Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 12, No. 1, 2025

Multimodal Road Traffic Detection Algorithm based on Improved YOLOv8

Xuanning Wei, Jianhan Zhou *

School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, Liaoning, 114000, China
*** Corresponding author:** Jianhan Zhou (Email: zjhzjhan@163.com)

Abstract: In road traffic detection, traditional unimodal object detection methods exhibit certain limitations in adapting to environmental variations. Moreover, in complex road conditions, mutual occlusion between targets and their confusion with the background pose significant challenges in feature extraction for multi-scale objects and the detection of densely distributed small targets. To address these challenges, this paper proposes a multi-modal object detection algorithm, CD-MMNet, based on YOLOv8. Firstly, the backbone network adopts a dual-branch structure to perform intermediate fusion of features from two modalities—visible light and infrared images—thereby leveraging their complementary characteristics to dynamically select optimal feature extraction in a targeted manner. Secondly, the CBAM attention mechanism is introduced to dynamically adjust the importance of each channel and spatial position in the feature maps, enhancing key regional features while suppressing background noise, thus improving the model's feature extraction capability. Finally, the DBB module is incorporated, utilising a diversified branch network to enhance the model's adaptability to feature maps of varying scales. Experimental results demonstrate that the proposed algorithm outperforms the original YOLOv8 and other mainstream algorithms on the M3FD dataset, achieving a 4.0% improvement in mAP@0.5~0.95 compared to the baseline YOLOv8. This effectively enhances object detection performance in challenging environments such as adverse weather conditions and traffic congestion.

Keywords: Multimodal; Target Detection; Feature Fusion.

1. Introduction

In recent years, automatic driving technology [1] has played an important role in reducing traffic accidents, alleviating traffic pressure, and promoting traffic universality. And target detection, as its important branch, has received attention and favour from many researchers. Target detection technology enables self-driving vehicles to accurately identify and classify various objects in the surrounding environment, including other vehicles, pedestrians and obstacles. This real-time sensing capability is the basis for vehicles to make safe decisions, especially in complex and dynamic traffic environments, where fast and accurate target recognition is the key to ensure safe driving. However, the traditional visible light-based target detection method mainly relies on the shape, size, colour and other physical characteristics of the object under visible light conditions, which makes it more limited in some specific situations. When the lighting conditions are insufficient, such as in bad weather, the visibility of the target decreases significantly, which leads to a serious impact on the accuracy and robustness of the detection system. In addition, when the targets are densely packed and occlude each other, it is difficult for conventional methods to identify and localise these targets effectively, which further limits their application in complex environments. To overcome this deficiency, the application of multimodal target detection has emerged. It fuses the information characteristics of many different modalities to provide accurate information in an all-round way, which helps to improve the speed and accuracy of target detection in various complex environments.

Target detection as one of the core in the field of computer vision is divided into two categories: detection based on traditional methods and detection based on deep learning. The deep learning based detection is further divided into two-stage and one-stage models based on whether or not a region proposal network (RPN) is used to generate candidate regions. The two-stage model divides the target detection task into two consecutive stages, with the first stage generating the RPN first, and the second stage performing classification and positional regression on the generated candidate regions to locate and segment the object more easily. Such models are led by R-NCC [2], the first deep learning model to use convolutional neural networks for target detection proposed by Girshick et al. in 2014. In the first stage, the model does not use RPNs and performs the prediction of target categories and bounding box locations directly on the input image or feature map. Redmon et al. were the first to propose the YOLO [3-5] (You Only Look Once) family of algorithms in 2016. Subsequently, liu et al. proposed the SSD [6-8] model, which utilises a single deep neural network to detect targets in an image, enhancing the detection of targets of different sizes. This type of model skips the candidate region generation stage, reduces intermediate steps, optimises computational efficiency and reduces processing time. Due to its efficiency and real-time nature, one-stage detection models are more suitable for use in target recognition for unmanned vehicles.

