Spatially Adaptive and Cross-Modal Differential Enhancement Network for Multispectral Object Detection

Miao Li[1], Xiao Huang[1(✉)], Jinlai Zhang[2(✉)], Mingchao Xiang[1]

[1] Elite Engineering School, Changsha University of Science and Technology, Changsha, 410114, Hunan, China
[2] College of Mechanical and Vehicle Engineering, Changsha University of Science and Technology, Changsha, 410114, Hunan, China
jinlai.zhang@csust.edu.cn

Abstract. Multispectral object detection in complex urban environments remains a challenging task due to severe thermal diffusion, geometric variations, and modality-specific background noise. To address these issues, we propose the Spatially Adaptive and Cross-Modal Differential Enhancement Network (SACDNet), a novel framework featuring three core components: Spatially Adaptive Modulated Convolution (SAMC), Cross-Modal Differential Enhancement (CMDE), and Illumination-Guided Geometric Loss (IGGL). SAMC effectively adapts to target deformations by dynamically adjusting the shape of the receptive field. CMDE utilizes cross-modal channel attention to adaptively emphasize reliable structural representations. Furthermore, IGGL modulates the geometric penalty based on environmental illumination to mitigate the influence of noisy predictions. Experimental evaluations on the FLIR-align dataset demonstrate the effectiveness of our approach. Compared to the state-of-the-art (SOTA) methods, SACDNet achieves a significant improvement, reaching a mean Average Precision (mAP) of 37.2%.

Keywords: Multispectral Object Detection, Spatially Adaptive Convolution, Cross-Modal Fusion.

1 Introduction

Object detection in complex urban environments is a fundamental task in computer vision [24-27], with broad applications in autonomous driving, pedestrian safety monitoring, and intelligent transportation systems [12]. Single-modality detectors based on visible cameras are highly sensitive to illumination variations and suffer severe performance degradation at night or under adverse weather conditions. Thermal infrared cameras, by contrast, capture emitted heat radiation and remain functional regardless of ambient lighting. Multispectral object detection, which fuses visible and thermal infrared information, has therefore attracted considerable research

attention as a means of achieving robust detection across diverse real-world conditions [7].

Existing multispectral fusion methods can be broadly categorized into early fusion, late fusion, and feature-level fusion. Early fusion methods [15] concatenate raw inputs from both modalities before the backbone network. While straightforward, this strategy cannot bridge the substantial domain gap between visible and infrared images, leading to feature interference. Late fusion methods [10] maintain independent processing streams and combine predictions at the output stage, which precludes deep cross-modal interaction. Feature-level fusion methods [11] have therefore become the dominant paradigm, as they allow the network to exchange complementary information at intermediate feature representations. However, most existing feature-level approaches share three fundamental limitations. First, standard backbone networks employ fixed-grid convolutions that cannot adapt to the irregular target shapes and blurry boundaries caused by thermal diffusion in infrared images. Second, existing fusion modules treat features from both modalities with equal importance, without accounting for the asymmetric reliability of the two sensors under varying illumination conditions. Third, standard bounding box regression losses apply uniform geometric penalties to all training samples, ignoring the fact that visible image quality degrades significantly at night, which introduces unreliable supervision signals during training.

To address these limitations, we propose the Spatially Adaptive and Cross-Modal Differential Enhancement Network (SACDNet), a novel multispectral object detection framework comprising three core components. To handle thermal diffusion and geometric deformation, we introduce the Spatially Adaptive Modulated Convolution (SAMC), which extends deformable convolution [23] by jointly learning spatial offsets and per-point modulation scalars, enabling the receptive field to dynamically conform to irregular target shapes while suppressing background noise. To achieve reliable cross-modal feature integration under varying illumination, we propose the Cross-Modal Differential Enhancement (CMDE) module, which constructs an explicit inter-modal differential representation and applies channel attention [16] to adaptively reweight the contribution of each modality. To improve bounding box regression under degraded lighting conditions, we design the Illumination-Guided Geometric Loss (IGGL), which modulates the geometric regression penalty according to an estimated illumination score, reducing the influence of unreliable visible-domain annotations during nighttime training [22].

We evaluate SACDNet on the FLIR-align dataset [3], a benchmark of strictly aligned RGB-thermal image pairs collected from complex urban driving scenarios. Experimental results demonstrate that SACDNet achieves a mean Average Precision (mAP) of 37.2%, representing a significant improvement over the basic dual-stream baseline. Ablation studies confirm the individual contribution of each proposed component. The main contributions of this paper are as follows:

- We propose SACDNet, a multispectral object detection framework that integrates three complementary components to address thermal diffusion, illumination-dependent modality reliability, and unreliable nighttime regression supervision.

- We propose SAMC, a spatially adaptive modulated convolution that dynamically adjusts the receptive field shape and the importance of each sampling point, effectively addressing thermal diffusion and geometric deformation in multispectral urban detection.
- We propose CMDE, a cross-modal differential enhancement module that constructs an explicit inter-modal difference representation and applies cross-modal channel attention to adaptively balance the contribution of each modality under varying illumination conditions.
- We propose IGGL, an illumination-guided geometric loss that modulates the bounding box regression penalty according to the estimated lighting quality, improving training stability and detection accuracy across day-night transitions.

2 Related Work

Multispectral object detection has been extensively studied in recent years. This section reviews the relevant literature from three perspectives closely related to the proposed method: multispectral fusion methods, deformable convolutional networks, and loss functions for bounding box regression.

