Improved Infrared Road Object Detection Algorithm Based on Attention Mechanism in YOLOv8

Zilong Luo, Ying Tian

Abstract—**In Currently, research in the field of infrared road object detection is primarily focused on enhancing model performance and robustness to address the challenges posed by complex real-world driving scenarios. In response to these challenges, this paper proposes an infrared road object detection algorithm based on an attention mechanism. By incorporating the CPCA module, which utilizes attention mechanisms, into the YOLOv8s model, the algorithm enhances the model's focus on unobstructed areas and highly illuminated sections, extracting crucial feature information to improve both accuracy and robustness. Additionally, the original model's downsampling layer is replaced with the Context Grided Network Block Downsampling (CGBD) module, which not only preserves feature edge information but also effectively handles local and contextual features, thereby enhancing the overall feature capturing capabilities of the model. To address the issue of equal aspect ratios in the model's original loss function, the proposed algorithm adopts the superior Weighted Intersection over Union (WIoU). This not only addresses the shortcomings of the original loss function (CIoU) but also demonstrates increased sensitivity in classification tasks. Experimental results show that the improved algorithm, compared to YOLOv8s, achieves a 1.4% increase in mean average precision (mAP), along with notable improvements in precision and recall. Furthermore, when compared to mainstream model algorithms, the enhanced model significantly outperforms in infrared road object detection tasks, providing validation of its effectiveness.**

Index Terms—**Deep Learning, Infrared Images, Object Detection, YOLOv8**

I. Introduction

Infrared road object detection is a crucial application of infrared technology in the fields of traffic management and intelligent driving. In the face of increasingly complex urban traffic scenarios, infrared road object detection captures the thermal radiation properties of moving vehicles, pedestrians, obstacles, and other objects. This process facilitates vehicle identification and tracking to ensure safety during the journey. In comparison to traditional visible light camera technologies, infrared technology offers unique advantages, particularly in conditions such as nighttime, adverse weather, or low lighting, providing viable technical means for enhancing the robustness and reliability of traffic monitoring systems.

In the current research on infrared road target detection, various advanced image-processing techniques and machine-learning algorithms have been introduced. These include feature extraction, target detection, and object tracking based on infrared images. As an important branch of machine vision, visual object tracking integrates the related technologies of image detection and image processing, which has important research significance and great challenge [1]. The rise of deep learning technology in recent years has brought breakthroughs to infrared road target detection, with deep learning models demonstrating remarkable performance improvements in accuracy and real-time capabilities.

The development of infrared road target detection not only contributes to elevating the level of traffic safety but also establishes a solid foundation for the further advancement of intelligent traffic systems and autonomous driving technologies [2]. By leveraging the unique advantages of infrared technology, research, and applications in this field are poised to provide innovative solutions for constructing safer and more efficient road traffic environments in the future.

The target detection technology is mainly divided into single-stage algorithms and two-stage algorithms. Two-stage algorithms include R-CNN [3], Fast R-CNN [4], Faster R-CNN [5], etc. The characteristic of such algorithms is to divide the target detection task into stages. In the first stage, candidate boxes are generated, and these boxes attempt to include the target as much as possible in the input target image. Candidate boxes can be generated based on a Region Proposal Network (RPN), either through sliding windows or using deep learning models. The second stage of two-stage algorithms refines and adjusts the generated candidate boxes to determine the final target position and category. This stage typically utilizes convolutional neural networks or other deep learning frameworks to extract image features. It then performs classification regression on the generated candidate boxes to determine whether they contain the target, calibrate the position of the candidate boxes, and better fit the target. Two-stage algorithms perform well in various target detection tasks, making system optimization and adjustment easier, and demonstrating good performance.

In addition, single-stage algorithms have also developed rapidly, including YOLO [6-9] series algorithms, SSD [10], RetinaNet [11], etc. Unlike two-stage algorithms, single-stage algorithms directly complete the object detection task with a single end-to-end model. They do not need to go

Manuscript received November 23, 2023; revised March 28, 2024.

Zilong Luo is a postgraduate student majoring in software engineering at School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Liaoning 114051, China. (e-mail: 1753372894@qq.com).

Ying Tian is a Professor of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China. (corresponding author to provide phone: +8613898015263; fax: 0412-5929818; e-mail: astianying@126.com).

through the two-stage process of candidate box generation and classification regression, simplifying the process and improving the real-time performance of detection. The single-stage algorithm transforms the object detection task into predicting the location and class of the object on the image.

Instead of just generating some candidate boxes, they can simultaneously predict multiple locations in the image. The single-stage algorithm can also perform object detection through multiscale feature maps to capture objects of different sizes, which improves the detection performance of the algorithm for objects of various sizes. Single-stage algorithms are mainly applied in scenarios where real-time processing is required. Compared to two-stage algorithms, they are usually more concise, more computationally efficient, and more suitable for domains such as autonomous driving, surveillance recognition, and face detection, providing high detection performance while completing the task more quickly.

