LiteFusion: A Lightweight Multimodal Feature Fusion Model for Adverse Driving Conditions

1st Zhicheng Qian
FuYang Normal University
FuYang, China
3278028423@qq.com

2nd Hongzhi Zhou*
FuYang Normal University
FuYang, China
40696670@qq.com

3rd Nina Ling
FuYang Normal University
FuYang, China
1956393642@qq.com

***Abstract*—With the rapid advancement of intelligent driving, object detection technology faces robustness challenges in complex environments, particularly the loss of RGB image information caused by low-light and adverse weather conditions. To address this, we propose the LiteFusion model—a novel lightweight multimodal object detection framework based on mid-term feature fusion. It effectively combines visible and infrared modalities to enhance detection robustness in complex driving scenarios. First, the model introduces the lightweight LFNet architecture, employing a dual-branch heterogeneous and hierarchical gated fusion strategy to achieve refined integration of visible and infrared features. To further improve localization accuracy, the WIoU v3 loss is integrated into the detection head. Its dynamic non-monotonic focusing mechanism emphasizes high-quality samples, enhancing bounding box precision and convergence speed, thereby boosting detection robustness. Finally, the hybrid convolution module C3k2_SCConv is introduced. By embedding spatial and channel reconstruction convolutions into the bottleneck structure, it enables adaptive re-calibration of features across both spatial and channel dimensions, mutually enhancing convolution efficiency and fusion performance. Experimental results demonstrate that LiteFusion achieves lightweight detection while improving performance, validating its practical value in intelligent driving scenarios.**

***Index Terms*—Multimodal Object Detection, Intelligent Driving, LiteFusion, Loss Function Optimization, Spatial-Channel Reconstruction Convolution**

I. Introduction

In recent years, with the rapid advancement of artificial intelligence and intelligent transportation technologies [1], the automotive industry has long embarked on a transformation toward intelligent systems [2]. Intelligent driving technology has emerged as the core driving force behind modern transportation innovation [3]. According to statistics from relevant authorities, the majority of traffic accidents are caused by human driving errors, posing a serious threat to public safety [4]. By integrating perception, decision-making, and execution into a closed-loop paradigm [5] [6], intelligent driving technology can effectively reduce such human driving errors, thereby decreasing traffic accidents and enhancing overall transportation efficiency. As a core component of intelligent driving technology, object detection technology plays a crucial role in monitoring surrounding road conditions and providing reliable environmental information for the decision-making module [7].

Despite significant progress in recent years, object detection still faces major challenges in complex driving environments. Experiments on public test datasets show that detection performance at night consistently lags behind daytime scenarios. This degradation primarily stems from the heavy reliance of most mainstream detection models on visible spectrum images [8] [9]. While these models perform well under well-lit conditions, their detection capabilities severely decline in low-light environments and adverse weather conditions such as rain, snow, and fog. In adverse driving conditions, visible light image quality suffers severe degradation, leading to reduced detection accuracy and increased risks of missed detections and false positives [10] [11]. This significantly limits the practical application of detection models in intelligent driving scenarios. Infrared images, however, remain unaffected by lighting variations and harsh weather conditions, enabling them to provide high-quality image information even in low-light, nighttime, or inclement weather environments [12]. However, infrared images typically exhibit low spatial resolution, limited texture information, and sensitivity to thermal noise, making them insufficient for standalone object detection [13]. Consequently, relying solely on either visible light or infrared imagery proves inadequate for achieving reliable object detection in complex driving environments [14].

To overcome the limitations of single-modal approaches, multimodal object detection has garnered increasing attention in recent years [15]. Multimodal detection methods enhance performance in complex environments by leveraging the complementary characteristics of visible light and infrared modalities. Among various multimodal strategies, feature-level fusion has emerged as an effective paradigm for capturing multimodal correlations and improving representational capabilities [16]. However, most current multimodal fusion approaches incur substantial computational overhead, posing significant challenges for deployment and application in resource-constrained intelligent driving scenarios [17]. Thus, achieving a balance between detection performance and computational efficiency remains a critical issue.

To address these challenges, we propose LiteFusion (Lightweight Fusion), a lightweight multimodal feature fusion object detection model tailored for intelligent driving scenarios. The main contributions of this study are summarized as follows:

1) We introduce LFNet (Lightweight Fusion Network), a lightweight mid-level feature fusion framework employing a dual-branch heterogeneous architecture. By leveraging the complementary characteristics of visible light and infrared modalities, it enhances detection robustness under low-light and adverse weather conditions.

2) We introduce the WIoU v3 loss to improve bounding box regression accuracy in multimodal detection tasks, mitigating edge blurring and noise interference caused by feature inconsistencies across modalities.

3) We propose the hybrid convolution module C3k2_SCConv, designed to enhance spatial and channel feature representations while maintaining low computational complexity, achieving a favorable balance between detection performance and model efficiency.