2.1 Multispectral Object Detection

Multispectral object detection [28-31] methods that jointly exploit visible and thermal infrared imagery have attracted considerable attention due to their robustness under varying illumination conditions. Early work by Hwang et al. [19] demonstrated that fusing RGB and thermal modalities significantly improves pedestrian detection at night. Subsequent methods can be broadly categorized by their fusion strategy. Early fusion methods [15] concatenate raw inputs from both modalities before the backbone network, which is straightforward but fails to account for the substantial domain gap between visible and infrared images. Late fusion methods [9] maintain independent processing streams and combine predictions at the output stage, which limits the depth of cross-modal interaction. Feature-level fusion methods have therefore become the dominant paradigm, and the two-stream Faster R-CNN framework [12] has been widely adopted as a strong baseline for this line of work, owing to its flexible dual-backbone design that naturally accommodates independent feature extraction from each modality.

Liu et al. [10] proposed a halfway fusion strategy that combines intermediate feature maps from two independent streams, demonstrating that mid-level fusion outperforms both early and late fusion. Zhang et al. [19] proposed a weakly aligned multispectral pedestrian detector that handles spatial misalignment between modalities through a region feature alignment module. GAFF [18] employs guided attentive feature fusion to adaptively integrate cross-modal features by learning channel-wise attention weights. CFT [11] applies a cross-modal Transformer to model long-range dependencies between visible and infrared feature maps. Channel attention mechanisms, such as SENet [6] and CBAM [16], have also been incorporated into

fusion modules to selectively emphasize informative channels, providing a lightweight yet effective means of cross-modal feature recalibration. Despite these advances, most existing methods treat features [36-38] from both modalities with equal importance regardless of the current illumination condition, and do not explicitly address the geometric deformation caused by thermal diffusion in infrared images.

2.2 Deformable Convolutional Networks

Standard convolutional neural networks [32-35] apply fixed-grid sampling patterns, which limits their ability to model geometric transformations such as scale variation, rotation, and irregular shape deformation. Dai et al. [2] introduced Deformable Convolutional Networks (DCN), which augment standard convolution with learnable spatial offsets, allowing the receptive field to adapt to the geometric structure of objects. Zhu et al. [23] further proposed DCNv2, which introduces per-point modulation scalars in addition to spatial offsets, enabling the network to suppress irrelevant background regions by down-weighting their contributions.

These deformable convolution operations have been widely adopted in object detection backbones to improve robustness to geometric variation. In the context of multispectral detection, however, the additional challenge of thermal diffusion in infrared images, which causes blurry and irregular target boundaries, has not been explicitly addressed by existing deformable convolution designs. The proposed Spatially Adaptive Modulated Convolution (SAMC) builds upon DCNv2 and introduces a modulation mechanism specifically tailored to suppress infrared background noise while adapting to thermally diffused target shapes.

2.3 Bounding Box Regression Loss

Accurate bounding box regression is essential for precise object localization. Early detectors employed smooth l_1 loss [4] to regress box coordinates independently. IoU loss [17] was subsequently proposed to directly optimize the Intersection over Union metric, providing a more geometrically meaningful training signal. GIoU [13] extended IoU loss to handle non-overlapping boxes by incorporating the area of the smallest enclosing box. DIoU and CIoU [22] further introduced a center distance term and an aspect ratio consistency penalty to accelerate convergence and improve localization accuracy.

These geometric loss functions have consistently improved detection performance across standard benchmarks. However, they apply a uniform penalty to all training samples regardless of the sensor modality or environmental condition, which is problematic in multispectral datasets where the reliability of visible image annotations degrades significantly at night. The proposed Illumination-Guided Geometric Loss (IGGL) addresses this limitation by modulating the geometric regression penalty according to an estimated illumination score, reducing the influence of unreliable annotations under poor lighting conditions.

3 Methodology

In this section, we first provide an overview of our proposed detection framework. Following this, we detail the Spatially Adaptive Modulated Convolution (SAMC) for spatial feature extraction, the Cross-Modal Differential Enhancement (CMDE) for feature fusion, and finally, the Illumination-Guided Geometric Loss (IGGL) for bounding box regression. The overall architecture of the proposed framework is illustrated in
Fig. 1.

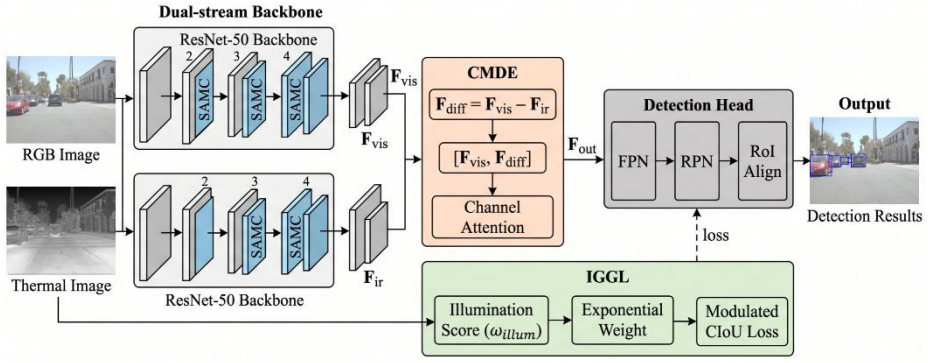

Fig. 1. Overview of the proposed multispectral object detection framework. The dual-stream ResNet-50 backbone extracts features from RGB and thermal infrared inputs independently. SAMC replaces standard convolutions in stages 2, 3, and 4 to handle thermal diffusion and geometric deformation. CMDE fuses the two-stream features via a cross-modal differential map and channel attention. The detection head produces final bounding boxes supervised by the proposed IGGL.

