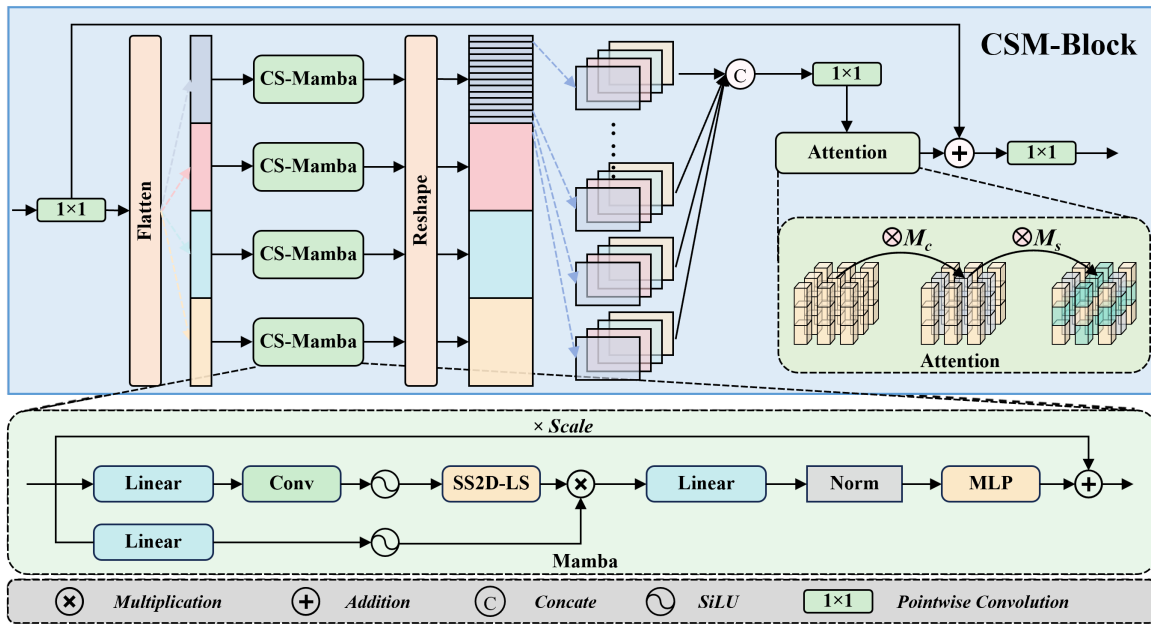


Figure 3. Architecture of the Channel-Split Mamba Block (CSM-Block). The input features are first processed by a 1×1 convolution, then flattened and split into multiple channel groups. Each group is independently processed by a CS-Mamba module. The outputs are reshaped and concatenated, followed by a 1×1 convolution. An attention mechanism with channel attention map M_c and spatial attention map M_s is applied for feature recalibration. The bottom panel shows the internal structure of the CS-Mamba module, which consists of a dual-branch architecture with Linear projection, depthwise convolution, SiLU activation, SS2D-LS (Selective Scan 2D with Local Scan), element-wise multiplication, Linear projection, Layer Normalization, and MLP with residual connection. The design balances global modeling capability and computational efficiency.

3.2.1. Channel Splitting and Parallel Processing

Given the concatenated RGB and IR features $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$ at a specific scale, the CSM-Block first applies a 1×1 convolution for channel projection:

$$\mathbf{F}' = \text{Conv}_{1 \times 1}(\mathbf{F}) \in \mathbb{R}^{C \times H \times W}. \quad (8)$$

The projected features are then flattened to $\mathbf{F}' \in \mathbb{R}^{C \times N}$ where $N = H \times W$, and split along the channel dimension into G groups:

$$\{\mathbf{F}'_1, \mathbf{F}'_2, \dots, \mathbf{F}'_G\} = \text{Split}(\mathbf{F}'), \quad \mathbf{F}'_g \in \mathbb{R}^{(C/G) \times N}. \quad (9)$$

This channel partitioning strategy enables parallel processing of feature subspaces, reducing per-branch computational burden while preserving expressive diversity. Each group is independently processed by a CS-Mamba module:

$$\hat{\mathbf{F}}_g = \text{CS-Mamba}(\mathbf{F}'_g), \quad g = 1, 2, \dots, G. \quad (10)$$

The parallel design improves computational scalability and facilitates efficient implementation on modern GPU architectures.

3.2.2. CS-Mamba Module

The CS-Mamba module is built upon the selective state space mechanism [13], which provides linear-time sequence modeling with hardware-aware characteristics. For each channel group $\mathbf{F}'_g$, the module employs a dual-branch architecture:

$$\mathbf{z}_g = \text{Linear}_z(\mathbf{F}'_g), \quad (11)$$

$$\mathbf{h}_g = \text{SiLU}\left(\text{DWConv}\left(\text{Linear}_h(\mathbf{F}'_g)\right)\right), \quad (12)$$

$$\mathbf{y}_g = \text{SS2D-LS}(\mathbf{h}_g) \odot \mathbf{z}_g, \quad (13)$$

where Linear_z and Linear_h are linear projection layers, DWConv is a depthwise convolution, and SS2D-LS denotes the Selective Scan 2D with Local Scan operation. This gated interaction mechanism enables effective cross-token modulation while preserving computational efficiency.

The SS2D-LS operation extends the 1D selective scan to 2D feature maps through four directional scans (left-to-right, right-to-left, top-to-bottom, bottom-to-top) combined with local scanning patterns. Compared with global self-attention, the directional scanning strategy avoids quadratic pairwise interactions while still enabling long-range information propagation. For each direction d , the selective state space dynamics are defined as:

$$\mathbf{s}_t^{(d)} = \bar{\mathbf{A}}^{(d)} \mathbf{s}_{t-1}^{(d)} + \bar{\mathbf{B}}^{(d)} \mathbf{h}_t^{(d)}, \quad (14)$$

$$\mathbf{o}_t^{(d)} = \mathbf{C}^{(d)} \mathbf{s}_t^{(d)}, \quad (15)$$

where $\mathbf{s}_t^{(d)}$ is the hidden state, $\bar{\mathbf{A}}^{(d)}$ and $\bar{\mathbf{B}}^{(d)}$ are the discretized state transition and input matrices, $\mathbf{C}^{(d)}$ is the output matrix, and $\mathbf{h}_t^{(d)}$ is the input at position t along direction d . The discretization is performed using the zero-order hold (ZOH) method:

$$\bar{\mathbf{A}} = \exp(\Delta \mathbf{A}), \quad \bar{\mathbf{B}} = (\Delta \mathbf{A})^{-1}(\exp(\Delta \mathbf{A}) - \mathbf{I}) \cdot \Delta \mathbf{B}, \quad (16)$$

where Δ is the input-dependent step size that enables selective processing and adaptive receptive behavior.

The outputs from all directions are merged:

$$\text{SS2D-LS}(\mathbf{h}_g) = \sum_{d=1}^4 \mathbf{o}^{(d)}. \quad (17)$$

This aggregation preserves global contextual cues while maintaining linear computational complexity.

The final output of the CS-Mamba module includes normalization and MLP:

$$\hat{\mathbf{F}}_g = \mathbf{F}'_g + \text{MLP}(\text{LN}(\text{Linear}(\mathbf{y}_g))), \quad (18)$$

where LN denotes Layer Normalization [37]. The residual formulation stabilizes training and facilitates efficient deep stacking in practical perception pipelines.

3.2.3. Attention-Based Recalibration

After parallel processing, the group outputs are concatenated and reshaped:

$$\hat{\mathbf{F}} = \text{Reshape}(\text{Concat}(\hat{\mathbf{F}}_1, \hat{\mathbf{F}}_2, \dots, \hat{\mathbf{F}}_G)) \in \mathbb{R}^{C \times H \times W}. \quad (19)$$

To further enhance discriminative capability without introducing substantial computational overhead, a lightweight dual-attention mechanism is applied for feature recalibration. The channel attention map M_c and spatial attention map M_s are computed as:

$$M_c = \sigma(\text{Conv}_{1 \times 1}(\text{Concat}(\text{AvgPool}(\hat{\mathbf{F}}), \text{MaxPool}(\hat{\mathbf{F}})))) \in \mathbb{R}^{C \times 1 \times 1}, \quad (20)$$

$$M_s = \sigma(\text{Conv}_{1 \times 1}(\text{Concat}(\text{AvgPool}_c(\hat{\mathbf{F}}), \text{MaxPool}_c(\hat{\mathbf{F}})))) \in \mathbb{R}^{1 \times H \times W}, \quad (21)$$

where σ denotes the sigmoid function, AvgPool and MaxPool are global average and max pooling along spatial dimensions, and AvgPool_c and MaxPool_c are average and max pooling along the channel dimension. The use of pooling-based statistics enables efficient global context encoding with minimal additional parameters.