2. Related Work

With the rapid development of computer technology, scholars at home and abroad have done a lot of research on improving the accuracy of target detection. In order to fuse multimodal data and improve the detection effect, researchers have proposed a variety of multimodal image fusion methods. Zhou [9] et al. proposed Modal Balanced Network (MBNet), which is able to achieve adaptive feature fusion through the introduction of differential modal sensing module and light sensing alignment module. By optimising the information

balance between modalities, MBNet significantly improves the accuracy of pedestrian detection in complex environments, especially in low-light or modal imbalance situations. Ma [10] et al. proposed a generative adversarial network (GAN) architecture and designed a dedicated loss function for it, aiming to improve the fusion of infrared images with visible light images. gan GAN can effectively improve the quality of fused images and the accuracy of target detection, especially in dynamic scenes, by making the fusion of different modal images more natural through adversarial training.Chen [11] et al. proposed an end-to-end multiscale Fully Convolutional Network (FCN), which takes advantage of the properties of fully convolutional networks and discards the fully connected layer of traditional convolutional neural networks (CNN), making it capable of accepting arbitrary sizes. structure of traditional convolutional neural network (CNN), so that it can accept input images of arbitrary size and output corresponding feature maps, compared with the U-Net model, the mAP of this method is improved by 6.5%. Liu [12] et al. proposed a road detection method based on multiscale convolutional neural network (MSCNN), where the network extracts features at different scales by multi-level convolutional operations, which enhances the accuracy of road target recognition, and the mAP of the method is improved by 7.8% compared to FCN.Zhou [13] et al. proposed Dual-Attention Network (DAN), which is specifically designed for detecting small targets in traffic scenes. By combining the spatial and channel attention mechanisms, the model can focus on small targets more accurately and filter out background noise, which improves the mAP by 7.2% on the ApolloScape dataset.

Although the above research has addressed many difficulties in the field of target detection, it still faces many substantial challenges. Therefore, this paper proposes a multimodal target detection algorithm CD-MMNet (CBAM_DBB-MultiModalNet) fusing visible and infrared images based on YOLOv8 [14] algorithm as the base network, firstly, a two-branch structure is used in the backbone network to intermediate fusion the features of the two modalities of the visible and infrared images respectively, and secondly , the CBAM [15] attention mechanism is introduced to enhance the feature extraction ability of the model, and finally, the DBB [16] module is introduced to replace the conv module of the backbone network, and the diversified branching network is used to enhance the adaptability of the model to different sizes of feature maps. The improved network structure is shown in Fig. 1.