3.1 Spatially Adaptive Modulated Convolution

In urban multispectral object detection, finding the exact location of targets is often a difficult task. This difficulty comes from several physical factors. For pedestrians, their body heat can cause thermal diffusion in infrared images, making their edges blurry. For vehicles, changing viewpoints can cause large changes in their apparent shape. Additionally, complex urban backgrounds, such as roads and buildings heated by the sun, create interference. Standard Convolutional Neural Networks (CNNs) use fixed grid sampling locations. This fixed grid design limits their ability to adapt to targets that have extreme scale changes or irregular shapes. As a result, the standard feature extraction process often pulls in a large amount of irrelevant background noise. This noise makes it harder for the model to clearly draw the outlines of pedestrians and vehicles.

To solve these physical limitations, we introduce the Spatially Adaptive Modulated Convolution (SAMC) module. Building upon the concept of deformable

convolutions, SAMC allows the network to dynamically change both the spatial shape of its receptive field and the importance of each individual sampling point.

Given an input feature map $\mathrm{X} \in \mathbb{R}^{C\times H\times W}$ from a previous layer, a standard 2D convolution samples data over a regular grid. For a convolution kernel with K sampling locations, let ω_k represent the weight and p_k represent the fixed spatial offset for the *k*-th location. SAMC introduces two new dynamic variables for each location: a spatial offset Δp_k and a modulation scalar Δm_k. The output feature *y(p)* at spatial location *p* is calculated as follows:

$$y(p) = \sum_{K=1}^{K} \omega_k \bullet x(p + p_k + \Delta p_k) \bullet \Delta m_k \tag{1}$$

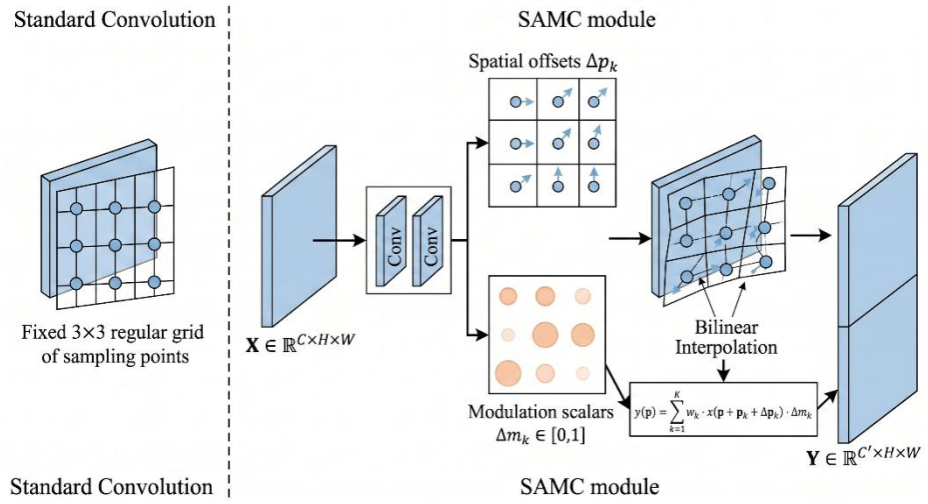

Fig. 2. Conceptual comparison between a standard 3×3 convolution (left) and the Spatially Adaptive Modulated Convolution (right). Unlike the fixed grid used in standard convolution, SAMC introduces dynamic spatial offsets (Δp_k) and amplitude modulation (Δm_k) to adjust the sampling locations and their corresponding weights, effectively suppressing background noise.

The spatial offset Δp_k is a 2D vector with real numbers. It allows the sampling points to move away from the fixed grid locations. This flexibility helps the receptive field to closely match the specific shapes of targets, like the thin outlines of walking pedestrians or the flat sides of cars. Because the newly calculated sampling position, $p + p_k + \Delta p_k$, will usually fall on fractional coordinates instead of exact pixels, we use bilinear interpolation to calculate the exact feature values $x(\bullet)$. Furthermore, the modulation scalar Δm_k, which takes a value between 0 and 1, controls the weight of the sampled features. In crowded street images, SAMC can reduce the effect of

infrared background noise by setting the Δm_k values of background regions closer to zero.

For the network architecture, we design SAMC to be easily integrated. Using ResNet-50 [5] as the main backbone network, we only replace the standard 3×3 convolutional layers with SAMC in the deeper stages (specifically Stage 2, Stage 3, and Stage 4). This setup allows the network to keep high-resolution basic textures in the early layers, while building a flexible, target-focused spatial representation in the later layers before the multi-modal fusion step.

3.2 Cross-Modal Differential Enhancement

Infrared images suffer from thermal diffusion, which blurs target boundaries and makes it difficult to separate foreground objects from heated backgrounds. While the visible modality provides richer texture and edge information during the day, its reliability drops significantly at night. A simple concatenation of features from both modalities treats them with equal importance and cannot adapt to these illumination-dependent quality changes.

To address this, we propose the Cross-Modal Differential Enhancement (CMDE) module. The key idea is to explicitly construct a cross-modal differential map by subtracting the infrared feature from the visible feature, which highlights the complementary information between the two modalities and suppresses their shared background noise. A channel attention mechanism then adaptively reweights the combined representation based on the current reliability of each modality.

