Research on lightweight pedestrian detection based on improved YOLOv5

Yunfeng Jin[1], Zhizhan Lu[2], Ruili Wang[3], Chao Liang[4]
School of Civil Engineering and Transportation, Beihua University, Jilin, China
[4]Corresponding author
E-mail: [1]*324656658@qq.com*, [2]*luzhizhan0826@163.com*, [3]*krystalpolly@163.com*, [4]*398503241@qq.com*

Received 21 October 2023; accepted 6 November 2023; published online 23 November 2023
DOI https://doi.org/10.21595/mme.2023.23719

Abstract. Aiming at the problems of low detection accuracy and the large size of the pedestrian detection algorithm, to improve the edge intelligent recognition capability of the terminal, this paper proposes a lightweight pedestrian detection scheme based on the improved YOLOv5. In this paper, the algorithm first takes the original YOLOv5 as the basic framework and uses the Ghost Bottleneck module to replace the C3 module in the original YOLOv5 network to reduce the number of parameters, eliminate redundant features, and obtain a more lightweight model. Then the attention mechanism CBAM module is added to improve the feature extraction capability and detection accuracy of the algorithm. After experimental verification, the improved lightweight YOLOv5 algorithm significantly reduces the model size and computational cost while guaranteeing accuracy, which is suitable for deployment in edge devices.

Keywords: pedestrian detection, YOLOv5, lightweight, attention mechanism.

1. Introduction

Pedestrian detection is an important research in the field of target detection [1], [2]. It aims to find all possible pedestrians in the input image and output the location of pedestrians in the image. Pedestrian detection can be widely used in fields such as safety monitoring and autonomous driving.

Pedestrian detection techniques have evolved from traditional human-assisted feature detection to deep learning-based feature detection. Traditional pedestrian detection algorithms require the manual design of filters and features based on the designer's statistical or a priori knowledge. Cheng et al. proposed a pedestrian detection method using a sparse Gabor filter, which is designed based on texture features learned from some manually selected typical pedestrian images [3]. Ddlal et al. proposed a pedestrian detection method using edge features extracted from the HOG extracted edge features for pedestrian detection, which is obtained by computing and counting the HOG of some manually selected local image regions. Traditional pedestrian detection algorithms are time-consuming and labor-intensive due to manual intervention, with relatively low detection accuracy and efficiency [4].

With the development of Convolutional Neural Networks, deep learning-based pedestrian detection algorithms have pushed the effectiveness of pedestrian detection to an unprecedented level. Modern pedestrian detection algorithms based on deep learning can autonomously learn and extract target features with high detection accuracy and efficiency [5]. Zhang et al. solved the small-scale pedestrian detection problem with non-time symmetric multi-stage CNN, but its shortcoming is that it does not perform accurately on the discontinuous information in a small range [6]. Xu et al. solved the efficiency problem of pedestrian detection by model reconstruction and pruning of the YOLOv3 network but did not consider this special case of pedestrians in gauge frequency due to the high number or in a dense state, and there will be a high leakage rate in detection [7]. Liu et al. addressed the effectiveness of pedestrian detection in hazy weather with a weighted combination layer that combines multi-scale feature maps with squeezing and excitation blocks, but the additional computation does not apply to embedded devices and the detection

frame rate is low [8].

Aiming at the problems of low accuracy of traditional target detection algorithms and large volume of deep convolutional neural networks, to further reduce the model volume and improve the detection speed while ensuring high detection accuracy, this paper proposes a transmission line defect detection scheme based on the lightweight improved YOLOv5. Firstly, the Ghost Bottleneck module is used to replace the C3 module in the original YOLOv5 network to reduce the number of parameters and eliminate redundant features, and then the attention mechanism CBAM module is added to improve the algorithm's feature extraction capability and detection accuracy.

2. Traditional YOLOv5 algorithm

YOLOv5 has four network models of different sizes, s, m, l, and x. Among them, YOLOv5s is the smallest network model, and all the other models continuously increase the network depth and width based on it. To achieve a lightweight model and make it easier to port to edge devices, YOLOv5s is chosen as the base model in this paper. The following is a schematic diagram of the overall network structure of YOLOv5s [9].