II. Related work

To enhance robustness in low-light and adverse weather conditions, current research primarily leverages the complementary characteristics of visible and infrared images to achieve multimodal object detection. Zhang et al. [18] proposed a dual-stream architecture-based Guided Attention Feature Fusion Network, employing a modality-aware attention mechanism to adaptively integrate RGB and infrared features. Their approach significantly improved pedestrian detection performance in low-light conditions. However, the use of multi-level attention modules and heavyweight backbones substantially increased computational complexity, limiting its applicability in autonomous driving scenarios. Jiang et al. [19] introduced M2FNet, a multimodal fusion framework combining convolutional feature extractors with cross-attention mechanisms to simulate modal interactions between visible light and thermal imaging. Experimental results demonstrate M2FNet consistently outperforms single-stage detectors across multiple public datasets. However, its extensive reliance on transformer-based attention blocks incurs substantial computational overhead, limiting deployment on resource-constrained platforms. Guo et al. [20] introduced DAMSDet, a dynamic adaptive multispectral detection transformer that combines competitive query selection with deformable cross-attention to enhance cross-modal feature aggregation. While DAMSDet improves detection accuracy and robustness against modal misalignment, its transformer-based architecture incurs inference latency and increased model complexity, rendering it unsuitable for lightweight intelligent driving applications. Wu et al. [21] proposed FreDFT, which fuses visible and infrared features in the frequency domain through a transformer-based frequency attention mechanism. By leveraging the complementary spectral characteristics of different modalities, FreDFT achieves strong detection performance under low-light conditions. However, the additional frequency-domain attention further increases computational cost, posing challenges for real-time deployment.

Feature-level fusion has become the mainstream paradigm for RGB and infrared object detection, as it enables fine-grained interactions between modality-specific representations while preserving spatial semantics. Shen et al. [22] proposed ICAFusion, an iterative cross-attention-guided feature fusion network that progressively refines multimodal representations through multiple interactive stages. Their approach demonstrates enhanced discriminative power in complex scenes, but the iterative attention mechanism requires repeated feature transformations, increasing network depth and inference latency. Gao et al. [23] developed a dual-branch detection network employing multi-scale attention-guided fusion to achieve adaptive weighting of modal features. While this approach enhances detection robustness in low-light conditions, its heavy reliance on repeated cross-feature-scale attention incurs substantial computational overhead. These studies indicate that although feature-level fusion effectively enhances multimodal expressiveness, it introduces complexity costs that hinder practical implementation in intelligent driving scenarios.

To address this limitation, research efforts have focused on lightweight multimodal detection frameworks that maintain robust detection performance while reducing computational costs. Hou et al. [24] designed a dual-modal detection framework based on convolutional feature fusion, enhancing inference efficiency by eliminating transformer-based modules. Although this strategy lowers computational complexity, its relatively shallow fusion mechanism restricts multimodal interactions, particularly for small or heavily occluded targets. Consequently, existing lightweight approaches face detection performance challenges in low-light and adverse weather conditions.

In summary, most existing methods rely on computationally expensive attention mechanisms or Transformer-based fusion schemes, limiting their applicability in resource-constrained intelligent driving scenarios. Therefore, designing an efficient multimodal detection framework that balances lightweight computation with effective feature fusion remains an urgent challenge. To address this challenge, we propose the LiteFusion model. By adopting an intermediate fusion strategy and lightweight convolutional modules, it achieves remarkable detection performance in complex driving environments.

III. Proposed model

A. Network Architecture of LiteFusion

The overall architecture of our proposed LiteFusion framework is shown in Fig. 1. LiteFusion is a lightweight RGB-infrared multimodal detection model specifically designed for resource-constrained scenarios such as intelligent driving. The framework adopts a dual-branch heterogeneous architecture, extracting specific modal features from visible and infrared images respectively. To balance detection accuracy and computational efficiency, LiteFusion employs an intermediate feature fusion strategy. This strategy integrates enhanced infrared features with visible features through a lightweight channel gating mechanism. The fused features undergo refinement via an efficient neck network before being fed into the detection head to produce the final detection results.

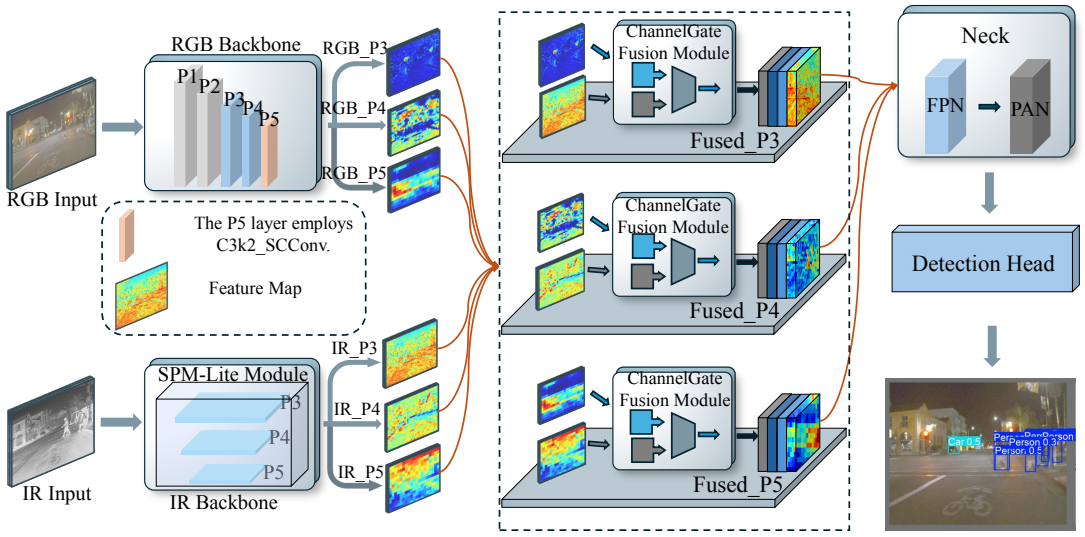


Fig. 1. Network Framework Diagram of the LiteFusion Model, where SPM-Lite is a Lightweight Spatial Prior Module, not a Complete IR Backbone.