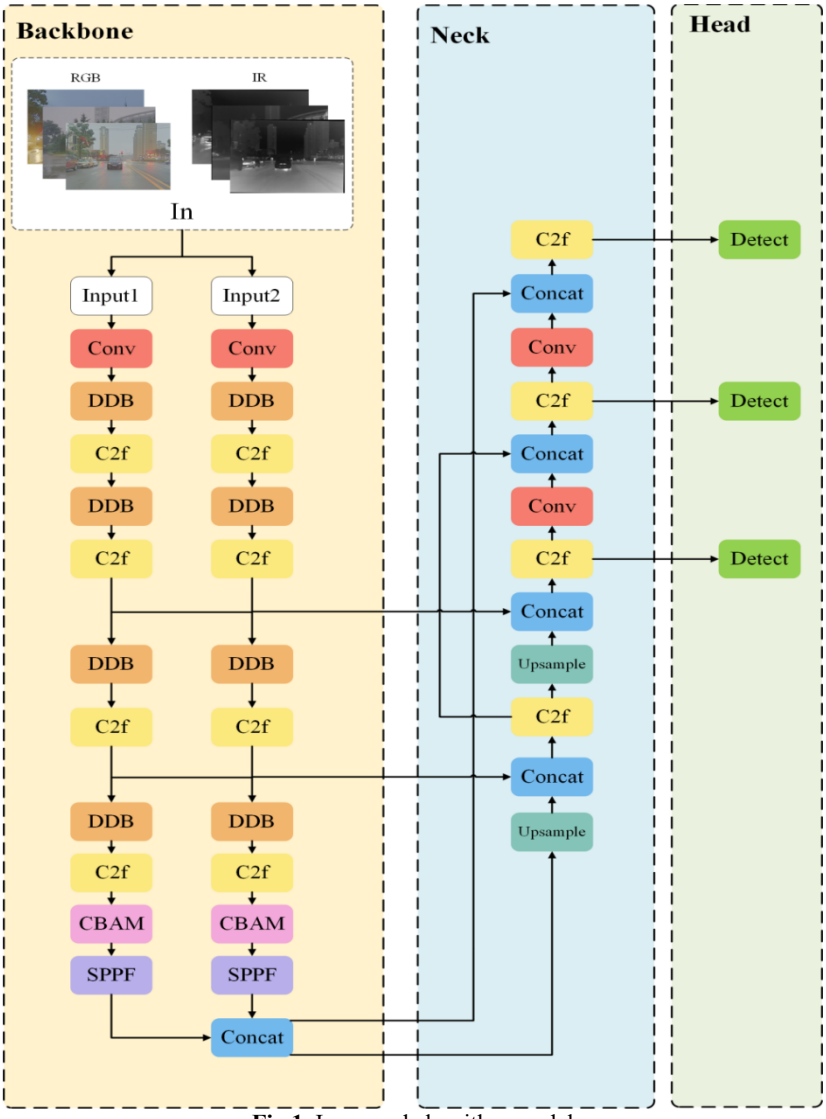

Fig 1. Improved algorithm model

3. Improved YOLOv8 Algorithm

3.1. Multimodal Intermediate Fusion Module

Unimodal-based target detection has limitations in recognising only one type of information data. For example, using visible light detection can only provide fine visual information when there is sufficient light, and when lighting conditions are insufficient, the distinction between the target and the background in a visible image often becomes difficult. And using infrared thermal image detection can only provide a clear outline when the thermal contrast between the figure and the background is large, but when faced with a complex background of thermal radiation, the target details may be masked by the complex thermal signals of the background, and identifying the target is equally difficult. This shows that using only one type of information data for recognition detection has a large error. To address this challenge, this study proposes a dual-path feature fusion mechanism, which achieves efficient integration and complementarity of information by dynamically aggregating effective features from multiple sources at the critical level of the detection network, thus providing a more discriminative feature representation for target recognition in complex scenes.

Firstly, the modular network receives images from both visible and infrared sensors as inputs, and the input feature tensor contains information about both modalities. At the initial stage of the network, these inputs are subjected to separate and independent feature extraction through a two-branch network, specifically, the input feature tensor is split into two parts: the first three channels are used as visible features, and the last three channels are used as infrared features. Each of these two branches progressively extracts features at different scales through successive convolution and downsampling operations, ensuring that the information of each modality can be adequately learnt and expressed. This design enables the network to extract independent feature representations for the characteristics of visible and infrared images, avoiding mutual interference between the two modalities.

Next, in a two-branch network, the feature maps in each branch are passed through multiple convolutional layers and pooling operations, respectively, to further optimise and refine the feature information. In this stage, the network performs multi-scale and hierarchical processing on features of different modalities, thus extracting low-level edge and texture information as well as high-level semantic features, which enhances the discriminative ability of each modal feature. This allows the network to improve feature learning while maintaining computational efficiency. This process ensures that visible and infrared features can be learnt and represented independently, providing richer and more representative features for the final multimodal fusion.

At a later stage, feature maps from different scales are spliced and feature maps from the visible and infrared branches are merged in the channel dimension, thus fusing useful information from both modalities. This fusion mechanism aims to take full advantage of the complementary nature of visible and infrared features to generate a more comprehensive feature representation that helps the model capture richer target information. The fused features are then further processed through a convolutional layer to achieve the aggregation of features at different scales, culminating in target classification and bounding box regression prediction.