Specifically, let $\boldsymbol{F}_{vis}$ and $\boldsymbol{F}_{ir} \in \mathbb{R}^{C \times H \times W}$ denote the feature maps extracted from the visible and infrared streams, respectively. We first compute the inter-modal differential feature:

$$\boldsymbol{F}_{diff} = \boldsymbol{F}_{vis} - \boldsymbol{F}_{ir} \tag{2}$$

This differential map encodes the discrepancy between the two modalities, emphasizing regions where visible and infrared responses diverge, such as object boundaries and thermally active areas. We then concatenate the visible feature and the differential feature along the channel dimension:

$$\boldsymbol{F}_{concat} = [\boldsymbol{F}_{vis}, \boldsymbol{F}_{diff}] \in \mathbb{R}^{2C \times H \times W} \tag{3}$$

A channel attention vector $\boldsymbol{A}_c \in \mathbb{R}^{2C \times 1 \times 1}$ is then computed to reweight the concatenated features:

$$\boldsymbol{A}_c = \sigma(\boldsymbol{W}_2 \delta(\boldsymbol{W}_1 AvgPool(\boldsymbol{F}_{concat}))) \tag{4}$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{\frac{2C}{R} \times 2C}$ and $\boldsymbol{W}_2 \in R^{2C \times \frac{2C}{R}}$ are the weights of a two-layer MLP with reduction ratio r, δ denotes ReLU, and σ denotes Sigmoid. The enhanced output is

obtained by element-wise multiplication followed by a 1×1 convolution to restore the original channel dimension:

$$\boldsymbol{F}_{out} = Conv_{1\times1}(\boldsymbol{A}_c \odot \boldsymbol{F}_{concat}) \quad (5)$$

The overall architecture of CMDE is illustrated in Fig. 3. When the visible image quality is high, the differential map $\boldsymbol{F}_{diff}$ carries strong boundary information, and the attention mechanism assigns higher weights to the corresponding channels. At night, when the visible features become noisy, the differential map loses its discriminability, and the attention naturally suppresses those channels. This allows CMDE to achieve illumination-adaptive fusion without any explicit illumination estimation.

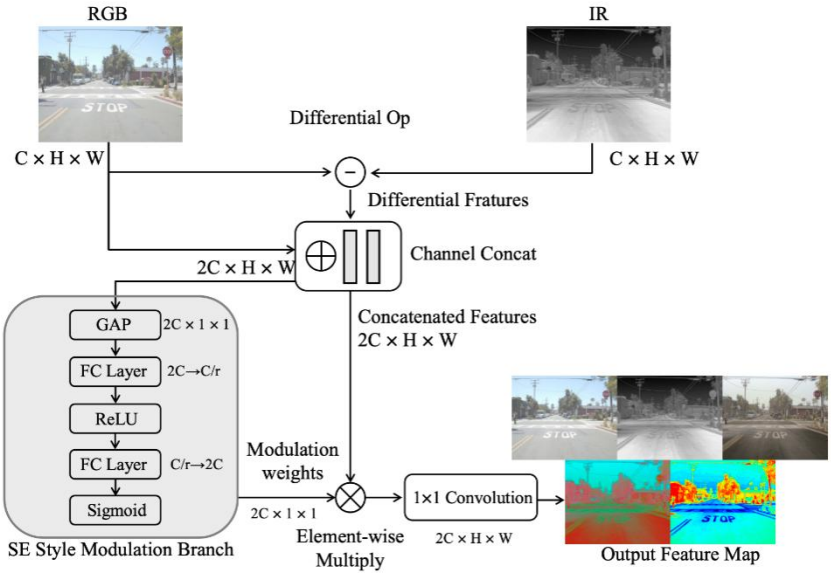

Fig. 3. Architecture of the Cross-Modal Differential Enhancement (CMDE) module. The visible feature $\boldsymbol{F}_{vis}$ and infrared feature $\boldsymbol{F}_{ir}$ are used to compute an inter-modal differential map $\boldsymbol{F}_{diff}$, which is concatenated with $\boldsymbol{F}_{vis}$ and passed through a channel attention mechanism to produce the enhanced output $\boldsymbol{F}_{out}$.

3.3 Illumination-Guided Geometric Loss

The overall training loss of the proposed framework is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{rpn\_cls} + \mathcal{L}_{rpn\_reg} + \mathcal{L}_{cls} + \mathcal{L}_{IGGL} \quad (6)$$

where $\mathcal{L}_{rpn\_cls}$ and $\mathcal{L}_{rpn\_reg}$ are the standard RPN classification and regression losses, and $\mathcal{L}_{cls}$ is the cross-entropy classification loss of the detection head. The core innovation lies in $\mathcal{L}_{IGGL}$, which replaces the standard uniform bounding box regression penalty in the detection head.

Standard geometric regression losses such as CIoU apply a uniform penalty to all training samples regardless of environmental lighting conditions. In multispectral datasets such as FLIR, the visible camera provides reliable boundary annotations during the day but produces noisy localizations at night. Forcing the network to fit these unreliable annotations with a uniform penalty degrades regression accuracy.