B. Asymmetric Backbone Design with Lightweight Infrared Encoding

To achieve an effective balance between detection accuracy and computational efficiency in multimodal object detection, LFNet employs an asymmetric backbone architecture where RGB and infrared modalities are processed through structurally distinct feature extraction pathways. This design leverages the fundamentally different imaging characteristics of the two modalities, aiming to reduce redundant computations while preserving complementary information.

For the RGB modality, LFNet employs a hierarchical convolutional feature extractor to encode rich appearance information, including texture, edge, and color details. The RGB branch performs progressive spatial downsampling, generating multi-scale feature representations across three semantic levels P3, P4, and P5, widely applicable for detecting objects at varying spatial scales. Designed for compactness and stability, the RGB backbone ensures reliable feature extraction and stable optimization behavior while enabling the overall network to focus on enhancing multimodal representation and fusion efficiency.

Infrared images primarily convey thermal radiation information, providing strong target saliency under low-light and adverse weather conditions. However, infrared images typically suffer from insufficient texture detail and low spatial resolution, rendering them unsuitable for deep hierarchical feature extraction using traditional symmetric backbone architectures. To address this issue, LFNet adopts a lightweight infrared feature encoding strategy that emphasizes efficient aggregation of thermal prior information rather than dense texture modeling. This strategy is realized through a dedicated infrared feature extraction pathway centered on the proposed SPM-Lite (Spatial Prior Module-Lite).

SPM-Lite generates multi-scale infrared feature representations through a single forward pass, thereby circumventing the pyramid structure commonly found in traditional multi-scale encoders. As shown in Fig. 2, SPM-Lite directly outputs feature maps at scales P3, P4, and P5, enabling efficient alignment with the corresponding RGB feature hierarchy. For an input infrared image $I_{IR} \in \mathbb{R}^{B\times3\times H\times W}$, SPM-Lite first applies a compact Stem module [25]: this module consists of consecutive 3×3 convolutional layers followed by spatial downsampling, efficiently extracting low-level thermal responses while reducing spatial redundancy. A progressive downsampling strategy is then employed to construct high-level semantic representations, enabling SPM-Lite to capture scale-sensitive thermal information with minimal parameter overhead. Unlike traditional infrared encoders that rely on deep convolutional stacks to model fine textures, SPM-Lite emphasizes preserving spatial prior information and enhancing target saliency, aligning more closely with the inherent characteristics of infrared images. By aggregating multi-scale thermal information through lightweight methods, SPM-Lite ensures target recognizability even under low-light conditions.

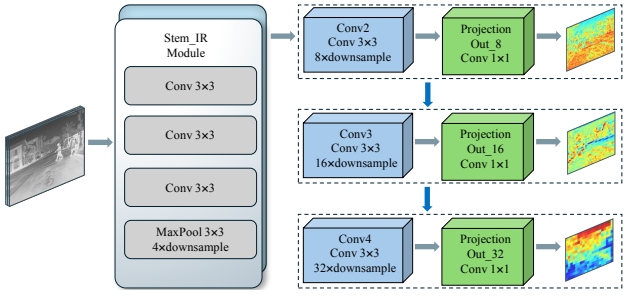


Fig. 2. Structure Diagram of the SPM-Lite Module, Adopting Progressive Downsampling.

From a computational perspective, SPM-Lite significantly reduces the number of parameters and computational com-

plexity by integrating multi-scale feature extraction into a unified module. Through its asymmetric backbone network design and the introduction of SPM-Lite, LFNet achieves an effective balance between feature representation capability and computational efficiency. The RGB branch employs a compact hierarchical encoder to preserve fine-grained appearance information, while the infrared branch focuses on lightweight extraction of thermal prior information. This complementary design lays a solid foundation for the LiteFusion model.

C. ChannelGate: Hierarchical Channel-wise Multimodal Feature Fusion

Although asymmetric backbone designs can efficiently extract features from RGB and infrared modalities, effectively fusing these multimodal features is crucial to fully leverage their complementary characteristics. Therefore, LFNet introduces the ChannelGate module, enabling adaptive channel-level fusion of RGB and infrared features across multiple semantic layers. ChannelGate is deployed independently at feature scales P3, P4, and P5, enabling hierarchical fusion to accommodate semantic features across different scales. Compared to simple fusion strategies like channel concatenation or element-wise addition, ChannelGate learns channel-level weighting mechanisms to dynamically suppress redundant information while enhancing complementary modal features.

ChannelGate is based on the SE (Squeeze-and-Excitation) [26] concept and extends it to multimodal fusion scenarios. For the RGB feature map $F_{RGB} \in \mathbb{R}^{B\times C\times H\times W}$ and the aligned infrared feature map $T_{IR} \in \mathbb{R}^{B\times C\times H\times W}$, each modality undergoes independent aggregation. Global Average Pooling (GAP) is applied to compress the spatial dimensions of both RGB and IR features, as expressed by the following formulas:

$$g_{\text{RGB}} = \text{GAP}(F_{\text{RGB}}) \quad , \quad g_{\text{IR}} = \text{GAP}(T_{\text{IR}}) \tag{1}$$

where $g_{\text{RGB}}, g_{\text{IR}} \in \mathbb{R}^{B\times C\times 1\times 1}$.

To enable cross-modal interaction between RGB and infrared modalities at the channel level, the two global descriptors are concatenated along the channel dimension, as expressed by the following formula:

$$g_{\text{cat}} = \text{Concat}([g_{\text{RGB}},\ g_{\text{IR}}]) \in \mathbb{R}^{B\times 2C\times 1\times 1} \tag{2}$$

After concatenation, a two-layer MLP network [27] learns nonlinear relationships between modalities. This network consists of 1×1 convolutions with a compression ratio r=0.5, followed by ReLU [28] and Sigmoid activation functions [29]. During processing, a channel-level gating vector $Z \in \mathbb{R}^{B\times C\times 1\times 1}$ is generated, with a range of [0, 1].

