Journal of Computer Technology and Software

ISSN:2998-2383

Vol. 3, No. 5, 2024

RT-DETR-Based Multimodal Detection with Modality Attention and Feature Alignment

Yujia Lou
University of Rochester, Rochester, USA
sharonlou25@gmail.com

Abstract: Infrared-visible fusion object detection plays a vital role in visual perception under complex environments. However, existing methods still face challenges in feature alignment, modality complementarity, and detection accuracy. To address these issues, this paper proposes a multimodal object detection method based on an improved RT-DETR. A dual-branch feature extraction network is designed to process infrared and visible images separately, and a modality attention mechanism is introduced to enhance cross-modal information interaction. In addition, a feature alignment loss is employed to optimize the fusion process and improve the model's adaptability to different modalities. Experimental results show that the proposed method achieves superior performance on multiple benchmark datasets. Compared to traditional single-modal approaches, the improved RT-DETR shows higher mAP@50 and mAP@95 scores and demonstrates greater robustness under challenging lighting conditions. Compared with existing multimodal detection methods, the proposed model maintains high detection accuracy while improving class discrimination and reducing false positives and missed detections, validating its effectiveness in multimodal visual perception tasks.

Keywords: Infrared and visible light fusion, target detection, RT-DETR, modal attention

1. Introduction

With the rapid development of computer vision and deep learning methods, object detection has become a core focus in image understanding. By automatically identifying and locating objects in complex environments, this technology is widely applied in surveillance, autonomous driving, and industrial inspection [1]. However, in traditional visible-light image-based detection, issues such as ambient lighting, shadows, and occlusion can dramatically degrade image quality, leading to reduced accuracy. In these circumstances, relying solely on visible-light images often falls short of achieving high reliability and robustness. As a result, multi-source information fusion — particularly the combination of visible-light and infrared images—has been attracting growing attention.

Infrared images capture the thermal radiation of objects under low or even no light conditions. They provide stable and continuous observations of object shapes and positions, which is especially advantageous for nighttime surveillance and low-visibility environments. Nevertheless, infrared images generally lack detailed textures and offer lower resolution than visible-light images. Depending solely on infrared data can thus result in reduced accuracy. Consequently, fusing the detailed features from visible-light images with the salient contours in infrared images can improve detection performance and localization accuracy in various complex scenarios. How to make full use of these multimodal datasets, so that the fused information delivers maximum value for object detection, remains a critical topic in both academic research and engineering practice [2].

To achieve real-time and precise multimodal object detection, numerous deep learning network designs have emerged in recent years. Among them, RT-DETR (Real-Time Detection Transformer) is known for balancing real-time performance with high accuracy. By employing a Transformer encoder-decoder structure and parallel computing strategies, it increases inference speed while maintaining satisfactory detection outcomes. However, directly applying RT-DETR to visible-light and infrared fusion still faces hurdles, including limited fusion mechanisms, insufficient feature extraction, and inadequate suppression of multimodal noise. Therefore, improving and optimizing RT-DETR to meet the requirements of multimodal imaging — such as feature alignment, detail extraction, and noise suppression — represents work of both scientific novelty and practical importance [3].

Such improvements not only involve the thoughtful design of fusion modules within the network architecture to facilitate interactive feature augmentation between infrared and visible-light data, but also require consideration of the differences in time, space, scale, and noise among various modalities [4]. By refining fusion strategies, the model can better adapt to diverse object shapes, background changes, and extreme lighting conditions. This leads to higher accuracy and faster detection, and enhances the robustness of deep learning models in real-world settings. Infrared-visible fusion detection can be broadly applied in national defense, aerospace, traffic safety, and emergency response, playing an indispensable role in public safety and critical infrastructure monitoring [5].

Building on this background and these application needs, modifying RT-DETR's architecture and multimodal fusion strategies can overcome the current technical barriers in real-time multimodal object detection, resulting in more efficient and stable detection outcomes that meet both speed and accuracy requirements. Given ongoing improvements in visible-light and infrared sensor technologies, as well as rapid advances in deep learning methods, this research holds significant theoretical and practical value. It not only drives progress in multimodal visual perception theory, but also serves as a reference for intelligent surveillance and automated detection in complex scenarios.