This multimodal fusion mechanism equips the network with dual perceptual capabilities: extracting the underlying visual features (e.g., edge contours and texture details) of the visible image on the one hand, and capturing the thermodynamic features (including the target temperature distribution and thermal radiation patterns) of the infrared image on the other. In deep feature learning, the bimodal information is complementarily fused to generate more discriminative high-level semantic representations (e.g., target category and geometric morphology). This architectural design retains the high-resolution advantage of visible light under good lighting conditions and incorporates the stable thermal characteristics of infrared in low-light environments, ultimately achieving robust target detection in complex scenes.

3.2. CBAM

In crowded areas in city centres, target detection faces many difficulties. Dense crowds cause objects to block each other, making the target often appear mutilated; when shooting from a distance, the boundary between the small target and the background is blurred, and the outline is difficult to be recognised. Dynamic targets, such as vehicles and pedestrians, have variable trajectories and are prone to sudden displacements or short-term occlusion; static targets, such as road signs, are often partially occluded. These situations lead to incomplete target features or fuzzy edges, which seriously affects the recognition and localisation accuracy of the detection algorithm. In order to solve this problem, this paper introduces a convolutional attention module CBAM, which enhances the representation ability of the network by focusing on important features to suppress unnecessary features, and its structure is shown in Fig. 2.

Since convolutional operations extract informative features by mixing cross-channel and spatial information, the module emphasises meaningful features along these two main dimensions. As shown in the figure, given the intermediate feature map $F \in R^{C\times H\times W}$ as input, CBAM sequentially infers a one-dimensional channel attention map $M_c \in R^{C\times 1\times 1}$ and a two-dimensional spatial attention map $M_s \in R^{1\times H\times W}$, and the overall attention process can be summarised as:

$$F' = M_c(F) \otimes F, \tag{1}$$

$$F'' = M_s(F') \otimes F', \tag{2}$$

where $\otimes$ denotes element multiplication. During the multiplication, the attention values are copied accordingly, the channel attention values are copied along the spatial dimension and the spatial attention values are copied along the channel dimension. F'' is the final refined output.

It makes sense to focus channel attention on the "what" of a given input image. CBAM first performs global averaging and max-pooling operations in the spatial dimension to condense global description vectors, which capture the global importance of each channel across the feature map. Subsequently, a small fully-connected neural network is used to learn and compute the channel importance weights, which are normalised by a Sigmoid function to output a weight

vector characterising the relative importance of each channel. Finally, the weight vectors are applied to the input feature maps by multiplication to obtain a once-refined feature map for feature response re-weighting. In this case, the channel attention is calculated as:

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) = \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{\max}))), \quad (3)$$

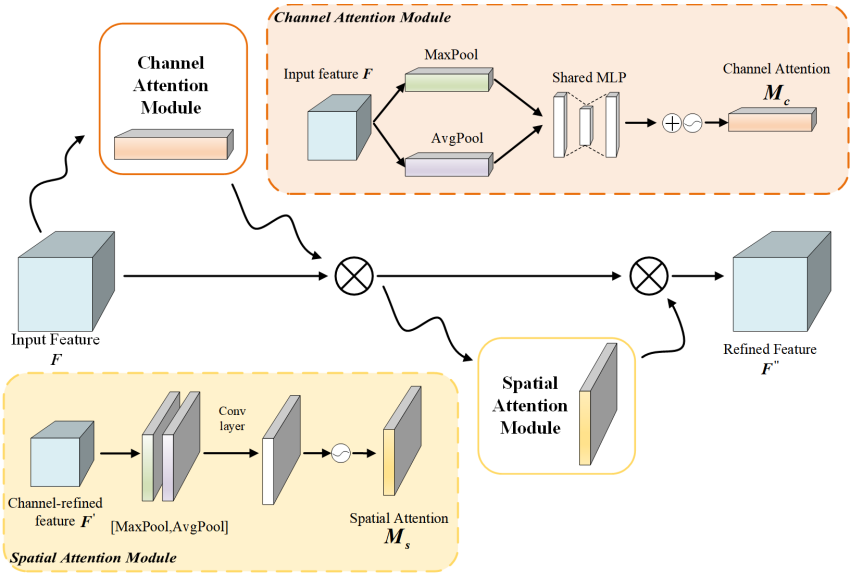

Fig 2. CBAM structure

where σ denotes the sigmoid function, $W_0 \in R^{C/r \times C}$, $W_1 \in R^{C \times C/r}$. Where the two input MLP weights W_0 and W_1 are shared and the ReLU activation function is followed by W_0. This process not only improves the sensitivity of the model to the key feature channels, but also helps the network to dynamically focus on the information-rich channels to enhance the model representation and detection performance.