The final recalibrated output is:

$$\tilde{\mathbf{F}} = \text{Conv}_{1 \times 1}((M_c \otimes \hat{\mathbf{F}}) \oplus (M_s \otimes \hat{\mathbf{F}})), \quad (22)$$

where $\otimes$ denotes broadcasting multiplication. This adaptive modulation refines both channel-wise and spatial responses while maintaining computational tractability for real-time deployment.

3.3. Global Multi-modal Interaction Network

While the CSM-Block captures long-range dependencies efficiently, fine-grained cross-modal alignment requires explicit modeling of the complementary and reliability-aware relationships between RGB and IR features. We propose the Global Multi-modal Interaction Network (GMIN) to achieve adaptive cross-modal modulation with limited additional computational cost. The architecture of GMIN is illustrated in Figure 4.

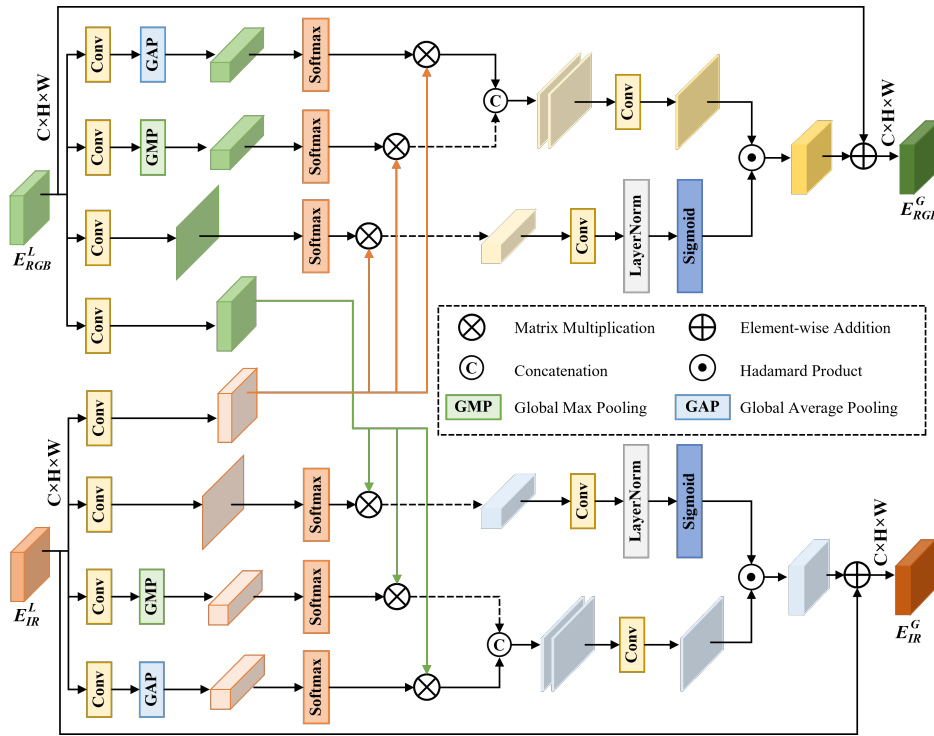


Figure 4. Architecture of the Global Multi-modal Interaction Network (GMIN). The module takes RGB features E_{RGB}^L and IR features E_{IR}^L as inputs and produces globally enhanced features E_{RGB}^G and E_{IR}^G . Each modality branch generates attention weights through Global Average Pooling (GAP) and Global Max Pooling (GMP) guided cross-attention. The cross-modal interaction is achieved through matrix multiplication ($\otimes$) between attention maps from different modalities. Modality-specific gating signals are generated through concatenation, convolution, LayerNorm, and Sigmoid activation to produce the final fused features via Hadamard product ($\odot$) and element-wise addition ($\oplus$). The design emphasizes adaptive modality weighting and computational efficiency.

3.3.1. Dual-Branch Cross-Modal Attention

Given the local features from the RGB and IR modalities at scale l , denoted as $E_{RGB}^L \in \mathbb{R}^{C \times H \times W}$ and $E_{IR}^L \in \mathbb{R}^{C \times H \times W}$, GMIN first generates multiple attention representations for each modality. The pooling-guided design provides compact global descriptors that encode modality saliency with negligible computational overhead.

For the RGB branch, three parallel convolution paths produce query, key, and value-like representations:

$$\mathbf{Q}_{rgb} = \text{Conv}_q(E_{RGB}^L), \quad \mathbf{K}_{rgb}^{\text{avg}} = \text{GAP}(\text{Conv}_k(E_{RGB}^L)), \quad \mathbf{K}_{rgb}^{\text{max}} = \text{GMP}(\text{Conv}_k(E_{RGB}^L)), \quad (23)$$

where $\text{GAP}(\cdot)$ and $\text{GMP}(\cdot)$ denote Global Average Pooling and Global Max Pooling, respectively. Similarly, for the IR branch:

$$\mathbf{Q}_{ir} = \text{Conv}_q(E_{IR}^L), \quad \mathbf{K}_{ir}^{\text{avg}} = \text{GAP}(\text{Conv}_k(E_{IR}^L)), \quad \mathbf{K}_{ir}^{\text{max}} = \text{GMP}(\text{Conv}_k(E_{IR}^L)). \quad (24)$$

The attention weights are computed through softmax normalization:

$$\mathbf{A}_{rgb}^{\text{avg}} = \text{Softmax}(\mathbf{K}_{rgb}^{\text{avg}}), \quad \mathbf{A}_{rgb}^{\text{max}} = \text{Softmax}(\mathbf{K}_{rgb}^{\text{max}}), \quad (25)$$

$$\mathbf{A}_{ir}^{\text{avg}} = \text{Softmax}(\mathbf{K}_{ir}^{\text{avg}}), \quad \mathbf{A}_{ir}^{\text{max}} = \text{Softmax}(\mathbf{K}_{ir}^{\text{max}}). \quad (26)$$

3.3.2. Cross-Modal Interaction

The key innovation of GMIN lies in its cross-modal interaction mechanism. Instead of computing self-attention within each modality, we perform cross-attention by using the attention weights from

one modality to modulate the features of the other. This asymmetric modulation enables each modality to emphasize informative cues from its complementary counterpart while suppressing unreliable responses:

$$\mathbf{V}_{\text{rgb} \leftarrow \text{ir}} = \mathbf{Q}_{\text{rgb}} \otimes \mathbf{A}_{\text{ir}}^{\text{avg}} + \mathbf{Q}_{\text{rgb}} \otimes \mathbf{A}_{\text{ir}}^{\text{max}}, \quad (27)$$

$$\mathbf{V}_{\text{ir} \leftarrow \text{rgb}} = \mathbf{Q}_{\text{ir}} \otimes \mathbf{A}_{\text{rgb}}^{\text{avg}} + \mathbf{Q}_{\text{ir}} \otimes \mathbf{A}_{\text{rgb}}^{\text{max}}, \quad (28)$$

where $\otimes$ denotes matrix multiplication. The use of pooled attention descriptors avoids full pairwise token interactions, thereby maintaining favorable computational characteristics.

The cross-modal interaction features are concatenated and processed:

$$\mathbf{V}_{\text{cross}} = \text{Conv}(\text{Concat}(\mathbf{V}_{\text{rgb} \leftarrow \text{ir}}, \mathbf{V}_{\text{ir} \leftarrow \text{rgb}})). \quad (29)$$

3.3.3. Modality-Specific Gating

To adaptively control the contribution of cross-modal information to each modality under varying sensing conditions, we generate modality-specific gating signals:

$$\mathbf{G}_{\text{rgb}} = \sigma(\text{LN}(\text{Conv}(\mathbf{V}_{\text{cross}}))), \quad (30)$$

$$\mathbf{G}_{\text{ir}} = \sigma(\text{LN}(\text{Conv}(\mathbf{V}_{\text{cross}}))), \quad (31)$$

where σ is the sigmoid function and LN is Layer Normalization. The gating signals act as modality reliability estimators and modulate the original features through Hadamard product and residual connection:

$$E_{\text{RGB}}^G = \mathbf{G}_{\text{rgb}} \odot \text{Conv}(\mathbf{V}_{\text{rgb} \leftarrow \text{ir}}) \oplus E_{\text{RGB}}^L, \quad (32)$$

$$E_{\text{IR}}^G = \mathbf{G}_{\text{ir}} \odot \text{Conv}(\mathbf{V}_{\text{ir} \leftarrow \text{rgb}}) \oplus E_{\text{IR}}^L. \quad (33)$$

The residual formulation preserves modality-specific information while selectively incorporating complementary cues. The globally enhanced features E_{RGB}^G and E_{IR}^G are then concatenated to form the fused multi-modal features at each scale.

