Journal of Artificial Intelligence Practice (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2025.080116
ISSN 2371-8412 Vol. 8 Num. 1

Research on the application of artificial intelligence and multi-scale image fusion technology to pedestrian detection in complex street view

Gong Li

Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia

Keywords: Pedestrian detection, Computer vision, Multiscale networks, YOLOv10s

Abstract: With the increasing face imaging data and the advancement of artificial intelligence (AI) technology, computer-aided monitoring systems are crucial for pedestrian detection in dense street view. However, due to occlusion and small pedestrian scale, pedestrian false alarms and missed detection problems become more and more serious. Therefore, this paper proposes a pedestrian detection model, YOLOv10s-pedestrian. Firstly, CA attention is introduced to redesign the MBConv module, resulting in an efficient MB-CANet backbone for pedestrian feature extraction, enhancing the accurate localization of densely occluded pedestrians. Secondly, a novel C2FN structure was created to reduce the number of parameters while improving the model's accuracy. Additionally, inspired by the BiFPN feature fusion concept, a Bi-C2FN-FPN network structure is proposed to effectively fuse features from different depth sources, strengthening feature fusion and improving pedestrian detection accuracy. Finally, the MPDIOU loss function replaces the original CIoU loss function to enhance anchor box localization. Experimental results demonstrate that the proposed model achieves a mAP50 of 95.6% on the WiderPerson pedestrian detection dataset, which is a 6.1% improvement over the original model, with a recall rate of 86.2%, showcasing excellent detection performance. Compared to several mainstream object detection models, the proposed model also exhibits superior performance.

1. Introduction

As society develops and populations grow, urban security management faces unprecedented challenges. The deployment of road surveillance equipment has significantly improved urban security monitoring. Pedestrian detection, which aims to accurately locate pedestrians in images or videos, has gained considerable research attention due to its importance in traffic safety and urban management [11]. Pedestrian detection is widely used in traffic monitoring, intelligent security [6], and autonomous driving [1].

Deep learning-based pedestrian detection algorithms are generally classified into two-stage and one-stage methods. Two-stage methods, such as R-CNN, Fast R-CNN, and Faster R-CNN, first generate candidate regions and then refine the results. One-stage methods, like YOLOv9, YOLOv8, YOLOv7, YOLOv5, and SSD, directly predict object categories and locations, balancing speed and accuracy. While two-stage methods were historically more accurate but slower, YOLOv5 and its successors have outperformed them in both accuracy and speed.For example,in general urban street scenes, object detection algorithms have already shown excellent performance[14].

Despite strong performance in general street scenes [12], challenges remain in complex scenarios [3], such as dense pedestrian occlusion and small target detection [4], which lead to false positives and missed detections. Liu et al.proposed the MCF-CP-NET model with multi-scale fusion and channel attention to improve occluded pedestrian detection, though it lacks real-time capability. Guo et al. [7] used the AdaBoost algorithm and cascade method for monocular pedestrian detection, while Hsu et al. [9] enhanced YOLOv4 with an MsSE-SR method to improve low-resolution detection. Yuan et al.[5] developed a closed-loop detection network to address overlapping pedestrians, and Jain et al. [10] proposed the RMPD-DCNN-EL model using EfficientNet for robust detection in multimodal scenarios, though performance may degrade in complex intersections.In summary, scale variations, small targets, and occlusions remain key challenges in pedestrian detection. Further research is needed to enhance accuracy and real-time performance in complex street scenes. To address these issues, further in-depth research is required. The main contributions of this paper are as follows:

This paper proposes the YOLOv10s-Pedestrian model for dense pedestrian occlusion and small target detection in complex street scenes. The model enhances the capture of pedestrian features under dense occlusion by constructing a new MB-CANet backbone network.

To improve YOLOv10s's ability to detect distant, blurred, and small pedestrians in complex environments, the C2FB module was proposed to enhance the network's capability to learn small targets.

To address the deficiency of the original YOLOv10s algorithm's feature pyramid network in lacking original information in feature fusion structures, this paper proposes the Bi-C2FN-FPN network structure by drawing on the feature fusion concept of BiFPN. This structure fully utilizes deep, shallow, and original features, strengthens feature fusion between different parts, reduces feature information loss during the convolution process, and improves the detection accuracy of the algorithm.

To enhance YOLOv10s' ability to capture pedestrian features in densely occluded scenes, a dynamic non-monotonic focal loss function MPDIOU is introduced to precisely predict the position of bounding boxes.

2. Related work

2.1. YOLOv10

The YOLOv10n model, proposed by Tsinghua University in 2024, is the latest version in the YOLO series, introducing a consistent dual assignment strategy for NMS-free training, enabling efficient end-to-end detection. It includes six scalable versions (YOLOv10n, YOLOv10s, YOLOv10m, YOLOv10l, YOLOv10x, YOLOv10b) and demonstrates strong performance in pedestrian detection in complex scenes. YOLOv10s, with its well-designed backbone network and effective multi-scale feature fusion, accurately handles targets of various sizes. This paper uses YOLOv10s as the benchmark model, enhancing its detection accuracy and efficiency in occluded pedestrian scenarios, thereby improving performance in complex environments.

2.2. Efficient backbone network

CNNs have relatively low computational efficiency when capturing long-distance dependencies. Without excessive repetition of convolution operations and stacking of multiple layers, the interaction of distant semantic information cannot be achieved. The network, based on CNN and EfficientNet, employs a concurrent structural design that integrates multi-resolution local and global features. This allows the model to learn more target weights, resulting in excellent performance in multi-scale object detection. Inspired by the above, this paper introduces MBConv convolution into the YOLOv10s network to enable the network to effectively capture object features.