Fig. 1 shows the overall block diagram of the YOLOv5s target detection algorithm. For a target detection algorithm, we can usually divide it into four generic modules, specifically: the input, the reference network, the Neck network, and the Head output, corresponding to the four red modules in Fig. 1 [10].

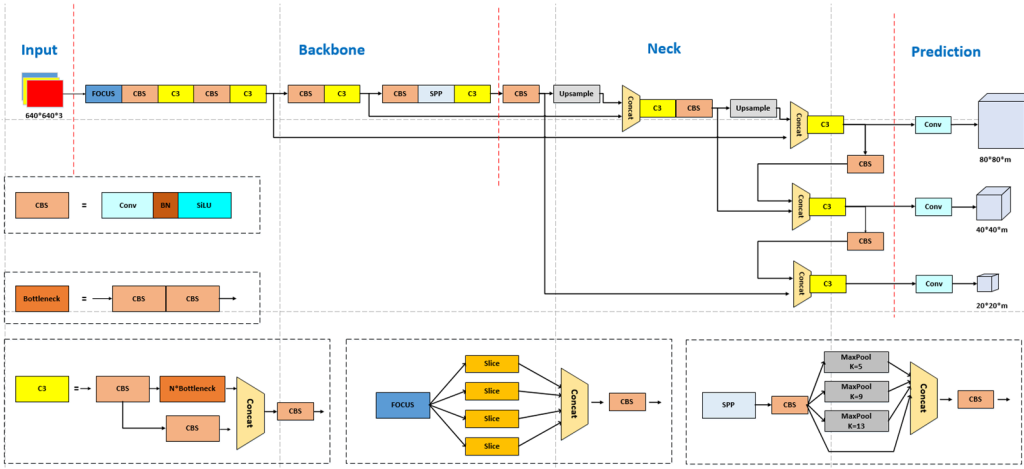


Fig. 1. YOLOv5s network structure

Input part uses Mosaic data enhancement, adaptive anchor box calculation and adaptive image scaling, etc. [11]. Mosaic data enhancement randomly combines 4 images at a time to form new image data. Adaptive anchor frame computation can adaptively compute the best anchor frame values in different training sets during training. Adaptive image scaling can adaptively add the least amount of black edges to the original input image, significantly reducing the amount of inference computation.

Backbone part is used to extract image features, mainly composed of Focus, C3, SPP, and other modules. The focus module crops the input image by slicing, reducing the amount of computation while increasing the sensory field to avoid the loss of the original information [12]. C3 mainly composed of SPP is a spatial pyramid pooling layer, which uses three types of pooling kernels of sizes 5, 9, and 13 to maximally pool the input image, effectively improving the network's receptive field.

Neck is the fusion part of the network, which uses the structure of feature pyramid networks

(FPN) plus pyramid attention networks (PAN). The FPN layer passes and fuses the feature information from the higher and lower layers by up-sampling, whereas the PAN layer splices the lower features with the higher ones so that the features with high resolution in the lower layers are passed on to the upper layers. The combination of the two enhances the fusion effect of features at different scales and effectively solves the multi-scale problem [13].

Head is the prediction part of the network, outputting three sets of vectors containing the prediction frame categories, confidence and coordinate positions. GIOU_Loss is used in Yolov5 as the loss function for target frame regression using non-maximal suppression (NMS) [14].

3. Improved lightweight YOLOv5s model

3.1. Adding the Ghost Bottleneck module

Ghost convolution is a lightweight convolution module proposed by Huawei's Noah's Ark Lab in 2020, whose core idea is to obtain more feature maps by performing another linear convolution on top of the feature maps obtained by a small number of non-linear convolutions, as a way to achieve the elimination of redundant features, and to obtain a more lightweight model, which reduces the cost of computation and computational resources [15]. The difference between normal convolution and Ghost convolution is shown in Fig. 2.