To address this, we estimate the lighting condition of each sample from the visible image I_{vis} (pixel values in [0, 255]). The illumination score $\omega_{illum} \in [0,1]$ is defined as:

$$\omega_{illum} = \frac{1}{CHW} \sum_{c,h,w} \frac{I_{vis}^{(c,h,w)}}{255} \tag{7}$$

A score approaching 1 indicates a well-lit daytime scene; a score approaching 0 indicates a dark nighttime environment. The IGGL is then formulated as:

$$\mathcal{L}_{IGGL} = exp\ (-\tau(1-\omega_{illum})) \cdot \mathcal{L}_{CIoU}(\boldsymbol{B}, \boldsymbol{B}^{gt}) \tag{8}$$

where $\boldsymbol{B}$ and $\boldsymbol{B}^{gt}$ are the predicted and ground-truth bounding boxes, and τ is a temperature parameter controlling the sensitivity of the weighting curve. We set $\tau = 2.0$ in all experiments.

When the scene is well-lit ($\omega_{illum} \to 1$), the weight approaches 1.0 and IGGL reduces to standard CIoU, fully utilizing high-quality visible annotations. When the scene is dark ($\omega_{illum} \to 0$), the weight drops to $exp(-\tau) \approx 0.135$, suppressing erroneous gradient updates from degraded visible features. This design requires no additional network branch and adds negligible computational overhead.

4 Experimental Results

4.1 Datasets

We evaluate the proposed method on the FLIR-align dataset [3], a widely used benchmark for multispectral object detection.

FLIR-align. The dataset consists of strictly aligned RGB and thermal infrared image pairs captured from complex urban driving scenarios. It covers three object categories: car, person, and bicycle. Following the standard split, we use 4,129 image pairs for training and 1,013 pairs for evaluation.

4.2 Experimental Details

Platform. All experiments were conducted on a single NVIDIA Tesla V100 GPU with PyTorch and the MMDetection toolbox [1].

Data Preprocessing. Paired RGB and infrared images were resized to 640×512 and padded to ensure spatial dimensions divisible by 32. Random horizontal flipping with a probability of 0.5 was applied synchronously to both modalities during training.

Table 1. Comparison with SOTA RGB+IR methods on FLIR-aligned dataset.

Method	Backbone	mAP_{50}	mAP_{75}	mAP
Halfway Fusion [10]	VGG-16	71.5	-	35.8
ICAFusion [14]	ResNet-50	72.0	-	-
TINet [20]	ResNet-50	76.1	-	36.5
YOLOv5 [8]	CSPD-53	70.1	29.9	34.9
RSDet [21]	ResNet-50	71.0	28.1	34.4
SACDNet(ours)	**ResNet-50**	**74.5**	**31.5**	**37.2**

Hyperparameters. The dual-stream backbone is ResNet-50 [5] pre-trained on ImageNet. Standard 3×3 convolutions in stages 2, 3, and 4 are replaced by SAMC, and cross-modal features are aggregated by CMDE. The RPN generates anchors with scales [4, 8] and aspect ratios [0.5, 1.0, 2.0]. Bounding box regression is supervised by the proposed IGGL with temperature parameter $\tau = 2.0$. The network is optimized end-to-end with SGD (learning rate 0.001, momentum 0.9, weight decay 0.0001) for 16 epochs with batch size 2. The learning rate decays by a factor of 0.1 at epochs 11 and 14, with a linear warmup over the first 500 iterations.

Evaluation Metrics. We adopt the standard COCO evaluation protocol, reporting mAP averaged over IoU thresholds from 0.50 to 0.95, as well as mAP50 and mAP75.

4.3 Comparison with SOTA Methods

Table 1 presents a comprehensive comparison of our proposed method against fundamental baselines (e.g., YOLOv5 [8], RSDet [21]) and established multispectral detection methods on the FLIR-align dataset.

The evaluation results demonstrate three key insights. First, multi-spectral fusion methods consistently outperform single-modality baselines in challenging environments. This variance confirms that integrating visible and thermal features can provide critical complementary cues to overcome adverse illumination and sensory degradation.

Second, within the multi-spectral fusion category, early naive dual-stream models like Halfway Fusion [10] and our baseline RSDet [21] achieve limited improvement, primarily because they lack effective cross-modal interaction.

Most importantly, our proposed method achieves the best mAP performance among all compared methods, delivering a prominent score of 0.372 mAP, 0.745

mAP50, and 0.315 mAP75. Compared with the pure YOLOv5 baseline [8], a gain of 2.3 percentage points in mAP and a 4.4% improvement in mAP50. Furthermore, compared with the RSDet baseline [21], the full model improves mAP from 0.344 to 0.372, a gain of 2.8%. This robust improvement proves that our adaptive design excels at aligning fine-grained cross-modal features while successfully suppressing non-aligned noise, justifying the superiority of our approach for multi-spectral object detection.

4.4 Ablation Study

To evaluate the contribution of each proposed component, we conducted ablation experiments on the FLIR-align dataset. The results are summarized in Table 2.

Starting from the two-stream baseline (mAP 0.344), adding each module individually yields consistent improvements. SAMC alone improves mAP to 0.352, confirming that spatially adaptive sampling helps the backbone better capture thermally diffused target shapes. CMDE alone achieves 0.354, demonstrating that the cross-modal differential enhancement effectively suppresses background noise during feature fusion. IGGL alone reaches 0.355, indicating that illumination-guided loss weighting reduces the impact of unreliable visible annotations at night.

When combining two modules, SAMC+IGGL achieves the highest mAP of 0.365 among the two-module combinations, suggesting that adaptive spatial representation and illumination-aware training are particularly complementary. The full model with all three components achieves the best performance across all metrics (mAP 0.372, mAP50 0.745, mAP75 0.315), confirming that SAMC, CMDE, and IGGL address complementary aspects of the multispectral detection problem and work best in combination.