Based on the learned gate vector Z, ChannelGate performs weighted feature fusion by summing RGB and IR features with weighting, as expressed by the following formula:

$$F_{\text{fusion}} = (1 - Z) \odot F_{\text{RGB}} + Z \odot T_{\text{IR}} \tag{3}$$

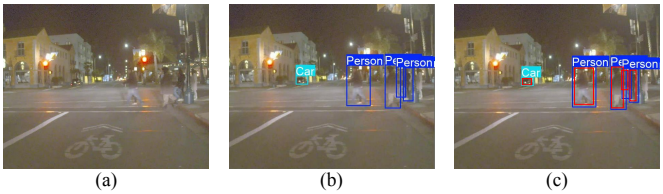
where $\odot$ denotes element-wise multiplication, and Z is expanded to the spatial dimension through broadcasting.

This fusion strategy essentially constitutes a soft mode selection process: when the gating weight approaches 1, the network tends toward an infrared-dominant fusion representation, which holds advantages in low-light and adverse weather conditions; when the gating weight approaches 0, the network leans toward an RGB-dominant fusion representation, making it more suitable for texture-rich scenes. The fusion process is repeated at three scales P3, P4, and P5 to generate multi-scale fused features $\{F_{\text{fuse}}^{P3}, F_{\text{fuse}}^{P4}, F_{\text{fuse}}^{P5}\}$. These features are then input to the neck network and detection head for multi-scale object prediction.

By integrating ChannelGate into the LFNet architecture, we achieve hierarchical, adaptive, and computationally efficient multimodal feature fusion. This model explicitly captures channel-level interactions between RGB and infrared modalities, enabling the network to dynamically balance texture information and thermal feature saliency across different semantic levels. This design effectively mitigates feature redundancy while enhancing complementary representation learning capabilities, laying a robust foundation for precise multimodal object detection in intelligent driving scenarios.

D. Optimization Objective with WIoU v3 Loss

In multimodal object detection, bounding box regression accuracy directly impacts overall detection performance. Extensive experiments conducted within the LFNet framework reveal a notable phenomenon when compared to mainstream monomodal detection models: LFNet typically generates larger predicted bounding boxes, a tendency that becomes particularly pronounced in crowded scenes or occlusion-prone environments, as illustrated in Fig. 3.



(a) (b) (c)

Fig. 3. Figure (a) shows the original image from the dataset. Figure (b) displays the image inferred under the LFNet framework. Figure (c) compares the results of the LFNet framework with those of unimodal inference. The comparison reveals that the prediction boxes generated by the LFNet framework are larger.

This phenomenon stems from the multimodal feature fusion mechanism in LFNet. By adaptively injecting infrared thermal signals at P3, P4, and P5 feature scales, the model enhances target saliency, activating pedestrians and vehicles on the fused feature map. This enables a large number of predicted bounding boxes to rapidly achieve high IoU values with ground truth bounding boxes. However, due to physical characteristics of infrared images—such as thermal halo effects, thermal reflections, and target edge diffusion—predicted bounding boxes often exceed actual bounding boxes. Consequently, even high IoU values may still indicate positioning errors. Under high IoU conditions, traditional IoU-based regression losses (including IoU, GIoU, DIoU, and CIoU) [30] become less effective. When IoU exceeds a threshold, the gradients provided by these loss functions sharply decay, causing the

model to prematurely declare localization convergence and halt further refinement of the bounding box. For LFNet, the multimodal fusion-enhanced heat features increase the probability of the model falling into a state of high IoU but inaccurate localization, thereby exacerbating the problem of over-expanded bounding boxes.

To address this issue, we adopt the WIoU v3 loss proposed by Tong et al. [31] as the bounding box regression objective. WIoU v3 introduces a dynamic non-monotonic focusing mechanism that adaptively redistributes gradient contributions based on overlap quality, maintaining effective optimization signals even in high-IoU regions. When combined with LFNet, this mechanism suppresses excessive bounding box expansion while enabling continuous fine-grained boundary adjustments. Thus, LFNet and WIoU v3 form a complementary optimization strategy: LFNet enhances target saliency through efficient multimodal feature fusion, while WIoU v3 improves localization accuracy via intelligent gradient allocation. This synergy significantly boosts bounding box regression precision and overall detection performance in complex intelligent driving scenarios.

The overall training objective of LFNet is defined as a weighted combination of classification loss and bounding box regression loss:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{WIoU-v3}} \tag{4}$$

where λ is the balancing coefficient used to control the relative weight of localization and classification during training.

E. Lightweight Feature Recalibration with C3k2_SCConv

Although LFNet effectively integrates complementary visible and infrared information through mid-term multimodal fusion, the generated feature representations suffer from redundancy and inconsistent responses. While enhancing salient regions, multimodal fusion introduces irrelevant activations due to thermal diffusion, background interference, and modality-specific noise, negatively impacting subsequent detection heads—particularly in crowded or occluded driving scenarios. Traditional convolutional operations face limitations when processing such multimodal features, as spatial and channel information are modeled in a coupled manner. This coupling makes it challenging to adaptively recalibrate the importance of different modal channels while selectively suppressing redundant spatial responses. Consequently, additional feature refinement mechanisms must be introduced prior to detection to reorganize the fused features.