YOLOv8 was chosen as the base model algorithm to better implement the deployment of an infrared road object detection system in real-world applications. YOLOv8 is the latest version of the YOLO series of algorithms. Compared to the previous YOLOv7 [12] model, it significantly reduces parameters and computational complexity, achieving faster detection speeds and better suitability for real-life deployment. The YOLOv8 algorithm introduces a new SOTA model, including target detection networks with resolutions of 640 and 1280, and the instance segmentation model YOLACT. Its backbone network and Neck section are designed based on the ELAN concept from YOLOv7. Unlike YOLOv5, YOLOv8 replaces the C3 structure with the C2f structure, providing a richer gradient flow for enhanced feature extraction. The Head section separates the classification and detection heads, adopting a decoupled head structure and incorporating data augmentation strategies from YOLOX [13] into the YOLOv8 model. In summary, compared to other target detection algorithms, YOLOv8 performs excellently in terms of performance, accuracy, and computational efficiency. It can identify object classes and bounding boxes in an image with just one detection, making it suitable for various real-world applications. However, in the test task of infrared road target detection, the original YOLOv8 algorithm performs poorly. This is due to the lower recognition, higher complexity, and tendency to miss small targets in infrared images compared to visible light images. To address these issues, an improved algorithm based on the YOLOv8s model, named YOLOv8-ITA (Infrared Target Attention), is proposed for the first time.

