

# Representation Space Constrained Learning with Modality Decoupling for Multimodal Object Detection

YiKang Shao, Tao Shi\*

**Abstract**—Multimodal object detection has attracted significant attention in both academia and industry for its enhanced robustness. Although numerous studies have focused on improving modality fusion strategies, most neglect fusion degradation, and none provide a theoretical analysis of its underlying causes. To fill this gap, this paper presents a systematic theoretical investigation of fusion degradation in multimodal detection and identifies two key optimization deficiencies: (1) the gradients of unimodal branch backbones are severely suppressed under multimodal architectures, resulting in under-optimization of the unimodal branches; (2) disparities in modality quality cause weaker modalities to experience stronger gradient suppression, which in turn results in imbalanced modality learning. To address these issues, this paper proposes a Representation Space Constrained Learning with Modality Decoupling (RSC-MD) method, which consists of two modules. The RSC module and the MD module are designed to respectively amplify the suppressed gradients and eliminate inter-modality coupling interference as well as modality imbalance, thereby enabling the comprehensive optimization of each modality-specific backbone. Extensive experiments conducted on the FLIR, LLVIP, M3FD, and MFAD datasets demonstrate that the proposed method effectively alleviates fusion degradation and achieves state-of-the-art performance across multiple benchmarks. The code and training procedures will be released at <https://github.com/yikangshao/RSC-MD>.

**Index Terms**—Object Detection, Multimodal learning, Visible and infrared, Modality Under-optimization.

## I. INTRODUCTION

MULTIMODAL object detection[1] has gained significant traction among researchers and engineers in recent years. By incorporating information from multiple modalities and establishing commonalities or complementarities between them, it achieves objectives unattainable through unimodal approaches or addresses novel challenges, garnering attention across diverse fields. Examples include object perception in autonomous driving[2], video surveillance and detection for urban security[3], and vehicle and object recognition in drone aerial imaging[4]. Although significant progress has been made in unimodal object detection, relying solely on visible (VIS) light for detection still suffers from insufficient robustness and generalization under adverse imaging conditions such as weather changes[5], [6]. Infrared (IR) imaging is formed through thermal radiation, and its images are characterized by limited color diversity and low resolution. Using only IR images therefore greatly restricts the representational capacity

of the model[7],[8]. Current studies combine visible and infrared images for object detection. By learning multimodal features, models can capture shared characteristics and fuse complementary semantic information, thus enhancing their representational capacity. For example, when visible images are degraded by adverse weather, infrared contours can supplement missing details, while clear textures in visible images can compensate for indistinct infrared features, leading to more robust and generalizable detection performance[6], [7].

In recent years, this research area has witnessed remarkable progress[9],[10],[11]. Existing approaches can be broadly categorized into two groups: two-stage and one-stage methods. Two-stage methods first perform image-level fusion of multiple modalities[12],[13],[14] and subsequently conduct object detection on the fused images. However, these methods suffer from an inherent limitation—independent optimization of fusion and detection leads to semantic misalignment and potential inconsistency between feature representations.

Researchers have proposed one-stage frameworks that integrate image fusion and object detection into a unified learning process. Recent studies[15],[16],[17] have shown that feature-level multimodal fusion achieves superior performance compared with image-level or decision-level fusion. One-stage methods mainly focus on improving multimodal fusion strategies to enhance detection performance[18],[19],[8] or employ sophisticated attention mechanisms to better exploit cross-modal complementarity[7],[20],[21]. Although these approaches have achieved notable results, they overlook a fundamental yet counterintuitive issue—certain objects that can be accurately detected by unimodal detectors fail to be detected by multimodal ones. The study[5] investigated this phenomenon and attributed it to insufficient learning within unimodal branches during multimodal training, referring to it as the fusion degradation phenomenon.

Although study[5] attempted to mitigate this issue by introducing additional knowledge distillation[22] to strengthen unimodal learning, it did not fundamentally reveal the underlying cause of unimodal learning insufficiency, nor did it provide a theoretical explanation for how unimodal degradation introduces optimization defects in multimodal models. Some studies have also addressed similar issues of multimodal optimization defects. Study[23] suggested that large variations in image acquisition conditions or technical challenges leading to modality degradation can result in extreme modality imbalance, thereby impairing model performance. Studies[24], [25] proposed that incompatible information across modalities may

\* Corresponding author: Tao Shi. Yikang Shao, Tao Shi are with the school of reliability and systems engineering, Beihang University, Beijing, 100191, China. (email: shaoyikang@buaa.edu.cn).

cause fusion conflicts, which constrain model optimization. However, these analyses focus only on modality fusion or data quality and still do not identify the fundamental architectural causes of multimodal optimization defects.

To address this issue, this study provides a theoretical analysis of fusion degradation in multimodal detection, demonstrating that current multimodal object detection architectures exhibit optimization defects. Even well-performing single-modality branches are affected and cannot achieve the performance of models trained on single-modality data.

Specifically, this study theoretically demonstrates two optimization defects in multimodal object detection frameworks. **First**, the gradients of unimodal branch backbones are excessively suppressed by the fusion module, resulting in under-optimization of the unimodal backbones. **Second**, due to varying quality among modalities, this gradient suppression exhibits an amplification effect across modalities: weaker modalities experience stronger gradient suppression, causing the model to overly rely on stronger modalities while neglecting weaker ones, which leads to optimization imbalance among the unimodal branches.

To remedy the optimization deficiencies identified through theoretical analysis, this paper proposes a Representation Space Constrained Learning with Modality Decoupling (RSC-MD) method. The proposed method consists of two major components: the Representation Space Constraint (RSC) module and the Modality Decoupling (MD) module.

Specifically, the RSC module imposes an auxiliary representational constraint on each backbone network to amplify the gradients suppressed by the fusion module, thereby promoting sufficient learning within unimodal backbones. Furthermore, the MD module aims to eliminate cross-modal competitive learning and interference induced by modality coupling. By decoupling the backbone networks of different modalities and enabling their independent optimization, the MD module prevents the optimization deficiencies that arise from gradient suppression and inter-modality competition between strong and weak modalities, thereby mitigating the modality imbalance problem during multimodal learning.

In summary, the contributions of this paper are as follows:

- **This paper theoretically demonstrates the existence of unimodal under-optimization in multimodal object detection:** the multimodal fusion module hinders the optimization of each modality-specific backbone network, resulting in under-optimization of the unimodal branches.
- **This paper theoretically demonstrates the imbalanced optimization defect in multimodal object detection:** weaker modalities experience greater optimization suppression, causing the model to prioritize dominant modalities while neglecting weaker ones, thereby failing to effectively leverage the complementary advantages of multimodality.
- This paper proposes a Representation Space Constraint (RSC) module for unimodal backbone networks. By imposing auxiliary representational learning constraints on each modality backbone, the module amplifies suppressed gradients and promotes sufficient learning within unimodal branches.

- This paper proposes a Modality Decoupling (MD) module for multimodal detection. By employing a modality decoupling strategy, the MD module enables independent optimization of each modality, thereby eliminating inter-modal conflicts.

## II. RELATED WORK

### A. Multimodal Object Detection

In recent years, numerous studies in the field of multimodal object detection have achieved remarkable success. These advancements have been successfully applied across multiple domains, including autonomous driving, robotics engineering, and satellite remote sensing imagery[26],[27],[28],[29], thereby attracting increasing attention to multimodal object detection. According to prevailing taxonomies, multimodal object detection methods are primarily categorized into three classes based on fusion stages: early fusion, intermediate fusion, and late fusion—also referred to as pixel-level, feature-level, and decision-level fusion, respectively. Certain early fusion approaches are termed two-stage detection methods in multimodal object detection. These methods first fuse multimodal images and then perform object detection on the fused result[12],[13],[14],[30],[31],[32],[33],[34]. However, pixel-level fusion generally incurs high computational costs and large model sizes, while often failing to achieve satisfactory performance and inference speed[35]. Moreover, two-stage methods are widely recognized to suffer from optimization conflicts caused by the decoupling of fusion and detection processes[8].

Late-stage or decision-level fusion methods[36] aim to perform detection using independent detectors for each modality and to combine their outputs in order to enhance the robustness of the final results. However, decision-level fusion is constrained by conflicts and imbalanced dependencies among the independent detectors, resulting in performance inferior to that of early and middle fusion approaches.

An increasing number of studies have demonstrated that feature-level fusion methods generally outperform the other two paradigms and have been extensively employed in multimodal object detection research[9],[37],[38]. In this study[9], a Transformer-based architecture was adopted to achieve mid-level feature fusion, and in a subsequent study[38], a cross-modal attention fusion module was introduced to enhance modality fusion and thereby improve detection performance. Reference[39] proposed an uncertainty-aware fusion approach to address calibration errors and modality discrepancies within paired images. With the widespread adoption of Transformer architectures, attention-based fusion methods have garnered considerable research interest. By constructing sophisticated attention mechanisms, these approaches aim to optimize the fusion of visible and infrared modalities, thereby further improving detection accuracy[40],[41],[42],[43],[44],[45],[46]. Similar to Transformer-based designs, Mamba-based modality fusion frameworks[47],[29],[48] have also achieved promising results in multimodal detection. Furthermore, the successful application of knowledge distillation[22] in object detection has inspired subsequent research efforts. Distillation-based frameworks that guide networks to more effectively

extract modality-specific representations and facilitate cross-modal feature fusion have become one of the prevailing research directions in this field[5],[49],[50].