Table 2. Ablation study of the proposed modules.

Baseline	SAMC	CMDE	IGGL	mAP	mAP_{50}	mAP_{75}
✓				0.344	0.710	0.281
✓	✓			0.352	0.726	0.289
✓		✓		0.354	0.728	0.294
✓			✓	0.355	0.729	0.300
✓	✓	✓		0.357	0.733	0.290
✓	✓		✓	0.365	0.740	0.303
✓		✓	✓	0.361	0.743	0.296
✓	✓	✓	✓	0.372	0.745	0.315

4.5 Qualitative Results

To further demonstrate the superiority of our approach, we provide a detailed visual comparison against the baseline methods across various challenging scenarios. Blue, orange, and red boxes denote true positive detections (correct positives), false negative detections (incorrect negatives), and false positive detections (incorrect positives), respectively.

First, Fig. 4 and Fig. 5 illustrate the detection performance under challenging nighttime illumination conditions. In these scenarios, the RGB modality suffers from severe visual degradation, street lighting interference, and extreme vehicle taillight glare. Consequently, both YOLOv5 and RSDet exhibit obvious false negatives (orange boxes) for pedestrians in the dark, as well as severe false positives (red boxes) triggered by the glare. For instance, in Fig. 5, the intense taillight overexposure causes the baseline models to either mislocalize the vehicle or generate redundant ghost boxes, while entirely missing the pedestrian on the dark sidewalk. In contrast, driven by our cross-modal feature extraction modules, our model effectively leverages robust thermal cues from the IR modality. This mechanism perfectly suppresses background interference from extreme glare and successfully captures the visually degraded pedestrian, resulting in precise and clean bounding boxes in both night scenes.

Second, Fig. 6 highlights the robustness of our model in resolving complex instance occlusion. As seen in the street view where a pedestrian heavily overlaps with a vehicle, the baselines struggle significantly. YOLOv5 completely misses the occluded pedestrian (orange box), whereas RSDet hallucinates multiple redundant bounding boxes (red boxes) on the right side of the street due to feature confusion. Thanks to our precise spatial alignment and multi-spectral fusion, our method clearly separates the highly overlapped instances, accurately detecting both targets without mutual suppression or false positive clutter.

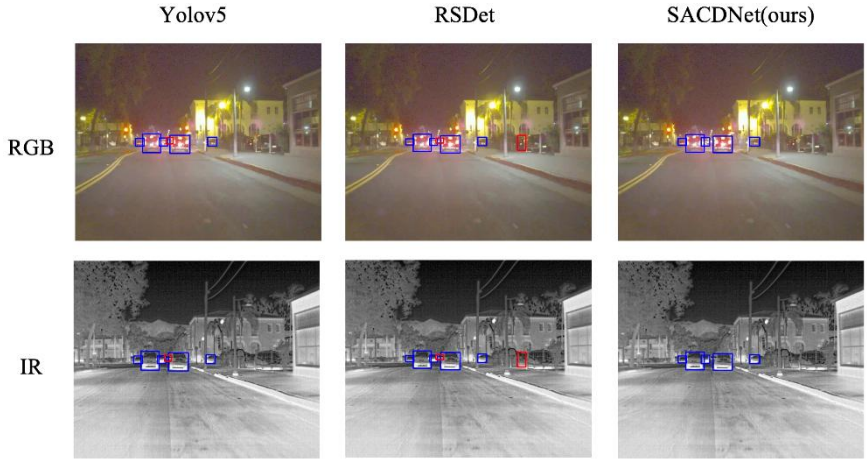

Fig. 4. Visual comparison of object detection under general nighttime illumination. From left to right: YOLOv5, RSDet, and SACDNet(ours). Our method effectively detects distant vehicles despite the interference of streetlights and low visibility.

Fig. 5. Visual comparison under extreme nighttime conditions with severe vehicle taillight glare. From left to right: YOLOv5, RSDet, and SACDNet(ours). Our proposed method successfully suppresses the extreme glare to accurately bound the vehicle, while also detecting the pedestrian hidden in the dark background on the right.

Fig. 6. Visual comparison in scenes with severe instance occlusion. From left to right: YOLOv5, RSDet, and SACDNet(ours). Our method accurately resolves overlapping targets without generating ghost predictions.

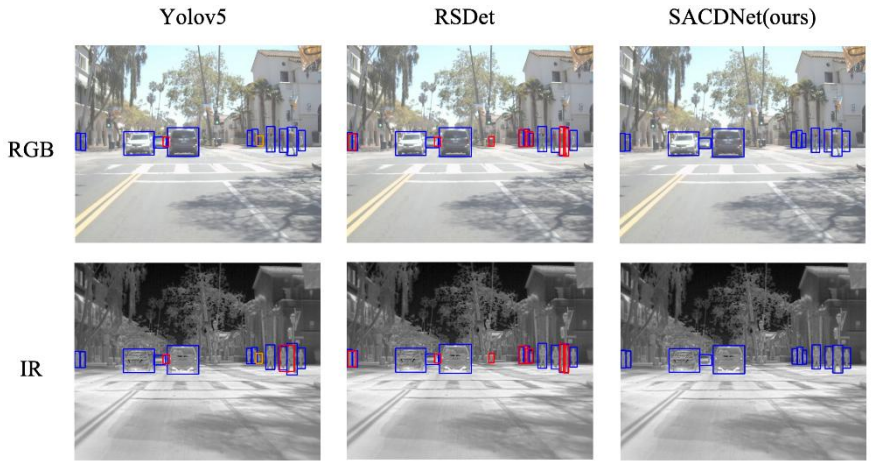

Fig. 7. Visual comparison in distant and dense multi-object scenarios. From left to right: YOLOv5, RSDet, and SACDNet(ours). Our method produces cleaner bounding boxes and eliminates clustered false positives.