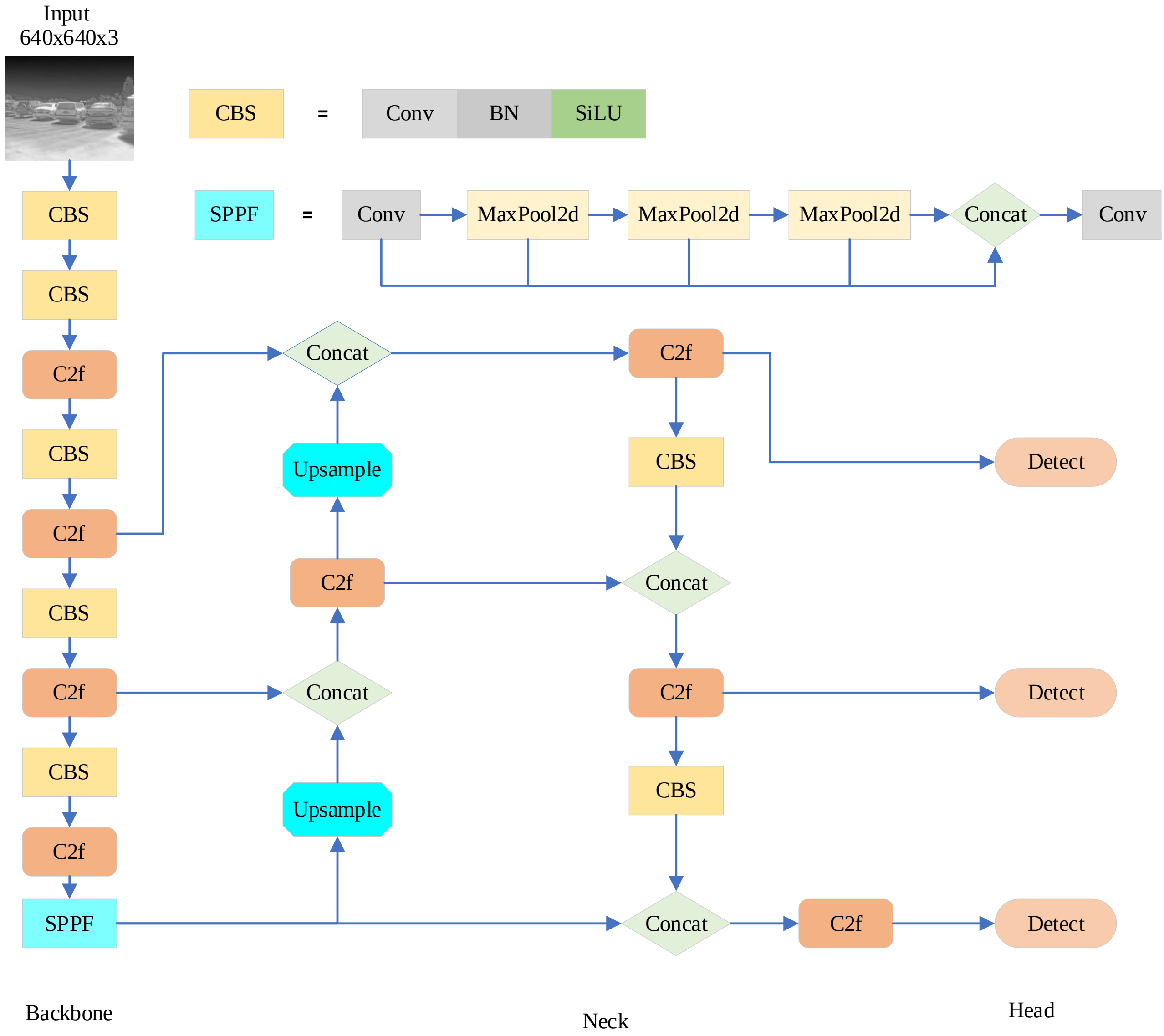


Fig. 1. YOLOv8-ITA network architecture

II. Improved Model

In this paper, we introduce enhancements based on the YOLOv8s base model that involve the replacement of specific modules to improve detection accuracy, specifically targeting the requirements of infrared road target tasks for better detection results. To augment the perception of specific areas in infrared images, the CPCA [14] attention mechanism has been incorporated into the model's neck section. Additionally, the original convolutional layers have been substituted with ContextGuidedBlock [15] modules, capable of capturing spatial feature correlations from three levels. The CIoU [16] loss function has been replaced with a monotonic focus mechanism for cross-entropy, WIoU v2 [17], to mitigate the contribution of simple tasks to the loss value, thereby enhancing the performance of the classification task. The overall model structure, depicted in the figure, is divided into three parts. The initial part constitutes the backbone network, responsible for the primary feature extraction of input images. Extracted features are then fed into the subsequent neck section for feature fusion before being transmitted to the head section for classification and detection, as illustrated in Fig. 1.

A. *CPCA attention mechanism*

In infrared road images, targets may exhibit different temperature characteristics or varying brightness and reflections in the infrared spectrum due to changes in lighting conditions. By employing an attention mechanism, the model can more selectively focus on regions with crucial temperature or brightness features. It also aids the model in concentrating on areas that are not obscured or disturbed, thereby enhancing the accuracy and robustness of target detection. The Channel Prior Convolutional Attention (CPCA) mechanism enables the dynamic distribution of attention weights across channels and spatial dimensions. Through the utilization of a multi-scale deep convolutional module, the model can effectively capture relationships at different spatial scales when processing input data. This module's design enables the network to simultaneously learn features with distinct abstraction levels, preserving channel priors for the input data. This contributes to improving the model's understanding of local and global structures in the input data, thereby enhancing its performance in handling complex tasks. The CPCA structure is illustrated in Fig. 2.

The channel attention of CPCA inherits the approach of CBAM [18] attention mechanism, which utilizes max pooling and average pooling to extract spatial information, followed by feeding it into an MLP, and finally applying a sigmoid function. The calculation formula is as follows:

$$AP(F)=MLP(AvgPool(F)) \tag{1}$$

$$MP(F)=MLP(MaxPool(F)) \tag{2}$$

$$CA(F)=\sigma(AP(F)+MP(F)) \tag{3}$$

The spatial attention of CPCA is achieved by applying depthwise separable convolution, allowing us to effectively capture spatial relationships between features, and ensuring the preservation of correlations between channels while simultaneously reducing computational complexity. The introduction of a multi-scale structure design further enhances the convolution operation's ability to capture spatial relationships. Finally, by employing a 1×1 convolutional layer, a mixture of channel features is achieved, thereby further enhancing the network's performance in feature extraction and representation learning. The formula is as follows:

$$DC(F)=\sum_{i=0}^{3} Branch_i(DwConv(F)) \tag{4}$$

$$SA(F)=Conv_{1\times1}(DC(F)) \tag{5}$$

DWConv stands for depthwise convolution, and the branch represents the i-th branch in it. Therefore, the overall structure of the CPCA attention mechanism consists of calculating channel attention first, followed by obtaining spatial attention. The formulas are as follows:

$$F_c=CA(F)\otimes F \tag{6}$$

$$\hat{F}=SA(F_c)\otimes F_c \tag{7}$$

Furthermore, compared to other attention mechanisms such as SE [19], which focuses more on channel attention but lacks spatial dimension information capture, and CBAM, which captures both channel and spatial information but compresses channel calculations, resulting in a consistent spatial attention weight distribution for each channel during element-wise multiplication with input features.