However, these studies primarily focus on exploring more effective modality fusion strategies to enhance the performance of multimodal detection models, while overlooking the phenomenon of fusion degradation inherent in current feature-level fusion methods and the underlying architectural deficiencies behind it.

### B. Modality Conflict and Modality Imbalance in Multimodal Detection

Recent studies have begun to address the issues of modality imbalance and fusion degradation in multimodal detection methods[5],[51]. However, most studies still attribute this problem to fusion conflicts caused by discrepancies among modalities[29]. Reference[52] points out that most existing studies rely on integrating complementary information from different modalities while overlooking the semantic conflicts induced by their inherent discrepancies. To mitigate this issue, it introduces a modality conflict correction approach. The work in[24] argues that indistinguishable intra-modal features can cause single-modality interference and weaken the dominant modality's representation. To address this, a confidence-based strategy is proposed to eliminate such interference. The study in[25] attributes inter-modal heterogeneity to differences in task-related information content within each modality, suggesting that a significant imbalance exists in the amount of information each modality carries. It proposes a dynamic modality information balancing method to alleviate this issue. According to[53], feature-level fusion methods inherently suffer from modality imbalance, and dynamic dropout with threshold masking is introduced to compensate for this problem. Reference[46] also identifies imbalance in multimodal detection and proposes the use of indicative illumination signals to guide the attention computation for mitigation.

Recent studies have identified differences in modality information content and feature representations across modalities, attributing them to the causes of fusion conflicts and imbalanced modality learning. These studies primarily focus on improving fusion strategies to achieve more effective modality integration, thereby alleviating conflicts and imbalance during multimodal learning. Although such efforts have indeed enhanced overall model performance, they have largely overlooked the intrinsic architectural deficiencies of existing detection frameworks. References[5], [54] employ knowledge distillation techniques to strengthen modality-specific feature extraction and mitigate the under-optimization of weaker modalities. However, they fail to provide a theoretical explanation for the underlying cause of such under-optimization. Similarly, the works in[55],[56],[35] also recognize the presence of modality imbalance, but their approaches essentially address the issue by handling data noise or by adaptively adjusting fusion mechanisms according to modality contributions. Studies[57],[58],[51] build on imbalance-related research in the multimodal domain and reveal that an intrinsic disparity exists between strong and weak modalities within multimodal architectures. However, they fail to provide theoretical

evidence from the perspective of the multimodal detection architecture itself.

In summary, recent studies concerning modality conflicts and imbalance have predominantly focused on refining fusion mechanisms, while few have investigated the inherent deficiencies within the architectural design of multimodal detection frameworks.

## III. METHOD

In this section, this paper first provides a theoretical analysis to demonstrate two optimization deficiencies inherent in current multimodal object detection frameworks and subsequently proposes a Representation Space Constrained Learning with Modality Decoupling (RSC-MD) approach to address these deficiencies.

### A. Theoretical Analysis of Defects in Modality Optimization for Multimodal Object Detection

To address defects such as fusion degradation in multimodal detection, this study employs theoretical analysis to elucidate both the origin of these deficiencies and their impact on the overall model. For a given sample  $x_i$ , a multimodal object detection model accepts inputs from two modalities,  $m_1$  and  $m_2$ , such that the sample can be represented as  $x_i = (x_i^{m_1}, x_i^{m_2})$ . In alignment with the current state-of-the-art object detection model YOLO, this work utilizes different layers of the feature pyramid as detection features, commonly including the outputs of layers  $P3$ ,  $P4$ ,  $P5$ , which are collectively abstracted here as  $B_i$ . Similarly, the structure of the feature extraction network is abstracted uniformly; for the input modality  $x_i^{m_1}$ , the resulting feature representation can be expressed as:

$$f_i^{m_1} = B_1(x_i^{m_1}; \theta_1) \quad (1)$$

where,  $f_1^{m_1}$  represents the image features of modality  $m_1$  extracted by the feature extraction layer  $backbone_1$ , and  $\theta_1$  denotes the parameters of  $B_1$ , where the outputs of layers  $P3$ ,  $P4$ , and  $P5$  are included within  $B_1$ .

Similarly, the modal representation of  $x_i^{m_2}$  can be obtained:

$$f_i^{m_2} = B_2(x_i^{m_2}; \theta_2) \quad (2)$$

Consistent with the majority of existing studies, this work employs feature-level intermediate fusion as the modality feature fusion method. Accordingly, the fusion module can be abstractly represented by the following formulation:

$$z_i = W_k^{m_1} f_i^{m_1} + W_k^{m_2} f_i^{m_2} \quad (3)$$

Here,  $f_1^{m_1}$  and  $f_2^{m_2}$  denote the outputs of the aforementioned feature extraction networks, which serve as the inputs to the fusion module. Similar to the backbone networks, the detection network module is abstracted as a complex composite function comprising multiple hierarchical levels. Although different versions and architectures of object detection algorithms may employ distinct computational procedures, its forward computation can be abstractly expressed, in a general sense, as follows:

$$C_i = \Phi(z_i; \theta_\phi) \quad (4)$$

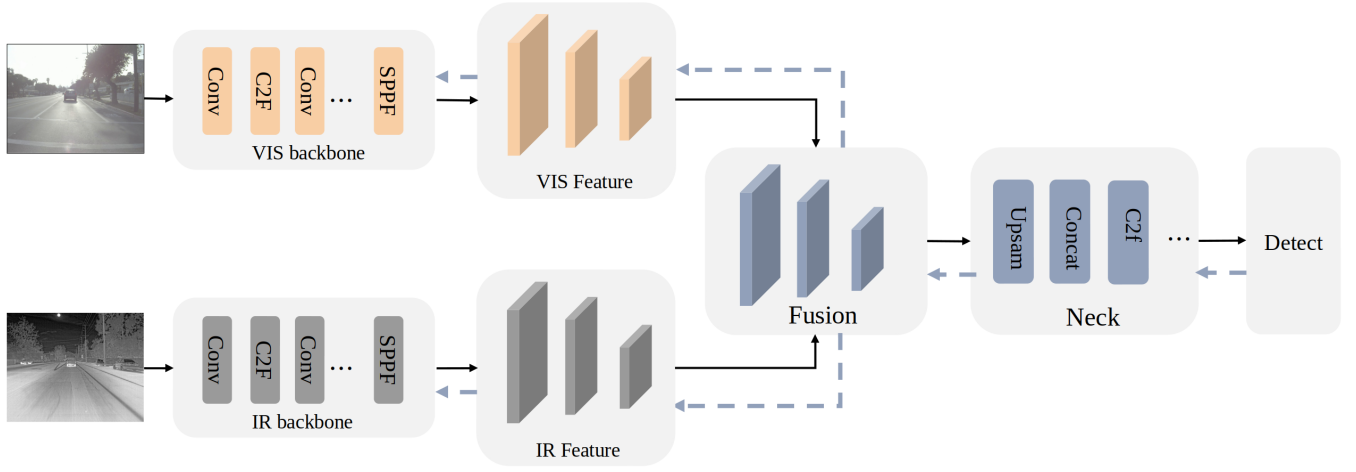


Fig. 1. Architectural Diagram of a Dual-Modal Object Detection Framework Using Naive Addition for Feature Fusion.

where,  $\theta_\phi$  denotes the parameters of the detection module function.

For the loss function, considering the currently widely adopted YOLO model, its loss function can be expressed as follows:

$$L = \lambda_{box} L_{box} + \lambda_{cls} L_{cls} + \lambda_{dfl} L_{dfl} \quad (5)$$

Among these, the classification loss is expressed as:

$$L_{cls} = -\frac{1}{n} \sum_{i=0}^n [y_i \log \sigma(\hat{y}_i) + (1 - y_i) \log(1 - \sigma(\hat{y}_i))] \quad (6)$$

According to the chain rule of gradients, the gradient of the classification loss in multimodal object detection backpropagated to  $backbone_1$  can be expressed as follows:

$$g_{B_1} = \frac{\partial L_{cls}}{\partial f(z)} \frac{\partial f(z)}{\partial z} \frac{\partial z}{\partial f_1} \quad (7)$$

According to the above formula, the gradient propagated backward to  $backbone_1$  can be calculated as:

$$g_{B_1} = \left( \frac{1}{1 + e^{-(W_k^{m1} f_i^{m1} + b_1) - (W_k^{m2} f_i^{m2} + b_2)}} - 1 \right) \theta_\phi, \quad \text{positive} \quad (8)$$

$$g_{B_1} = \left( \frac{1}{1 + e^{-(W_k^{m1} f_i^{m1} + b_1) - (W_k^{m2} f_i^{m2} + b_2)}} \right) \theta_\phi, \quad \text{negative} \quad (9)$$

Here,  $W_k^{m1}$  and  $W_k^{m2}$  represent the corresponding parameters in the fusion module at an abstract mathematical level. The terms positive and negative denote the cases of positive and negative samples, respectively. Correspondingly, the gradient for the unimodal model, denoted as  $g_{Uni}$ , can be expressed as follows:

$$g_{Uni} = \left( \frac{1}{1 + e^{-(W_k^{m1} f_i^{m1} + b_1)}} - 1 \right) \theta_\phi, \quad \text{positive} \quad (10)$$

$$g_{Uni} = \left( \frac{1}{1 + e^{-(W_k^{m1} f_i^{m1} + b_1)}} \right) \theta_\phi, \quad \text{negative} \quad (11)$$