To address this issue, we propose the C3k2_SCConv hybrid convolution module, which integrates Spatial and Channel Reconstruction Convolution (SCConv) [32] into the C3k2 bottleneck architecture. This design retains the advantages of the original cross-stage partial architecture while enhancing its multi-modal feature recalibration capabilities. SCConv decouples spatial and channel modeling by integrating a Spatial Reconstruction Unit (SRU) and a Channel Reconstruction Unit (CRU). This enables independent suppression of spatial redundancy and adaptive weighting of channel responses. This feature is particularly suited for multimodal feature processing, where infrared signals cause excessive spatial activation amplification while channel contributions across different modalities require balanced adjustment. C3k2_SCConv achieves an optimal balance between representational capacity and computational efficiency. The C3k2_SCConv module is illustrated in Fig. 4.

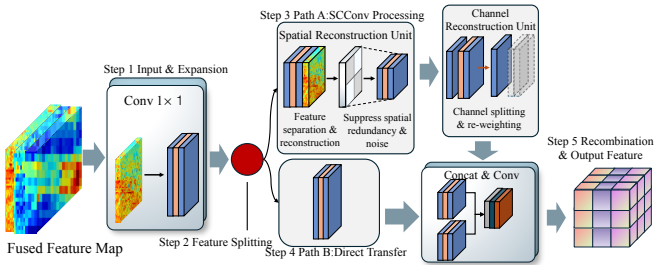


Fig. 4. Integrate SCConv into the C3k2 hybrid convolution module.

The proposed C3k2_SCConv module can be expressed as follows:

$$\mathbf{F}_{\text{out}} = \text{C3k2\_SCConv}(\mathbf{F}_{\text{in}}) \tag{5}$$

where F_{in} and F_{out} denote the input feature map and the refined feature map, respectively.

As the final feature refinement stage preceding the detector head, C3k2_SCConv reorganizes the fused multimodal features into a more compact representation with enhanced discriminative power. This mechanism effectively suppresses redundant activations while amplifying location-sensitive responses, thereby boosting the robustness and accuracy of object detection in complex driving environments.

IV. Experiments

A. Dataset

We evaluate the proposed method on the widely used multimodal object detection benchmark datasets FLIR_aligned and M3FD_split, both of which are tailored for intelligent driving scenarios and low-light conditions.

FLIR_aligned is the manually aligned version of the FLIR ADAS dataset. The original version contained significant misalignment issues between RGB and infrared images. Zhang et al. [33] manually aligned the image pairs in their 2020 ICIP paper, creating the FLIR_aligned version. This dataset is now widely used in multimodal object detection, particularly for nighttime and low-light driving scenarios. The dataset is split into a training set and a validation set at a 4:1 ratio, containing three categories: Person, Car, and Bicycle.

M3FD_split is a specific partitioned version of the M3FD (Multi-scene Multi-modality Fusion Detection) dataset, released by Liu et al. [34] in their 2022 CVPR paper. The dataset provides high-resolution synchronized infrared and visible light image pairs. It is partitioned into training, validation, and test sets at an 8:1:1 ratio, covering six categories: People, Car, Bus, Motorcycle, Lamp, and Truck.

B. Experimental Environment

All experiments were conducted on an Ubuntu 22.04 operating system equipped with an Intel i5-1135G7 CPU, 60GB of system memory, and an NVIDIA RTX-4090 graphics card with 24GB of VRAM. The proposed model was implemented using Python version 3.10.15, based on the PyTorch 2.3.0 deep learning framework, and accelerated via CUDA 12.1 for GPU [35] processing. To ensure fairness and consistency in experimental evaluation, no pre-trained weights were used during training. The model underwent 100 epochs with a batch size of 8, while all other hyperparameters were set to their default values.

C. Ablation Study

To validate the effectiveness of the improvements made to the proposed multimodal feature fusion intelligent driving detection model, four distinct experimental settings were designed, as shown in TABLE I. The first experiment group directly trained the backbone network using additive fusion of raw aligned visible and infrared images, serving as the baseline model. The second experiment group built upon the first by incorporating a three-scale channel gating mechanism and introducing the SPM-Lite lightweight pyramid module for multi-scale feature fusion. The third experiment builds upon the second by integrating the WIoU v3 loss to enhance bounding box regression accuracy and achieve synergistic effects with the LFNet lightweight framework. The fourth experiment combines C3k2 with SCConv convolutions on top of the third experiment to address feature redundancy and semantic mismatch issues. These four experimental setups collectively constitute the ablation studies for each improvement in the LiteFusion model, with strict control over model parameters and experimental environments across all four groups. As shown in TABLE I, the LFNet architecture, WIoU v3, and C3k2_SCConv collectively contribute to the outstanding performance of the LiteFusion model.

To investigate the individual and combined contributions of different modalities, we conducted modal ablation experiments on the FLIR_aligned dataset, as shown in TABLE I. Three experimental settings were configured: visible light images only (RGB-only), infrared images only (Thermal-only), and a multimodal setting utilizing both modalities simultaneously. Experimental results indicate that detection performance using only RGB images is limited, particularly under low-light conditions. While thermal-only detection leverages infrared features to improve recall, relying solely on thermal information still results in insufficient localization accuracy. In contrast, the proposed multimodal model consistently outperforms unimodal approaches across all evaluation metrics. This demonstrates that effective multimodal feature fusion is crucial for achieving object detection in complex driving scenarios, further validating the proposed framework's efficacy.