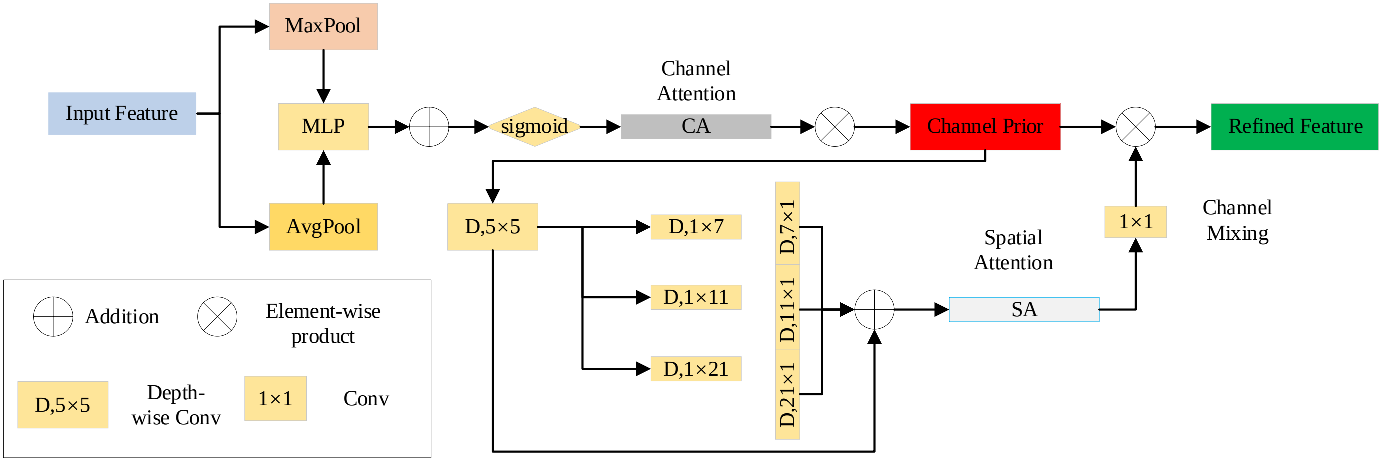


Fig. 2. Diagram of the CPCA module

This limitation restricts the adaptive capability of the attention mechanism because spatial attention weights cannot dynamically adjust based on the specific characteristics of each channel. The CPCA attention mechanism effectively addresses these issues and demonstrates significant improvements in infrared road target images.

B. CGBD module

In the YOLOv8-ITA model, the CGBD downsampling module is employed to replace the original convolutional module in the YOLOv8 network. Unlike a conventional downsampling module, the CGBD module excels in preserving edge information, and effectively handling local and contextual features. Additionally, the design of this network module allows global contextual features to permeate the entire network, spanning from low-level (spatial level) to high-level (semantic level), not confined to capturing contextual features solely after the encoding stage. Overall, in the task of infrared road target detection, this network module achieves a better balance between local and contextual features, enhances model performance by introducing global contextual information, and better preserves edge information by reducing the number of downsampling layers.

This module consists of four parts: f_loc, which extracts local feature information constructed by regular convolution; f_sur, which extracts surrounding contextual feature information implemented through a dilated convolution module; f_joi, responsible for joint feature extraction used to concatenate the subsequent BN layer and the PReLU activation function; and finally, the global feature extractor f_glo, which introduces a global pooling layer followed by two fully connected layers for feature extraction. The outputs of these two fully connected layers form a weight vector, which is then utilized to guide the fusion of joint features. The CGBD module is illustrated in Fig. 3.

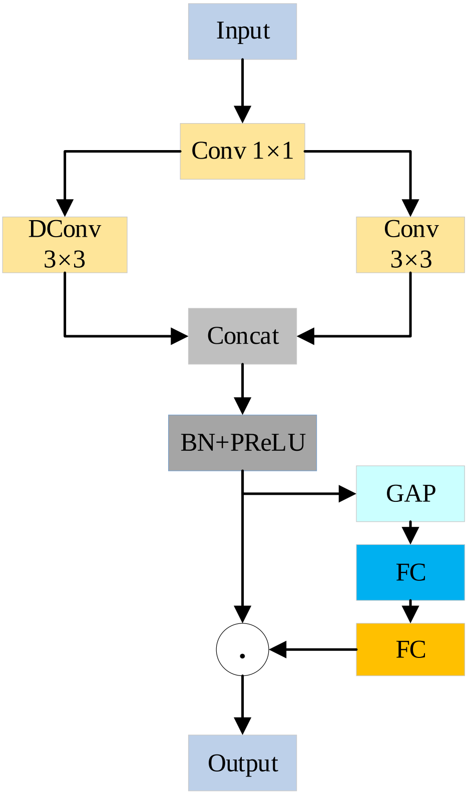


Fig. 3. Diagram of the CGBD module

In the overall process, the input image undergoes processing through a 1×1 convolutional layer. Subsequently, the processed image is passed through two 3×3 convolutional layers: one is a standard convolutional layer (referred to as f_loc), and the other is a dilated convolutional layer (referred to as f_sur). These two convolutional layers are employed to extract local features and the surrounding features, respectively. After obtaining the local and surrounding features, these two features are fused through feature concatenation. The fused features then undergo normalization (BN) and processing through the PReLU activation function to obtain the final fused features. Subsequently, the fused features are subjected to channel-wise global feature processing (f_glo), contributing to a more global capture of feature information. The ultimate output represents the result obtained through the channel-wise global feature processing, signifying the comprehensive understanding of the input image by the CGBD module.

C. WIoU loss function

The original YOLOv8 model utilizes the CIoU loss function, which builds upon DIoU [16] by introducing a scale loss related to the detection box dimensions, thereby incorporating additional loss related to aspect ratio information. This design aims to stabilize the regression of the target box, preventing divergence issues during training, as observed in IoU and GIoU [20]. By considering comprehensive scale information of the target box, the CIoU loss contributes to the overall stability of the training of object detection models.

However, in CIoU, if the aspect ratios of the predicted and ground truth boxes are the same, the aspect ratio penalty remains zero, which is logically unreasonable. The gradients of width (w) and height (h) relative to the overall loss (v) in CIoU are found to be opposite. This implies that w and h cannot increase or decrease simultaneously, which is also deemed illogical.