Spatial attention plays an important role in deciding "where to focus". CBAM performs global averaging and maximum pooling in the channel dimension, focuses on capturing the potential information of the target region, compresses the spatial information by 7×7 convolution, and normalises it using the Sigmoid function to generate a single-channel spatial attention map, and then multiplies the spatial attention map with the primary refined feature map and weights the features again to obtain the secondary refined feature map. spatial attention map is multiplied with the primary refined feature map and weighted features again to obtain the secondary refined feature map. Where spatial attention is calculated as:

$$M_s(F) = \sigma(f^{7\times7}([AvgPool(F); MaxPool(F)])) = \sigma(f^{7\times7}([F^s_{avg}; F^s_{\max}])), \quad (4)$$

where σ denotes the sigmoid function and $f^{7\times7}$ denotes the convolution operation with a filter size of 7×7. This spatial refinement improves the sensitivity of the network to key spatial locations, so that it focuses on the target area and enhances the target perception capability of the network.

In the slight occlusion scene, the spatial attention can locate the boundary information of the target object more accurately, enhance the feature representation of the target region, and reduce the feature blurring due to occlusion; in the severe occlusion scene, the channel attention can highlight the key feature channels in the unoccluded part, providing strong support for the subsequent inference, while effectively suppressing the interference of background noise on detection. The introduction of the CBAM module in the Backbone of the YOLOv8 network structure optimises the distribution of the attention of the feature map, mitigates the interference of noise, enhances the model's ability to represent the features of the occluded objects, and helps the model to further understand the image contextual information, and then achieve the effective detection and identification of the occluded objects.

3.3. DBB

In complex and changing scenes, the visual feature recognition performance of the network will be significantly reduced. Taking traffic peaks or holidays as an example, objects of different scales in dense crowds intertwine with each other to form intricate spatial relationships, which greatly increases the difficulty of target recognition. To solve this problem, we introduce a multivariate branching block DBB, which is used to replace the conv operation to enrich the feature space by combining different branches with different scales and complexities, and its structure is shown in Fig. 3.

DBB, short for Diverse Branch Block, improves performance by increasing the complexity and multi-scale representation capability of the network during training, and reduces the network complexity through transformation operations during inference to achieve performance improvement without additional inference cost. This multi-branch topology using path operations with different acceptance domains and different levels of complexity can enrich the image features. In the training phase, the input feature maps are first fed into multiple branches in the DBB, and each branch performs a different operation. The standard convolutional branch performs regular K×K convolution to obtain feature information; the sequence convolutional branch employs 1×1 convolution, which is used to reduce the number of parameters while maintaining the expressive power of the network; the average pooling branch employs average pooling with 1×1 convolution, which is used to

capture a wider range of context information; and the multi-scale convolutional branch employs K×1 convolution, which is used for the multi-scale convolutional branch. scale convolution branch uses K×K convolution and 1×1 convolution to capture features at different scales. Finally, the features of each branch are fused to form the final output. The application of diversified branching structure can greatly enhance the feature extraction capability of the model. In the inference stage, the DBB module performs a series of equivalent transformations for parameter recomputation through the linear property of convolution operations, and uses structural reparameterisation to merge multiple branches into a single standard convolutional layer.