D. Comparison With the State-of-the-Art Models

To validate the improvement of the proposed model, we first selected several mainstream and representative unimodal network models, such as the YOLO series, for comparative experiments. The experiments selected precision, recall, mAP50, and mAP as the primary evaluation metrics for model accuracy. Subsequently, the performance of the LiteFusion algorithm was compared with advanced multimodal algorithms to validate the extent of model improvement. The experiments selected Params, FLOPs, mAP50, and mAP as the primary evaluation metrics for model accuracy. All experiments were trained on the FLIR_aligned dataset, with parameters and experimental environments strictly controlled for consistency across experiments.

The experimental results are shown in TABLE II. Our proposed model achieves a remarkable detection accuracy of 73.4% on the dataset, surpassing the highest detection accuracy of the YOLO series, YOLO12, by 5.9%. Compared with recently proposed multimodal models, LiteFusion achieves comparable or even superior detection accuracy while reducing computational complexity and parameter count, thus striking a good balance between performance and efficiency.

To further validate the generalization capability of the proposed model across different scenarios, we also conducted comparative experiments on the M3FD_split dataset. The experimental results are shown in TABLE III. The experimental results demonstrate that among all compared models, the LiteFusion model achieves the best overall performance. Compared with unimodal detection models, the proposed multimodal framework significantly improves detection accuracy and recall, effectively demonstrating the effectiveness of visible and infrared information fusion. Compared with existing multimodal fusion methods, LiteFusion exhibits comparable or even superior performance, achieving a balance between detection performance and computational efficiency.

In summary, LiteFusion effectively leverages complementary multimodal information through lightweight feature fusion, achieving robust detection performance without introducing excessive computational overhead. This makes it well-suited for practical deployment in resource-constrained intelligent driving scenarios [36].

V. Conclusion

This paper addresses the issue of insufficient target detection accuracy in intelligent driving systems under low-light and adverse weather conditions by proposing the lightweight multimodal feature fusion model LiteFusion. To achieve adaptive fusion of visible and infrared features with a lightweight design, the LFNet architecture is introduced. To effectively enhance its capabilities, the WIoU v3 loss function replaces traditional loss functions, and the C3k2_SCConv hybrid convolutional module is introduced for joint utilization. Comparisons with multiple state-of-the-art models across various datasets demonstrate the highest gains, with its outstanding performance primarily attributed to the proposed architecture and components. Although LiteFusion still faces limitations in multimodal feature fusion detection, the LiteFusion framework holds broad application prospects not only in intelligent driving but also in security surveillance and other fields.

TABLE I

ROWS 2 TO 5 PRESENT THE ABLATION EXPERIMENTS FOR EACH MODULE BASED ON THE FLIR_ALIGNED DATASET; ROWS 5 TO 7 SHOW THE ABLATION EXPERIMENTS FOR MODAL ABLATION BASED ON THE FLIR_ALIGNED DATASET.

Baseline	LENet	WIoU v3	C3k2_SCConv	RGB-only	Thermal-only	P	R	mAP_{50}	mAP	FLOPs	Params
✓	×	×	×	×	×	69.9%	60.2%	66.4%	34.3%	29.6G	5.0M
✓	✓	×	×	×	×	71.8%	61.6%	67.8%	35.4%	24.8G	**3.8M**
✓	✓	✓	×	×	×	72.5%	62.5%	68.6%	35.5%	24.8G	**3.8M**
✓	✓	✓	✓	✓	✓	**73.4%**	**63.9%**	**70.4%**	**37.1%**	**24.1G**	4.0M
✓	✓	✓	✓	✓	×	58.5%	43.4%	47.1%	22.2%	24.8G	4.2M
✓	✓	✓	✓	×	✓	71.9%	60.5%	68.6%	36.1%	24.8G	4.2M

TABLE II

COMPARISON OF DETECTION MODEL RESULTS BASED ON THE FLIR_ALIGNED DATASET, WHERE ROWS 2–6 REPRESENT SINGLE-MODAL DETECTION MODELS AND ROWS 7–13 REPRESENT MULTI-MODAL DETECTION MODELS.

Model	P	R	mAP_{50}	mAP	FLOPs	Params
YOLOv8 [37]	65.2%	52.3%	58.3%	27.6%	6.8G	2.7M
YOLOv10 [38]	67.0%	48.4%	54.8%	26.6%	6.5G	**2.3M**
YOLO11 [37]	65.8%	52.1%	56.7%	26.9%	6.3G	2.6M
YOLO12 [39]	67.5%	48.1%	54.9%	26.5%	5.8G	2.5M
YOLO13 [40]	65.8%	50.0%	56.0%	26.8%	6.1G	2.5M
YOLO26 [37]	67.8%	57.0%	63.3%	33.4%	**5.6G**	2.4M
hyper-yolo-Midfusion [41]	72.9%	62.9%	**70.7%**	36.3%	15.1G	5.7M
YOLO13-Scorefusion [41]	73.1%	60.1%	67.8%	35.6%	15.0G	5.6M
YOLO13-Share [41]	72.9%	61.0%	67.9%	35.2%	15.5G	5.8M
YOLO13-Midfusion [41]	73.1%	60.3%	67.8%	35.2%	15.4G	6.8M
RT-DETR-Midfusion [42]	65.6%	57.3%	62.1%	30.8%	194.0G	66.3M
RetinaNet-Earlyfusion [43]	66.9%	52.1%	58.8%	28.7%	66.7G	36.4M
LiteFusion(ours)	**73.4%**	**63.9%**	70.4%	**37.1%**	24.1G	4.0M

TABLE III

COMPARISON OF DETECTION MODEL RESULTS BASED ON THE M3FD_SPLIT DATASET, WHERE ROWS 2–3 REPRESENT THE MODALITY ABLATION EXPERIMENTS FOR THIS DATASET, ROWS 4–7 REPRESENT SINGLE-MODALITY DETECTION MODELS, AND ROWS 8–13 REPRESENT MULTI-MODALITY DETECTION MODELS.