2.3. YOLOv10s

Similar to the human perception process, the attention mechanism aims to focus more on informative areas while paying less attention to non-essential areas, demonstrating excellent performance in various image processing tasks. The attention mechanism is typically inserted after convolutional blocks to enhance the network's ability to handle short-term and long-term dependencies, thereby improving the learning capability of pedestrian features in complex scenes [2]. Therefore, this paper introduces a powerful CA[8] attention mechanism in the proposed network. This mechanism maintains high internal resolution and integrates a SoftMax-Sigmoid combination only within the channel and spatial attention blocks. By integrating the attention mechanism into the convolutional neural network, CA can automatically learn the importance of different regions in an image, thereby extracting image features more effectively and improving classification and detection accuracy.

3. YOLOv10s-Pedestrian

As displayed in Figure 1, YOLOv10s-Pedestrian is composed of the MB-CANet structure. This enhances performance in feature extraction, multi-feature fusion, and prediction output. To balance detection accuracy and speed, this paper introduces the MBConv module into the baseline YOLOv10s. The MB-CANet network structure is constructed by incorporating the CA attention mechanism into the MBConv structures at the 6th and 12th layers of the backbone network. Additionally, the Bi-C2FN-FPN structure is proposed, which is constructed from C2FN and BiFPN.

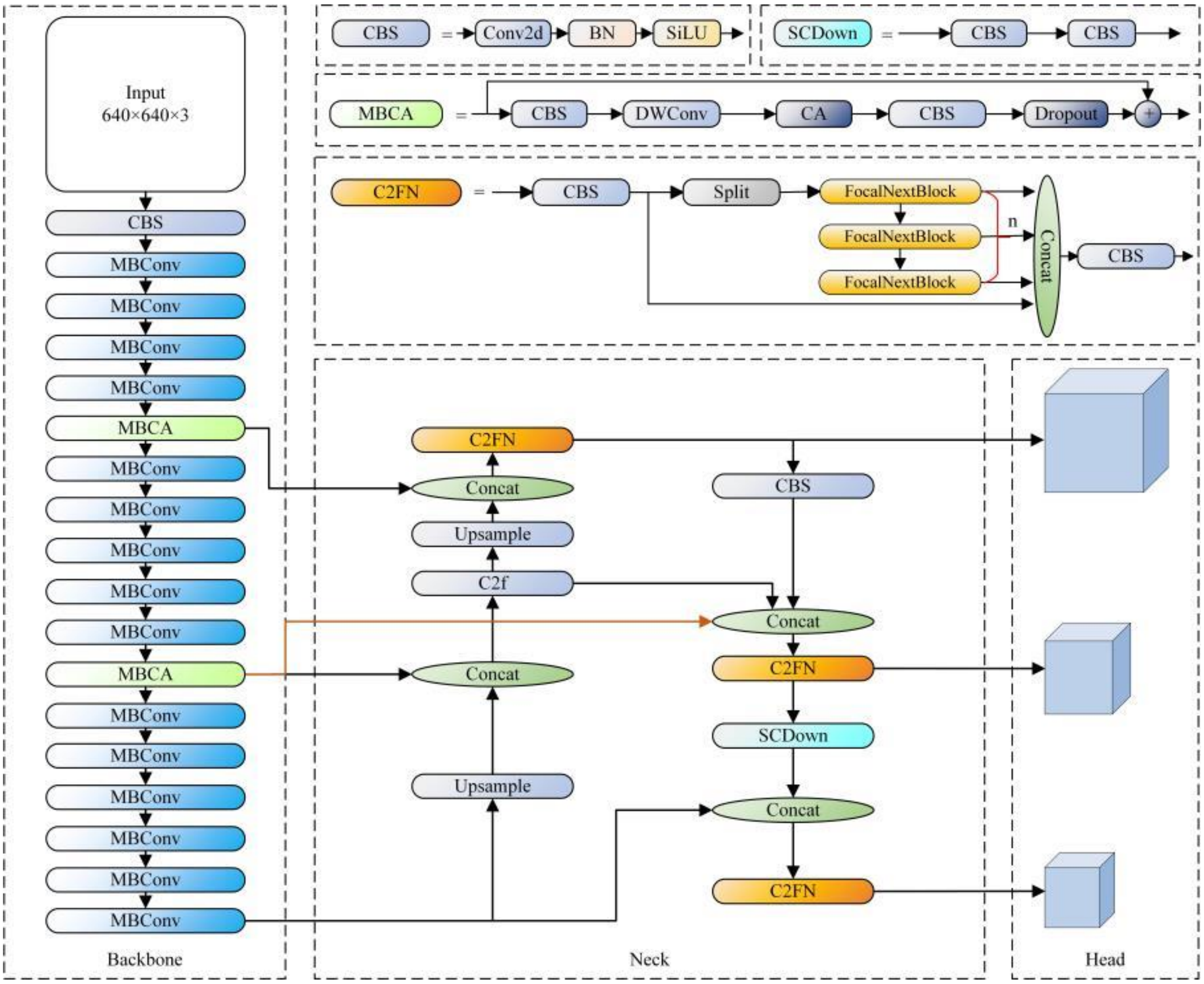


Fig. 1. YOLOv10s-Pedestrian network structure.

3.1. MB-CANet backbone network

EfficientNet, proposed by Google, is a neural network architecture designed to maintain high accuracy while reducing model parameters and computational complexity. It improves the ability to extract features from targets in images with complex dense occlusions by adjusting the number of convolutional modules. Additionally, it effectively avoids issues such as gradient vanishing and high computational costs, achieving optimal detection performance. Therefore, this paper draws on EfficientNet to propose an MB-CANet backbone network with a hybrid scaling method to obtain the optimal scaling factors for network width, depth, and input image resolution. These scaling factors are used to scale the network's width, depth, and input image resolution. The MB-CANet network structure comprises a convolution module and multiple MBConv and MBCA modules. The MB-CANet backbone network replaces the original backbone network of YOLOv10s, as shown in Figure 2. Compared to the original YOLOv10s backbone network, the proposed backbone network better balances training speed and accuracy, enhancing the network's detection performance.