3.4. Cross-Modal Feature Fusion

After obtaining the enhanced features from the CSM-Block and GMIN at multiple scales, we perform Cross-scale Channel Feature Fusion (CCFF) to generate the final multi-scale fused representations. The GMIN modules operate at three scales (P_3, P_4, P_5), and the fused features are concatenated:

$$\mathbf{F}_l^{\text{fused}} = \text{Concat}(E_{\text{RGB},l}^G, E_{\text{IR},l}^G), \quad l \in \{3, 4, 5\}. \quad (34)$$

The multi-scale fused features $\{\mathbf{F}_3^{\text{fused}}, \mathbf{F}_4^{\text{fused}}, \mathbf{F}_5^{\text{fused}}\}$ are further processed through the Efficient Transformer Encoder with self-attention (SA) and cross-attention (CA) mechanisms to capture intra-scale and cross-scale relationships. The encoder is configured to maintain a favorable balance between representational power and computational overhead for real-time perception pipelines.

3.5. Uncertainty-Minimal Query Selection

In DETR-based detectors, the quality of object queries significantly impacts detection performance [8]. When dealing with multi-modal inputs, inconsistency between modalities may introduce prediction variance that affects query reliability. We propose an Uncertainty-minimal Query Selection (UMQS) strategy that explicitly favors candidates exhibiting both high confidence and cross-modal consistency.

Given the fused encoder features, we first generate candidate queries through a scoring mechanism:

$$s_i = \text{MLP}_{\text{cls}}(\mathbf{f}_i) + \lambda \cdot \text{MLP}_{\text{loc}}(\mathbf{f}_i), \quad (35)$$

where $\mathbf{f}_i$ is the i -th candidate feature, MLP_{cls} and MLP_{loc} predict classification and localization scores, and λ is a balancing coefficient.

The uncertainty of each candidate is estimated by measuring the prediction variance across the two modalities:

$$u_i = \text{Var}(\text{MLP}_{\text{cls}}(\mathbf{f}_i^{\text{rgb}}), \text{MLP}_{\text{cls}}(\mathbf{f}_i^{\text{ir}})) + \text{Var}(\text{MLP}_{\text{loc}}(\mathbf{f}_i^{\text{rgb}}), \text{MLP}_{\text{loc}}(\mathbf{f}_i^{\text{ir}})). \quad (36)$$

This variance-based metric provides a lightweight estimation of cross-modal disagreement without requiring additional forward passes. The final query selection criterion combines the score and uncertainty:

$$q_i = s_i - \beta \cdot u_i, \quad (37)$$

where β controls the penalty for high-uncertainty candidates. The top- K candidates with the highest q_i values are selected as object queries for the decoder, thereby promoting stable and reliable detection initialization.

3.6. Separable Dynamic Decoder

For the detection decoder, we adopt the Separable Dynamic Decoder proposed by Hu et al. [38], which replaces the standard multi-head cross-attention in DETR decoders with an efficient dynamic convolution attention mechanism. This design further reduces quadratic attention overhead and improves scalability for high-resolution multi-modal feature maps. The architecture is shown in Figure 5.

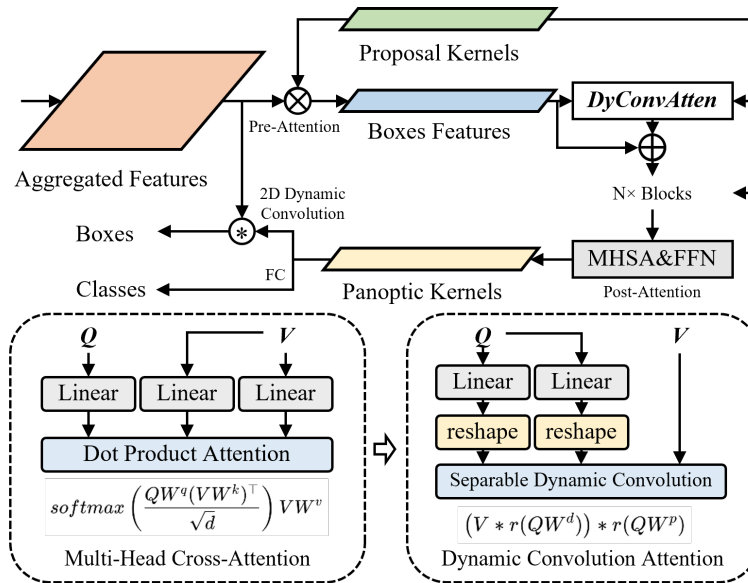


Figure 5. Architecture of the Separable Dynamic Decoder [38]. The decoder takes aggregated features and proposal kernels as inputs. Pre-Attention generates box features through matrix multiplication. The DyConvAtten module applies dynamic convolution attention with residual connection, followed by N blocks of Multi-Head Self-Attention (MHSA) and Feed-Forward Network (FFN) in the Post-Attention stage. The bottom panels compare the standard Multi-Head Cross-Attention (left) with the proposed Dynamic Convolution Attention (right), where the attention computation $\text{softmax}\left(\frac{QW^q(VW^k)^T}{\sqrt{d}}\right)VW^v$ is replaced by the separable dynamic convolution $(V * r(QW^d)) * r(QW^p)$. The separable formulation improves computational efficiency while preserving modeling flexibility.

The standard multi-head cross-attention computes:

$$\text{MHCA}(Q, K, V) = \text{softmax}\left(\frac{QW^q(KW^k)^T}{\sqrt{d}}\right)VW^v, \quad (38)$$

which has $\mathcal{O}(N^2d)$ complexity. The Separable Dynamic Decoder replaces this with dynamic convolution attention:

$$\text{DyConvAttn}(Q, V) = (V * r(QW^d)) * r(QW^p), \quad (39)$$

where $*$ denotes convolution, $r(\cdot)$ is a reshape operation, W^d and W^p are learnable weight matrices that generate depth-wise and point-wise dynamic convolution kernels from the queries, respectively. This formulation reduces the computational complexity to $\mathcal{O}(Ndk)$ where k is the kernel size, enabling a more favorable accuracy–latency trade-off compared with standard cross-attention.

The decoder iteratively refines the selected queries through N blocks of dynamic convolution attention and self-attention:

$$\mathbf{q}^{(n)} = \text{FFN}\left(\text{MHSA}\left(\text{DyConvAttn}(\mathbf{q}^{(n-1)}, \mathbf{V})\right)\right), \quad n = 1, 2, \dots, N, \quad (40)$$

where $\mathbf{q}^{(0)}$ represents the initial queries from UMQS and $\mathbf{V}$ represents the aggregated multi-scale features. The iterative refinement improves localization precision while maintaining computational tractability.

The final detection outputs are produced by applying classification and regression heads to the refined queries:

$$\hat{c}_i = \text{FC}_{\text{cls}}(\mathbf{q}_i^{(N)}), \quad \hat{b}_i = \text{FC}_{\text{box}}(\mathbf{q}_i^{(N)}), \quad (41)$$

where $\hat{c}_i$ and $\hat{b}_i$ are the predicted class and bounding box for the i -th query.

3.7. Loss Function

Following the DETR paradigm [30], we employ the Hungarian algorithm [39] for bipartite matching between predictions and ground truth. The total loss function is defined as:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{cls}} \mathcal{L}_{\text{cls}} + \lambda_{\text{box}} \mathcal{L}_{\text{box}} + \lambda_{\text{giou}} \mathcal{L}_{\text{giou}}, \quad (42)$$

where $\mathcal{L}_{\text{cls}}$ is the focal loss [40] for classification, $\mathcal{L}_{\text{box}}$ is the L1 loss for bounding box regression, and $\mathcal{L}_{\text{giou}}$ is the Generalized IoU loss [41]. The loss weights are set as $\lambda_{\text{cls}} = 2.0$, $\lambda_{\text{box}} = 5.0$, and $\lambda_{\text{giou}} = 2.0$, which follow common DETR-based configurations.

4. Experiments

4.1. Datasets

We evaluate the proposed CMAFNet on two widely-used multi-modal object detection benchmarks for autonomous driving and real-world perception evaluation.

M3FD [4]. The Multi-scenario Multi-Modality Fusion Dataset (M3FD) is a challenging benchmark containing 4,200 pairs of well-aligned RGB and infrared images captured under diverse driving scenarios. The dataset covers six object categories: *People*, *Car*, *Bus*, *Motorcycle*, *Lamp*, and *Truck*. The images are collected across multiple scenarios including daytime, nighttime, overexposure, and adverse conditions (rain, fog, snow), which are critical for robustness assessment in automotive perception systems. Following the standard protocol, we use 2,940 image pairs for training and 1,260 for testing.

FLIR-Aligned [3]. The FLIR ADAS dataset provides paired RGB and thermal infrared images for autonomous driving perception. The aligned version contains 5,142 well-registered image pairs. The dataset includes three main categories: *Person*, *Car*, and *Bicycle*. We follow the standard split with 4,129 pairs for training and 1,013 pairs for testing. This dataset further evaluates the generalization capability of multi-modal detectors across diverse traffic environments.