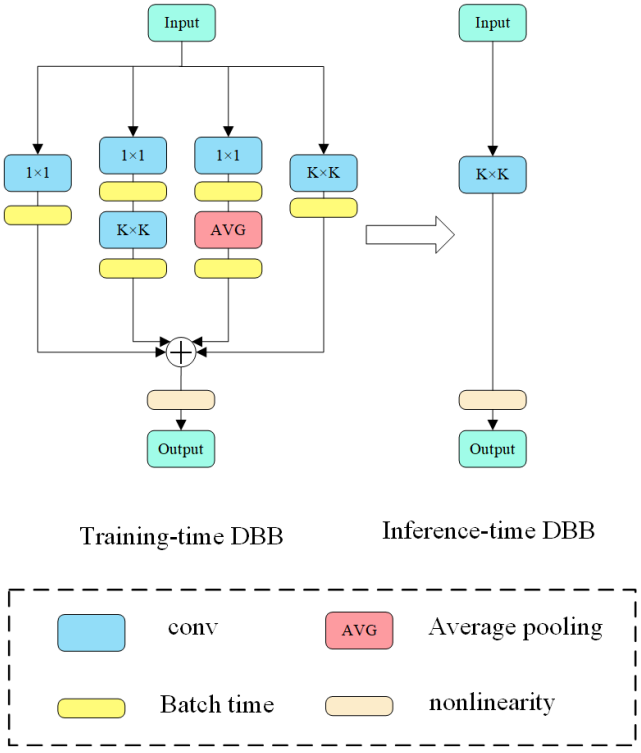

Fig 3. DBB structure diagram

By adding this module, the model can be trained to reach a higher performance level, and then converted to the original inference time structure for inference, so that the model can be more flexible and effective in dealing with targets at different scales, especially in the face of complex backgrounds and occlusions, and can better extract effective features. Meanwhile, the DBB module adopts a more efficient convolutional structure, which reduces redundant computation and significantly improves the computational efficiency of the model while ensuring the feature extraction capability.

4. Experiments and Analysis of Results

4.1. Data Sets

This experiment uses the M3FD [17] dataset. The M3FD dataset was collected using a simultaneous imaging system constructed from well-calibrated infrared and optical sensors covering a wide range of scenarios including illumination, seasonal, and weather conditions involving different types of objects. It labels 33,603 objects in six categories: people, cars, buses, streetlights, motorbikes, and trucks.M3FD accumulates a total of 4,200 image pairs, of which 3,360 pairs are designated for training and 840 pairs for testing, with an image resolution of 1024 × 768.The large size and rich diversity of the dataset make it an ideal basis for learning and evaluating target detection using fused images an ideal basis for learning and evaluating target detection using fused images.

4.2. Experimental Environment and Parameters

In order to better demonstrate the performance of the CD-MMNet model as well as to ensure the fairness of the experiments, the experiments are conducted in the same environment, and the specific experimental parameters are shown in Table 1.

4.3. Evaluation Indicators

In order to effectively evaluate the target detection capability of the CD-MMNet model, this paper uses precision, recall, mAP@0.5~0.95 as the evaluation indexes with the following formulas:

$$P = \frac{TP}{TP + FP} 100\% \tag{5}$$

$$R = \frac{TP}{TP + FN} 100\% \tag{6}$$

$$\mathrm{m}AP = \frac{\sum_{n=1}^{N} \int_0^1 p(r)dr}{N} \tag{7}$$

where P is the correct rate, TP is the number of samples predicted to be correct positive samples in the target detection task, FP is the number of samples that are negative but predicted to be positive samples, and R is the regression rate, where FN is the number of positive samples that are predicted

to be negative samples. The field of target detection generally uses P-R curves to describe the performance of target detection. In addition, mAP is a composite measure of average accuracy across multiple categories to measure the performance of a target detection model. mAP's threshold (IOU) can be set based on the degree of overlap of the boxes. mAP@0.5~0.95 is the average accuracy when the degree of overlap between the predicted box and the target box is greater than 0.5 and less than 0.95.