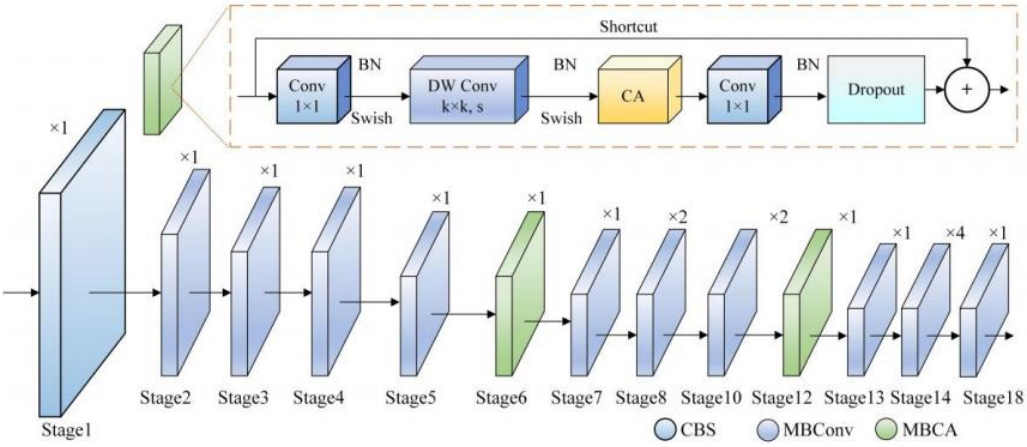


Fig. 2. MB-CANet structure.

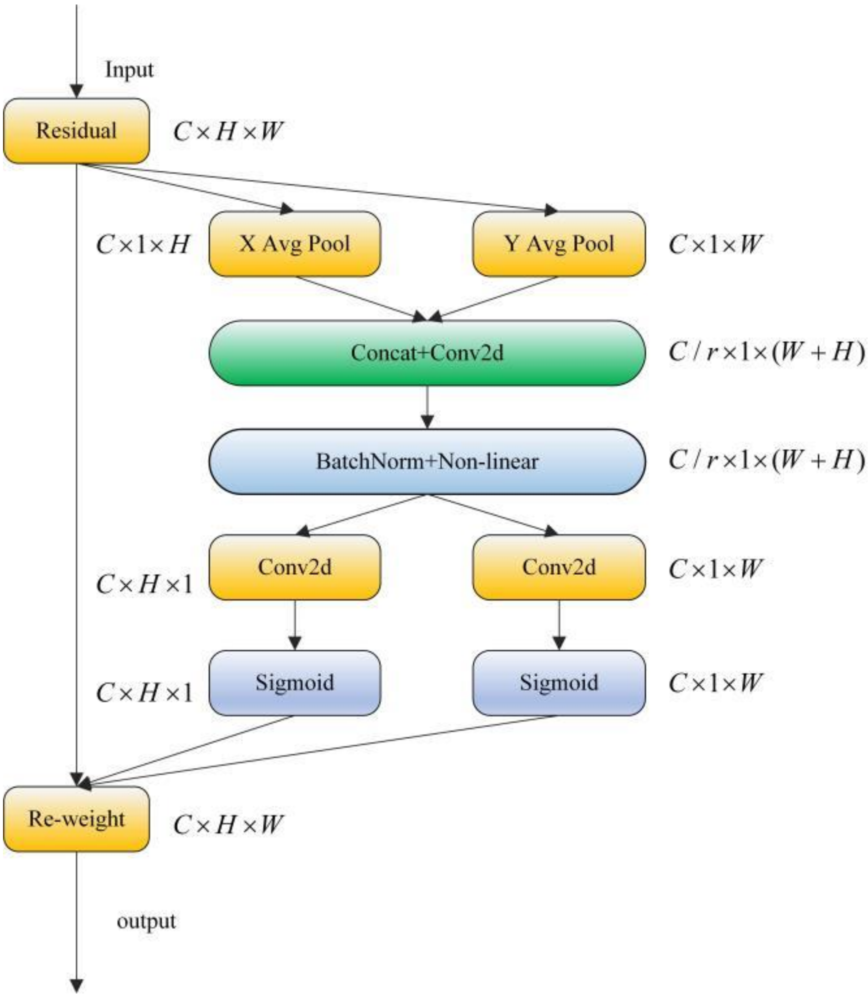


Fig. 3. Structure of the CA module.

The MB-CANet backbone network consists of 18 stages, with Stage 1 as a convolutional layer and Stages 2 to 18 using MBConv and MBCA structures to improve feature extraction by increasing network width and depth. As shown in Figure 3, Stage 6 (5×5 convolution + CA module) outputs large-scale features from Stages 1 to 5, while Stage 12 (5×5 convolution + CA module) outputs medium-scale features from Stages 7 to 11. Stage 18 (3×3 convolution) outputs small-scale feature maps. The CA mechanism in Stages 6 and 12 enhances focus on pedestrian features in dense environments by refining feature maps along spatial and channel dimensions, improving target detection. Replacing the original YOLOv10s backbone with MB-CANet reduces computation while maintaining high accuracy, improving robustness and generalization for pedestrian detection in complex street scenes.

3.2. C2FN module

In pedestrian detection on complex streets, the YOLOv10s network is prone to interference from obstacles when extracting image feature information, causing the model to focus only on local pixel positions. To effectively utilize context for capturing target information, it is necessary to stack convolutional layers multiple times. However, directly stacking these layers leads to low computational efficiency and difficulties in optimizing the model. To address this issue, this paper constructs a new C2FN module using a cross-layer connection approach. The C2FN module consists of the C2f module and the FocalNextBlock module. These modules significantly enhance the model's ability to capture contextual information. The C2FN structure is illustrated in Figure 4.