Since the model employs the *SiLU* activation function, the output feature values can be approximately regarded as non-negative. Its formulation is given as follows:

$$SiLU(x) = x * \frac{1}{1 + e^{-x}} \quad (12)$$

It can be inferred that the range of the *logits* values after passing through the activation function is given by:

$$\begin{cases} SiLU(x) > 0, & x > 0 \\ SiLU(x) \approx 0, & x \leq 0 \end{cases} \quad (13)$$

Based on the above results, it can be approximately assumed that the value of  $e^{-(W_k^{m2} f_i^{m2})}$  is less than or equal to 1. By applying this conclusion to the gradient computations propagated back to the backbone in both multimodal and unimodal architectures, the following expressions can be obtained:

$$e^{-(W_k^{m1} f_i^{m1} + b)} \geq e^{-(W_k^{m1} f_i^{m1} + W_k^{m2} f_i^{m2} + b)} \quad (14)$$

Based on the foregoing analysis, the representation for positive samples can be derived as follows:

$$\begin{aligned} \frac{1}{1 + e^{-(W_k^{m1} f_i^{m1} + b)}} - 1 &< \\ \frac{1}{1 + e^{-(W_k^{m1} f_i^{m1} + b) - (W_k^{m2} f_i^{m2})}} - 1 &< 0 \end{aligned} \quad (15)$$

Due to the substantial suppression of gradients in the multimodal backbone networks compared with the unimodal case, the performance of the multimodal detection backbones converges more slowly and less effectively than that of unimodal networks, which severely limits the optimization of the model. Consequently, we identify the first optimization deficiency.

**Optimization Deficiency (1):** In the multimodal architecture, the gradients propagated from the fusion detection module back to the backbone are significantly smaller than those in the unimodal case, resulting in under-optimization of the unimodal branches within the multimodal detection framework.



Due to differences in modality quality, the ease of learning varies across modalities. A widely acknowledged consensus in multimodal learning is that the model tends to prioritize the easier-to-learn modalities, falling into a predicament in which it focuses on the dominant modalities while neglecting the weaker ones, thereby substantially limiting overall model performance. In extreme cases, the performance of a multimodal detection model may even be inferior to that of a unimodal model, which contradicts the original intention of multimodal learning to effectively exploit information from multiple modalities and leverage complementary features.

Further analysis reveals that, under the current multimodal architecture, the gradients applied by the fusion detection module to the backbone are identical across modalities, which leads to imbalance in the optimization of the two modalities. Compared with the unimodal architecture, the gradients propagated to the backbone in multimodal detection include additional modality-specific terms, the magnitude of which can be expressed as follows:

$$e^{-(W_k^{m_2} f_i^{m_2})} < e^{-(W_k^{m_1} f_i^{m_1})}, m_1 = \text{weakModality} \quad (16)$$

$$e^{-(W_k^{m_2} f_i^{m_2})} > e^{-(W_k^{m_1} f_i^{m_1})}, m_2 = \text{weakModality} \quad (17)$$

where,  $m_i = \text{weakmodality}$  denotes that modality  $m_i$  represents the weaker modality, which is more difficult to learn. In contrast, the other modality serves as the dominant modality, being relatively easier to learn and represent. If  $m_2$  corresponds to the easier-to-learn modality, the relationship  $W_k^{m_1} f_i^{m_1} < W_k^{m_2} f_i^{m_2}$  generally holds during representation learning. This is because the dominant modality converges faster and more effectively during optimization, causing its feature weight vector to be closer to the class center, thereby resulting in a larger inner product value[59], [60].

By integrating the above formulations, the expression for the gradient difference between the multimodal and unimodal cases can be derived as follows:

$$\frac{1}{1 + e^{-(W_k^{m_1} f_i^{m_1} + b) - (W_k^{m_2} f_i^{m_2})}} - \frac{1}{1 + e^{-(W_k^{m_1} f_i^{m_1} + b)}} > \frac{1}{1 + e^{-(W_k^{m_2} f_i^{m_2} + b) - (W_k^{m_1} f_i^{m_1})}} - \frac{1}{1 + e^{-(W_k^{m_2} f_i^{m_2} + b)}}, m_1 = \text{weakModality} \quad (18)$$

$$\frac{1}{1 + e^{-(W_k^{m_1} f_i^{m_1} + b) - (W_k^{m_2} f_i^{m_2})}} - \frac{1}{1 + e^{-(W_k^{m_1} f_i^{m_1} + b)}} < \frac{1}{1 + e^{-(W_k^{m_2} f_i^{m_2} + b) - (W_k^{m_1} f_i^{m_1})}} - \frac{1}{1 + e^{-(W_k^{m_2} f_i^{m_2} + b)}}, m_2 = \text{weakModality} \quad (19)$$

Based on the above theoretical analyses, it can be concluded that the weak modality suffers from more pronounced gradient suppression compared with the strong modality, resulting in insufficient optimization of its feature extraction capability.

**Optimization Deficiency (2):** The weak modality undergoes greater gradient suppression than the strong modality, resulting in imbalanced learning across unimodal branches. Consequently, the model tends to prioritize the dominant

modality at the expense of the weaker one, thereby failing to effectively exploit the complementary advantages of multimodal information.

The same optimization deficiencies are present for negative samples. Recent studies in object detection[61], have demonstrated that during the optimization of detection models, there exists an imbalance between positive and negative samples. Existing optimization methods primarily aim to emphasize positive samples, treating them as the dominant contributors. However, according to the aforementioned theoretical analysis, the gradients applied by multimodal detection models for negative samples behave in a manner entirely opposite to that for positive samples. Specifically, in multimodal object detection, negative samples contribute larger gradients compared to unimodal models. This observation contradicts the principle of weighted positive and negative samples and instead imposes additional detrimental effects on model optimization. Therefore, the conclusions drawn for positive samples also hold for negative samples: the increased negative sample gradients similarly impair the proper optimization of the model.

In summary, the above theoretical analysis clearly identifies the origins of fusion degradation in multimodal detection: optimization conflicts exist inherently among modules in multimodal object detection methods. The gradients propagated to the backbone networks of multimodal detection models are smaller than the corresponding unimodal gradients, and modalities of differing learning difficulty are subject to varying degrees of gradient suppression. This results in imbalanced learning across multiple modalities, further constraining the overall optimization of the detection task.

### B. Representation Space Constrained Learning with Modality Decoupling

By comparing the mathematical expressions derived from the theoretical analysis, it can be observed that the interference causing the optimization deficiencies originates from the other modality. This interference does not exist in unimodal learning but persists in multimodal learning due to improper operations involving logarithmic and exponential transformations, which prevent its correct elimination. Specifically, the term  $W_k^{m_2} f_i^{m_2}$  from the other modality is not properly canceled during the derivative computation of  $e^{-(W_k^{m_2} f_i^{m_2})}$  resulting in the two modalities being coupled during training and interfering with the optimization of their respective backbone networks.

Motivated by this observation, this study proposes an architectural innovation that decouples the coupled modalities, allowing each modality to learn representations independently without mutual interference. This study proposes a Representation Space Constrained Learning with Modality Decoupling (RSC-MD) method, as illustrated in Figure 2, which consists of two main components. The first component enhances the feature extraction capability of each unimodal branch backbone by imposing representational learning constraints within the multimodal architecture, wherein each backbone is equipped with an independent detection head and corresponding loss supervision. The second component mitigates optimization deficiencies caused by inter-modality coupling

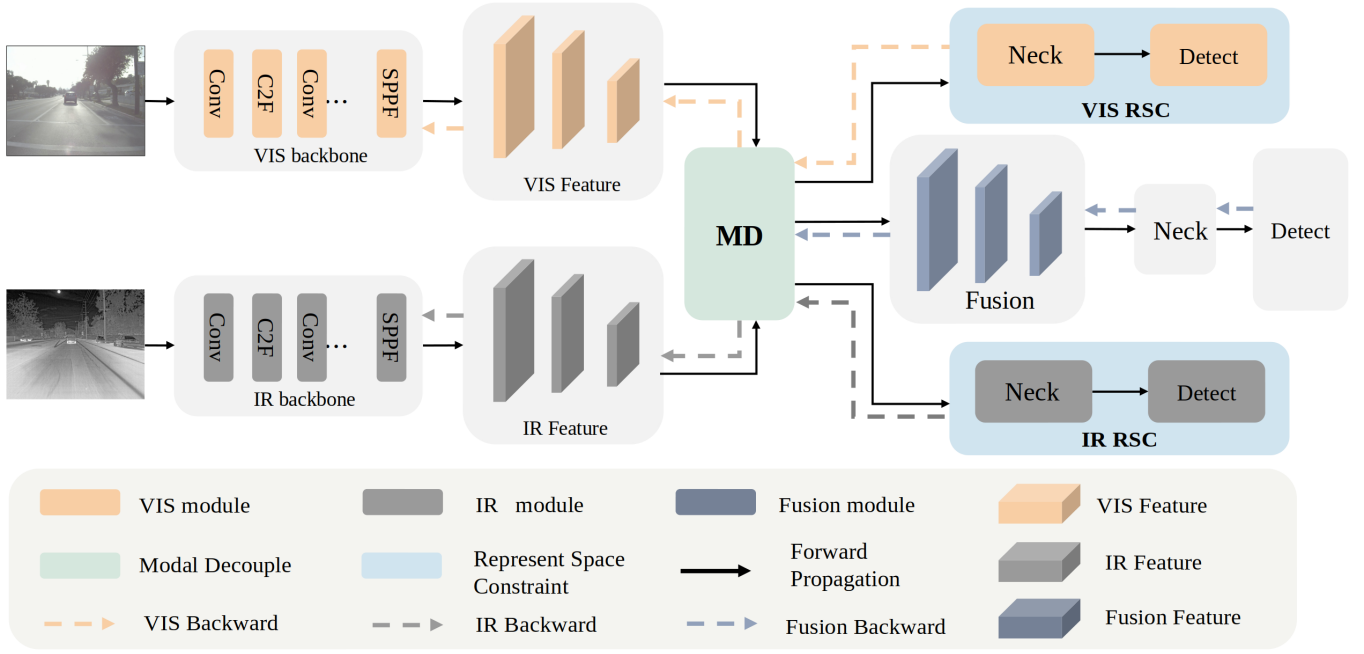


Fig. 2. Architecture Diagram of Representation Space Constrained Learning with Modality Decoupling Framework.