Table 1. Experimental environment configuration

		Name	Parameter
Hardware	PC	CPU	AMD EPYC 9654
		GPU	RTX 4090
		Graphics card	24G
		Operation system	Ubuntu 22.04
Software	Deep Learning framework		Pytorch 2.3.0
	Programming languages		Python 3.12
	CUDA		12.1
hyperparameter	Epochs		150
	Batch size		64
	Picture size		640×640
	Initial learning rate		0.01
	IoU		0.7

4.4. Ablation Experiments

In order to prove the effectiveness of the target detection effect of the CD-MMNet model, ablation experiments are conducted on the M3FD dataset, and the consistency of the input images and training parameters is maintained during the experiments, and the experimental results are shown in Table 2. Among them, the first group is the evaluation results of the original YOLOv8 model; the second group adds the intermediate fusion module; the third group of experiments adds the CBAM module again; the fourth group is the CD-MMNet model proposed in this paper, which improves the precision, recall, mAP@0.5%, and mAP@0.5~0.95% by 2.9%, 3.7%, 3.5%, and 4.0, respectively. ablation Experiments show that the improved model can effectively improve the precision of the algorithm.

Table 2. Results of ablation experiments

batch number	FMN	CBAM	DBB	Precision%	recall%	mAP@0.5%	mAP@0.5~0.95%
1	-	-	-	79.60	68.60	75.70	48.70
2	√	-	-	80.80	70.70	77.30	51.40
3	√	√	-	84.00	69.70	78.50	51.50
4	√	√	√	**82.50**	**72.30**	**79.20**	**52.70**

4.5. Comparative Experiments

Under consistent experimental conditions, the improved model proposed in this study was compared with other models, as shown in Table 3. The results demonstrate that the model achieves a significantly higher mAP@0.5~0.95% on the M3FD dataset than competing approaches, exhibiting superior precision and robustness in object detection.

Table 3. Comparative experimental results of models

modelling	Precision%	recall%	mAP@0.5~0.95%
YOLOv8n	79.60	68.60	48.70
YOLOv5n	80.30	66.90	47.00
YOLOv6n	80.90	61.40	45.10
YOLOv9t	81.90	64.60	47.70
YOLOv10n	76.90	64.10	46.90
GAN-based[17]	-	-	50.20
CD-MMNet	82.50	72.30	52.70

4.6. Visualisation Analysis

In order to visually compare the detection effects of the YOLOv8 model and the CD-MMNet model, images of small multi-scale dense targets with poor lighting conditions and occlusion were identified in the M3FD dataset to visualise and analyse the results of the models, as shown in Fig. 4. (a) shows the detection results of unimodal visible light, (b) shows the detection results of unimodal infrared light, and (c) shows the detection results of CD-MMNet model. As can be seen from the figure, the CD-MMNet model significantly improves the occurrence of leakage and misdetection, and improves the detection accuracy of the target.

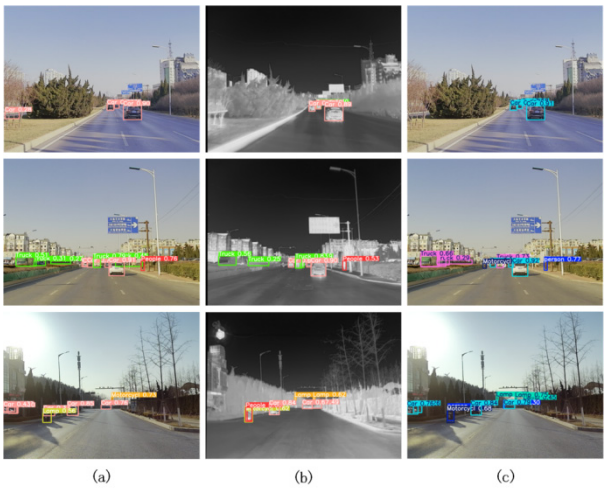

Fig 4. Comparison of detection results under different conditions

5. Conclusion

In summary, this paper introduces improvements to certain modules of the YOLOv8 backbone network. The proposed CD-MMNet model employs a dual-branch structure to dynamically select optimal features from visible and infrared images, thereby overcoming environmental limitations. Additionally, the CBAM attention mechanism is integrated into the backbone network, enabling dual channel and spatial attention weighting to enhance focus on target regions. Finally, the DBB module replaces conventional convolutional structures, leveraging multi-branch convolution operations to improve the network's adaptability to multi-scale features. Experiments on the M3FD dataset demonstrate that the CD-MMNet model significantly outperforms existing state-of-the-art algorithms, achieving a 4.0% improvement in mAP@0.5~0.95 compared to the original YOLOv8. It exhibits particularly strong robustness and detection accuracy in challenging scenarios such as adverse weather and complex traffic conditions. However, the model's increased parameter count results in slower detection speeds. Consequently, future work will focus on model lightweighting to enhance inference speed while preserving accuracy and robustness, ensuring the model is better suited for efficient real-time processing demands.