Finally, Fig. 7 demonstrates the performance in dense scenarios with distant, small-scale pedestrians. Observing the clustered pedestrians on the right sidewalk, YOLOv5 suffers from missed detections (orange boxes), while RSDet generates a severe cluster of false positives (red boxes) due to the lack of discriminative feature representation for grouped small targets. Unlike the baselines, our proposed method generates sharper and more concentrated feature activations. This precision results in tighter bounding boxes and naturally eliminates duplicate predictions during the NMS process, providing a remarkably accurate set of true positive predictions.

4.6 Parameter Sensitivity Analysis

We further investigate the sensitivity of CMDE to the reduction ratio r in the channel attention bottleneck. As shown in Fig. 8 and Table 3, the model achieves the best mAP at $r = 16$. Smaller values ($r = 4,8$) provide wider bottlenecks but yield slightly lower performance due to overfitting on the limited training data. A larger value ($r = 32$) compresses the channel information too aggressively, resulting in a noticeable drop in mAP. We therefore adopt $r = 16$ in all experiments.

Table 3. Sensitivity analysis of the reduction ratio r in CMDE.

r	mAP	mAP_{50}	mAP_{75}
4	0.369	0.735	0.319
8	0.371	0.737	0.315
16	0.372	0.745	0.315
32	0.366	0.735	0.302

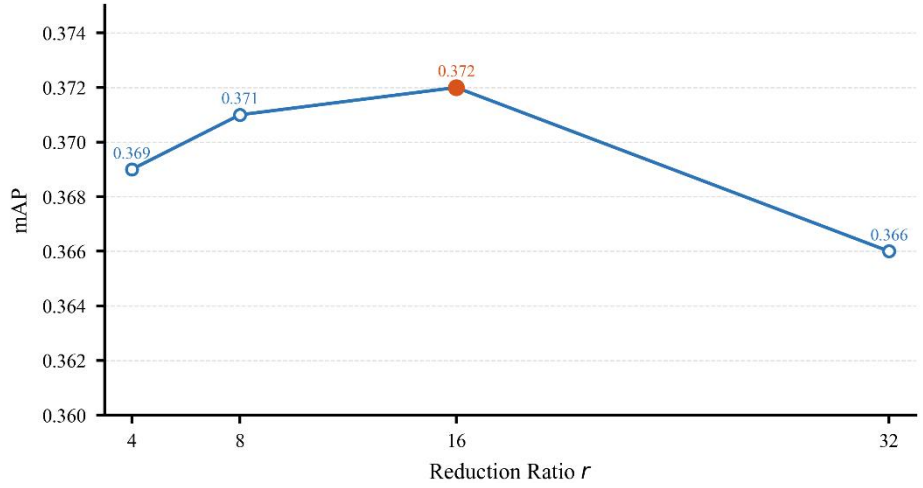

Fig. 8. Effect of the reduction ratio r on detection performance. The model achieves the best mAP at $r=16$.

5 Conclusion

In this paper, we propose a multispectral object detection framework for complex urban scenarios, comprising three components. First, we introduce SAMC, which extends deformable convolution by jointly learning spatial offsets and modulation scalars, enabling the backbone to adapt to thermally diffused target shapes in infrared images. Second, we propose CMDE, which constructs an explicit cross-modal differential map and applies channel attention to adaptively balance the contribution of each modality under varying illumination conditions. Third, we design IGGL, which modulates the bounding box regression penalty according to the estimated illumination score, reducing the influence of unreliable visible annotations during nighttime training. Experiments on the FLIR-align dataset demonstrate that the proposed method achieves an mAP of 0.372, outperforming the two-stream baseline by 2.8% and YOLOv5 by 2.3%. Ablation studies confirm the individual contribution of each component, with the full model achieving the best performance across all metrics.

In future work, we plan to evaluate the proposed method on additional multispectral datasets and explore more lightweight fusion architectures to improve inference efficiency for real-time deployment.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (No. 62403076 and 52472399), the Humanities and Social Science Fund of Ministry of Education (No. 24YJCZH416)