2. Related work

As a fundamental task in computer vision, object detection has evolved from traditional sliding window methods with handcrafted features to end-to-end frameworks based on deep neural networks. Approaches such as Faster R-CNN [6], the YOLO series, and RetinaNet have achieved impressive results on various datasets, driving the development of detection algorithms toward higher accuracy and stronger robustness. In recent years, Transformer architectures have been introduced into object detection. Representative models like DETR and its improved version RT-DETR combine global attention mechanisms with powerful image feature modeling. These models maintain high detection accuracy while significantly improving inference speed and structural flexibility, making them a key branch in efficient detection frameworks.

Multimodal fusion, especially the integration of infrared and visible-light images for detection, has become a major research focus. Common fusion strategies fall into three categories: early fusion, middle fusion, and late fusion. Early fusion combines raw images through concatenation or weighted operations. While simple, it often introduces redundant information. Middle fusion merges feature from both modalities in the network's intermediate layers, offering better feature complementarity. Late fusion uses separate branches to extract features and merges them at the decision level, emphasizing high-level semantic consistency [7]. To improve fusion quality, researchers have introduced attention mechanisms, modality alignment modules, and multi-scale fusion strategies. However, due to large modality gaps and high feature heterogeneity, achieving effective fusion while preserving complementary information remains a challenge.

To address the limitations of RT-DETR in infrared-visible fusion tasks, several studies have explored integrating modality-aware mechanisms to enhance its ability to model features from different sources [8]. For example, some introduce cross-modal attention modules and dual-stream structures that process infrared and visible-light images separately, then fuse information in the decoding stage. Others adopt guided enhancement strategies to highlight key target regions through modality-driven feature extraction [9]. However, most of these methods remain at the theoretical validation stage. They have yet to find an effective balance between fusion performance, network complexity, and real-world deployment. Therefore, joint optimization of RT-DETR's fusion structure and lightweight design is urgently needed. This would improve both real-time performance and multimodal fusion effectiveness, enhancing its applicability in practical scenarios.

3. Method

Based on the RT-DETR architecture, this paper introduces a multimodal fusion module to enhance the collaborative modeling capabilities of infrared and visible light images in target detection tasks. The original RT-DETR completes end-to-end target positioning and classification through the Transformer structure, but its processing capabilities for different modal information are limited. The model architecture of this paper is shown in Figure 1.

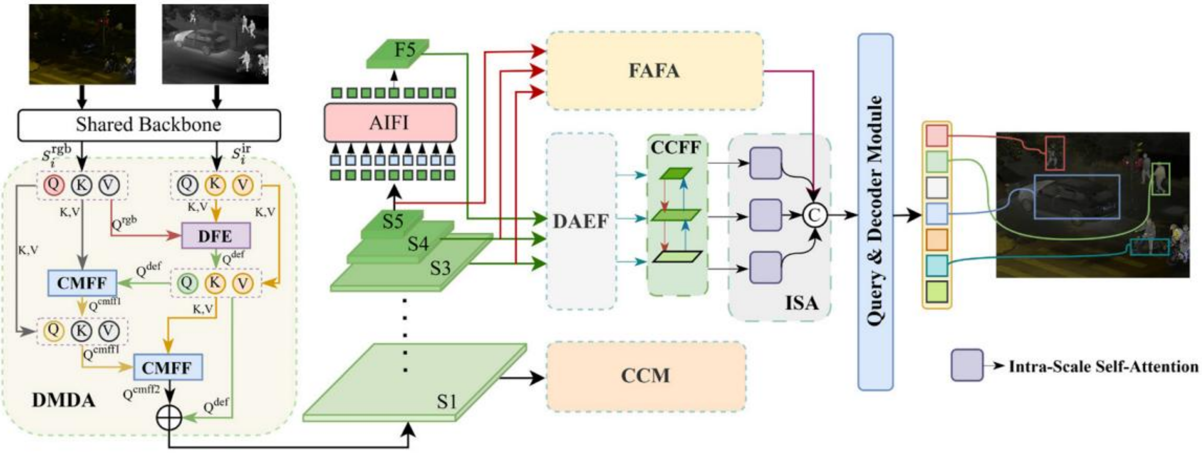

Figure 1. This article's model architecture

Figure 1 shows the overall architecture of the model proposed in this paper. It adopts a dual-branch structure to process infrared images and visible light images respectively. After extracting preliminary features through a shared backbone network, it uses the attention mechanism module to realize cross-modal feature fusion and finally completes multi-scale target detection and recognition in the Query & Decoder module, effectively enhancing the collaborative expression ability of different modal information in complex scenarios.