In contrast, WIoU (Weighted Intersection over Union) introduces a weighted consideration of the area between the predicted and ground truth boxes, providing a solution to potential biases in traditional IoU evaluation. By incorporating weights, WIoU more comprehensively considers the overlapping region of the target box, thereby enhancing evaluation accuracy. WIoU v2 adopts a consistent monotonic focus mechanism with CIoU and SIoU v2 [21]. This customized cross-entropy monotonic focus mechanism significantly reduces the impact of simple examples on loss values. This allows the model to focus more on challenging examples, considering the complexity of the infrared road target scenarios, thereby improving the model's classification performance and, consequently, detection accuracy.

WIoU first calculates the Intersection over Union (IoU) between the predicted and ground truth boxes, which measures the degree of overlap in object detection tasks. The formula for the anchor frame and target box is as follows:

$$\vec{\mathrm{B}} = [\mathrm{x}, y, w, h] \tag{8}$$

$$\vec{\mathrm{B}}_{\mathrm{gt}} = [\mathrm{x}_{\mathrm{gt}}, y_{\mathrm{gt}}, w_{\mathrm{gt}}, h_{\mathrm{gt}}] \tag{9}$$

The degree of overlap between the predicted box and the ground truth box is illustrated in Fig. 4. The formula for calculating their union is as follows:

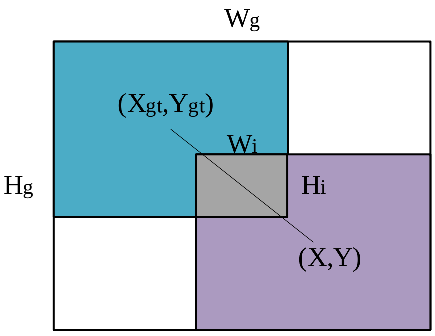


Fig. 4. The area of the union

$$S_u = wh + w_{gt}h_{gt} - W_iH_i \quad (10)$$

$$L_{IoU} = 1 - \frac{W_iH_i}{S_u} \quad (11)$$

R_{WIoU} significantly amplifies the anchor box L_{IoU}, and the formulation is constructed as follows:

$$R_{WIoU} = \exp\left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{(W_g^2 + H_g^2)^*}\right) \quad (12)$$

The "*" in the formula is used to detach W_g and H_g from the computation graph when R_{WIoU} hinders the convergence of gradients. In the end, $L_{WIoU\,v1}$ is obtained through R_{WIoU}:

$$L_{WIoU\,v1} = R_{WIoU}L_{IoU} \quad (13)$$

What we obtain here is WIoU v1 with two layers of attention mechanisms. To obtain the monotonic focusing mechanism for WIoU v2, we still need the monotonic focusing coefficient $L_{IoU}^{\gamma*}$ for $L_{WIoU\,v1}$, then introduce the mean of L_{IoU} for normalization, and finally obtain the:

$$L_{WIoU\,v2} = \left(\frac{L^*_{IoU}}{L_{IoU}}\right)^{\gamma} L_{WIoU\,v1} \quad (14)$$

In WIoU, the weight value for each bounding box is determined based on its overlap with the ground truth annotation box. When the overlap between the bounding box and the ground truth box is high, the weight value is correspondingly high; conversely, when the overlap is low, the weight value is low. Through this mechanism, in the infrared road target detection task, WIoU can more effectively assess the detection results, providing more accurate evaluations even in the presence of size-imbalanced objects.

III. Experimental Results And Discussion

A. Datasets

In Revised sentence: For this experiment, we intentionally selected the FLIR_ADAS_v2 infrared dataset recently released by FLIR Systems [22]. This updated version not only expanded the number of labels to 15 categories but also introduced video data, resulting in a total of 26,442 annotated frames - a 1% increase from the original version. All images included in this dataset are labeled with these 15 categories. We chose to use 10,467 infrared images for our dataset processing and partitioned them into training (7,326), testing (2,094), and validation sets (1,047) at a ratio of 7:3:1. To ensure sufficient training and testing for our chosen categories - person, bike, car, bus, light and sign - we carefully selected these six categories while maintaining balance within the dataset. Our selection strategy aims to improve model generalization ability and better adapt it to real-world scenarios in infrared object detection tasks.

B. Experimental Environment

The model was developed using the Python programming language and implemented with the PyTorch deep learning framework, specifically using PyTorch version 1.8.1. The training process was conducted on hardware based on the GeForce GTX 1080ti with a VRAM size of 11,178MB. For model training, the input image size was set to 640 x 640 pixels. The optimizer used was the Stochastic Gradient Boosting (SGB) function. The training process consisted of 300 epochs with a batch size of 8. The momentum and decay parameters were set to 0.937 and 0.0005, respectively. The initial learning rate was set to 0.01, and a cosine annealing algorithm was applied. Additionally, the mosaic augmentation technique was employed in the last 10 epochs of training to enhance model performance.