Model	P	R	mAP_{50}	mAP	FLOPs	Params
RGB-only	73.5%	56.4%	63.3%	38.2%	24.8G	4.2M
Thermal-only	76.4%	55.1%	61.7%	38.9%	24.8G	4.2M
YOLO11 [37]	78.8%	59.8%	68.6%	44.2%	6.4G	2.6M
YOLO12 [39]	77.3%	56.4%	64.7%	41.1%	6.3G	2.6M
YOLO13 [40]	75.7%	57.5%	64.7%	40.3%	6.1G	2.5M
YOLO26 [37]	70.4%	57.4%	64.7%	41.0%	**5.7G**	**2.4M**
hyper-yolo [41]	80.3%	62.7%	71.1%	45.6%	10.8G	4.0M
YOLO11-Earlyfusion [41]	79.2%	60.5%	68.7%	43.7%	6.4G	2.6M
YOLO11-Midfusion [41]	80.4%	64.9%	71.6%	46.1%	11.0G	5.0M
YOLO13-Midfusion [41]	80.8%	66.3%	72.5%	46.7%	15.4G	6.8M
YOLO13-Scorefusion [41]	78.7%	64.3%	72.5%	47.1%	15.0G	5.6M
RT-DETR-Midfusion [42]	76.3%	64.4%	70.9%	42.7%	194.0G	66.3M
LiteFusion(ours)	**81.3%**	**67.6%**	**73.5%**	**47.4%**	24.8G	3.8M

ACKNOWLEDGMENT

This work was supported by the Postgraduate Innovation and Entrepreneurship Competition Project (No. 2024cxcyjs034) and the Professional Degree Teaching Case Library Project (No. 2025zyxwjxalk317). The authors would like to thank Prof. Hongzhi Zhou for his valuable guidance and constructive comments on this work.

REFERENCES

[1] Z. Zhao and J. Chen, "Application of artificial intelligence technology in the economic development of urban intelligent transportation system," PeerJ Computer Science, vol. 11, p. e2728, 2025.

[2] B. Sun, S. Yang, Y. Wang, J. Lu, Z. Pang, X. Feng, H. Guang, and Y. Cao, "A fusion safety and security analysis framework for intelligent and connected vehicles," PLoS One, vol. 20, no. 9, p. e0332050, 2025.

[3] D. Kowald, S. Scher, V. Pammer-Schindler, P. M¨ullner, K. Waxnegger, L. Demelius, A. Fessl, M. Toller, I. G. Mendoza Estrada, I. ˇSimi´c et al., "Establishing and evaluating trustworthy AI: overview and research challenges," Frontiers in Big Data, vol. 7, p. 1467222, 2024.

[4] S. Cociu, O. Ioncu, D. Ciobanu, and S. Cebanu, "Road safety knowledge and attitudes among drivers," One health & risk management, vol. 4, no. 2, p. 25, 2023.

[5] C. Jiang, S. Yu, J. Fu, C.-C. Lin, H. Zhu, X. Ma, M. Li, and L. Guo,"Behaviors speak more: Achieving user authentication leveraging facial activities via mmwave sensing," in Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems, 2024, pp. 169–183.

[6] S. Yu, J. Fu, C. Jiang, C. Lin, Z. Zhang, L. Cheng, M. Li, X. Zhang,and L. Guo, "Freeem: Uncovering parallel memory emr covert communication in volatile environments," in Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services, 2024, pp. 372–384.

[7] C. Sun, Y. Chen, X. Qiu, R. Li, and L. You, "Mrd-yolo: A multispectral object detection algorithm for complex road scenes," Sensors, vol. 24, no. 10, p. 3222, 2024.

[8] H. Xu, X. Wang, and J. Ma, "Drf: Disentangled representation for visible and infrared image fusion," IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1–13, 2021.

[9] H. Li, N. Hussin, D. He, Z. Geng, and S. Li, "Design of image segmentation model based on residual connection and feature fusion," PloS one, vol. 19, no. 10, p. e0309434, 2024.

[10] F. Huang, J. Zheng, X. Liu, Y. Shen, and J. Chen, "Polarization of road target detection under complex weather conditions," Scientific Reports, vol. 14, no. 1, p. 30348, 2024.

[11] Q. Guo, Y. Wang, Y. Zhang, M. Zhao, and Y. Jiang, "Aie-yolo: Effective object detection method in extreme driving scenarios via adaptive image enhancement," Science progress, vol. 107, no. 3, p. 00368504241263165, 2024.

[12] Z. Yang, Y. Li, X. Tang, and M. Xie, "Mgfusion: a multimodal large language model-guided information perception for infrared and visible image fusion," Frontiers in Neurorobotics, vol. 18, p. 1521603, 2024.

[13] S. Motayyeb, F. Samadzedegan, F. D. Javan, and H. Hosseinpour, "Fusion of uav-based infrared and visible images for thermal leakage map generation of building facades," Heliyon, vol. 9, no. 3, 2023.

[14] W. Wang, J. Xu, and R. Zhang, "Optimized small object detection in low resolution infrared images using super resolution and attention based feature fusion," PLoS One, vol. 20, no. 7, p. e0328223, 2025.