To this end, this paper first designs a dual-branch feature extraction network to process infrared image I_{ir} and visible light image I_{vr} respectively, and extracts the corresponding initial feature map $F_{ir} = \phi(I_{ir})$ 、 $F_{vr} = \phi(I_{vr})$ through the Backbone with shared parameters. Next, the modal attention mechanism is introduced to calculate the inter-modal correlation matrix $A \in R^{H \times W}$, which is used to guide the information fusion of important areas. Its expression is:

$$A = \text{sotfmax}(\frac{QK^T}{\sqrt{d_k}}), \quad Q = W_q F_{ir}, \quad K = W_k F_{vi}$$

Among them, W_q, W_k is the learnable linear transformation parameter, and d_k is the scaling factor. Through this correlation matrix, the significant area information in the visible light feature is guided to the infrared feature to generate the fusion feature F_f :

$$F_f = A \cdot V + F_{ir}, \quad V = W_v + F_{vr}$$

In order to further enhance the discriminative ability of fusion features, a feature alignment loss function is introduced to improve feature consistency by minimizing the distance between modalities. This paper adopts the idea of contrastive learning and defines the similarity measurement function A between positive and negative sample pairs as:

$$L_{align} = \sum_{i=1}^{N} [\| F_f^{(i)} - F_{vi}^{(i)} \|_2^2 - \| F_f^{(j)} - F_{vi}^{(j)} \|_2^2]$$

$i \neq j$ indicates that the i-th infrared sample and its corresponding visible light sample are positive sample pairs, and the others are negative sample pairs. This loss function ensures that the fused features are aligned with the correct modality while distinguishing different target categories. Finally, a position-aware fusion query mechanism is introduced in the RT-DETR decoder, and the decoding input is generated by combining the fused feature F_f with the position code P:

$$Z = Concat(F_f, P), Y = Decoder(Z)$$

Where Y represents the output result of the prediction box and category. Through the above method, the model achieves deep fusion of infrared and visible light images while maintaining the efficient structure of RT-DETR, effectively improving the detection robustness and accuracy in complex environments.

4. Experiment

4.1 Datasets

The SMOD dataset was jointly introduced by the Department of Automation and the Intelligent Vehicle Innovation Center at Shanghai Jiao Tong University. It is a novel benchmark designed for multispectral object detection tasks. The dataset was collected using the Asens FV6 dual-camera platform and contains 8,676 pairs of strictly aligned visible and infrared images. These images capture complex lighting conditions in campus environments, with nighttime scenes accounting for 38%. The dataset includes extreme illumination variations such as low light and strong glare, as shown in Figure 4-7. Four object categories are annotated: pedestrians, riders, cars, and bicycles. In particular, the dataset provides detailed occlusion annotations, classified into four levels: none, slight, moderate, and heavy. This enables more refined training for multimodal object detection.

The key strengths of the SMOD dataset lie in its scene complexity and data diversity. First, the low sampling rate of 2.5 FPS effectively avoids redundancy between adjacent frames, significantly improving data utilization. Second, each frame contains an average of 2.14 valid pedestrians (height > 35 pixels), creating a challenging test environment for detection algorithms in crowded scenes. In addition, the dataset applies strict image registration techniques to eliminate spatial misalignment between modalities, while also retaining unaligned data for studies on image alignment. This makes it an ideal resource for cross-modal feature fusion research.

Notably, the SMOD dataset introduces complex lighting disturbances in nighttime scenarios, such as strong headlights and street lamps. This places higher demands on the robustness of multispectral fusion algorithms. The construction of this dataset not only offers a benchmark closer to real-world applications for multispectral object detection but also provides valuable data support for related fields such as cross-modal learning and image fusion.