interference through modality decoupling, ensuring that the optimization of each unimodal backbone proceeds independently without interference from other modalities, while enabling the fusion module to effectively integrate multimodal features for joint optimization.

1) *Representation Space Constrained Learning*: To address the under-optimization of unimodal branches in multimodal detection models, this work introduces a Representation-Constrained Supervision (RCS) module, which imposes representation learning constraints on each unimodal backbone. RCS is designed to amplify the backpropagated gradients to strengthen the feature extraction capability of each modality and to ensure that the learned representations remain aligned with the original optimization direction of the corresponding unimodal branch. Specifically, RCS employs two additional detection heads that respectively receive the multi-scale feature maps produced by the backbones of the two modalities. The process can be formulated as follows:

$$\begin{cases} Aux^{m_1} = H(f_i^{m_1}; \theta_{a_1}) \\ Aux^{m_2} = H(f_i^{m_2}; \theta_{a_2}) \end{cases} \quad (20)$$

where,  $Aux^{m_1}$  and  $Aux^{m_2}$  denote the outputs of auxiliary detection heads, while  $H$  represents an abstract detection head module—a multi-level composite function.  $\theta_{a_1}$  and  $\theta_{a_2}$  denote their internal parameters. After the auxiliary heads  $Aux^{m_1}$  and  $Aux^{m_2}$  independently complete their forward computations, the backpropagated gradients are transmitted solely to the corresponding modality-specific backbones. In this way, each backbone receives a targeted representation constraint, guiding its feature space to progressively approach the representation space obtained under unimodal training.

Accordingly, two auxiliary loss functions are added for the two auxiliary detection heads, expressed as:

$$\begin{cases} L_{A_1} = Loss(Aux^{m_1}, Y) \\ L_{A_2} = Loss(Aux^{m_2}, Y) \end{cases} \quad (21)$$

where,  $Loss(\bullet)$  denotes the model loss computation. For consistency, it follows exactly the same formulation as the loss function used in the unimodal architecture.

The total model loss is expressed as:

$$L_{total} = \alpha L_{fusion} + \beta L_{A_1} + \gamma L_{A_2} \quad (22)$$

where,  $\alpha, \beta, \gamma$  denote the constraint coefficients applied to different modalities, and they serve as hyperparameters of the entire model.

The RCS module imposes directional constraints on each unimodal branch backbone through auxiliary detection heads, guiding each backbone network to approach the representation space corresponding to its respective unimodal training. Consistent with prior studies such as [5], linear probing evaluations are conducted on the backbone networks of each modality. The results, shown in Figure 5 and Figure 6, indicate that although the auxiliary detection heads increase the previously suppressed gradients and improve overall model performance, the optimization of each unimodal branch backbone is still influenced by the presence of the other modality. Consequently, some unimodal branch backbones within the multimodal architecture fail to reach the performance levels obtained under unimodal training. This indicates that, although the gradient suppression deficiency is partially alleviated, the optimization deficiencies caused by the imbalance between strong and weak modalities, as well as the excessive negative-sample gradients, remain unaddressed.

To overcome this limitation, a modality decoupling method is proposed to eliminate these optimization deficiencies, ensur-

ing that the backbone networks used for feature extraction in multimodal object detection are fully optimized and capable of providing more effective features for multimodal feature fusion.

2) *Modality-Decoupled Learning*: The purpose of the MD module is to eliminate the residual terms  $e^{-(W_k^{m_2} f_i^{m_2})}$  or  $e^{-(W_k^{m_1} f_i^{m_1})}$  in the polynomial  $\frac{1}{1+e^{-(W_k^{m_1} f_i^{m_1}+b)}-(W_k^{m_2} f_i^{m_2})}$  which, due to modality coupling, are not correctly eliminated during backbone optimization in multimodal learning. In the aforementioned analysis, even after introducing additional representation constraints, optimization interference between modalities still persists. Although auxiliary detection heads can partially mitigate the gradient suppression on positive samples, the optimization deficiencies arising from imbalanced modality learning and excessive negative-sample gradients that undermine the dominance of positive samples remain unresolved. At this stage, the backbone networks of different modalities are jointly constrained by the detection head of the fusion module and the newly introduced auxiliary detection heads. As a result, the model suffers from both imbalanced modality learning and interference from negative samples that weakens the dominance of positive samples, causing each unimodal backbone to remain unable to match or surpass the performance of single-modality models.

Building upon the RSC module, this study proposes a modality decoupling (MD) method, which is applied at the interface between the modality backbone networks and the fusion network to correctly map or directly discard backpropagated gradients. As illustrated in Figure 2, the MD module enables decoupled representation learning for each backbone network, thereby eliminating inter-modality interference. Specifically, by integrating with the auxiliary detection heads in the RSC module, the MD module ensures that each unimodal branch backbone network receives only the optimization gradient signals corresponding to its designated representation space. This achieves modality-decoupled optimization of the backbone networks and removes the interference caused by  $e^{-(W_k^{m_i} f_i^{m_i})}$ .

The detailed procedure of the MD module can be expressed as follows:

$$\begin{cases} \frac{\partial \text{MD}^{(j)}}{\partial \text{MD}_{(i)}} = 0, i \neq j \\ \frac{\partial \text{MD}^{(j)}}{\partial \text{MD}_{(i)}} = 1, i = j \end{cases} \quad (23)$$

Here,  $\text{MD}_{(i)}$  denotes the  $i$ -th input branch of the MD module during the network forward computation, and  $\text{MD}^{(j)}$  denotes the  $j$ -th output branch of the MD module. Specifically,  $i = 0, 1$  correspond to the feature map inputs of the two modalities, while  $j = 0, 1, 2$  correspond to the two auxiliary detection heads  $Aux^{m_1}$ ,  $Aux^{m_2}$ , and the fusion module along with its detection head after passing through the MD module. Through the gradient mapping mechanism of the modality decoupling module, the gradients backpropagated from the newly introduced independent auxiliary detection heads are individually mapped to the corresponding feature-extraction backbones. This enforces a representation learning constraint consistent with unimodal learning, while simultaneously elim-

inating gradient interference from the fusion module and unrelated detection heads via gradient masking. Consequently, decoupled training of the modality-specific backbones is achieved, ensuring that each backbone can independently optimize towards its optimal representation. Furthermore, modality decoupling mitigates the imbalanced optimization between strong and weak modalities, preventing optimization conflicts arising from competitive learning among modalities.

In summary, under the proposed architecture, the multimodal object detection framework adopts the following paradigm: the loss generated by the fusion module and its detection head constrain the representation learning of the fusion subnetwork, whereas the representation learning of the unimodal branch backbones used for modality-specific feature extraction is supervised by the additional independent auxiliary detection heads. This design effectively prevents modality optimization conflicts and imbalanced learning among unimodal branches caused by modality coupling, thereby enhancing the generalization capability and robustness of the model.

#### IV. EXPERIMENTS AND ANALYSIS

##### A. Public Datasets and Evaluation Metrics

1) *FLIR*: The FLIR[62] dataset is a challenging multimodal object detection benchmark encompassing a variety of scenarios, including dark environments, heavy fog, smoke, adverse weather conditions, and glare. It provides both visible and infrared multimodal images, with the primary aim of encouraging researchers to exploit infrared or LiDAR features to complement and enhance visible-spectrum image representations. Although the dataset contains over 20,000 images, some are misaligned; therefore, following[8], we adopt a filtered, aligned subset. This aligned version comprises 5,142 image pairs, partitioned according to the official split, with 4,129 pairs allocated for training and 1,013 pairs for validation. Consistent with prior studies, categories with very few instances, such as dogs, are excluded, and only the people, car, and bicycle classes are retained for the experiments conducted in this work.

2) *LLVIP*: The LLVIP[3] dataset is a manually aligned visible-infrared multimodal object detection dataset. It is specifically designed for low-light detection tasks, and consequently, the majority of images in this dataset exhibit low illumination and predominantly dark environmental conditions. The dataset comprises 15,488 image pairs, of which 12,025 pairs are allocated for training and 3,463 pairs for testing.