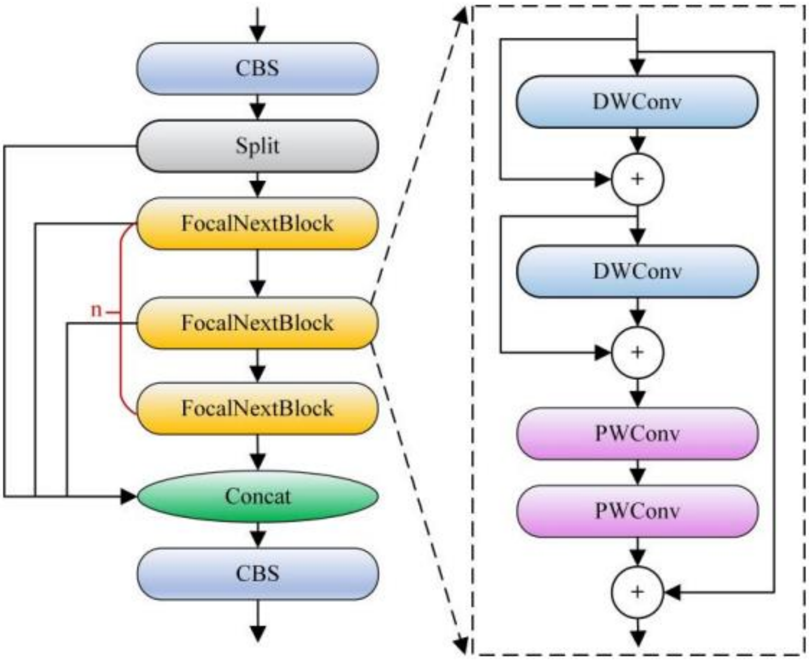


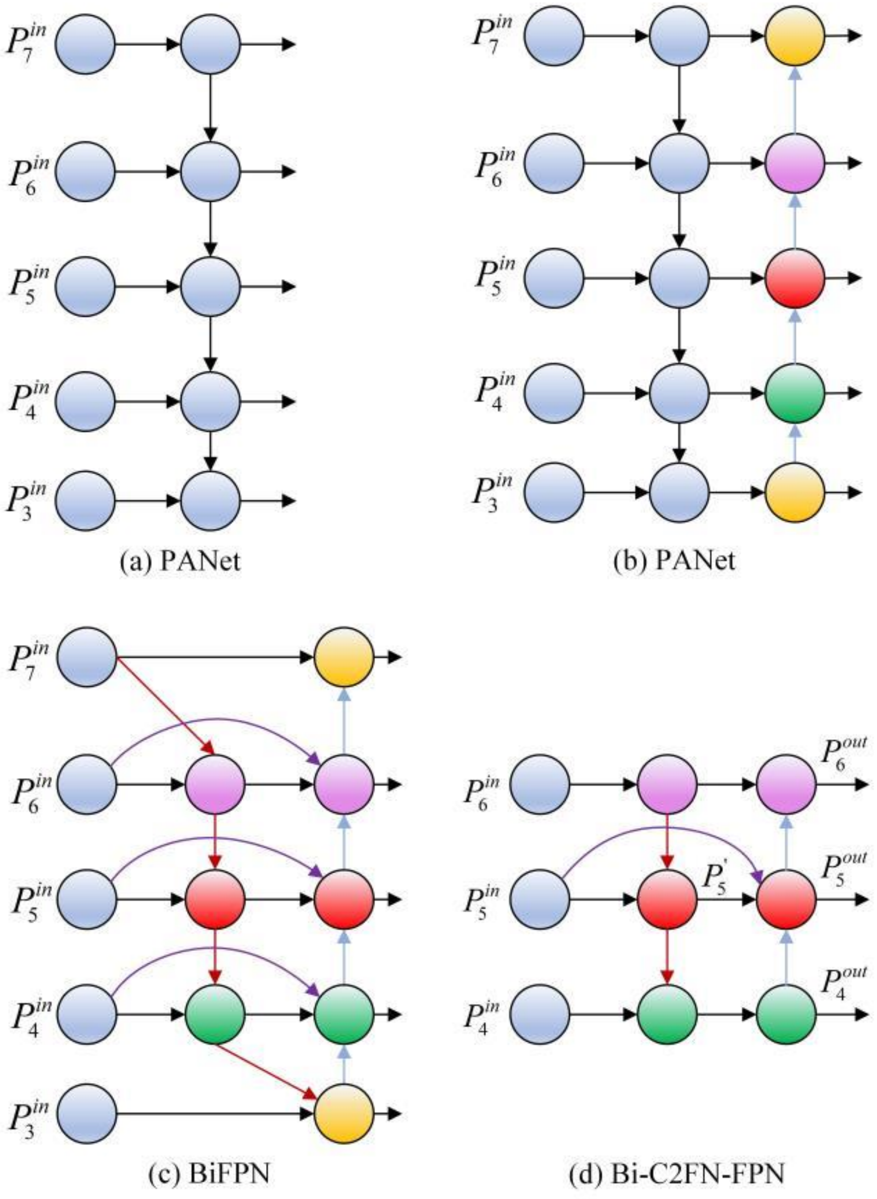
Fig. 4. C2FN structure.

The FocalNextBlock module includes a token mixer for spatial feature interaction and two pointwise convolution modules. In this paper, the lightweight 7×7 depth-separable convolution from ConvNeXt serves as the default token mixer for spatial feature interaction. The FocalNextBlock module incorporates a 7×7 extended depth-separable convolution along with two skip connections, as illustrated in Figure 4.

By integrating the extended depthwise separable convolution module, FocalNextBlock can capture long-range dependencies between spatial features, transforming the lowest resolution features within each channel of the image. The C2f module integrates the FocalNextBlock module, creating the C2FN module, which has enhanced feature extraction capabilities, efficiency, and multiscale characteristics. Therefore,incorporating C2FN as a new module structure in the neck network of YOLOv10s enriches feature extraction, meeting the demand for high- precision target recognition.