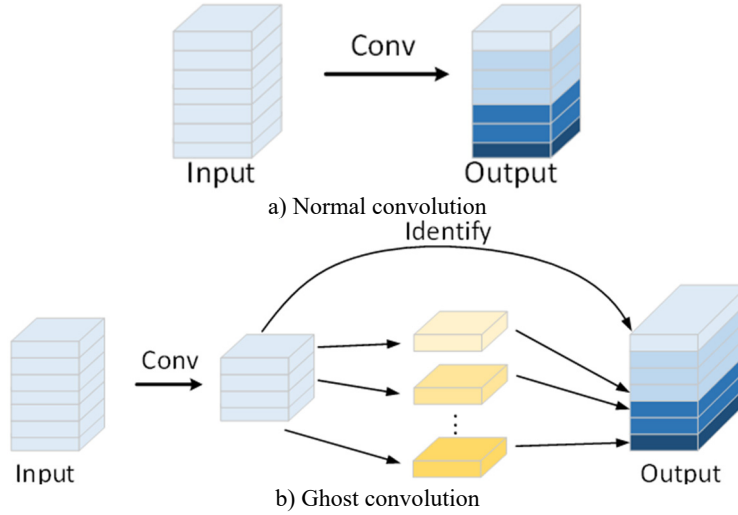


Fig. 2. Comparison of the principles of normal convolution and Ghost convolution

Assuming that the convolution kernel inputs a feature map of size $c \cdot h \cdot w$, respectively, the input channel, feature map height and width, and outputs a feature map of size $c' \cdot h' \cdot w'$ after one convolution, with regular convolution kernel of size k, and linear transformation convolution kernel of size d. After s transformations, it can be deduced that the computation of regular convolution is $c' \cdot h' \cdot w' \cdot c \cdot k \cdot k$, and the computation of Ghost convolution is $\frac{c'}{s} = h' \cdot w' \cdot c \cdot k \cdot k + (s-1) \cdot \frac{c'}{s} \cdot h' \cdot w' \cdot d \cdot d$. The comparison of the two computations is shown in Eq. (1):

$$r = \frac{c' \cdot h' \cdot w' \cdot c \cdot k \cdot k}{\frac{c'}{s} \cdot h' \cdot w' \cdot c \cdot k \cdot k + (s-1) \cdot \frac{c'}{s} \cdot h' \cdot w' \cdot d \cdot d} = \frac{c \cdot k \cdot k}{\frac{1}{s} \cdot c \cdot k \cdot k + \frac{s-1}{s} \cdot d \cdot d} \approx \frac{s \cdot c}{s + c - 1} \approx s. \quad (1)$$

From the results, it can be seen that the computational cost of the regular convolution is s times that of the Ghost convolution.

The Ghost Bottleneck module replaces traditional convolution layers with a small number of conventional convolutions and a lightweight redundant feature generator. This effectively reduces the computational complexity of convolutional neural networks, making them more lightweight and easier to deploy on edge devices. The specific structure of the Ghost Bottleneck module is shown in Fig. 3.

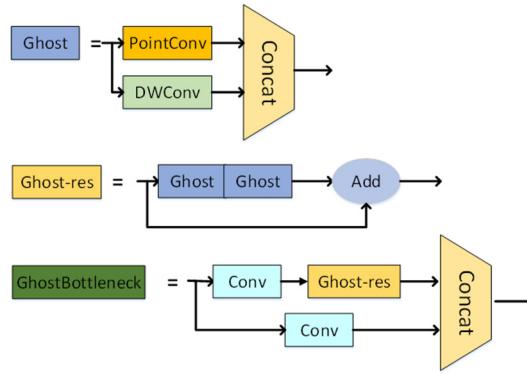


Fig. 3. Ghost Bottleneck module

The Ghost Bottleneck module has two structures. One is the Ghost Bottleneck with a stride of 1, primarily composed of two stacked Ghost modules. The other is the Ghost Bottleneck with a stride of 2, where two Ghost modules are connected by a depth-wise convolution layer with a stride of 2, as shown in Fig. 4.