3) *M3FD*: The M3FD[64] dataset is a multimodal object detection dataset captured using a binocular optical camera and a binocular infrared sensor, with the aligned visible and infrared images having a resolution of  $1024 \times 768$ . The dataset encompasses multiple scenarios, including daytime, cloudy, and nighttime conditions, and contains a total of 4,200 VIS-IR image pairs with six object categories: person, car, bus, motorcycle, truck, and Lamp, totaling 33,603 instances. However, the dataset does not provide predefined training and testing splits. For fairness, this work adopts the same training-testing partition as reported in this study[8], and all subsequent comparative evaluations are conducted based on the data splits from this reference.

TABLE I  
PERFORMANCE COMPARISON OF DIFFERENT METHODS ON FLIR DATASET.

Modality	Methods	AP50(%)			mean AP50(%)	mean AP75(%)	mean AP50-95(%)
		Person	Car	Bicycle			
VIS	Faster R-CNN	-	-	-	65.0	22.8	30.2
VIS	YOLOV5	-	-	-	67.8	25.9	31.8
VIS	SSD	-	-	-	52.2	-	21.8
IR	Faster R-CNN	-	-	-	74.4	32.5	37.6
IR	YOLOV5	-	-	-	73.9	35.7	39.5
IR	SSD	-	-	-	65.5	32.4	29.6
VIS + IR	CFT [9]	80.4	90.2	61.4	72.9	30.9	37.3
VIS + IR	ICAfusion [21]	81.6	89	66.9	79.2	36.9	40.8
VIS + IR	UniRGB-IR [63]	-	-	-	81.4	40.2	44.1
VIS + IR	EI <sup>2</sup> Det [8]	84.9	89.4	66.3	80.2	-	-
VIS + IR	LIF [5]	-	-	-	-	-	45.2
VIS + IR	RCS-MD (ours)	85.4	91.3	67.7	<b>81.5</b>	<b>47.2</b>	<b>47.8</b>

TABLE II  
PERFORMANCE COMPARISON OF DIFFERENT METHODS ON LLVIP DATASET.

Modality	Methods	mean AP50(%)	mean AP75(%)	mean AP50-95(%)
VIS	Faster R-CNN	91.4	48.0	49.2
VIS	SSD	82.6	31.8	39.8
VIS	YOLOV7 X	90.1	52.7	50.6
VIS	YOLOV10 L	87.1	53.9	50.5
IR	Faster R-CNN	96.1	68.5	61.1
IR	SSD	90.2	57.9	53.5
IR	YOLOV7 X	90.1	52	1
IR	YOLOV10 L	95.1	72.0	63.3
VIS + IR	CFT [9]	97.5	72.9	63.6
VIS + IR	ICAfusion [21]	97.4	70.9	62.7
VIS + IR	UniRGB-IR [63]	96.1	72.2	63.2
VIS + IR	EI <sup>2</sup> Det [8]	98	73.2	63.9
VIS + IR	LIF [5]	-	-	67.9
VIS + IR	RCS-MD (ours)	97.7	<b>81.7</b>	<b>69.5</b>

4) *MFAD*: The MFAD[8] dataset contains visible-infrared image pairs captured under diverse weather conditions and encompasses a wide range of scene categories, including roads, tunnels, overpasses, parks, and parking lots. It covers images acquired under various visibility conditions, such as sunny, cloudy, foggy, and rainy weather, providing richer feature diversity. The dataset consists of 12,370 VIS-IR image pairs and provides an official training and testing split, with 9,879 pairs allocated for training and the remaining 2,473 pairs for testing. Six object categories are annotated in the dataset: car, bus, truck, pedestrian, EbikeRider (electric bicycle riders), and cyclist.

5) *Mean Average Precision(mAP)*: mAP is one of the most representative metrics for evaluating model performance in object detection. It assesses model capability based on classification accuracy of predicted results and the Intersection over Union (IOU) between predicted bounding boxes and ground truth boxes. This paper employs  $mAP_{50}$  and  $mAP_{50-95}$  for evaluation, where  $mAP_{50}$  denotes the average AP across all classes at an IOU threshold of 0.5, while  $mAP_{50-95}$  denotes the average mAP calculated at IOU thresholds ranging from 0.5 to 0.95 in 0.05 increments. Higher mAP values indicate superior model performance.

## B. Implementation Details

This paper conducts optimization using the SGD optimizer, with a default initial learning rate of  $1 \times e^{-2}$ . The momentum is set to 0.937, the learning rate gradually decays to  $1 \times e^{-6}$ , and the weight decay is set to  $1.0 \times e^{-5}$ . Consistent with[8], the training image input size is set to  $640 \times 640$ , and the testing image input size is also set to  $640 \times 640$ . Data augmentation techniques include mosaic data augmentation and random flipping. To ensure fair comparison with current SOTA models, minor parameter adjustments were made across datasets. These adjustments will be detailed in the results comparison section. Unless otherwise specified, the implementation details outlined above are used by default. The implementation details of each module in the network are same as that for the YOLOv8 network.

## C. Comparison with the current state-of-the-art Methods

1) *Results on FLIR*: On the FLIR dataset, we compare the proposed method with five representative prior approaches as well as the latest state-of-the-art (SOTA) models, including both multimodal and unimodal detectors, as summarized in Table I. The proposed RCS-MD achieves SOTA performance on FLIR, obtaining 47.8%  $mAP_{50-95}$  and 81.5%  $mAP_{50}$ , which represents an improvement of approximately 2.6% over



TABLE III  
PERFORMANCE COMPARISON OF DIFFERENT METHODS ON M3FD DATASET.

Modality	Methods	AP50(%)						mean AP50(%)	mean AP75(%)	mean AP50-95(%)
		People	Car	Bus	Motorcycle	Lamp	Truck			
VIS	YOLO_v5 L	72.2	90.5	91.2	73.1	82.9	86.1	82.7	54.9	52.5
VIS	YOLO_v7 X	75.6	91.2	92.0	75.3	84.7	88.4	84.9	56.0	53.2
VIS	YOLO_X L	70.3	88.6	89.0	71.2	79.1	82.9	80.3	53.5	51.3
IR	YOLO_v5 L	81.3	87.2	86.8	73.4	70.5	82.5	80.3	53.2	50.7
IR	YOLO_v7 X	83.9	88.7	87.1	74.0	61.2	86.7	80.5	50.7	49.9
IR	YOLO_X L	79.6	85.3	84.9	67.8	60.1	79.5	74.6	51.3	49.1
VIS + IR	CFT [9]	82.2	91.3	91.6	73.6	85.1	86.1	85.0	57.4	54.5
VIS + IR	ICAFusion [21]	83.0	91.0	92.3	76.5	85.0	88.9	85.1	56.4	53.5
VIS + IR	MMI-Det [18]	80.6	90.7	89.5	70.7	83.3	85.9	83.5	54.6	51.9
VIS + IR	EI <sup>2</sup> Det [8]	82.8	91.4	91.6	77.2	84.7	89.3	86.2	59.2	55.5
VIS + IR	RCS-MD (ours)	<b>85.2</b>	<b>92.4</b>	<b>91.7</b>	73.2	82.3	86.2	85.2	<b>64.7</b>	<b>59.5</b>

the most recent SOTA method[5]. RSC-MD delivers the best AP across all categories (person, car, and bicycle). Specifically, using a model configured with a parameter scale comparable to YOLOv8-M, RSC-MD attains an  $AP_{50}$  of 85.4% for the person class, an  $AP_{50}$  of 91.3% for the car class, and an  $AP_{50}$  of 67.7% for the bicycle class. Although our model maintains the same parameter scale as YOLOv8-M, it surpasses transformer-based detectors with substantially larger parameter capacities. When evaluating the stricter  $meanAP_{75}$  metric, RSC-MD exhibits an even more pronounced performance advantage over all comparison methods.

2) *Results on LLVIP*: On the LLVIP dataset, Table II presents a performance comparison between the proposed RSC-MD method and several recent multimodal approaches, with the best results highlighted in bold. RSC-MD likewise achieves state-of-the-art performance on LLVIP, surpassing the previous SOTA by approximately 1.6% and reaching 69.5% in terms of  $mAP_{50-95}$ . These results consistently exceed those of existing multimodal learning methods, confirming that the optimization deficiencies identified in this work indeed suppress model performance. In terms of  $AP_{50}$ , RSC-MD attains 97.7%, which is slightly lower than the result reported in[8]. This minor discrepancy arises because the study[8] employs a more complex modality fusion mechanism, whereas the proposed method adopts only a naive element-wise addition for feature fusion. Nonetheless, the difference between the two models in  $AP_{50}$  remains small.

3) *Results on M3FD*: Since the M3FD dataset does not provide an official division of the training and testing sets, this work adopts the same experimental setting as the latest study[8] to enable a fair comparison. In that study, 2,100 image pairs out of the 4,200 pairs were used for training, and the remaining 2,100 pairs were used for validation. As shown in Table III, the proposed RSC-MD method achieves state-of-the-art performance in terms of  $mAP_{50-95}$  metric, attaining a 4% improvement over the best existing model under the same setting. For the  $mAP_{50}$  metric, the proposed method obtains the highest performance for the person, car, and bus categories. Although it does not achieve the best performance for the motorcycle and lamp categories, careful analysis indicates that this is caused by two reasons. First, an analysis of the M3FD dataset reveals that its images were captured using an in-

vehicle camera under motion, resulting in substantial motion blur and other negative degradations that severely affect image features, which is unfavorable for multimodal fusion-based detection. Second, the modality fusion method adopted in this work is naive addition, which is the most basic and primitive fusion strategy. Therefore, it may not fully preserve effective features from all modalities, and the noise contained in the modality-specific features may be propagated to the fused representations.