4.2. Implementation Details

CMAFNet is implemented using PyTorch and trained on 4 NVIDIA A100 GPUs. The input images are resized to 640×640 for both modalities to maintain a balance between detection accuracy

and computational cost. The Dynamic Receptive Backbone uses a base channel dimension $D = 64$. The CSM-Block employs $G = 4$ channel groups with a state dimension of 16 for the selective state space model, which ensures linear-complexity modeling without excessive memory overhead. The GMIN module uses $C = 256$ channels at each scale. The Separable Dynamic Decoder consists of $N = 6$ decoder layers with 300 object queries ($K = 300$), providing sufficient detection capacity while preserving real-time feasibility.

We use the AdamW optimizer [42] with an initial learning rate of 1×10^{-4} , weight decay of 1×10^{-4} , and a cosine annealing schedule [43]. The model is trained for 120 epochs with a batch size of 8. Data augmentation includes random horizontal flipping, random scaling ($0.5\text{--}1.5\times$), color jittering, and mosaic augmentation [44]. The backbone is initialized with COCO pre-trained weights [45]. The uncertainty penalty coefficient β is set to 0.5, and the score balancing coefficient λ is set to 1.0.

4.3. Evaluation Metrics

We adopt the standard COCO-style evaluation metrics [45]: mAP₅₀ (mean Average Precision at IoU threshold 0.50), mAP₇₅ (at IoU threshold 0.75), and mAP_{50:95} (averaged over IoU thresholds from 0.50 to 0.95 with step 0.05). We also report per-category AP₅₀ for detailed analysis. In addition to accuracy metrics, we report the number of parameters (Params), floating-point operations (FLOPs), and inference speed (FPS) measured under identical hardware settings to comprehensively evaluate computational efficiency and practical deployability.

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

We compare CMAFNet with a comprehensive set of state-of-the-art methods spanning multiple categories, including two-stage detectors, single-stage detectors, YOLO-based detectors, Transformer-based detectors, and dedicated multi-modal fusion methods. All comparison methods are evaluated under the same experimental settings to ensure a fair assessment of both detection accuracy and computational complexity.

4.4.1. Results on M3FD Dataset

Table 1 presents the overall comparison results on the M3FD dataset. CMAFNet achieves the highest mAP₅₀ of 83.9% and the highest mAP_{50:95} of 52.3%, while maintaining competitive parameter size and FLOPs compared with other high-performing methods.

Table 1. Comparison with state-of-the-art methods on the M3FD dataset. The best results are shown in **bold** and the second-best results are underlined. “Two-Stage”, “Single-Stage”, “YOLO”, “Transformer”, and “Multi-Modal” indicate the method category.

Category	Method	Params (M)	FLOPs (G)	mAP ₅₀ (%)	mAP _{50:95} (%)
Two-Stage	Faster R-CNN [46]	41.1	134.2	68.3	38.7
	Cascade R-CNN [47]	69.2	189.5	71.6	41.2
	MBNet [15]	43.8	142.7	74.2	43.5
Single-Stage	SSD [48]	24.4	62.8	63.5	34.1
	RetinaNet [40]	36.3	97.1	70.8	40.3
	FCOS [49]	32.1	88.6	72.1	41.8
YOLO-based	YOLOv5-L [50]	46.5	109.1	74.8	43.2
	YOLOX-L [51]	54.2	155.6	75.3	44.1
	YOLOv7 [52]	36.9	104.7	76.5	45.3
	YOLOv8-L [53]	43.7	165.2	77.2	46.1
	YOLOv9-E [54]	57.3	189.0	78.4	47.2
Transformer-based	DETR [30]	41.3	86.0	71.5	40.8
	Deformable DETR [31]	40.0	78.3	75.6	44.7
	DINO [8]	47.6	98.5	79.3	48.6
	RT-DETR-L [32]	32.0	103.4	80.1	49.3
Multi-Modal	Halfway Fusion [5]	38.5	112.3	73.8	42.6
	CFT [6]	44.2	125.8	78.6	46.8
	ICAFusion [17]	48.7	138.4	80.5	48.9
	EAEFNet [19]	42.3	118.6	81.2	49.5
	TFDet [20]	45.1	131.2	82.4	50.8
CMAFNet (Ours)		48.5	132.6	83.9	52.3

Among the two-stage detectors, MBNet achieves 74.2% mAP₅₀ by addressing modality imbalance, but its two-stage pipeline typically introduces higher inference latency. Single-stage detectors such as FCOS achieve 72.1% mAP₅₀ with improved efficiency but comparatively lower accuracy. YOLO-based methods show progressive improvements, with YOLOv9-E reaching 78.4% mAP₅₀ at increased computational cost. Transformer-based detectors demonstrate strong performance, with RT-DETR-L achieving 80.1% mAP₅₀ while emphasizing real-time optimization. Among multi-modal methods, TFDet achieves 82.4% mAP₅₀ through target-aware fusion. CMAFNet surpasses TFDet by 1.5% in mAP₅₀ and 1.5% in mAP_{50:95}, indicating that the proposed cross-modal alignment and uncertainty-aware query selection provide consistent performance gains under comparable model complexity.

Table 2 presents the per-category AP₅₀ results on M3FD.

Table 2. Per-category AP₅₀ (%) comparison on the M3FD dataset.

Method	People	Car	Bus	Motorcycle	Lamp	Truck	mAP ₅₀
Faster R-CNN [46]	62.1	78.5	72.3	58.4	65.7	72.8	68.3
YOLOv8-L [53]	72.8	85.3	81.6	68.5	73.2	81.8	77.2
YOLOv9-E [54]	74.2	86.1	82.8	70.3	74.5	82.5	78.4
DINO [8]	75.8	87.2	83.5	71.6	75.3	82.4	79.3
RT-DETR-L [32]	76.5	87.8	84.2	72.4	76.1	83.6	80.1
CFT [6]	74.6	86.5	82.4	70.8	74.8	82.5	78.6
ICAFusion [17]	77.2	88.1	84.6	73.1	76.8	83.2	80.5
EAEFNet [19]	78.1	88.6	85.3	73.8	77.5	83.9	81.2
TFDet [20]	79.5	89.3	86.2	75.2	78.4	85.8	82.4
CMAFNet (Ours)	81.3	90.5	87.8	76.8	79.6	87.4	83.9

CMAFNet achieves the best performance across all six categories. Notably, for the challenging *People* category, CMAFNet achieves 81.3% AP₅₀, surpassing TFDet by 1.8%. This improvement is particularly important for safety-critical autonomous driving scenarios, where pedestrian detection reliability directly affects collision avoidance systems. The consistent gains across all categories indicate that the proposed cross-modal interaction mechanism generalizes well across object scales and environmental conditions.

To provide a more intuitive comparison, Figure 6 presents a radar chart of the per-category AP₅₀ for the top-performing methods on M3FD.

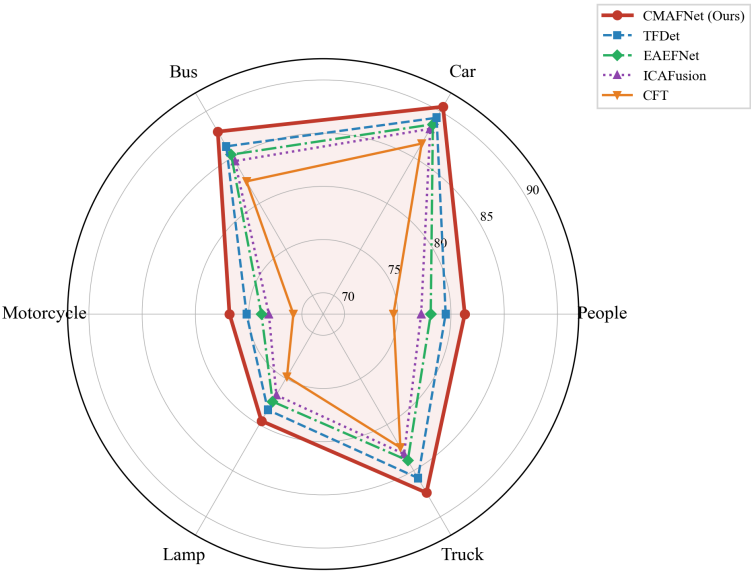


Figure 6. Radar chart comparing per-category AP₅₀ (%) of the top-performing methods on the M3FD dataset. CMAFNet (red solid line) consistently achieves superior balanced performance across all six object categories.

4.4.2. Results on FLIR-Aligned Dataset

Table 3 presents the comparison results on the FLIR-Aligned dataset. CMAFNet achieves the highest mAP₅₀ of 84.2% and the highest mAP_{50:95} of 53.1%, while maintaining competitive parameter size and FLOPs compared with other high-performing methods.

Table 3. Comparison with state-of-the-art methods on the FLIR-Aligned dataset.