3.3. Bi-C2FN-FPN Structure

The YOLOv10s model uses the PANet feature fusion structure, but its reliance on FPN processing causes loss of original feature information, reducing detection accuracy. To address this, this paper proposes a Bi-C2FN-FPN structure, which applies the BiFPN concept to YOLOv10s and replaces the C2f module with the C2FN module in the neck network. BiFPN enhances multi-level feature fusion, while C2FN, integrating FocalNextBlock, improves contextual perception and feature extraction. These adjustments enhance pedestrian detection accuracy and robustness in complex environments. Diagrams illustrating different feature pyramid network structures are shown in Figure 5.



(a) PANet

(b) PANet

(c) BiFPN

(d) Bi-C2FN-FPN

Fig. 5. Structure of pyramid network with different features.

Compared to the BiFPN structure, Bi-C2FN-FPN reduces the number of input nodes to accommodate three effective feature layers in a simple and efficient backbone network. Additionally, it introduces an additional edge using skip connections to fuse features from the feature extraction network with features of the same size from the bottom-up path. This approach aims to preserve detailed information from shallow layers while retaining essential semantic information from deeper layers.

In contrast to PANet, Bi-C2FN-FPN removes nodes with only one input edge where no feature fusion has occurred, minimizing their impact on the network. This simplification helps streamline the network. The purple sections added aim to maximize feature fusion within acceptable computational limits to improve accuracy. In traditional feature fusion methods, such as Concatenation or Shortcut operations, different feature maps contribute differently due to their varied resolutions. In Bi-C2FN-FPN, a straightforward and efficient weighted feature fusion mechanism is employed to address these differences, ensuring fast training speeds and high efficiency.

Specific examples of Bi-C2FN-FPN feature fusion, as shown in equations (1) and (2), illustrate the fusion of two features from the P5 level.

$$p_5^{td} = Conv(\frac{w_1 \times p_5^{in} + w_2 \times \mathrm{Re}\, size(p_6^{in})}{w_1 + w_2 + \varepsilon}) \tag{1}$$

$$p_5^{out} = Conv(\frac{w_1^{'} \times p_5^{in} + w_2^{'} \times p_5^{td} + w_3 \times \mathrm{Re}\, size(p_4^{out})}{w_1^{'} + w_2^{'} + w_3^{'} + \varepsilon}) \tag{2}$$

Where p_5^{td} is the input parameter of the middle node in the 5th level, p_5^{in} is the input of the first node in the 5th level, w is the learned weight parameter, Resize is the feature map sampling operation, and Conv is the convolution operation. In general, Bi-C2FN-FPN features a weighted feature fusion mechanism with repeated bidirectional cross-scale connections, which enhances robustness in detecting multi-scale pedestrian features.

3.4. MPDIOU loss function

Currently, many loss functions in bounding box regression tasks yield the same values for different prediction results, which limits convergence speed and accuracy. Inspired by the geometric properties of rectangles, SILIANG et al.[18] proposed a new loss function called MPDIOU. This loss function simplifies the comparison of similarity between boxes. The specific calculation formula is:

$$MPDIOU = \frac{A \cap B}{A \cup B} - \frac{d_1^2}{w^2 + h^2} - \frac{d_2^2}{w^2 + h^2} \tag{3}$$

$$d_1^2 = (x_1^B - x_1^A)^2 + (y_1^B - y_1^A)^2 \tag{4}$$

$$d_2^2 = (x_2^B - x_2^A)^2 + (y_2^B - y_2^A)^2 \tag{5}$$

In the formula, A and B represent any input shapes. W and h denote the width and height of the input image. $(x_1^A, y_1^A)$and(x_2^B, y_2^B) represent the coordinates of the top-left and bottom-right points of shape A. Similarly, $(x_1^B, y_1^B)$and (x_2^B, y_2^B)represen the coordinates of the top-left and bottom-right points of shape B. d1 and d2 represent the Euclidean distances between the top-left corners and bottom-right corners of shapes A and B, respectively.

MPDIOU directly measures the corner point distances between the predicted bounding box and the actual annotation box, enabling the algorithm to more accurately select the most suitable bounding box to locate the target. Specifically, in pedestrian detection tasks, when pedestrians are dense and there is occlusion overlap, the MPDIOU loss function can effectively reduce distortion of the detection boxes, decrease the miss detection rate, and improve overall detection performance.

4. Experiments

4.1. WiderPersondataset

This paper utilizes the publicly available WiderPerson [13] dataset, which consists of 13,382 images and contains approximately 400,000 targets for various occlusion scenarios. The WiderPerson dataset is divided into 8,000 training images, 1,000 validation images, and 4,382 test images.

4.2. Experimental setup

The experimental environment for this paper is an Ubuntu 18.04 64- bit operating system, with 64GB of RAM. The GPU used is an NVIDIA GeForce RTX 4090 with 24GB of VRAM, and the CPU is an Intel(R) Xeon(R) Platinum i9- 13900k. The model is built using the PyTorch deep learning framework, with CUDA version 11.1 and cuDNN version 8.0.4.

4.3. Model Training

In this experiment, the input image size is set to 640 × 640 with an initial learning rate of 0.01. The learning rate is adjusted using stochastic gradient descent (SGD), incorporating a momentum of 0.937 and a weight decay of 0.0005. To enhance the detection background, mosaic data augmentation is applied during training. This involves loading four images simultaneously, flipping, scaling, and stitching them together. To improve model generalization and mitigate overfitting, label smoothing is configured at 0.01. Both YOLOv10s-Pedestrian and the baseline models are trained for 100 epochs using a batch size of 32 and 32 workers.

4.4. Evaluation metrics

This paper evaluates model performance using Precision (P), Recall (R), mean Average Precision (mAP), model parameters (Params), and inference time (Time). Recall measures the percentage of correctly classified relevant results, while mAP averages the AP values across all categories for a comprehensive performance assessment. Model parameters reflect the model's size and complexity. The calculation formulas are as follows:

$$\text{Precision}=\frac{TP}{TP+FP} \tag{6}$$

$$\text{Recall}=\frac{TP}{TP+FN} \tag{7}$$

$$mAP=\frac{1}{m}\sum_{1}^{m}AP \tag{8}$$

By comprehensively considering these metrics, a more thorough understanding of the performance of roadside target detection methods can be achieved. This provides targeted guidance for improving the methods.