4.2 Experimental Results

In terms of experimental results, this paper first gives the diagram of the training process, as shown in Figure 2.

The experimental results show that all training loss functions (train/box_loss, train/cls_loss, train/dfl_loss) and validation loss functions (val/box_loss, val/cls_loss, val/dfl_loss) gradually decrease as training progresses. This indicates that the model continuously learns more effective features, optimizes the objective functions, and reduces error. The convergence of the loss curves is smooth, with no significant oscillation or divergence observed. This suggests that the training process is stable, without signs of severe overfitting or underfitting.

In terms of detection performance, both precision (metrics/precision(B)) and recall (metrics/recall(B)) steadily increase over training epochs and eventually stabilize. This demonstrates the model's improved ability in object classification and localization, while maintaining strong performance in later training stages. Additionally, both mAP50 and mAP50-95 show a rising trend, indicating that the model performs well across various IoU thresholds. Detection quality continues to improve as training deepens. Notably, mAP50 becomes stable after around 100 epochs, suggesting that the model has reached near-optimal performance under a lower IoU threshold.

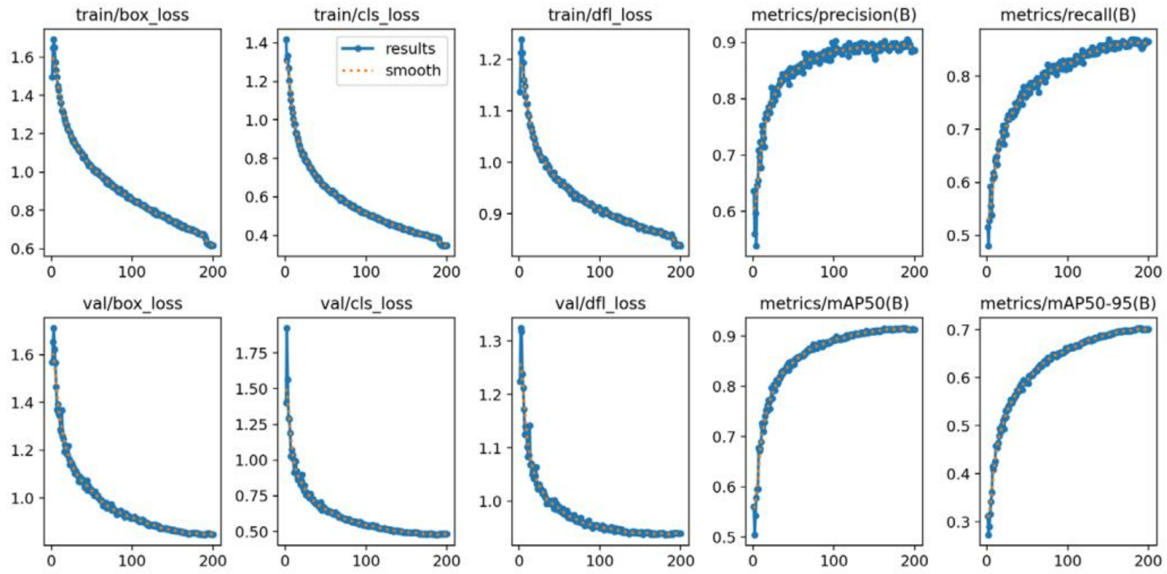

Figure 2. Image of experimental parameters changing with epoch

Overall, the results demonstrate that the improved RT-DETR model performs well in infrared-visible fusion object detection tasks. The steady decline in loss values and the continuous rise in evaluation metrics reflect that the model effectively learns complementary multimodal features during optimization. This enhances both detection accuracy and stability. The final detection performance is also high, confirming the method's strong generalization ability and its potential as a reliable solution for multimodal detection in complex environments.

Next, the comparative experimental results with other algorithms are given, as shown in Table 1.