References

1. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., et al.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
2. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: ICCV (2017)
3. FLIR Systems: FLIR thermal dataset for algorithm training. https://www.flir.com/oem/adas/adas-dataset-form/ (2019)
4. Girshick, R.: Fast R-CNN. In: ICCV (2015)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
6. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR (2018)
7. Hwang, S., Park, J., Kim, N., Choi, Y., Kweon, I.S.: Multispectral pedestrian detection: Benchmark dataset and baseline. In: CVPR, pp. 1037–1045 (2015)
8. Jocher, G., et al.: YOLOv5. https://github.com/ultralytics/yolov5 (2020)
9. König, D., Adam, M., Jarvers, C., Layher, G., Neumann, H., Teutsch, M.: Fully convolutional region proposal networks for multispectral person detection. In: CVPR Workshops (2017)
10. Liu, J., Zhang, S., Wang, S., Metaxas, D.N.: Multispectral deep neural networks for pedestrian detection. In: BMVC (2016)
11. Qingyun, F., Dapeng, H., Zhaokui, W.: Cross-modality fusion transformer for multispectral object detection. arXiv preprint arXiv:2111.00273 (2021)
12. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NeurIPS (2015)
13. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss. In: CVPR (2019)
14. Shen, J., Chen, Y., Liu, Y., Zuo, X., Fan, H., Yang, W.: ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection. Pattern Recognition 145, 109913 (2024)
15. Wagner, J., Fischer, V., Herman, M., Behnke, S.: Multispectral pedestrian detection using deep fusion convolutional neural networks. In: ESANN (2016)
16. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: Convolutional block attention module. In: ECCV (2018)
17. Yu, J., Jiang, Y., Wang, Z., Cao, Z., Huang, T.: UnitBox: An advanced object detection network. In: ACM MM (2016)
18. Zhang, D., Ye, L., Han, J., Li, L.: GAFF: Learning where to focus for efficient multispectral pedestrian detection. In: WACV (2021)
19. Zhang, L., Zhu, X., Chen, X., Yang, X., Lei, Z., Liu, Z.: Weakly aligned cross-modal learning for multispectral pedestrian detection. In: ICCV (2019)
20. Zhang, Y., Yu, H., He, Y., Wang, X., Yang, W.: Illumination-guided RGBT object detection with inter-and intra-modality fusion. IEEE Transactions on Instrumentation and Measurement 72, 1–13 (2023)
21. Zhao, T., Yuan, M., Jiang, F., Wang, N., Wei, X.: Removal then selection: A coarse-to-fine fusion perspective for RGB-infrared object detection. IEEE Transactions on Intelligent Transportation Systems 27(2), 2504–2519 (2026)
22. Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D.: Distance-IoU loss: Faster and better learning for bounding box regression. In: AAAI (2020)
23. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable ConvNets v2: More deformable, better results. In: CVPR (2019)

24. Zhou S, Yuan Z, Yang D, et al. Pillarhist: A quantization-aware pillar feature encoder based on height-aware histogram, Proceedings of the computer vision and pattern recognition conference. 2025: 27336-27345.
25. Zhou S, Li L, Zhang X, et al. LiDAR-PTQ: Post-Training Quantization for Point Cloud 3D Object Detection, The Twelfth International Conference on Learning Representations. 2026.
26. Zhou, S., Cao, Y., Nie, J., Fu, Y., Zhao, Z., Lu, X., & Wang, S. Comptrack: Information bottleneck-guided low-rank dynamic token compression for point cloud tracking. In Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 40, No. 16, pp. 13773-13781. 2026
27. Li, T., Wang, G., Yu, B., Liu, Y., & Sun, W. Don't let the information slip away. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8504-8513).2026.
28. Zhou, H., Tang, J., Li, Y., Xiao, C., Hou, L., Ke, Z., & Yao, J. Comem: Compositional concept-graph memory for vision–language adaptation. In The Fourteenth International Conference on Learning Representations. 2026.
29. Liu, H., Peng, N., Chen, Q., & Xiao, C. Affordance-first decomposition for continual learning in video-language understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3908-3919). 2026.
30. Xiao, C., Zhao, C., Ke, Z., & Shen, F. Curiosity-Driven Cooperation for Long-Tailed Multi-Label Learning. Neural Networks. 2026
31. Zhang, J., Song, X., Li, Y., Liang, D., Zhang, Z., & Cai, J. Adaptive Dual Cross-Attention Network for Multispectral Object Detection in Autonomous Driving. Expert Systems with Applications, 132012. 2026.
32. Zhang, J., Xiang, M., Hu, Y., Hao, W., Lei, L., & Yi, K. Multivariate feature learning and associative spatial information enhancement for snow object detection in autonomous driving. Engineering Applications of Artificial Intelligence, 175, 114672.2026.
33. Jiao, R., Zhang, J., Li, C., & Hu, L. Large-kernel spatially parallel feature fusion for monocular 3D perception in autonomous driving. Knowledge-Based Systems, 343, 115998. 2026.
34. Du, R., Li, Z., Zhang, J., Gao, K., & Hu, L.Point Cloud Mapping and Loop Closure Detection Using Superpoint Semantic Graph for Autonomous Driving. IEEE Transactions on Intelligent Transportation Systems. 2026.
35. Zhang, J., Tao, X., Wang, L., Gao, K., Liu, Z., & Hu, L. Predictive Local Multiple Attention For Real-time Risk Assessment in Autonomous Driving. IEEE Transactions on Vehicular Technology. 2026.
36. Zhao, F., Xu, D., Ren, Z., Shao, X., Wu, Q., Liu, Y., ... & Mizuno, K. Mamba-based super-resolution and semi-supervised YOLOv10 for freshwater mussel detection using acoustic video camera: A case study at Lake Izunuma, Japan. Ecological Informatics, 103324. 2025.
37. Zhao, F., Chen, Y., Xi, D., Liu, Y., Wang, J., Tabeta, S., & Mizuno, K. (2025). Enhanced hermit crabs detection using super-resolution reconstruction and improved YOLOv8 on UAV-captured imagery. Marine Environmental Research, 210, 107313. 2025.
38. Zhao, F., Ren, Z., Wang, J., Wu, Q., Xi, D., Shao, X., ... & Mizuno, K. Smart UAV-assisted rose growth monitoring with improved YOLOv10 and Mamba restoration techniques. Smart Agricultural Technology, 10, 100730. 2025.