4.5. Experimental results and analysis

4.5.1 Ablation Study

This paper conducted comparative experiments on different attention mechanisms based on YOLOv10s. The default SENet attention mechanism in the original MBConv module was replaced with CBAM, SimAM, GAMAttention, and CA attention mechanisms, respectively. The experimental results are shown in Figure 6.

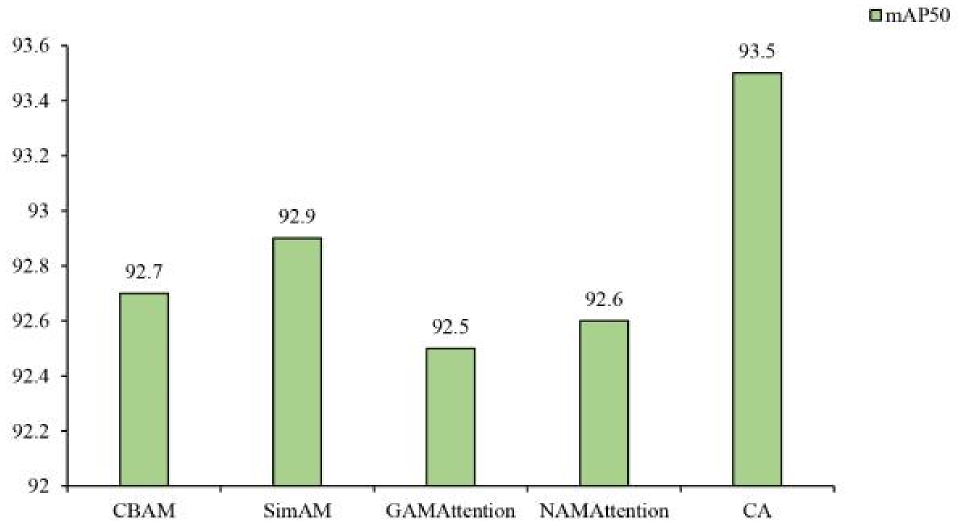


Fig. 6. mAP50 values for YOLOv10s using different attention mechanisms.

As shown in Figure 6, when YOLOv10s uses CBAM, SimAM, GAMAttention, and NAMAttention as attention mechanisms, they achieve mAP50 values of 92.7%, 92.9%, 92.5%, and 92.6%,respectively. Although these attention mechanisms achieve good detection accuracy, there are still slight differences among them. When using the CA attention mechanism, the model achieves the highest mAP50 value of 93.5%.

In this paper, new improvement strategies are successively introduced based on YOLOv10s, resulting in the construction of three different models to verify the effectiveness of MBConv, CA, Bi-C2FN-FPN, and MPDIOU in pedestrian detection tasks. As depicted in Table 1, where √ indicates the introduction of the new module structure, and × indicates the new module structure was not used.

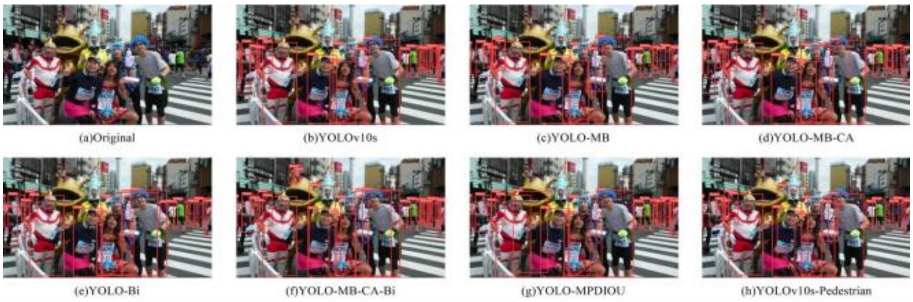
Table 1 Comparison of detection results under different improvements to the YOLOv10smodel.

Model	MBConv	CA	Bi-C2FN-FPN	MPDIOU	P(%)	R(%)	mAP50(%)	Params	Time(ms)
YOLOv10s	×	×	×	×	86.6	80.8	89.5	8,922,262	1.5
YOLO-MB	√	×	×	×	89.5	81.3	91.9	9,646,606	1.7
YOLO-MB-CA	√	√	×	×	91.6	82.7	93.5	9,708,526	1.7
YOLO-Bi	×	×	√	×	90.9	83.7	92.5	8,889,174	1.5
YOLO-MB-CA-Bi	√	√	√	×	93.4	85.4	95.0	9,642,670	1.9
YOLO-MPDIOU	×	×	×	√	88.4	82.5	91.5	8,922,262	1.5
YOLOv10s-Pedestrian	√	√	√	√	93.1	86.2	95.6	9,642,670	1.8

From Table 1, it is evident that in terms of detection accuracy, the proposed YOLO-MB model improves over the original YOLOv10s model with increases of 2.9% in Precision (P), 0.5% in Recall (R), and 2.4% in mAP50 (mean Average Precision at IoU 0.5). This indicates that the