C. Experimental Evaluation Metrics

In this study, three model evaluation metrics were employed, including precision, recall, and mean Average Precision (mAP) at a threshold of 0.5. Precision: Precision is a key metric in classification tasks, indicating the proportion of true positive predictions among all samples predicted as positive. The precision is calculated using the following formula:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (15)$$

Where TP represents the number of samples correctly classified as positive, and FP represents the number of samples incorrectly classified as positive. The precision reflects the model's ability to accurately identify targets in infrared images. A higher precision indicates that the model can precisely identify targets, helping to reduce false positives and false negatives, thus enhancing the reliability and safety of the driving assistance system.

Recall, another crucial evaluation metric, represents the proportion of correctly predicted samples among all true positive instances. The calculation formula is as follows:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (16)$$

Where TP represents the number of true positives, and FN represents the number of samples incorrectly classified as negatives. The recall rate reflects whether the model can effectively identify all true positive targets. A high recall rate indicates that the model can accurately identify as many positive targets as possible, reducing false negatives and enhancing the reliability and safety of the system.

mAP@0.5 (Mean Average Precision at IoU=0.5) is a commonly used performance evaluation metric in object detection tasks. It calculates the average precision for each class and then takes the mean. In object detection, if the overlap ratio between a detection box and the corresponding ground truth box is greater than 0.5 (i.e., IOU greater than 0.5), the detection box is considered a true positive (TP). mAP@0.5 represents the average precision across all classes under the condition of IOU=0.5, and the calculation formula is as follows:

$$\mathrm{mAP} = \frac{1}{n}\sum_{i=1}^{n}\int_{0}^{1}\mathrm{Precision(Recall)d(Recall)} \tag{17}$$

Here, n represents the number of classes, AP represents the average precision for each class. The calculation of mAP considers the average of the average precision for each class, providing a comprehensive evaluation of the model's overall performance across multiple classes.

D. Blation Experiment

To investigate the impact of different enhancement strategies on the final performance of the model, ablation experiments are conducted. The enhancements include adding the CPCA attention mechanism to the model, replacing the original downsampling with the CGBD downsampling module, and substituting the original CIoU with WIoU. The evaluation metrics for the model's final performance include precision, recall, and mAP. Each module is individually validated in the experiments using a controlled variable approach to isolate the specific impact of each module on the model. The results are summarized in Table I.

The table indicates that each module exhibits certain improvements compared to the original model. Particularly, the CPCA attention mechanism module shows an increase of 0.7% in mAP, along with growth in precision and recall. This demonstrates the effectiveness of the CPCA attention mechanism module in infrared target detection tasks, showcasing its ability to address challenges such as low resolution, complex environments, and significant interference in infrared road target detection. The CGBD and WIoU modules show notable improvements in recall and precision, aligning with the enhanced perception capability of the CGBD module for feature information and the pronounced target classification ability of the WIoU module. While there may be slight shortcomings in other aspects, there is still a substantial improvement in the mAP metric.

TABLE I
RESULTS OF ABLATION EXPERIMENTS

CPCA	CGBD	WIoU	P(%)	R(%)	mAP@0.5(%)
			84.0	68.2	76.8
✓			84.1	68.7	77.5
	✓		83.6	69.4	77.3
		✓	84.6	68.9	77.3
✓	✓		83.9	69.5	77.6
✓		✓	84.0	69.1	77.6
✓	✓	✓	84.2	70.2	78.2

To better visualize the comparative effectiveness between the final model and the original model, the actual images output by the two models are contrasted. The up image depicts the detection results of the YOLOv8s model, while the down image showcases the results of the improved model, as illustrated in Fig. 5.

The figures depict an improvement in object detection by the ITA model compared to the original model, demonstrating the practical performance of the ITA model. Compare the PR curves of the two models tested, as shown in Fig. 6. Compared to the original model, the improved model has shown overall growth, with improvements in most detection categories. However, accompanying challenges persist; despite an increase in accuracy, the model's performance on small objects is not optimal, indicating a potential area for further research. In summary, the overall model exhibits a 1.4% improvement compared to the original model, with each module contributing to small advancements when deployed independently.