However, these two points are not the focus of this study. The proposed RSC-MD method can be combined with any modality fusion strategy for detection, and thus we do not discuss these aspects in detail.

4) *Results on MFAD*: Table IV presents the detection results on the MFAD dataset. On this dataset, the proposed RSC-MD method also achieves state-of-the-art performance. While obtaining the best  $mAP_{50-95}$ , the method does not reach the highest performance for a few individual categories. Specifically, RSC-MD yields a 3.7% improvement in  $mAP_{50-95}$ , achieving a score of 57%. It attains the best  $AP_{50}$  results for large-object categories such as car, bus, and truck. Although it ranks second for the pedestrian and bicycle categories, the gap compared with the best result is only 0.1%. This slight discrepancy is consistent with the reasons analyzed for the previous datasets, and thus is not reiterated here.

Although the RSC-MD method incurs a slight increase in computational cost during training due to the introduction of two additional detection heads, its overall parameter count remains substantially smaller than that of complex distillation-based approaches and Transformer-based methods with large model capacities. Moreover, during inference, the auxiliary detection heads can be discarded, and only the fusion detection head is retained, resulting in a model whose parameter size is identical to that of the naive-addition baseline and introducing no additional parameters. Compared with complex distillation strategies and parameter-intensive Transformer architectures, the proposed method therefore achieves an efficient and lightweight parameter optimization.

#### D. Ablation Studies

The ablation studies are conducted on the FLIR and LLVIP datasets. We first examine the individual contributions of the

TABLE IV  
PERFORMANCE COMPARISON OF DIFFERENT METHODS ON MFAD DATASET.

Modality	Methods	AP50(%)						mean AP50(%)	mean AP75(%)	mean AP50-95(%)
		Car	BUs	Tru	Ped	Ebi	Cyc			
VIS	YOLOV7_X	89.5	84.9	82.6	65.8	71.9	42.4	72.9	52.7	48.8
VIS	YOLOV10_L	88.1	85.1	80.4	61.8	66.8	44.1	71.1	53.3	48.9
VIS	YOLOX_L	80.3	73.5	75.2	57.1	65.3	41.5	68.6	52.2	46.8
IR	YOLOV7_X	81.7	80.6	72.5	68.1	62.1	34.0	66.5	42.4	40.6
IR	YOLOV10_L	81.1	79.9	72.4	63.3	59.1	38.4	65.7	43.9	41.8
IR	YOLOX_L	79.6	76.4	71.3	61.1	56.6	37.6	63.2	41.5	39.2
VIS + IR	CFT [9]	89.5	89.0	84.5	75.9	75.3	51.0	77.8	56.8	52.5
VIS + IR	ICAFusion [21]	89.2	88.6	86.1	70.9	75.3	50.2	77.6	57.2	52.7
VIS + IR	TINet [65]	84.1	84.3	77.2	67.6	63.8	42.6	69.1	45.5	43.6
VIS + IR	MMI-Det [18]	89.6	88.2	86.0	76.9	75.2	45.2	76.9	55.9	51.4
VIS + IR	EI <sup>2</sup> Det [8]	89.8	88.8	85.7	78.6	75.5	53.3	79.0	58.0	53.3
VIS + IR	RCS-MD (ours)	<b>90.9</b>	<b>90.6</b>	<b>86.9</b>	78.2	<b>76.9</b>	53.2	<b>79.4</b>	<b>62.0</b>	<b>57.0</b>

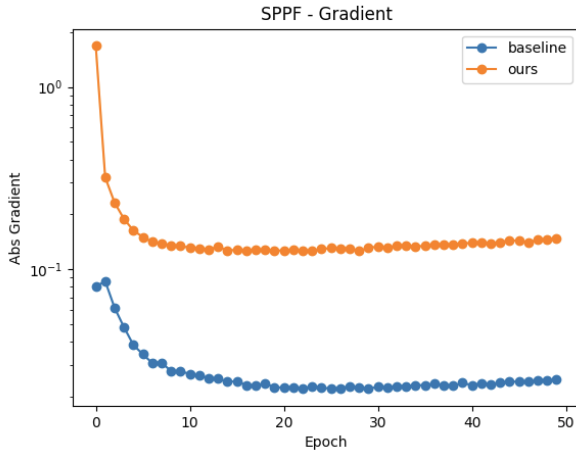


Fig. 3. Visual comparison of gradients in SPPF layers of visible modality.

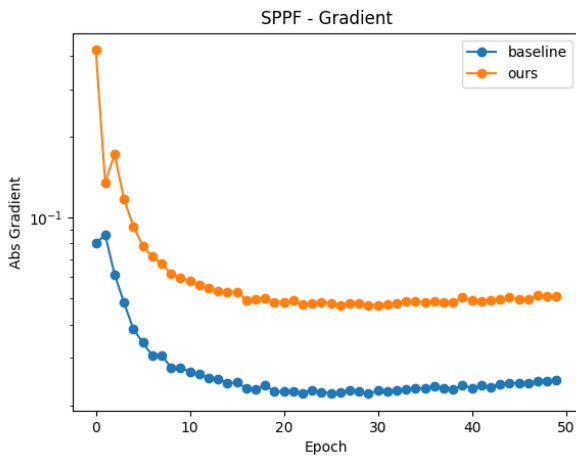


Fig. 4. Visual comparison of gradients in SPPF layers of infrared modality.

two modules and subsequently investigate the hyperparameters on the FLIR dataset. Consistent with widely adopted practice in prior studies, this work employs naive addition as the baseline, whose architecture is illustrated in Figure 1.

TABLE V  
ABLATION STUDY ON FLIR DATASET.

RSC	MD	mean AP50(%)	mean AP75(%)	mean AP50-95(%)
-	-	75.5	42.1	43.5
✓	-	79.9	45	45.9
✓	✓	<b>81.5</b>	<b>47.2</b>	<b>47.8</b>

TABLE VI  
ABLATION STUDY ON LLVIP DATASET.

RSC	MD	mean AP50(%)	mean AP75(%)	mean AP50-95(%)
-	-	96.9	75.8	65.9
✓	-	97.3	77.9	67.3
✓	✓	<b>97.7</b>	<b>81.7</b>	<b>69.5</b>

The RSC and MD modules are then progressively integrated into the baseline to assess the respective contributions of each component. It is noted that the modality decoupling module must be used in conjunction with the RSC module; when applied in isolation, the modality decoupling module yields zero gradients for the backbone networks and thus renders them untrainable. Therefore, the case of using the modality decoupling module alone is not considered. For clarity of presentation, the model using naive addition is referred to as the *baseline* model. When the RSC module is added to the baseline, the resulting model is denoted as *baseline<sub>RSC</sub>*. Finally, when the MD module is incorporated into the *baseline<sub>RSC</sub>* model, it is referred to as the *RSCMD* model.

1) *Ablation Study on the FLIR Dataset:* To evaluate the effectiveness of the RSC-MD method, the two modules were incrementally incorporated into the baseline model to verify their impact on model performance. As shown in Table V, the *baseline<sub>RSC</sub>* model, which includes the RSC module, achieves an improvement of 4.4% in  $mAP_{50}$ , 2.9% in  $mAP_{75}$ , and 2.4% in  $mAP_{50-95}$  compared with the *baseline* model. The *RSCMD* model, relative to the *baseline* model, attains increases of 6% in  $mAP_{50}$ , 5.1% in  $mAP_{75}$ , and 4.3% in  $mAP_{50-95}$ . Compared with the *baseline<sub>RSC</sub>* model, *RSCMD* exhibits gains of 1.6% in  $mAP_{50}$ , 1.8% in  $mAP_{75}$ ,

and 1.9% in  $mAP_{50-95}$ . These results indicate that the sequential addition of the RSC and MD modules progressively enhances the model performance, thereby validating the effectiveness of the proposed method.

2) *Ablation Study on the LLVIP Dataset*: Similar to the experiments conducted on the FLIR dataset, the two modules were incrementally incorporated into the baseline model to evaluate their effects. As shown in Table VI, the  $baseline_{RSC}$  model on the LLVIP dataset achieves a marginal improvement of 0.4% in  $mAP_{50}$  compared with the baseline model, while demonstrating increases of 2.1% in  $mAP_{75}$  and 1.4% in  $mAP_{50-95}$ . The  $RSCMD$  model, relative to the baseline model, attains improvements of 0.8% in  $mAP_{50}$ , 5.9% in  $mAP_{75}$ , and 3.6% in  $mAP_{50-95}$ . Compared with the  $baseline_{RSC}$  model,  $RSCMD$  further improves performance by 0.4% in  $mAP_{50}$ , 3.8% in  $mAP_{75}$ , and 2.2% in  $mAP_{50-95}$ . These results demonstrate that the sequential addition of the RSC and MD modules consistently enhances model performance on this dataset, thereby validating the effectiveness of the proposed approach.