Category	Method	Params (M)	FLOPs (G)	mAP ₅₀ (%)	mAP _{50:95} (%)
Two-Stage	Faster R-CNN [46]	41.1	134.2	70.5	39.8
	Cascade R-CNN [47]	69.2	189.5	73.2	42.5
	MBNet [15]	43.8	142.7	75.8	44.6
Single-Stage	SSD [48]	24.4	62.8	65.2	35.6
	RetinaNet [40]	36.3	97.1	72.4	41.5
	FCOS [49]	32.1	88.6	73.8	42.9
YOLO-based	YOLOv5-L [50]	46.5	109.1	76.2	44.8
	YOLOX-L [51]	54.2	155.6	76.8	45.3
	YOLOv7 [52]	36.9	104.7	78.1	46.5
	YOLOv8-L [53]	43.7	165.2	78.9	47.3
	YOLOv9-E [54]	57.3	189.0	79.6	48.1
Transformer-based	DETR [30]	41.3	86.0	73.1	42.2
	Deformable DETR [31]	40.0	78.3	77.2	46.1
	DINO [8]	47.6	98.5	80.5	49.8
	RT-DETR-L [32]	32.0	103.4	81.3	50.5
Multi-Modal	Halfway Fusion [5]	38.5	112.3	75.4	43.8
	CFT [6]	44.2	125.8	79.8	47.6
	ICAFusion [17]	48.7	138.4	81.6	49.7
	EAEFNet [19]	42.3	118.6	82.3	50.8
	TFDet [20]	45.1	131.2	83.1	51.6
CMAFNet (Ours)		48.5	132.6	84.2	53.1

Compared with TFDet, CMAFNet improves mAP₅₀ by 1.1% and mAP_{50:95} by 1.5%, indicating that the proposed cross-modal alignment and uncertainty-aware query selection mechanisms consistently enhance detection accuracy under comparable model complexity. The consistent performance gains across both M3FD and FLIR-Aligned benchmarks suggest that the framework generalizes well across datasets with different scene distributions and annotation characteristics.

Table 4 shows the per-category results on FLIR-Aligned.

Table 4. Per-category AP₅₀ (%) comparison on the FLIR-Aligned dataset.

Method	Person	Car	Bicycle	mAP ₅₀
Faster R-CNN [46]	65.8	80.2	65.5	70.5
YOLOv8-L [53]	75.6	87.4	73.7	78.9
YOLOv9-E [54]	76.3	88.1	74.4	79.6
DINO [8]	77.8	89.2	74.5	80.5
RT-DETR-L [32]	78.5	89.8	75.6	81.3
CFT [6]	76.5	88.3	74.6	79.8
ICAFusion [17]	78.8	89.6	76.4	81.6
EAEFNet [19]	79.5	90.2	77.2	82.3
TFDet [20]	80.6	91.1	77.6	83.1
CMAFNet (Ours)	82.1	92.3	78.2	84.2

CMAFNet achieves the best performance across all three categories. The improvement is particularly notable for the *Person* category (82.1% vs. 80.6%), which is directly related to pedestrian detection reliability in safety-critical autonomous driving scenarios. The gains in the *Car* and *Bicycle* categories further demonstrate that the proposed cross-modal interaction mechanism enhances detection consistency across object scales and thermal-appearance variations.

Figure 7 provides a cross-dataset comparison and component-wise ablation analysis.

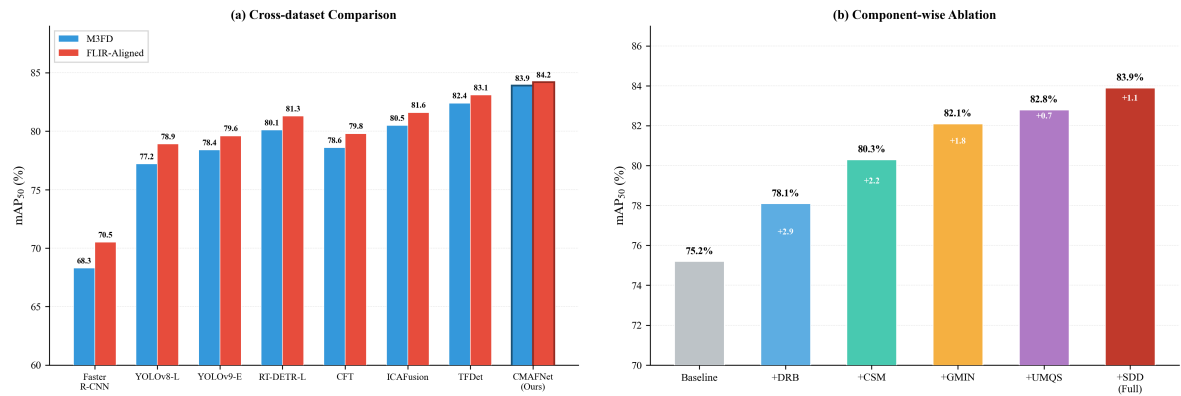


Figure 7. (a) Cross-dataset comparison of mAP₅₀ (%) on M3FD and FLIR-Aligned benchmarks. CMAFNet achieves the highest performance on both datasets. (b) Component-wise ablation on M3FD showing the progressive mAP₅₀ improvement as each module is added, with incremental gains of +2.9%, +2.2%, +1.8%, +0.7%, and +1.1%.

4.5. Ablation Studies

We conduct comprehensive ablation studies on the M3FD dataset to quantitatively analyze the contribution of each proposed component and to verify the rationality of the overall architectural design.

4.5.1. Component-wise Ablation

Table 5 presents the ablation results for each key component of CMAFNet.

Table 5. Component-wise ablation study on the M3FD dataset. “Baseline” uses a standard backbone with simple concatenation fusion and vanilla DETR decoder. DRB: Dynamic Receptive Backbone; CSM: Channel-Split Mamba Block; GMIN: Global Multi-modal Interaction Network; UMQS: Uncertainty-minimal Query Selection; SDD: Separable Dynamic Decoder.

DRB	CSM	GMIN	UMQS	SDD	mAP ₅₀ (%)	mAP _{50:95} (%)
					75.2	44.1
✓					78.1	46.5
✓	✓				80.3	48.2
✓	✓	✓			82.1	50.5
✓	✓	✓	✓		82.8	51.2
✓	✓	✓	✓	✓	83.9	52.3

The baseline model with simple concatenation fusion achieves 75.2% mAP₅₀. Replacing the standard backbone with DRB improves performance to 78.1% (+2.9%), indicating that adaptive receptive field modeling enhances multi-scale feature representation under multi-modal inputs.

Introducing the CSM-Block further increases performance to 80.3% (+2.2%), demonstrating the effectiveness of selective state space modeling for efficient long-range cross-modal dependency capture. Incorporating the GMIN module provides an additional 1.8% improvement (82.1%), confirming that explicit cross-modal interaction is more effective than implicit fusion alone.

The UMQS strategy contributes a further 0.7% gain (82.8%), suggesting that uncertainty-aware query initialization improves detection stability in multi-modal scenarios. Finally, replacing the vanilla DETR decoder with the Separable Dynamic Decoder yields the final 1.1% improvement (83.9%). The consistent and progressive gains across all components indicate that the overall performance improvement arises from complementary architectural design rather than isolated modifications.

The cross-dataset comparison and ablation results are jointly visualized in Figure 7.

4.5.2. Analysis of CSM-Block Design

Table 6 analyzes the impact of different design choices in the CSM-Block.

Table 6. Ablation study on CSM-Block design choices on the M3FD dataset. G: number of channel groups; “Attn. Recalib.”: attention-based recalibration.

Variant	G	Attn. Recalib.	mAP ₅₀ (%)	FPS
Standard Self-Attention	–	–	81.5	18.2
Standard Mamba (no split)	1	✓	81.8	28.5
CSM-Block (G = 2)	2	✓	82.6	30.1
CSM-Block (G = 4)	4	✓	83.9	29.3
CSM-Block (G = 8)	8	✓	83.2	29.8
CSM-Block (G = 4, no recalib.)	4	–	82.4	31.2

Compared with standard self-attention, the CSM-Block with $G = 4$ improves mAP₅₀ by 2.4% while significantly increasing inference speed (29.3 FPS vs. 18.2 FPS), highlighting the efficiency advantage of linear-complexity selective state space modeling over quadratic attention mechanisms.

The channel splitting strategy plays an important role. Using $G = 4$ groups outperforms the non-split variant ($G = 1$) by 2.1%, indicating that parallel group-wise modeling enhances representation diversity and cross-modal interaction. However, excessive splitting ($G = 8$) slightly reduces performance due to decreased per-group channel capacity, suggesting a trade-off between modeling granularity and feature richness.

The attention-based recalibration mechanism contributes an additional 1.5% improvement (83.9% vs. 82.4%), demonstrating that channel and spatial reweighting effectively refine the aggregated group-wise features.

4.5.3. Analysis of GMIN Design

Table 7 evaluates different fusion strategies in the GMIN module.

Table 7. Ablation study on GMIN design choices on the M3FD dataset.