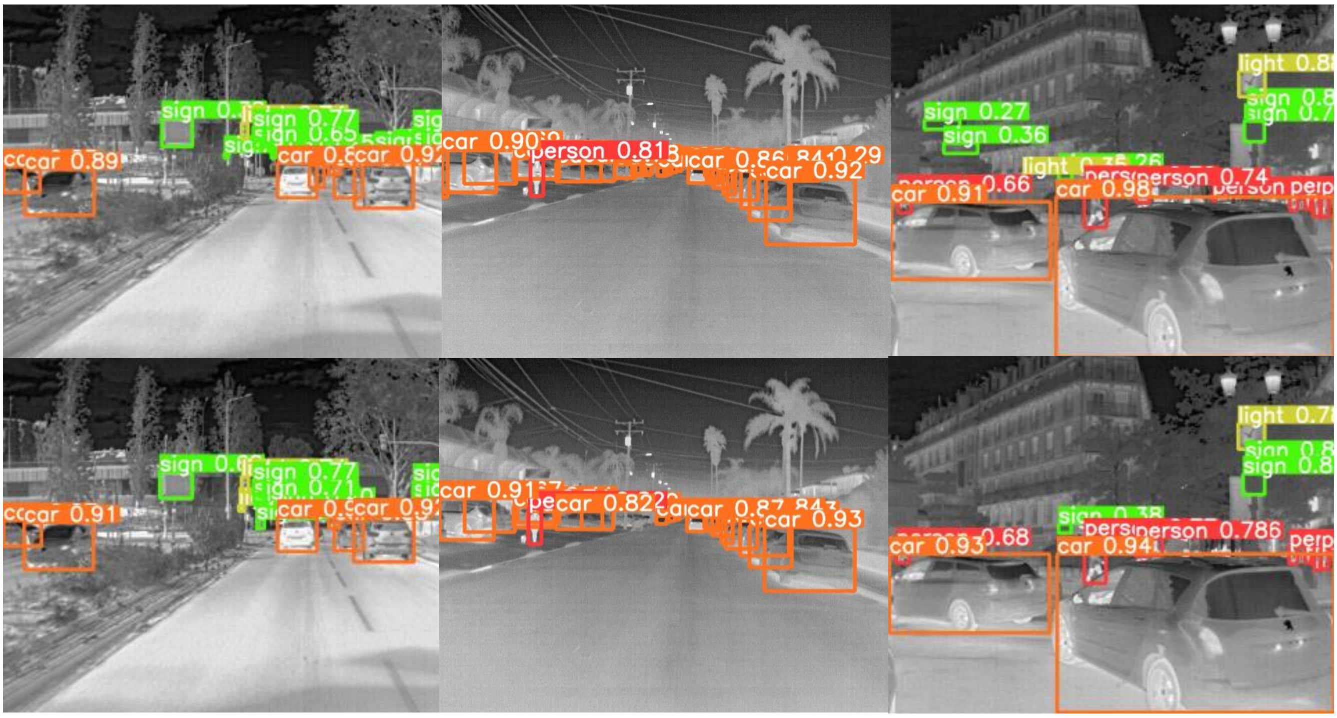
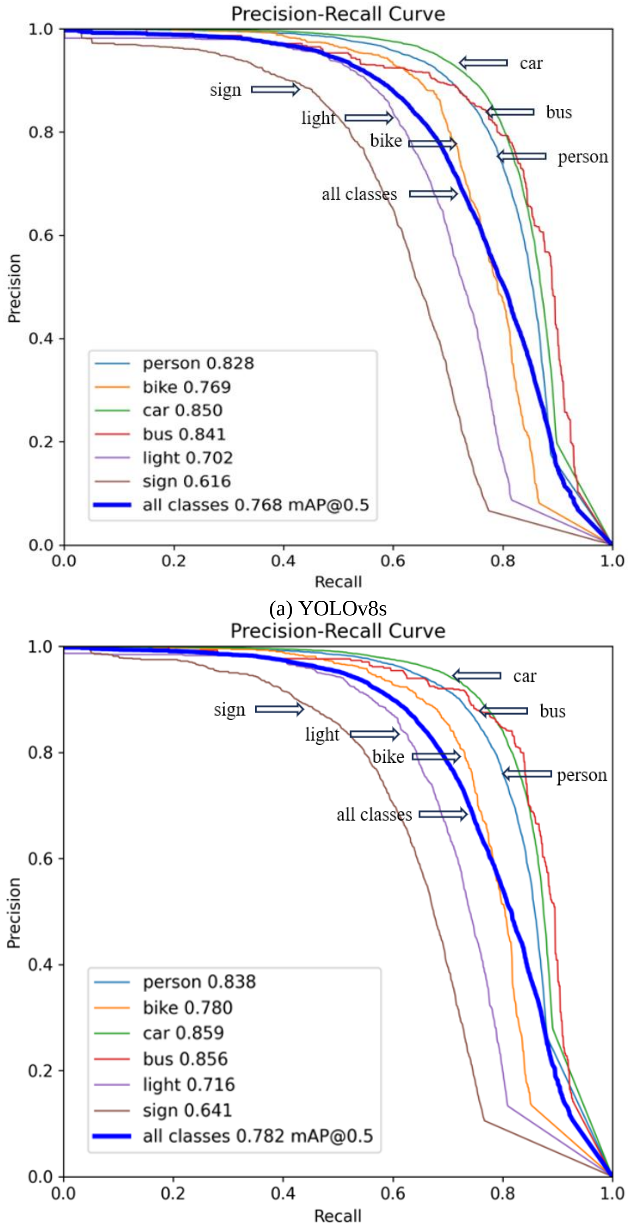


Fig. 5. YOLOv8s (up) and YOLOv8-ITA (down)



(a) YOLOv8s

(b) YOLOv8-ITA

Fig. 6. PR curve comparison

E. Comparative experiments

To provide a more intuitive comparison of the advantages of the improved model, a comprehensive evaluation was conducted by contrasting it with mainstream models, including YOLOv5m, YOLOv6n, YOLOv7-tiny, and YOLOv8s. The dataset, parameters, and evaluation metrics used in this process remained consistent with those employed for the improved model. Additionally, the parameter count and computational complexity were selected to be close to the architecture of the improved model. The final results are presented in Table II.