3) *Ablation Study of Hyperparameters*: This paper investigates the impact of different parameters on model performance within the FLIR dataset, as illustrated in the Table VII. This study finds that, unlike classification tasks where increasing the loss weight typically improves performance, excessively large hyperparameters in dense prediction tasks such as object detection can cause the optimization direction to deviate from the optimum, thereby degrading model performance. Analysis reveals that unlike the imbalance observed in classification tasks, the visible and infrared images studied here exhibit relatively smaller modal differences compared to typical classification tasks such as image and speech. The two modalities are comparatively similar. Consequently, significant differences exist in both data distribution and scale compared to classification tasks. Therefore, in the field of multimodal detection, it is appropriate to seek the optimal model performance through small adjustments of hyperparameters.

#### E. Regression verification of theoretical analysis

Through theoretical analysis, two optimization deficiencies are identified in this study: (1) the backbone of a unimodal branch within a multimodal detection model experiences substantial gradient suppression due to optimization conflicts, resulting in under-optimization of the unimodal branch; and (2) the weak modality is subjected to greater gradient suppression than the strong modality, leading to imbalanced optimization across unimodal branches.

In this section, this study experimentally validates the two defects to confirm the consistency between the experimental results and the theoretical derivations.

Regarding **Optimization Deficiency (1)**, this paper conducted a gradient visualization study at the SPPF layer, as shown in Figures 3 and 4. It can be observed from the figures that the gradient magnitude at the SPPF layer of the unimodal branch backbone, when employing the proposed method, is several times larger than that obtained using the baseline method with naive addition fusion. This observation is consistent with the conclusion of **Optimization Deficiency (1)**,

TABLE VII  
ABLATION STUDY OF HYPERPARAMETERS ON FLIR DATASET.

$\alpha$	$\beta$	$\gamma$	mean AP50(%)	mean AP75(%)	mean AP50-95(%)
1	1	1	81.5	47.2	47.8
1	2	1	81.4	47.1	47.5
1	2	2	81.7	46.0	47.4
1	4	4	81.6	46.3	47.4
1	5	5	80.2	46.6	47.1
2	2	3	81.2	46.4	47.3
3	1	1	81.9	45.9	47.3

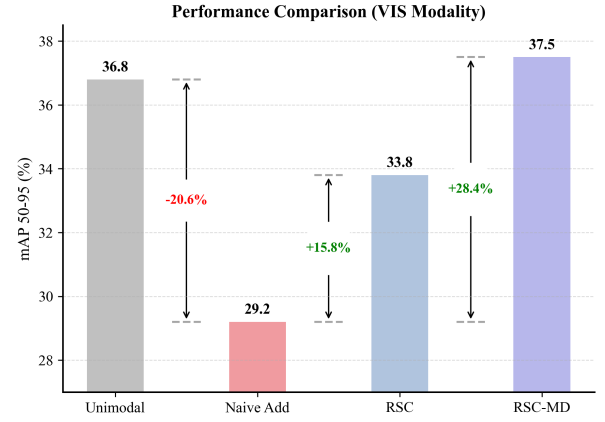


Fig. 5. Performance Comparison (VIS Modality).

wherein the unimodal branch backbone suffers from gradient suppression, resulting in gradients that are significantly smaller than those of a single-modality backbone.

For **Optimization Deficiency (2)**, this study employs a linear evaluation consistent with the study[5] and evaluates the performance of the corresponding unimodal branch. As shown in Figures 5 and 6, the VIS modality constitutes the weaker modality within the FLIR dataset. When naive addition is employed as the feature fusion method, its performance was reduced by 20.6% due to gradient suppression. The IR modality experienced a performance reduction of 10.5% due to gradient suppression. These results are consistent with our conclusions: the greater the gradient suppression experienced by the weaker modality, the more significant the impact on its performance.

Although certain studies[5] have conducted linear evaluations from a unimodal perspective, they only reveal the under-optimization of unimodal branches based on the naive addition baseline, without validating or comparing the performance improvements of unimodal branches under the improved methods. In contrast, this study not only identifies the root causes of unimodal learning deficiencies through theoretical analysis but also evaluates the post-improvement performance of both VIS and IR modalities.

As illustrated in Figures 5 and 6, the introduction of RSC and MD modules yields performance gains of 15.8% and 28.4% in the VIS modality, and 3.6% and 11.2% in the IR modality, relative to naive addition. This indicates that the VIS modality exhibits a larger improvement, consistent with the prior analysis that VIS experiences greater suppression under

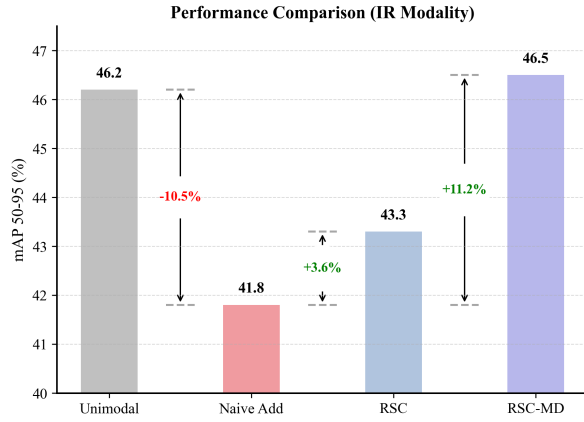


Fig. 6. Performance Comparison (IR Modality).

modal coupling and, consequently, benefits more substantially from the enhancements. Furthermore, the comparison between Figures 3 and 4 shows that the VIS modality in Figure 3 undergoes more severe gradient suppression than the IR modality in Figure 4, corroborating **Optimization Deficiency (2)**.

Additionally, Figures 5 and 6 demonstrates that the RSC module alone cannot completely eliminate interference arising from modal coupling. Specifically, for the IR modality branch trained on the FLIR dataset, the incorporation of RSC to amplify suppressed gradients results in a 3.6% performance improvement; however, due to interference from excessive gradients contributed under negative sample scenarios, the model still fails to reach the performance level obtained when trained exclusively on unimodal data. This observation aligns with the conclusions discussed in the RSC methodology.

## V. CONCLUSION

This study provides a theoretical analysis of feature degradation in multimodal detection and reveals two key optimization deficiencies: under-optimization of unimodal branches and imbalanced modality learning. To address these issues, we propose a representation space constrained learning with modality decoupling approach, comprising an RSC module and an MD module, which mitigate gradient suppression from modality coupling and imbalance, enabling full optimization of each modality backbone. Extensive experiments on the FLIR, LLVIP, M3FD, and MFAD datasets demonstrate that the proposed RSC-MD consistently achieves state-of-the-art performance, improving  $mAP_{50-95}$  by 2.6% on FLIR and 1.6% on LLVIP, while maintaining a lightweight parameter configuration. Importantly, the proposed method is independent of the specific modality fusion strategy, allowing seamless integration with different fusion approaches and offering broad applicability as well as substantial potential for further exploration in multimodal object detection.

## REFERENCES

[1] N. M. Sadic, W. A. Shalaby, S. El-Dolil, F. E. Abd El-Samie, M. I. Dessouky, and S. M. Elkaffas, "Utilization of infrared images for object detection: a survey," *Journal of Optics*, pp. 1–8, 2025.

[2] M. Person, M. Jensen, A. O. Smith, and H. Gutierrez, "Multimodal fusion object detection system for autonomous vehicles," *Journal of Dynamic Systems, Measurement, and Control*, vol. 141, no. 7, p. 071017, 2019.

[3] X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou, "Llvip: A visible-infrared paired dataset for low-light vision," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021, pp. 3489–3497.

[4] C. Rana *et al.*, "Artificial intelligence based object detection and traffic prediction by autonomous vehicles—a review," *Expert Systems with Applications*, vol. 255, p. 124664, 2024.

[5] T. Zhao, B. Liu, Y. Gao, Y. Sun, M. Yuan, and X. Wei, "Rethinking multi-modal object detection from the perspective of mono-modality feature learning," *arXiv preprint arXiv:2503.11780*, 2025.

[6] C. Xia, X. Wang, F. Lv, X. Hao, and Y. Shi, "Vit-comer: Vision transformer with convolutional multi-scale feature interaction for dense predictions," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 5493–5502.

[7] H. Li, L. Xiao, L. Cao, D. Wu, Y. Liu, Y. Li, Y. Zhang, and H. Bao, "Crossmodalnet: A dual-modal object detection network based on cross-modal fusion and channel interaction," *Expert Systems with Applications*, p. 129677, 2025.

[8] K. Hu, Y. He, Y. Li, J. Zhao, S. Chen, and Y. Kang, "Ei<sup>2</sup>det: Edge-guided illumination-aware interactive learning for visible-infrared object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 7, pp. 7101–7115, 2025.

[9] Q. Feng, D. Hu, and Z. Wang, "Cross-modality fusion transformer for multispectral object detection," *arXiv preprint arXiv:2111.00273*, 2021.

[10] J. S. Yun, S. H. Park, and S. B. Yoo, "Infusion-net: Inter-and intra-weighted cross-fusion network for multispectral object detection," *Mathematics*, vol. 10, no. 21, p. 3966, 2022.

[11] Y. Zhang, H. Yu, Y. He, X. Wang, and W. Yang, "Illumination-guided rgbt object detection with inter-and intra-modality fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–3, 2023.

[12] J. Yao, Y. Zhao, Y. Bu, S. G. Kong, and J. C.-W. Chan, "Laplacian pyramid fusion network with hierarchical guidance for infrared and visible image fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 4630–4644, 2023.