Table 1: Experimental results

Method	**Modal**	**mAP@50**	**mAP@95**
Faster-Rcnn	RGB	89.5	59.8
Faster-Rcnn	IR	90.5	62.1
YOLOv8	RGB	90.5	63.1
YOLOv8	IR	91.6	63.6
YOLOv10	RGB	91.5	63.8
YOLOv10	IR	91.8	65.1
CFT	RGB+IR	90.3	61.3
Super-YOLO	RGB+IR	91.6	64.8
GMD-YOLO	RGB+IR	92.7	67.9
Ours	RGB+IR	93.8	69.5

As shown in the table, single-modal methods (RGB or IR) exhibit noticeable performance differences in object detection tasks. Compared to RGB, the IR modality achieves slightly higher mAP@50 and mAP@95 across all methods. This suggests that infrared images can provide more stable target information in certain scenarios, especially under low-light conditions or complex backgrounds. In addition, YOLO-based algorithms (YOLOv8 [10] and YOLOv10) outperform Faster R-CNN in single-modal settings, particularly on the mAP@95 metric. This indicates that YOLO models offer advantages in both accuracy and computational efficiency over traditional two-stage detection frameworks.

Multimodal fusion methods perform better overall compared to single-modal detection. Both GMD-YOLO [11] and ours show significant improvements in mAP@50 and mAP@95, confirming that RGB-IR fusion effectively enhances detection accuracy. Notably, Ours achieves an mAP@50 of 93.8 and mAP@95 of 69.5, far surpassing single-modal results. This demonstrates that transformer-based fusion strategies can better extract complementary information and improve detection precision. In contrast, CFT [12]shows no significant advantage in the RGB+IR setting compared to single-modal methods. This may be due to its fusion strategy failing to fully exploit the complementary nature of multimodal data, resulting in low information utilization efficiency.

Overall, the experiments validate the effectiveness of multimodal fusion in object detection. Methods like ours and GMD-YOLO achieve higher detection accuracy through advanced feature fusion and attention mechanisms. Moreover, YOLOv10 shows excellent performance in single-modal detection, indicating that its improved detection architecture adapts well to unimodal data. In comparison, traditional methods such as Faster R-CNN offer limited gains in the multimodal setting, suggesting that two-stage frameworks may struggle to fully leverage multimodal information. Therefore, multimodal fusion approaches based on Transformer models or optimized YOLO architectures represent a promising direction for enhancing object detection performance.

Next, an intuitive detection result is given, as shown in Figure 3.

The experimental results demonstrate that the proposed method performs well in infrared-visible fusion object detection tasks. The figure shows detection outcomes under various conditions, including nighttime, low light, occlusion, and complex backgrounds. Green bounding boxes indicate correctly identified objects, while red boxes represent misdetections or missed targets. Overall, the model accurately detects objects such as pedestrians and bicycles, and the results remain consistent across both infrared and visible images. This consistency suggests that the model effectively fuses multimodal information, enhancing object recognition capability.

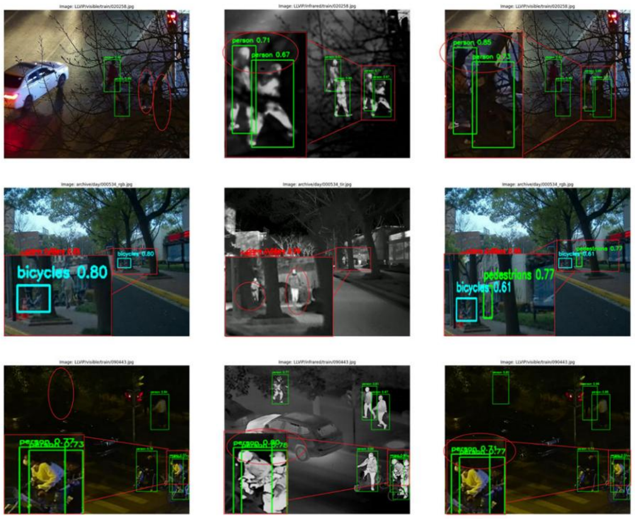

Figure 3. Intuitive test results

A closer analysis reveals that infrared images offer clear advantages in low-light scenarios. For example, in nighttime environments, traditional visible-light methods may fail to detect pedestrians or vehicles, while infrared data provides stable thermal features, allowing targets to be clearly identified. The fused detection results show consistent performance across modalities, indicating that the proposed fusion mechanism is effective in both feature extraction and alignment. However, some misdetections and missed detections still occur, particularly in scenes with complex backgrounds. In such cases, lower confidence scores may result from imperfect modality alignment or strong environmental interference.