Fusion Strategy	mAP ₅₀ (%)	mAP _{50:95} (%)
Concatenation only	80.3	48.2
Element-wise addition	80.8	48.6
Self-attention (intra-modal)	81.4	49.3
Cross-attention (standard)	82.2	50.1
GMIN (GAP only)	82.8	50.6
GMIN (GMP only)	82.5	50.3
GMIN (GAP + GMP, w/o gating)	83.1	51.4
GMIN (Full)	83.9	52.3

The full GMIN module, which integrates GAP- and GMP-guided cross-attention with modality-specific gating, achieves the best performance. Compared with simple fusion strategies such as concatenation and element-wise addition, explicit cross-modal interaction consistently improves detection accuracy, indicating that structured feature exchange is more effective than implicit blending.

Using only GAP or only GMP leads to 1.1% and 1.4% lower mAP₅₀, respectively, suggesting that average statistical responses and salient feature responses provide complementary cues for cross-modal alignment. The modality-specific gating mechanism further contributes 0.8% improvement over the variant without gating, demonstrating that adaptive control of modality contribution enhances robustness across object categories and illumination conditions.

Figure 8 jointly visualizes the fusion strategy comparison and the β sensitivity analysis. As shown in Figure 8(a), the full GMIN design consistently outperforms simplified variants across both mAP₅₀ and mAP_{50:95} metrics. Standard cross-attention yields moderate improvement, while the dual-pooling guided cross-attention design further enhances discriminative feature transfer between modalities.

Figure 8(b) reveals that the uncertainty penalty coefficient β in UMQS exhibits a clear inverted-U sensitivity pattern across all three evaluation metrics. When β is too small, uncertain candidates

are insufficiently penalized, leading to unstable query initialization. Conversely, excessively large β overly suppresses candidate diversity and may discard informative proposals. The optimal balance is achieved at $\beta^* = 0.5$, where the model attains peak performance of 83.9% mAP₅₀, 64.5% mAP₇₅, and 52.3% mAP_{50:95}.

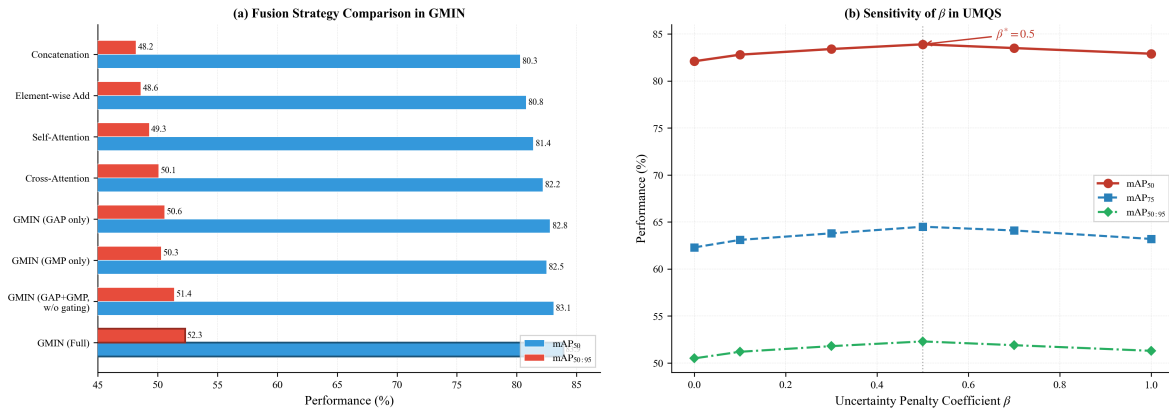


Figure 8. (a) Horizontal bar chart comparing different fusion strategies in the GMIN module on M3FD. The full GMIN design with both GAP and GMP guided cross-attention and modality-specific gating achieves the best performance. (b) Sensitivity analysis of the uncertainty penalty coefficient β in UMQS. All three metrics exhibit a consistent inverted-U pattern, with the optimal performance at $\beta^* = 0.5$.

4.5.4. Analysis of Uncertainty-Minimal Query Selection

Table 8 evaluates the UMQS strategy with different uncertainty penalty coefficients β .

Table 8. Ablation study on the uncertainty penalty coefficient β in UMQS on the M3FD dataset.

β	mAP ₅₀ (%)	mAP ₇₅ (%)	mAP _{50:95} (%)
0 (no uncertainty)	82.1	62.3	50.5
0.1	82.8	63.1	51.2
0.3	83.4	63.8	51.8
0.5	83.9	64.5	52.3
0.7	83.5	64.1	51.9
1.0	82.9	63.2	51.3

The UMQS strategy consistently improves performance over the baseline without uncertainty consideration ($\beta = 0$). Introducing moderate uncertainty penalization improves query reliability and stabilizes detection performance. The optimal $\beta = 0.5$ achieves 83.9% mAP₅₀, representing a 1.8% improvement over the no-uncertainty variant. Both overly small and overly large penalty coefficients result in reduced performance, indicating that balanced uncertainty control is essential for effective query selection.

The detailed β sensitivity results are jointly presented in Figure 8(b).

4.6. Efficiency Analysis

Table 9 compares the computational efficiency of CMAFNet with representative multi-modal detection methods.

Table 9. Computational efficiency comparison. FPS is measured on a single NVIDIA A100 GPU with input size 640×640 .

Method	Params (M)	FLOPs (G)	FPS	mAP ₅₀ (%)
CFT [6]	44.2	125.8	22.5	78.6
ICAFusion [17]	48.7	138.4	19.8	80.5
EAEFNet [19]	42.3	118.6	24.1	81.2
TFDet [20]	45.1	131.2	21.3	82.4
CMAFNet (Ours)	48.5	132.6	29.3	83.9

CMAFNet achieves the highest mAP₅₀ while maintaining the fastest inference speed (29.3 FPS) among the compared multi-modal methods. This efficiency advantage is primarily attributed to the linear-complexity CSM-Block and the dynamic convolution attention mechanism in the Separable Dynamic Decoder, which reduce the quadratic cost of conventional attention.

Compared with ICAFusion, CMAFNet is substantially faster while also achieving higher detection accuracy, demonstrating a favorable accuracy–efficiency trade-off.

Figure 9 provides a comprehensive visualization of the accuracy–speed trade-off and robustness under different environmental conditions.

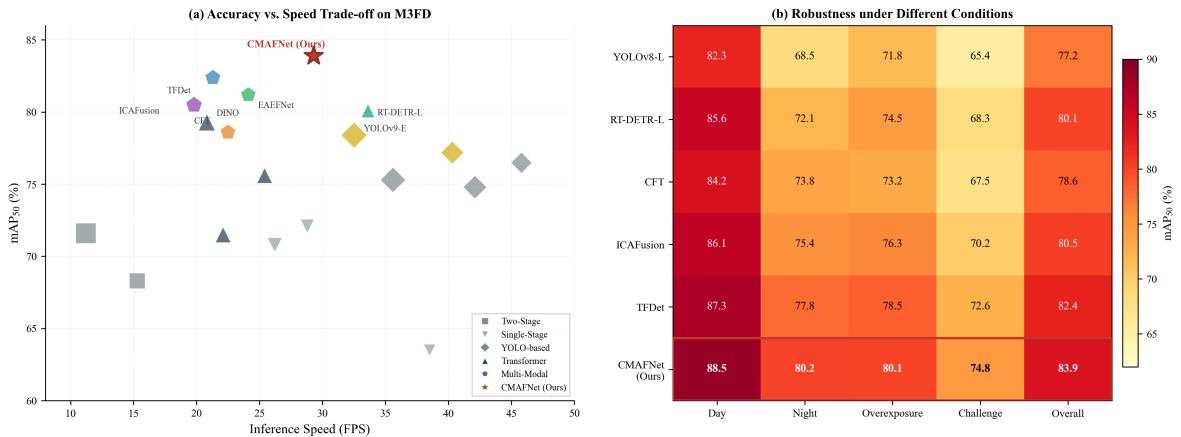


Figure 9. (a) Accuracy vs. speed trade-off on the M3FD dataset. Each point represents a detection method, with bubble size proportional to the number of parameters. CMAFNet achieves a strong balance between accuracy and inference speed. (b) Heatmap visualization of mAP₅₀ (%) under different environmental conditions. CMAFNet achieves consistently high performance across all conditions, including nighttime and challenging weather scenarios.

4.7. Qualitative Results

We present qualitative detection results to visually demonstrate the effectiveness of CMAFNet under various challenging real-world conditions.

4.7.1. Results on M3FD Dataset

Figures 10 and 11 show the detection results of CMAFNet on RGB and IR images from the M3FD dataset, respectively.



Figure 10. Detection results of CMAFNet on RGB images from the M3FD dataset. The results demonstrate robust detection across diverse scenarios including daytime, nighttime, rainy, and foggy conditions. CMAFNet accurately detects multiple object categories including People, Car, Motorcycle, Bus, and Truck with high confidence scores.