[15] L.-S. Guo, Y. An, Z.-Y. Zhang, C.-B. Ma, J.-Q. Li, Z. Dong, J. Tian, Z.-Y. Liu, and J.-G. Liu, "Exploring the diagnostic potential: magnetic particle imaging for brain diseases," Military Medical Research, vol. 12, no. 1, p. 18, 2025.

[16] Y. Song, A. Liu, X. Meng, Z. Liu, P. Liu, and X. Zhen, "A maturity classification model for winter jujubes based on dsaf-resnet," npj Science of Food, vol. 9, no. 1, p. 187, 2025.

[17] S. Liu, S. Zhang, F. Gao, and Y. Feng, "Highly condensed all-mlp architecture for long-term human motion prediction," IEEE Transactions on Neural Networks and Learning Systems, 2025.

[18] H. Zhang, E. Fromont, S. Lef'evre, and B. Avignon, "Guided attentive feature fusion for multispectral pedestrian detection," in Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2021, pp. 72–80.

[19] C. Jiang, H. Ren, H. Yang, H. Huo, P. Zhu, Z. Yao, J. Li, M. Sun, and S. Yang, "M2fnet: Multi-modal fusion network for object detection from visible and thermal infrared images," International Journal of Applied Earth Observation and Geoinformation, vol. 130, p. 103918, 2024.

[20] J. Guo, C. Gao, F. Liu, D. Meng, and X. Gao, "Damsdet: Dynamic adaptive multispectral detection transformer with competitive query selection and adaptive feature fusion," in European Conference on Computer Vision. Springer, 2024, pp. 464–481.

[21] W. Wu, X. Zhang, H. Yin, S. Dai, H. Zhang, and Y. Zhang, "Fredft: Frequency domain fusion transformer for visible-infrared object detection," arXiv preprint arXiv:2511.10046, 2025.

[22] J. Shen, Y. Chen, Y. Liu, X. Zuo, H. Fan, and W. Yang, "Icafusion: Iterative cross-attention guided feature fusion for multispectral object detection," Pattern Recognition, vol. 145, p. 109913, 2024.

[23] H. Gao, Y. Zhang, Z. Chen, and C. Li, "A multiscale dual-branch feature fusion and attention network for hyperspectral images classification," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 8180–8192, 2021.

[24] Z. Hou, C. Yang, Y. Sun, S. Ma, X. Yang, and J. Fan, "An object detection algorithm based on infrared-visible dual modal feature fusion," Infrared Physics & Technology, vol. 137, p. 105107, 2024.

[25] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in International conference on machine learning. PMLR, 2021, pp. 10 096–10 106.

[26] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[27] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit et al., "Mlp-mixer: An all-mlp architecture for vision," Advances in neural information processing systems, vol. 34, pp. 24 261–24 272, 2021.

[28] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in Proceedings of the 27th international conference on machine learning (ICML-10), 2010, pp. 807–814.

[29] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," nature, vol. 323, no. 6088, pp. 533–536, 1986.

[30] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-iou loss: Faster and better learning for bounding box regression," in Proceedings of the AAAI conference on artificial intelligence, vol. 34, no. 07, 2020, pp. 12 993–13 000.

[31] Z. Tong, Y. Chen, Z. Xu, and R. Yu, "Wise-iou: bounding box regression loss with dynamic focusing mechanism," arXiv preprint arXiv:2301.10051, 2023.

[32] J. Li, Y. Wen, and L. He, "Scconv: Spatial and channel reconstruction convolution for feature redundancy," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 6153–6162.

[33] H. Zhang, E. Fromont, S. Lefevre, and B. Avignon, "Multispectral fusion for object detection with cyclic fuse-and-refine blocks," in 2020 IEEE International conference on image processing (ICIP). IEEE, 2020, pp. 276–280.

[34] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo, "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, 2022, pp. 5802–5811.

[35] J. Fu and Y. Liu, "A comprehensive performance evaluation to gpgpu applications under stt-ram based hybrid cache architectures," in 2020X Brazilian Symposium on Computing Systems Engineering (SBESC). IEEE, 2020, pp. 1–8.

[36] J. Ji, G. Li, J. Fu, F. Afghah, L. Guo, X. Yuan, and X. Ma, "A single-step, sharpness-aware minimization is all you need to achieve efficient and accurate sparse training," Advances in Neural Information Processing Systems, vol. 37, pp. 44 269–44 290, 2024.

[37] R. Sapkota and M. Karkee, "Ultralytics yolo evolution: An overview of yolo26, yolo11, yolov8 and yolov5 object detectors for computer vision and pattern recognition," arXiv preprint arXiv:2510.09653, 2025.

[38] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han et al., "Yolov10: Real-time end-to-end object detection," Advances in Neural Information Processing Systems, vol. 37, pp. 107 984–108 011, 2024.

[39] Y. Tian, Q. Ye, and D. Doermann, "Yolov12: Attention-centric real-time object detectors," arXiv preprint arXiv:2502.12524, 2025.

[40] M. Lei, S. Li, Y. Wu, H. Hu, Y. Zhou, X. Zheng, G. Ding, S. Du, Z. Wu, and Y. Gao, "Yolov13: Real-time object detection with hypergraph-enhanced adaptive visual perception," arXiv preprint arXiv:2506.17733, 2025.

[41] D. Wan, R. Lu, Y. Fang, X. Lang, S. Shu, J. Chen, S. Shen, T. Xu, and Z. Ye, "Yolov11-rgbt: Towards a comprehensive single-stage multispectral object detection framework," arXiv preprint arXiv:2506.14696, 2025.

[42] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "Detrs beat yolos on real-time object detection," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 16 965–16 974.

[43] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Doll´ar, "Focal loss for dense object detection," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.