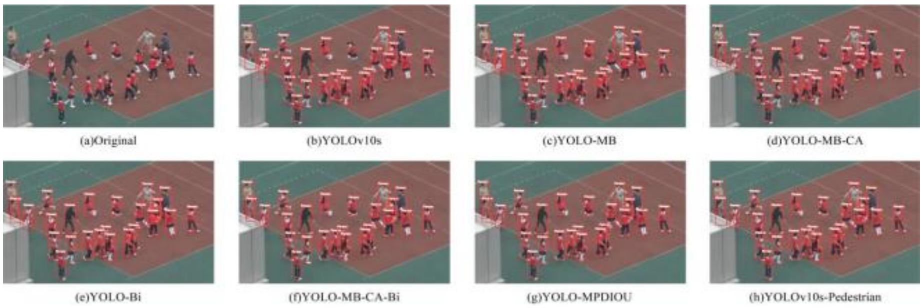
MBConv structure enhances the model's ability to extract spatial information about pedestrians, thereby improving detection performance by detecting more pedestrians. Further introduction of the CA (Channel Attention) mechanism in the YOLO-MB-CA model results in improvements of 5.0% in P, 1.9% in R, and 4.0% in mAP50. This demonstrates that the CA attention structure enables the model to focus more on the upper body regions of pedestrians, utilizing local features to distinguish pedestrians and thereby improving pedestrian detection performance. Comparing the performance of the YOLOv10s model and the YOLO-Bi model, the latter shows improvements of 4.3% in P, 2.9% in R, and 3.0% in mAP50. This suggests that the Bi-C2FN-FPN structure, compared to the PANet structure, more effectively integrates features of targets at different scales and expands the feature receptive fields. For the YOLO-MB-CA-Bi model compared to YOLOv10s, there are increases of 6.8% in P, 4.6% in R, and 5.5% in mAP50, illustrating that the combined integration of MBConv, CA, and Bi-C2FN-FPN significantly enhances the effectiveness of pedestrian detection. After introducing the MPDIOU loss function into YOLOv10s, the mAP50 of YOLO-MPDIOU reaches 91.5%, indicating that the MPDIOU loss function helps the model adjust the position and size of predicted boxes more accurately during training, thereby improving detection accuracy. Finally, YOLOv10s-Pedestrian achieves a good detection performance with P of 93.1%, R of 86.2%, and mAP50 of 95.6%.

Compared to the original YOLOv10s model, the parameter counts of YOLO-MB, YOLOv10s-2, and YOLOv10s-Pedestrian increased by 0.01MB, 0.27MB, and 0.21MB, respectively. YOLOv10s had the shortest detection time. However, after introducing the MBConv, CA, and Bi- C2FN-FPN structures, the detection time increased, but the model's accuracy improved significantly. Comparing the initial YOLOv10s model with all the improved models, the model parameters and detection speed showed a slight decrease. This is due to the newly proposed MB-CANet structure and Bi-C2FN-FPN structure, which deepen the network and increase model parameters, but also retain more high-level feature information, beneficial for target detection in complex environments. Overall, considering all factors, YOLOv10s-Pedestrian remains a highly effective detection method. The pedestrian detection performance in densely occluded and small distant instances is compared in Figures 7 and 8.


(a)Original (b)YOLOv10s (c)YOLO-MB (d)YOLO-MB-CA

(e)YOLO-Bi (f)YOLO-MB-CA-Bi (g)YOLO-MPDIOU (h)YOLOv10s-Pedestrian

Fig. 7. Dense occlusion detection image.


(a)Original (b)YOLOv10s (c)YOLO-MB (d)YOLO-MB-CA

(e)YOLO-Bi (f)YOLO-MB-CA-Bi (g)YOLO-MPDIOU (h)YOLOv10s-Pedestrian

Fig. 8. Small target detection image.

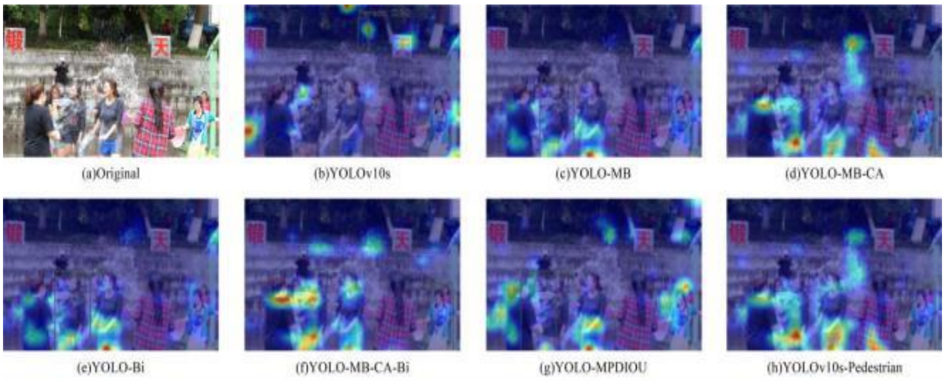


Fig. 9. Comparison of heat map detection performance.

Figures 7 and 8 show that YOLOv10s exhibits detection errors, including missed detections and misaligned bounding boxes. YOLOv10s-Pedestrian, however, successfully detects pedestrians in dense occlusion with higher accuracy. For example, YOLOv10s fails to detect a skipping girl (Figure 8b), while the YOLO-MB model (Figure 8c) improves detection accuracy, and the YOLO-MB-CA model (Figure 8d) further enhances feature learning, though some misalignment remains. YOLO-Bi (Figure 8e) shows a false detection, but YOLOv10s-Pedestrian (Figure 8h) achieves the highest accuracy, demonstrating the effectiveness of the Bi-C2FN-FPN structure in improving feature integration and detection accuracy.

Grad-CAM visualizations (Figure 9) confirm that YOLOv10s-Pedestrian enhances feature capture and focus accuracy. The MBConv structure (Figure 9c) improves pedestrian feature extraction, CA attention (Figure 9d) sharpens focus on pedestrian targets, and Bi-C2FN-FPN (Figure 9e) enhances multi-scale fusion. The combined MBConv, CA, and Bi-C2FN-FPN structures (Figure 9f) further improve detection, and the MPDIOU loss function (Figure 9g) reduces noise and sharpens focus. YOLOv10s-Pedestrian (Figure 9h) shows precise pedestrian localization and improved robustness in complex scenes.

4.5.2 Comparison with state-of-the-art methods

As shown in Table 2, to further validate the detection performance of YOLOv10s-Pedestrian, this paper compares YOLOv10s-Pedestrian with advanced models.

Table 2 Comparison with advanced models.