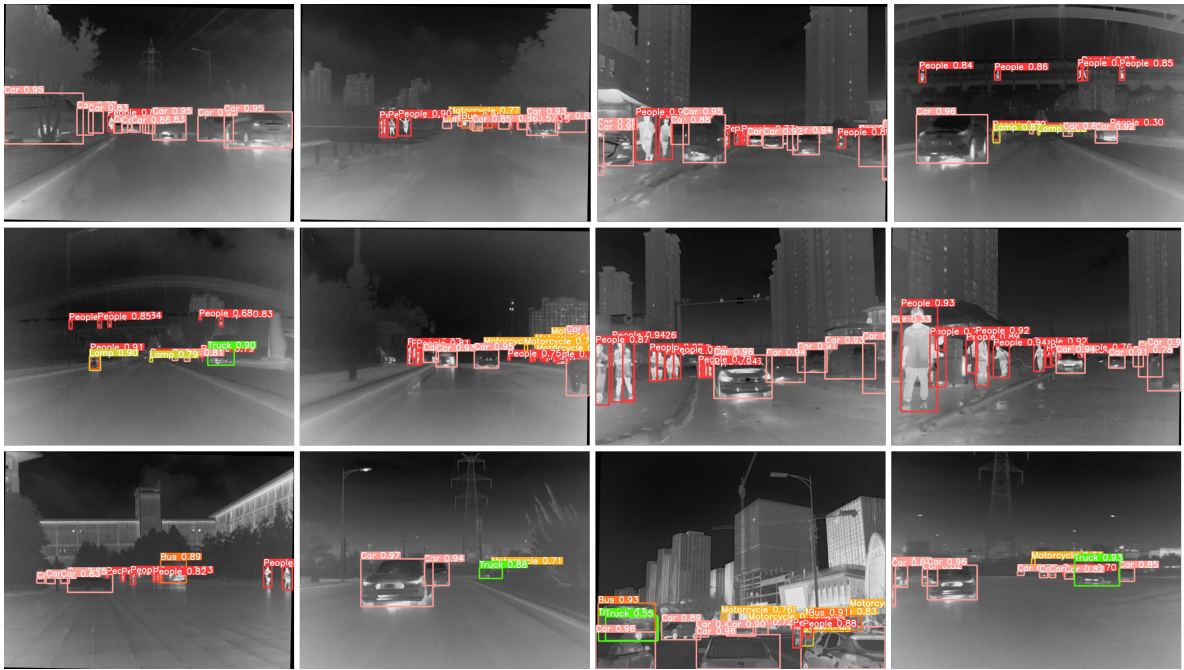


Figure 11. Detection results of CMAFNet on IR images from the M3FD dataset. The infrared modality provides clear thermal signatures of objects even in challenging visibility conditions. CMAFNet effectively leverages the complementary thermal information to maintain high detection accuracy.

As illustrated in Figure 10, CMAFNet maintains stable detection performance across diverse driving scenarios, including strong illumination, nighttime scenes, rain, and fog. The model accurately detects objects of varying scales and densities, from small distant pedestrians to large nearby vehicles, demonstrating robust multi-scale representation capability.

Figure 11 further shows that the infrared modality provides reliable thermal signatures even when RGB visibility is degraded. CMAFNet effectively integrates complementary thermal cues, leading to

accurate detection in low-contrast and low-illumination environments. These results visually confirm the advantage of cross-modal interaction under adverse conditions.

4.7.2. Results on FLIR-Aligned Dataset

Figures 12 and 13 present the detection results on the FLIR-Aligned dataset.



Figure 12. Detection results of CMAFNet on RGB images from the FLIR-Aligned dataset. The model demonstrates accurate detection of Person, Car, and Bicycle categories across various urban driving scenarios with different lighting conditions and traffic densities.



Figure 13. Detection results of CMAFNet on IR images from the FLIR-Aligned dataset. The thermal infrared images clearly reveal pedestrians and vehicles through their heat signatures, enabling reliable detection in nighttime and low-visibility conditions.

On the FLIR-Aligned dataset, CMAFNet demonstrates consistent detection quality across both modalities. As shown in Figure 12, the model accurately detects pedestrians, vehicles, and bicycles under varying lighting and traffic conditions. Figure 13 highlights the effectiveness of thermal cues for pedestrian detection, particularly in nighttime scenes where RGB information is limited. The qualitative results align with the quantitative improvements reported in Section 4.4, confirming that the proposed cross-modal fusion strategy enhances both robustness and discriminative capability.

4.8. Robustness Analysis under Different Conditions

To further evaluate robustness, we analyze the performance of CMAFNet under different environmental conditions on the M3FD dataset, which provides scenario-level annotations.

As shown in Table 10, CMAFNet achieves the best performance across all environmental conditions. The gains are particularly notable under challenging scenarios. Compared with TFDet, CMAFNet improves performance by 2.4% in nighttime conditions and 2.2% in challenge scenarios (rain, fog, snow). These improvements indicate that the proposed cross-modal interaction and uncertainty-aware query selection mechanisms effectively enhance robustness when one modality is partially degraded.

Table 10. Performance comparison under different environmental conditions on the M3FD dataset (mAP₅₀, %).

Method	Day	Night	Overexposure	Challenge	Overall
YOLOv8-L [53]	82.3	68.5	71.8	65.4	77.2
RT-DETR-L [32]	85.6	72.1	74.5	68.3	80.1
CFT [6]	84.2	73.8	73.2	67.5	78.6
ICAFusion [17]	86.1	75.4	76.3	70.2	80.5
TFDet [20]	87.3	77.8	78.5	72.6	82.4
CMAFNet (Ours)	88.5	80.2	80.1	74.8	83.9

The robustness heatmap is jointly presented in Figure 9(b), where CMAFNet consistently achieves strong performance across different environmental categories, further confirming its stability and generalization capability.

5. Discussion

The experimental results demonstrate the effectiveness of CMAFNet for multi-modal object detection in autonomous driving. Beyond the quantitative improvements, several structural insights can be drawn from the proposed design.

Effectiveness of State Space Models for Cross-Modal Fusion. The CSM-Block demonstrates that selective state space models provide an efficient alternative to quadratic-complexity attention mechanisms for modeling long-range cross-modal dependencies. As shown in Table 6, the CSM-Block achieves 2.4% higher mAP₅₀ than standard self-attention while delivering 61% faster inference. This performance gain is not merely computational but structural: the selective scanning mechanism enables adaptive information propagation across spatial positions, while the channel-splitting strategy promotes representation diversity and reduces feature interference. These results suggest that state space modeling is particularly well-suited for multi-modal fusion, where efficient global context modeling is critical.

Importance of Fine-Grained Cross-Modal Interaction. The GMIN module addresses a common limitation in existing fusion strategies—namely, insufficient explicit alignment between modalities. The ablation study (Table 7) confirms that simple fusion operations such as concatenation or element-wise addition are inadequate for capturing complex complementary relationships between RGB and IR features. By integrating GAP- and GMP-guided cross-attention, GMIN captures both global statistical responses and salient discriminative cues. The modality-specific gating mechanism further stabilizes fusion by dynamically regulating the contribution of each modality. This structured interaction is particularly beneficial under distribution shifts such as illumination variation.

Robustness under Adverse Conditions. The robustness analysis (Table 10) shows that CMAFNet exhibits the largest performance gains under nighttime and challenging weather conditions. This behavior reflects the adaptive nature of the proposed fusion strategy. When one modality is degraded, the gating mechanism increases reliance on the complementary modality, effectively mitigating information loss. Such adaptive cross-modal balancing is essential for safety-critical autonomous driving systems operating under unpredictable environmental conditions.

Efficiency–Accuracy Trade-off. Although CMAFNet introduces additional fusion modules, it maintains competitive real-time performance (29.3 FPS) as shown in Table 9. This efficiency is largely attributed to the linear-complexity state space modeling in the CSM-Block and the dynamic convolution attention in the Separable Dynamic Decoder [38]. The results indicate that improved cross-modal interaction does not necessarily require heavy attention-based architectures; carefully designed linear mechanisms can achieve a favorable accuracy–efficiency balance.

Limitations and Future Work. Despite achieving state-of-the-art results on RGB–IR benchmarks, several limitations remain. First, the current framework focuses on dual-modality fusion; extending CMAFNet to incorporate additional sensing modalities such as LiDAR or radar may further enhance robustness. Second, the method assumes well-aligned RGB–IR pairs; developing alignment-robust or alignment-free fusion strategies is an important practical direction. Third, while the current model achieves real-time inference on high-end GPUs, further model compression and architectural simplification would be valuable for deployment on edge platforms in resource-constrained autonomous vehicles.

6. Conclusions

In this paper, we proposed CMAFNet, a Cross-Modal Alignment and Fusion Network for robust multi-modal object detection in autonomous driving. The proposed framework integrates three key technical contributions: (1) the Channel-Split Mamba Block, which leverages selective state space

modeling to efficiently capture long-range cross-modal dependencies with linear complexity; (2) the Global Multi-modal Interaction Network, which performs fine-grained cross-modal alignment via dual-branch attention guided by global pooling; and (3) the Uncertainty-minimal Query Selection strategy, which improves detection reliability by selecting object queries based on cross-modal prediction consistency.