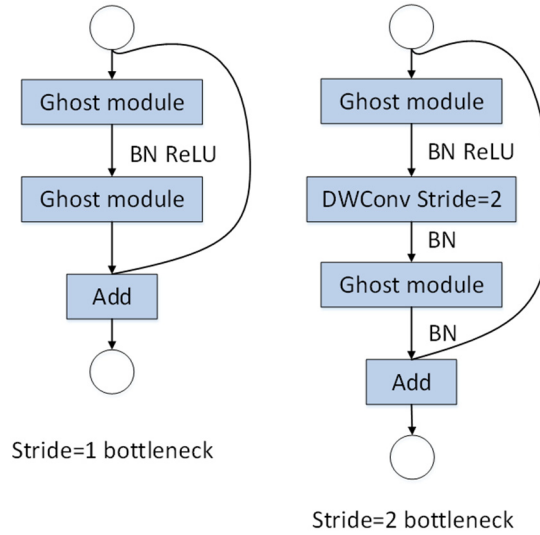


Fig. 4. Two structures of Ghost Bottleneck module

In the YOLOv5s network, the Bottleneck structure applies convolution operations to the input feature map using 32 1×1 convolution kernels followed by 64 3×3 convolution kernels. Therefore, in this paper, we replace the original C3 module in the YOLOv5s network with a Ghost Bottleneck module with a stride of 1 and replace the CBS module with a Ghost Bottleneck module with a stride of 2 to reduce the model's computational complexity. The YOLOv5s model incorporating Ghost Bottleneck modules is referred to as Ghost-YOLOv5s, and its structure is depicted in Fig. 5.

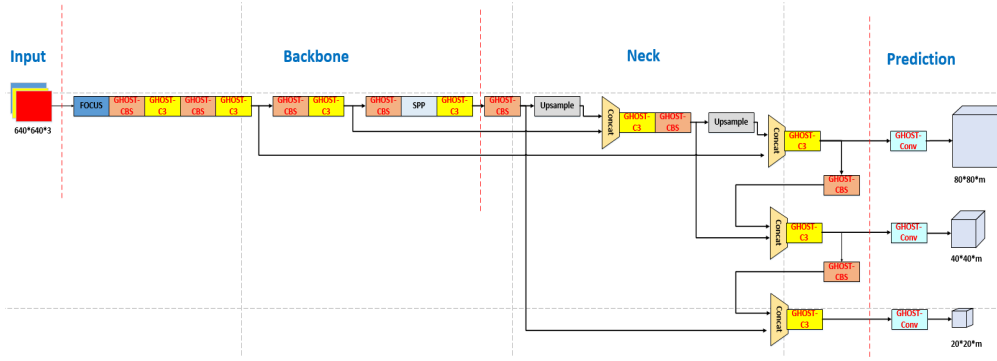


Fig. 5. Structure of Ghost-YOLOv5s

3.2. Add attention mechanism CBAM module

To allocate computational resources to more crucial tasks in situations where neural network computational capabilities are limited, attention mechanisms mimic human visual perception. They first scan the entire global image to identify target areas that require significant attention. Then, these attention resources are focused more intensively on these specific regions to gather detailed information related to the targets while filtering out irrelevant information from other regions. Therefore, attention mechanisms effectively address the problem of information overload by rapidly selecting high-value information from vast amounts of data using limited attention resources, thereby enhancing the efficiency and accuracy of task processing.

In this study, an attention mechanism module, CBAM (Convolutional Block Attention Module), is introduced into the YOLOv5s model [16]. CBAM consists of a Channel Attention Module (CAM) and a Spatial Attention Module (SAM), incorporating both spatial and channel attention, creating a sequential attention structure from spatial to channel dimensions. This design achieves a balance between lightweight architecture and significantly improved detection performance. The CBAM model structure is illustrated in Fig. 6.