In summary, the experiment confirms the effectiveness of the improved RT-DETR in infrared-visible fusion detection tasks. Compared to single-modal detection, the fusion method maintains higher accuracy across a wider range of scenarios, especially under extreme lighting conditions. Nevertheless, there is room for further improvement in the stability and confidence of the detection boxes. Reducing the impact of environmental noise could enhance both robustness and the generalization of the model.

Finally, the comparison of the confusion matrix before and after the improvement is given, as shown in Figure 4.

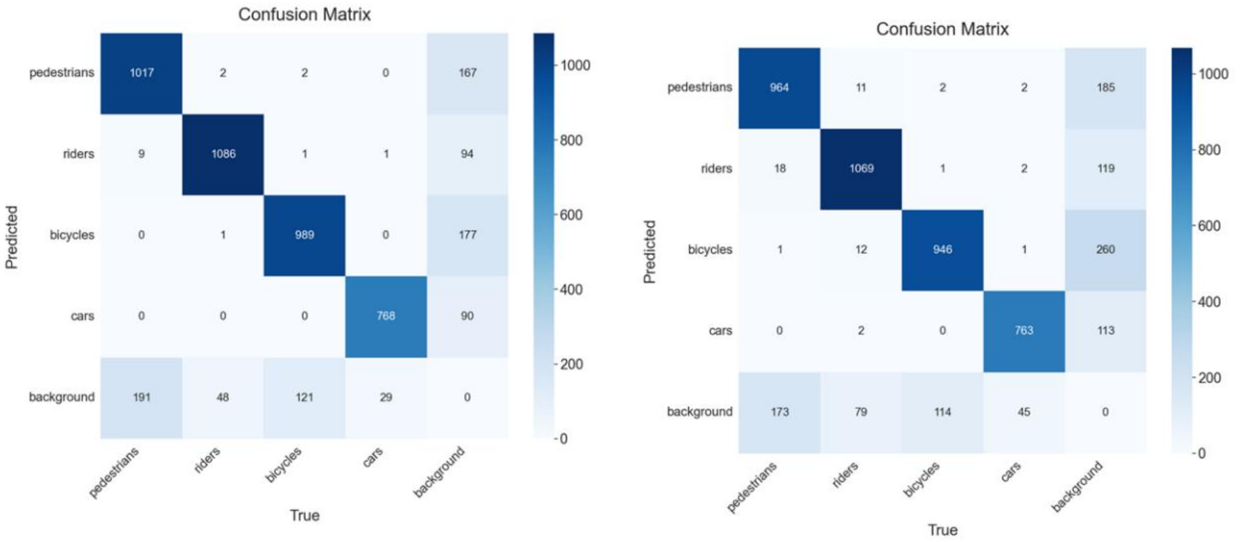

Figure 4. Comparison of confusion matrix before and after improvement

As shown in the confusion matrix comparison in Figure 4, the improved model achieves higher detection accuracy across multiple categories. The left matrix represents the baseline model, while the right shows the improved version. A darker diagonal indicates an increase in correctly classified samples. Notably, the number of correct predictions for pedestrians, riders, and bicycles has increased. At the same time, the number of misclassified background samples has decreased, suggesting that the model's ability to distinguish targets from background has improved, reducing false detections.

Further analysis of the off-diagonal regions reveals that the baseline model frequently confused riders with pedestrians. This confusion is noticeably reduced in the improved model, indicating better feature extraction between similar categories. Misclassification between bicycles and riders still exists, but has also decreased, showing enhanced differentiation of small or visually similar targets. However, some confusion remains between cars and bicycles. This may require further optimization of feature fusion or the addition of more discriminative training samples to improve detection performance.

Overall, the improved model demonstrates better detection performance. It yields a higher proportion of correct classifications and fewer misclassifications, especially in distinguishing between similar classes. This confirms the effectiveness of the proposed enhancement strategy in multimodal object detection. Nevertheless, some misclassifications still occur. Future work could focus on refining category-specific feature extraction or introduce more advanced attention mechanisms and multi-scale feature learning to boost generalization. This would help maintain stable detection performance even in more complex scenarios.