Model	P(%)	R(%)	mAP50(%)	Params(M)	Time(ms)
YOLOv10s	86.6	80.8	89.5	8,922,262	1.5
YOLOv3	82.3	69.6	79.9	8,666,692	1.3
YOLOv5-Lite	88.5	76.6	85.8	4,366,230	3.7
YOLOv5n	86.2	75.9	86.8	1,760,518	6.1
YOLOv7-tiny	87.1	79.5	88.7	6,007,596	3.1
YOLOv8n	85.5	80.4	88.5	3,005,843	0.6
YOLOv9-c	86.7	79.7	88.7	3,602,406	8.6
YOLOv10ss-Pedestrian	93.1	86.2	95.6	9,642,670	1.8

As shown in Table 2, the YOLOv10s-Pedestrian model has a slight increase in parameter count but achieves a significant improvement in mAP50, reaching 95.6%, which is 6.1% higher than the original YOLOv10s model's 89.5%. Compared to other single-stage object detection networks, such as YOLOv3, YOLOv5-Lite, YOLOv5n, YOLOv7-tiny, YOLOv8n, and YOLOv9-c, the YOLOv10s-Pedestrian model demonstrates higher mAP50 by 15.7%, 9.8%, 8.8%, 6.9%, 7.1%, and

6.9% respectively. Furthermore, in detection speed, the YOLOv10s- Pedestrian model performs excellently, with an inference time of just 1.8ms, which is faster than the YOLOv5-Lite, YOLOv5n, YOLOv7-tiny, and YOLOv9-c models by 1.9ms, 4.3ms, 1.3ms, and 6.8ms respectively. These characteristics make the improved YOLOv10s model more suitable for pedestrian detection tasks.

5. Conclusion

This paper presents a pedestrian detection model for densely occluded scenarios, named YOLOv10s-Pedestrian, based on the YOLOv10s model. By introducing MBConv and incorporating the CA attention mechanism into the YOLOv10s backbone, we propose the MB-CANet network structure, which effectively enhances the model's ability to extract pedestrian features in dense occlusion environments. Additionally, the BiFPN network is integrated into the neck network structure, and combined with the C2FN module, we propose the Bi-C2FN-FPN neck network structure. This enables the collection of richer semantic information, achieving more efficient feature fusion, and increasing focus on pedestrian detection areas, thus addressing the issues of missed detections and false positives for small targets. Finally, the MPDIOU loss function is introduced to reduce the positionalbias of the predicted boxes. Overall, the YOLOv10s-Pedestrian detection model achieves an mAP50 of 95.6%, with a model parameter count of approximately 9.64M, meeting the precision requirements for pedestrian detection in complex environments. In future research, efforts should focus on achieving a more lightweight model while ensuring accuracy, and applying the results to pedestrian detection in real complex street scenes.

References

[1] Azam, S., Munir, F., Kyrki, V., Kucner, T.P., Jeon, M., Pedrycz, 2024. Exploring Contextual Representation and Multi- modality for End-to-end Autonomous Driving. 135, 108767, Engineering Applications of Artificial Intelligence.

[2] Bai, S., Wang, Y., Luo, Z., Tian, 2024. DriveCP: Occupancy-Assisted Scenario Augmentation for Occluded Pedestrian Perception Based on Parallel Vision. IEEE,Journal of Image and Graphics.

[3] Bar-Joseph, M., Ezra, D., Licciardello, G., Catara, A., 2023. Science and Tradition. Springer, pp. 145-215.Diseases of Etrog Citron and Other Citrus Trees, The Citron Compendium: The Citron (Etrog) Citrus medica L.

[4] Gao, H., Huang, S., Li, M., Li, 2024. Multi-scale Structure Perception and Global Context-aware Method for Small-scale Pedestrian Detection. IEEE,Towson University Journal of International Affairs.

[5] Yuan, Q., Huang, G., Zhong, G., Yuan, X., Tan, Z., Lu, Z., Pun, C., Measurement, 2023. Triangular Chain Closed-Loop. Detection Network for Dense Pedestrian Detection.Transactions on Instrumentation and Measurement.

[6] Gong, W., Yang, S., Guang, H., Ma, B., Zheng, B., Shi, Y., Li, B., & Cao, Y. (2024). An intrusion detection scheme based on multi-order feature interaction to enhance cybersecurity in intelligent connected vehicles. Engineering Applications of Artificial Intelligence, 135, 108815.

[7] Guo, L., Ge, P.-S., Zhang, M.-H., Li, L.-H., & Zhao, Y.-B. (2012). Pedestrian detection in intelligent transportation systems using a combination of AdaBoost and support vector machines. Expert Systems with Applications, Elsevier.

[8] Hou, Q., Zhou, D., & Feng, J. "Designing Efficient Mobile Networks with Coordinate Attention". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2021.

[9] Hsu, W.-Y., & Yang, P.-Y. (2023). Multi-scale structure-enhanced super-resolution for pedestrian detection. IEEE Transactions on Intelligent Transportation Systems.

[10] Jain, D. K., Zhao, X., Garcia, S., & Neelakandan, S. (2024). A robust deep convolutional neural network-based multimodal pedestrian detection model using ensemble learning. Expert Systems with Applications, Elsevier.

[11] Jiang, H., Liao, S., Li, J., Prinet, V., & Xiang, S. "Semantic Modulation for Urban Scene-Based Pedestrian Detection". Neurocomputing, vol. 474, pp. 1-12, 2022.

[12] Li, J., Bi, Y., Wang, S., Li, Technology, S.f.V., 2023. CFRLA-Net: A Context-Aware Feature Representation Learning Anchor- Free Network for Pedestrian Detection. IEEE.Transactions on Circuits and Systems for Video Technology.

[13] Zheng, A., Wang, H., Wang, J., Huang, H., He, R., Hussain, 2023. Diverse features discovery transformer for pedestrian attribute recognition. 119, 105708.Engineering Applications of Artificial Intelligence.

[14] Kilicarslan, M., Zheng., 2022. DeepStep: Direct detection of walking pedestrian from motion by a vehicle camera. IEEE .Transactions on Intelligent Vehicles