Acknowledgments

This study was funded by Liaoning Province Graduate Education Teaching Reform Research Project (LNYJG2024092) and Liaoning University of Science and Technology College Students' Innovation and Entrepreneurship Training Programme Project Funding Support No: S20211146002X.

References

[1] X. Zhang, L. Zhang, C. Liu, X. Yang, and Q. Tian, "Vehicle Detection and Classification in Traffic Surveillance Using Convolutional Neural Networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 6, pp. 1749–1759, Jun. 2017, doi: 10.1109/TITS.2017.2755953.

[2] REDMON J, DIVVALA S, GIRSHICK R, et al. "You only look once: unified, real-time object detection." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: pp. 779-788.

[3] REDMON J, FARHADI A. YOLO9000: Better, faster, stronger. Pro-ceedings of the lEEE Conference on Computer Vision and PatternRecognition.Hawaii: EEE,2017:7263-7271.

[4] REDMON J,FARHADI A.Yolov3: An incremental improvement. ArXi Preprint, 2018, arXiv: 1804.02767.

[5] LIU W, ANGUELOV D, ERHAN D, et al. "SSD: Single shot multibox detector." European Conference on Computer Vision. Cham: Springer, 2016: pp. 21-37.

[6] FU C Y, LIU W, RANGA A, et al. DSSD: Deconvolutional SingleShot Detector. ArXiv Preprint, 2017, arXiv: 1701.06659.

[7] LI Z, ZHOU F. FSSD: Feature fusion single shot multibox detector. ArXi Preprint, 2017, arXiv:1712.00960.

[8] Zhou K, Chen L, Cao X. Improving multispectral pedestrian detection by addressing modality imbalance problems [C]// Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23 – 28, 2020, Proceedings, Part XVIII 16. Springer International Publishing, 2020: 787-803.

[9] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. Fusiongan: A generative adversarial network for infrared and visible image fusion. Information Fusion, 48:11– 26, 08 2019. 1, 6, 8.

[10] Chen, Y., Liu, J., & Zhao, Z. (2018). End-to-End Multi-Scale Road Detection with Fully Convolutional Networks.

Proceedings of the IEEE International Conference on Computer Vision (ICCV), 3627-3635.

[11] Liu, J., Zhang, J., & Li, M. (2019). Multi-Scale Convolutional Neural Networks for Robust Road Detection. Journal of Visual Communication and Image Representation, 64, 1-9.

[12] Zhou, C., & Li, Z. (2023). Small object detection in traffic scenes using dual-attention networks. Journal of Machine Learning Research, 24(95), 1-15.

[13] Jocher, G., Chaurasia, A., Qiu, J.: YOLO by Ultralytics (version 8.0.0). GitHub (2023).

[14] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[M]‖Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018.Lecture notes in computer science. Cham: Springer,2018, 11211: 3-19.

[15] DING X H, ZHANG X Y, HANJG, et al. Diverse Branch Block: Building a Convolution as an Inception-Like Unit[C]‖ Proceedings of the IEEE Conference on Computer Visionand Pattern Recognition. 2021:10886-10895.

[16] LIU J, FAN X, HUANG Z, et al. Target-aware dual adversarial learn-ing and a multi-scenario multi-modality benchmark to fuse infraredand visible for object detection, Proceedings ofthe IEEE/CVF Con-ference on Computer Vision and Pattern Recognition (CVPR). NewOrleans: IEEE,2022:5802-5811.

[17] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. 2022. Target-aware dual adversarial learning and a multiscenario multi-modality benchmark to fuse infrared and visible for object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 5802–5811.