[13] W. Tang, F. He, Y. Liu, Y. Duan, and T. Si, "Datfuse: Infrared and visible image fusion via dual attention transformer," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 7, pp. 3159–3172, 2023.

[14] X. Li, Y. Li, H. Chen, Y. Peng, and P. Pan, "Ccafusion: Cross-modal coordinate attention network for infrared and visible image fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 2, pp. 866–881, 2024.

[15] C. Li, D. Song, R. Tong, and M. Tang, "Multispectral pedestrian detection via simultaneous detection and segmentation," *arXiv preprint arXiv:1808.04818*, 2018.

[16] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," *arXiv preprint arXiv:1611.02644*, 2016.

[17] A. Wolpert, M. Teutsch, M. S. Sarfraz, and R. Stiefelhausen, "Anchor-free small-scale multispectral pedestrian detection," *arXiv preprint arXiv:2008.08418*, 2020.

[18] Y. Zeng, T. Liang, Y. Jin, and Y. Li, "Mmi-det: Exploring multi-modal integration for visible and infrared object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 11, pp. 11 198–11 213, 2024.

[19] F. Yang, B. Liang, W. Li, and J. Zhang, "Multidimensional fusion network for multispectral object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[20] Y. Cao, J. Bin, J. Hamari, E. Blasch, and Z. Liu, "Multimodal object detection by channel switching and spatial attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 403–411.

[21] 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.

[22] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[23] C. Tian, C. Yang, G. Zhu *et al.*, "Learning a robust rgb-thermal detector for extreme modality imbalance," *Pattern Recognition Letters*, 2025.

[24] Y. Wang, S. Wei, S. Xu *et al.*, "Confidence-driven unimodal interference removal for enhanced multimodal object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.



- [25] H. Wang, S. Qu, Z. Qiao *et al.*, “Kcdnet: Multimodal object detection in modal information imbalance scenes,” *IEEE Transactions on Instrumentation and Measurement*, 2024.
- [26] N. Lyu, J. Zhao, P. Liu *et al.*, “Rcqvfusion: Radar and camera sensor fusion with joint quantization for robust object detection in autonomous driving,” *IEEE Transactions on Instrumentation and Measurement*, 2025.
- [27] D. Feng, A. Harakeh, S. L. Waslander *et al.*, “A review and comparative study on probabilistic object detection in autonomous driving,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 9961–9980, 2021.
- [28] S. Jain, J. Gallagher, T. Treat *et al.*, “Multispectral rgb-lwir fusion with yolo for autonomous object detection in low-cost mobile robotics under variable lighting conditions,” *Journal of Student-Scientists’ Research*, vol. 7, 2025.
- [29] M. Zhou, T. Li, C. Qiao *et al.*, “Dmm: Disparity-guided multispectral mamba for oriented object detection in remote sensing,” *IEEE Transactions on Geoscience and Remote Sensing*, 2025.
- [30] J. Zhang, J. Lei, W. Xie *et al.*, “Superyolo: Super resolution assisted object detection in multimodal remote sensing imagery,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.
- [31] W. Zhao, S. Xie, F. Zhao *et al.*, “afusion: Infrared and visible image fusion via meta-feature embedding from object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 955–13 965.
- [32] Y. Yang, J. Liu, S. Huang *et al.*, “Infrared and visible image fusion via texture conditional generative adversarial network,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 12, pp. 4771–4783, 2021.
- [33] Z. Wang, X. Li, S. Yu *et al.*, “Vsp-fuse: Multifocus image fusion model using the knowledge transferred from visual saliency priors,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 6, pp. 2627–2641, 2022.
- [34] J. Liu, X. Li, Z. Wang *et al.*, “Promptfusion: Harmonized semantic prompt learning for infrared and visible image fusion,” *IEEE/CAA Journal of Automatica Sinica*, 2024.
- [35] Z. Wang, X. Liao, J. Yuan *et al.*, “Emcformer: Equalized multi-modal cues fusion transformer for remote sensing visible-infrared object detection under long-tailed distribution,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.
- [36] J. Yao, Y. Zhang, F. Liu *et al.*, “Object detection based on decision level fusion,” in *2019 Chinese Automation Congress (CAC)*, 2019, pp. 3257–3262.
- [37] Z. Cao, H. Yang, J. Zhao *et al.*, “Attention fusion for one-stage multispectral pedestrian detection,” *Sensors*, vol. 21, no. 12, p. 4184, 2021.
- [38] Q. Feng and Z. Wang, “Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery,” *Pattern Recognition*, vol. 130, p. 108786, 2022.
- [39] J. U. Kim, S. Park, and Y. M. Ro, “Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1510–1523, 2022.
- [40] S. Hu, F. Bonardi, S. Bouchafa *et al.*, “Rethinking self-attention for multispectral object detection,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 11, pp. 16 300–16 311, 2024.
- [41] H. Fu, J. Yuan, G. Zhong *et al.*, “Cf-deformable detr: An end-to-end alignment-free model for weakly aligned visible-infrared object detection,” in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2024, p. 758.
- [42] X. Zeng, G. Liu, J. Chen *et al.*, “Efficient multimodal object detection via coordinate attention fusion for adverse environmental conditions,” *Digital Signal Processing*, vol. 156, p. 104873, 2025.
- [43] X. Huang and G. Ma, “Cross-modality object detection based on detr,” *IEEE Access*, 2025.
- [44] G. Li, G. Ren, J. Wang *et al.*, “Multimodal fusion transformer network for multispectral pedestrian detection in low-light condition,” *Scientific Reports*, vol. 15, no. 1, p. 18778, 2025.
- [45] Y. Zhang, H. Gao, F. Sohel *et al.*, “Multi-modal stream focusing salient object detection based on visible-infrared complementary fusion,” *IEEE Transactions on Instrumentation and Measurement*, 2025.
- [46] Z. Xiong, Z. Yao, X. Liu *et al.*, “Efficient multispectral object detection with attentive feature aggregation leveraging zero-shot implicit illumination guidance,” *Information Fusion*, vol. 118, p. 102939, 2025.
- [47] H. Zhu, W. Dong, L. Yang *et al.*, “Wavemamba: Wavelet-driven mamba fusion for rgb-infrared object detection,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 11 219–11 229.
- [48] C. Liu, X. Ma, X. Yang *et al.*, “Como: Cross-mamba interaction and offset-guided fusion for multimodal object detection,” *Information Fusion*, p. 103414, 2025.
- [49] T. Wang, H. Wang, Y. Zhu *et al.*, “Infrared-visible object detection via distillation-fermentation dual processing,” *IEEE Signal Processing Letters*, 2025.
- [50] Z. Chen, Y. Qian, X. Yang *et al.*, “Amfd: Distillation via adaptive multi-modal fusion for multispectral pedestrian detection,” *IEEE Transactions on Multimedia*, 2025.
- [51] H. R. Medeiros, D. Latortue, E. Granger *et al.*, “Mixed patch visible-infrared modality agnostic object detection,” in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025, pp. 9023–9032.
- [52] X. He, C. Tang, X. Zou *et al.*, “Multispectral object detection via cross-modal conflict-aware learning,” in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 1465–1474.
- [53] Z. Wang and Q. Zhang, “Real-time aerial multispectral object detection with dynamic modality-balanced pixel-level fusion,” *Sensors*, vol. 25, no. 10, p. 3039, 2025.
- [54] L. Zhang, J. Zhang, B. Wang *et al.*, “Yolo-mslite: Lightweight multi-spectral object detection algorithm with feature channel-wise knowledge distillation for autonomous vehicles,” *IEEE Transactions on Vehicular Technology*, 2025.
- [55] X. Han, Z. Qu, and S. Xia, “A method for noise-suppressed multimodal feature integration in urban scene detection,” *Information Processing & Management*, vol. 62, no. 6, p. 104290, 2025.
- [56] Y. Xiu and X. Tong, “Dual-layer cross-modal alignment recommendation based on the diffusion model,” *Information Fusion*, vol. 125, p. 103472, 2026.
- [57] S. Wang, Z. Xu, and Y. Lin, “Multi-stage training and fusion method for imbalanced multimodal uav remote sensing classification,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.
- [58] J. Jang, C. Park, H. Kim *et al.*, “Multispectral object detection enhanced by cross-modal information complementary and cosine similarity channel resampling modules,” in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025, pp. 9437–9446.
- [59] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning approach for deep face recognition,” in *European conference on computer vision*. Cham: Springer, 2016, pp. 499–515.
- [60] Wen, Yandong and Zhang, Kaipeng and Li, Zhifeng and Qiao, Yu, “A comprehensive study on center loss for deep face recognition,” *International Journal of Computer Vision*, vol. 127, no. 6, pp. 668–683, 2019.
- [61] Y. Cao, K. Chen, C. C. Loy, and D. Lin, “Prime sample attention in object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 583–11 591.
- [62] 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.
- [63] M. Yuan, B. Cui, T. Zhao, J. Wang, S. Fu, X. Yang, and X. Wei, “Unirgb-ir: A unified framework for visible-infrared semantic tasks via adapter tuning,” in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 2409–2418.
- [64] J. Liu, X. Fan, Z. Huang, G. Wu, L. Liu, and R. Zhong, “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.
- [65] Y. Zhang, H. Yu, Y. He, X. Wang, and W. Yang, “Illumination-guided rgbt object detection with inter-and intra-modality fusion,” *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–13, 2023.