TABLE II
COMPARATIVE EXPERIMENT RESULTS OF DIFFERENT MAINSTREAM MODELS

Approaches	P(%)	R(%)	mAP@0.5(%)
YOLOv5m	86.3	68.4	77.4
YOLOv6n	73.1	69.5	71.2
YOLOv7-tiny	78.5	62.9	70.6
YOLOv8s	84.0	68.2	76.8
YOLOv8-ITA	84.2	70.2	78.2

A comparative analysis between the YOLOv8-ITA model and the YOLOv5m model reveals a substantial improvement in recall rate, with a corresponding increase of 0.8% in the comprehensive metric mAP. When compared with YOLOv6 and YOLOv7-tiny models, the improved model exhibits a more significant enhancement in mAP. Therefore, the overall performance of the enhanced model surpasses these mainstream models, making it more robust and well-suited for applications in infrared target detection tasks.

IV. CONCLUSION

Addressing challenges in infrared road target tasks, such as complex environments, low visibility in the infrared spectrum, and issues related to recognition rates and small target detection, we propose an improved network model, YOLOv8-ITA, based on the attention mechanism. This model enhances precision by introducing the CPCA attention module, which extracts target features from two dimensions, focusing on local features to improve model accuracy. The addition of the CGBD downsampling module extracts edge features, balancing local and contextual features, while also incorporating global contextual information to enhance overall model performance. Replacing the original CIoU with the WIoU loss function evaluates target box coverage more accurately through weighting, improving evaluation precision. Experimental results demonstrate that the enhanced model exhibits improvements over the baseline. Applied in real-world scenarios, the improved model is better suited for infrared road target detection applications, including areas such as assisted driving and road monitoring platforms.

REFERENCES

[1] Zhangfang Hu, Hongling Yu, and Kehuan Linghu, “Siamese Network Tracker Based on Dynamic Convolution and Attention Fusion of Shallow and Deep Information,” Engineering Letters, vol. 32, no. 1, pp. 30-42, 2024.

[2] Zhixian Zhang, Wenhua Cui, Ye Tao, and Tianwei Shi, “Road Damage Detection Algorithm Based on Multi-scale Feature Extraction,” Engineering Letters, vol. 32, no. 1, pp. 151-159, 2024.

[3] R. Girshick, J. Donahue, T. Darrell and J. Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation,” IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 580-587.

[4] R. Girshick, “Fast R-CNN,” IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, Dec. 2015. pp. 1440-1448.

[5] S. Ren, K. He, R. Girshick and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017.

[6] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 779-788.

[7] J. Redmon and A. Farhadi, “YOLO9000: Better, Faster, Stronger,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, Jul. 2017.

[8] J. Redmon and A. Farhadi, “YOLOv3: An Incremental Improvement,” arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition, Apr. 2018.

[9] Bochkovskiy A, Wang C Y, Liao H Y M. “YOLOv4: Optimal Speed and Accuracy of Object Detection,” arXiv preprint arXiv:2004.10934, 2020.

[10] W. Liu et al., “SSD: Single Shot MultiBox Detector,” in Computer Vision - ECCV 2016, Lecture Notes in Computer Science, 2016, pp. 21-37.

[11] Lin, Tsung Yi, et al. “Focal Loss for Dense Object Detection.” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2017, pp. 2999-3007.

[12] C.-Y. Wang, A. Bochkovskiy, and H.-Y. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors." *arXiv preprint arXiv*, 2022.

[13] Ge, Zheng, et al. "YOLOX: Exceeding YOLO Series in 2021," arXiv preprint arXiv:2107.08430, 2021.

[14] H. Huang, Z. Chen, Y. Zou, M. Lu, and C. Chen, "Channel prior convolutional attention for medical image segmentation," Jun. 2023.

[15] T. Wu, S. Tang, R. Zhang, J. Cao, and Y. Zhang, "CGNet: A Light-Weight Context Guided Network for Semantic Segmentation," *IEEE Transactions on Image Processing*, Jan. 2021, pp. 1169-1179.

[16] Zheng, Zhaohui, et al. "Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression." *Proceedings of the AAAI Conference on Artificial Intelligence*, June 2020, pp. 12993-13000.

[17] Tong, Zanjia, et al. "Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism." arXiv preprint arXiv:2301.10051 (2023).

[18] Woo, Sanghyun, et al. "CBAM: Convolutional Block Attention Module." *Computer Vision – ECCV 2018, Lecture Notes in Computer Science*, 2018, pp. 3-19.

[19] Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-Excitation Networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7132-7141.

[20] Rezatofighi, Hamid, et al. "Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression." *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[21] Gevorgyan, Zhora. "SIoU loss: More powerful learning for bounding box regression." arXiv preprint arXiv:2205.12740 (2022).

[22] Venkataraman V, FAN G, FAN X. "Target tracking with online feature selection in FLIR imagery." *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2007: pp. 1-8.