Combined with a Dynamic Receptive Backbone and a Separable Dynamic Decoder, CMAFNet achieves state-of-the-art performance on both the M3FD (83.9% mAP₅₀) and FLIR-Aligned (84.2% mAP₅₀) benchmarks while maintaining competitive real-time efficiency. Extensive ablation studies validate the effectiveness and complementarity of each component, and robustness analysis demonstrates consistent improvements under adverse environmental conditions. These results suggest that structured cross-modal alignment, efficient state space modeling, and uncertainty-aware query selection together form a promising direction for reliable multi-modal perception in autonomous driving.

Author Contributions: Conceptualization, Z.-H.H. and M.-J.-S.W.; methodology, Z.-H.H. and C.-W.L.; software, Z.-H.H.; validation, Z.-H.H., C.-W.L. and M.-J.-S.W.; formal analysis, Z.-H.H. and C.-W.L.; investigation, Z.-H.H.; resources, M.-J.-S.W.; data curation, Z.-H.H.; writing—original draft preparation, Z.-H.H.; writing—review and editing, C.-W.L. and M.-J.-S.W.; visualization, Z.-H.H.; supervision, M.-J.-S.W.; project administration, M.-J.-S.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The M3FD dataset is publicly available at <https://github.com/JinyuanLiu-CV/TarDAL> (accessed on 3 March 2026). The FLIR ADAS dataset is publicly available at <https://www.flir.com/oem/adas/adas-dataset-form/> (accessed on 3 March 2026). All other data supporting the findings of this study are included within the article.

Acknowledgments: The authors would like to thank the anonymous reviewers for their constructive comments and valuable suggestions, which helped improve the clarity and quality of this manuscript.

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

Abbreviations

The following abbreviations are used in this manuscript:

Abbreviation	Definition
CMAFNet	Cross-Modal Alignment and Fusion Network
DRB	Dynamic Receptive Backbone
CSM-Block	Channel-Split Mamba Block
GMIN	Global Multi-modal Interaction Network
UMQS	Uncertainty-Minimal Query Selection
SDD	Separable Dynamic Decoder
SSM	State Space Model
IR	Infrared
mAP	Mean Average Precision
FPS	Frames Per Second
FLOPs	Floating-Point Operations
GAP	Global Average Pooling
GMP	Global Max Pooling
DETR	Detection Transformer

References

1. Zhang, Y.; Ding, W.; Xu, Z.; Zhao, D. Multimodal object detection in autonomous driving: A review. *Neurocomputing* **2024**, *594*, 127885.
2. Hwang, S.; Park, J.; Kim, N.; Choi, Y.; Kweon, I.S. Multispectral pedestrian detection: Benchmark dataset and baseline. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* **2015**, pp. 1037–1045.
3. Teledyne FLIR. Free FLIR Thermal Dataset for Algorithm Training. <https://www.flir.com/oem/adas/adas-dataset-form/>, 2021.
4. Liu, J.; Fan, X.; Huang, Z.; Wu, G.; Liu, R.; Zhong, W.; Luo, Z. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* **2022**, pp. 5802–5811.
5. Liu, J.; Zhang, S.; Wang, S.; Metaxas, D.N. Multispectral deep neural networks for pedestrian detection. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2016, pp. 73.1–73.13.
6. Qingyun, F.; Zhaokui, W. Cross-modality fusion transformer for multispectral object detection. *arXiv preprint arXiv:2111.00273* **2022**.
7. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.
8. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.Y. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.
9. Wang, M.; Yang, W.; Guo, Y.; Wang, S. Conditional fault tolerance in a class of Cayley graphs. *International Journal of Computer Mathematics* **2016**, *93*, 67–82.
10. Wang, S.; Wang, Z.; Wang, M.; Han, W. g-Good-neighbor conditional diagnosability of star graph networks under PMC model and MM* model. *Frontiers of Mathematics in China* **2017**, *12*, 1221–1234.
11. Wang, M.; Ren, Y.; Lin, Y.; Wang, S. The tightly super 3-extra connectivity and diagnosability of locally twisted cubes. *American Journal of Computational Mathematics* **2017**, *7*, 127–144.
12. Wang, M.; Lin, Y.; Wang, S. The connectivity and nature diagnosability of expanded k -ary n -cubes. *RAIRO-Theoretical Informatics and Applications-Informatique Théorique et Applications* **2017**, *51*, 71–89.
13. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752* **2024**.
14. Zhang, H.; Fromont, E.; Lefevre, S.; Avignon, B. Guided attentive feature fusion for multispectral pedestrian detection. *Pattern Recognition* **2021**, *118*, 108038.
15. Zhou, K.; Chen, L.; Cao, X. Improving multispectral pedestrian detection by addressing modality imbalance problems. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 787–803.
16. Zhang, L.; Liu, Z.; Zhang, S.; Yang, X.; Qiao, H.; Huang, K.; Hussain, A. Cross-modality interactive attention network for multispectral pedestrian detection. In *Proceedings of the Information Fusion*. Elsevier, 2019, Vol. 50, pp. 20–29.
17. Tang, S.; Yuan, Z.; Li, X. ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognition* **2024**, *145*, 109913.
18. Jung, H.; Yoo, J.; Kweon, I.S. BAANet: Learning bi-directional adaptive attention gates for multispectral pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems* **2024**, *25*, 2580–2592.
19. Liang, M.; Hu, J.; Bao, C.; Feng, H.; Deng, F.; Lam, T.L. Explicit attention-enhanced fusion for RGB-thermal perception tasks. *IEEE Robotics and Automation Letters* **2023**, *8*, 4060–4067.
20. Chen, X.; Luo, B.; Guo, D.; Du, B. TFDet: Target-aware fusion for RGB-thermal pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems* **2024**, *25*, 7038–7050.
21. Wang, S.; Wang, Y.; Wang, M. Connectivity and matching preclusion for leaf-sort graphs. *Journal of Interconnection Networks* **2019**, *19*, 1940007.
22. Wang, M.; Yang, W.; Wang, S. Conditional matching preclusion number for the Cayley graph on the symmetric group. *Acta Math. Appl. Sin.(Chinese Series)* **2013**, *36*, 813–820.
23. Wang, S.; Wang, M. The strong connectivity of bubble-sort star graphs. *The Computer Journal* **2019**, *62*, 715–729.
24. Wang, S.; Wang, M. A Note on the Connectivity of m -Ary n -Dimensional Hypercubes. *Parallel Processing Letters* **2019**, *29*, 1950017.
25. Gu, A.; Goel, K.; Ré, C. Efficiently modeling long sequences with structured state spaces. *International Conference on Learning Representations (ICLR)* **2022**.

26. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision Mamba: Efficient visual representation learning with bidirectional state space model. *International Conference on Machine Learning (ICML)* **2024**.
27. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Liu, Y. VMamba: Visual state space model. *Advances in Neural Information Processing Systems (NeurIPS)* **2024**.
28. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7132–7141.
29. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV). Springer, 2018, pp. 3–19.
30. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision (ECCV). Springer, 2020, pp. 213–229.
31. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable transformers for end-to-end object detection. In Proceedings of the International Conference on Learning Representations (ICLR), 2021.
32. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs beat YOLOs on real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16965–16974.
33. Zong, Z.; Song, G.; Liu, Y. DETRs with collaborative hybrid assignments training. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6748–6758.
34. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1251–1258.
35. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning (ICML), 2015, pp. 448–456.
36. Elfving, S.; Uchibe, E.; Doya, K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks* **2018**, *107*, 3–11.
37. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv preprint arXiv:1607.06450* **2016**.
38. Hu, J.; Cao, L.; Jin, X.; Zhang, S.; Ji, R. Universal Image Segmentation with Efficiency. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2025**.
39. Kuhn, H.W. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly* **1955**, *2*, 83–97.
40. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988.
41. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 658–666.
42. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. In Proceedings of the International Conference on Learning Representations (ICLR), 2019.
43. Loshchilov, I.; Hutter, F. SGDR: Stochastic gradient descent with warm restarts. In Proceedings of the International Conference on Learning Representations (ICLR), 2017.
44. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *Journal of Big Data* **2019**, *6*, 1–48.
45. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In Proceedings of the European Conference on Computer Vision (ECCV). Springer, 2014, pp. 740–755.
46. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2017**, *39*, 1137–1149.
47. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into high quality object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6154–6162.
48. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision (ECCV). Springer, 2016, pp. 21–37.
49. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9627–9636.
50. Jocher, G.; et al. YOLOv5 by Ultralytics. <https://github.com/ultralytics/yolov5>, 2022.
51. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO series in 2021. *arXiv preprint arXiv:2107.08430* **2021**.

52. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7464–7475.
53. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLOv8. <https://github.com/ultralytics/ultralytics>, 2023.
54. Wang, C.Y.; Liao, H.Y.M. YOLOv9: Learning what you want to learn using programmable gradient information. *European Conference on Computer Vision (ECCV)* **2024**, pp. 1–21.

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.