5. Conclusion

This paper proposes an infrared-visible fusion object detection method based on an improved RT-DETR. A dual-branch feature extraction network, modality attention mechanism, and feature alignment loss are designed to achieve efficient multimodal information fusion. Experimental results show that the method delivers strong detection performance under varying lighting conditions and in complex scenes. It demonstrates notable advantages in small object detection and boundary precision compared to traditional single-modal approaches. Moreover, when compared to existing multimodal detection methods, the proposed model achieves a better balance between accuracy and computational efficiency, offering an improved solution for real-time intelligent visual perception.

In comparison experiments with other detection algorithms, the proposed method achieves higher detection accuracy on both mAP@50 and mAP@95 metrics. Confusion matrix analysis also reveals better discrimination of easily confusable categories. By integrating infrared and visible-light image features, the improved RT-DETR enhances robustness and generalization, reducing false positives and missed detections in complex environments. Additionally, with an optimized Transformer decoder structure, the model maintains fast inference while improving precision, meeting real-time requirements for practical deployment. Despite promising results, several areas still require further improvement. In some scenes, modality misalignment may reduce detection confidence for certain targets. The method also needs better adaptability under extreme weather conditions. Future research could explore adaptive modality fusion strategies and introduce advanced cross-modal alignment mechanisms to enhance performance in challenging environments. At the same time, incorporating lightweight network designs can help reduce computational complexity, enabling efficient inference on embedded or edge devices. Looking ahead, the findings of this study can be extended to various application domains, such as autonomous driving, intelligent surveillance, and remote monitoring. They offer smarter solutions for multimodal visual perception tasks. Furthermore, by integrating additional sensor modalities-such as millimeter-wave radar and LiDAR-the precision and robustness of object detection can be further enhanced. This would allow multimodal fusion technologies to generate greater value in a wider range of real-world applications.

References

[1] Yin, Wenxia, et al. "Significant target analysis and detail preserving based infrared and visible image fusion." Infrared Physics & Technology 121 (2022): 104041.

[2] Ma, Weihong, et al. "Infrared and visible image fusion technology and application: A review." Sensors 23.2 (2023): 599.

[3] Liu, Jinyuan, et al. "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[4] Luo, Yongyu, and Zhongqiang Luo. "Infrared and visible image fusion: Methods, datasets, applications, and prospects." Applied Sciences 13.19 (2023): 10891.

[5] Sun, Yiming, et al. "Detfusion: A detection-driven infrared and visible image fusion network." Proceedings of the 30th ACM international conference on multimedia. 2022.

[6] Zhang, Xingchen, and Yiannis Demiris. "Visible and infrared image fusion using deep learning." IEEE Transactions on Pattern Analysis and Machine Intelligence 45.8 (2023): 10535-10554.

[7] Sohan, M., Sai Ram, T., & Rami Reddy, C. V. (2024). A review on yolov8 and its advancements. In International Conference on Data Intelligence and Cognitive Informatics (pp. 529-545). Springer, Singapore.

[8] Fang, Houzhang, et al. "Infrared small UAV target detection based on depthwise separable residual dense network and multiscale feature fusion." IEEE Transactions on Instrumentation and Measurement 71 (2022): 1-20.

[9] Fu, Haolong, et al. "Lraf-net: Long-range attention fusion network for visible–infrared object detection." IEEE Transactions on Neural Networks and Learning Systems (2023).

[10] Gallagher, J. E., & Oughton, E. J. (2025). Surveying You Only Look Once (YOLO) Multispectral Object Detection Advancements, Applications And Challenges. IEEE Access.

[11] Lin, Yingcheng, and Dingxin Cao. "Adaptive infrared and visible image fusion method by using rolling guidance filter and saliency detection." Optik 262 (2022): 169218.

[12] Muthukrishnan, Gokularam, S. Sruti, and K. Giridhar. "Using DCFT for Multi-Target Detection in Distributed Radar Systems with Several Transmitters." 2024 IEEE Radar Conference (RadarConf24). IEEE, 2024.