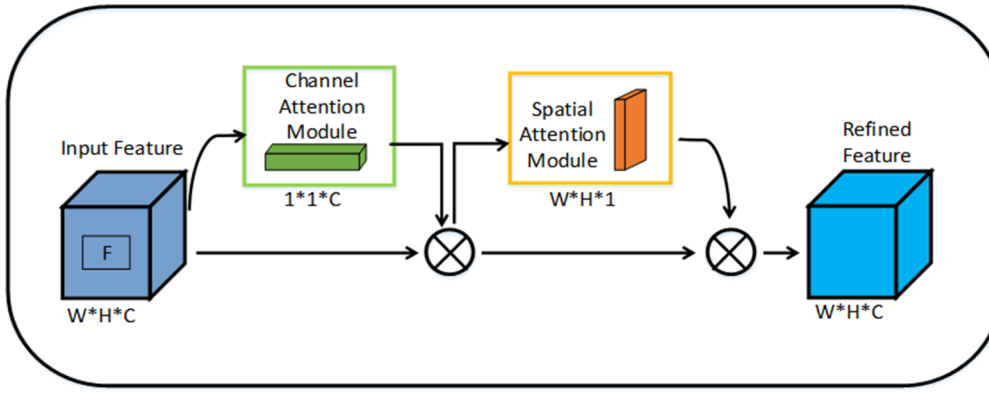


Fig. 6. CBAM module structure diagram

Given an input feature, the one-dimensional channel attention output is represented as in Eq. (2), and the two-dimensional spatial attention output is represented as in Eq. (3):

$$M_C(F) = \sigma\left(MLP\left(AvgPool(F)\right) + MLP\left(MaxPool(F)\right)\right), \tag{2}$$

$$M_S(F) = \sigma\left(f^{7\times7}([AvgPool(F); MaxPool(F)])\right). \tag{3}$$

ISSN PRINT 2351-5279, ISSN ONLINE 2424-4627

In the equations, σrepresents the sigmoid function, and MLP denotes a multi-layer perceptron comprising two fully connected layers with ReLU activation. $f^{7\times7}$ represents the convolution operation with a 7×7 kernel size.

The complete attention computation process is represented in Eq. (4) and Eq. (5):

$$F_C = M_C(F)\otimes F, \tag{4}$$
$$F_S = M_S(F)\otimes F. \tag{5}$$

where, ⊗ denotes element-wise multiplication.

This paper introduces the CBAM into the YOLOv5s model. A CBAM module is added to the C3 layer of the Backbone, transforming it into CBAM-C3, while keeping the remaining parameters of the model unchanged. The structure of CBAM-C3 is depicted in Fig. 7.

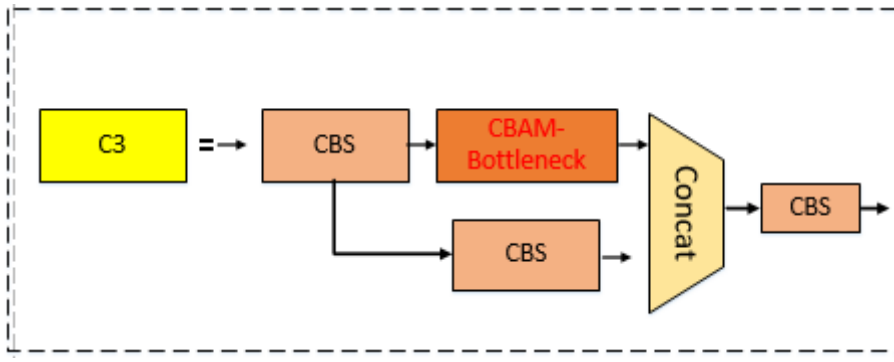


Fig. 7. Structure of CBAM-C3

4. Experiment and result analysis

4.1. Experimental environment and datasets

The dataset used in this study includes images of pedestrians captured under various natural conditions on roads. To train a robust detection model, the images obtained through data augmentation are divided into training, validation, and test sets in an 8:1:1 ratio.

The experimental environment is outlined in Table 1: The experiments are based on the TensorFlow open-source deep learning framework, with programming implemented in the Python language.

Table 1. Experimental platform

Category	Version
Operating System	Windows10
CPU	AMD Ryzen 7 5800H
GPU	NVIDIA GeForce RTX 3060
RAM	16Gb
Tensorflow-gpu	Tensorflow-gpu1.13.2
Python	Python3.8
TensorFlow	TensorFlow2.0
Ubuntul	Ubuntul16.04

The batch training data size is 32, the training momentum is set to 0.9, the initial learning rate is 0.001, the weight decay is 0.0005, and the training process continues for 200 batches. Stochastic Gradient Descent (SGD) is used as the optimization function to train the model 4.2 Evaluation indicators.

In the context of object detection, precision (P), recall (R), PR curve, average precision (AP),

and mean average precision (mAP) are commonly used metrics for evaluation. Precision and recall are defined as shown in Eq. (6) and Eq. (7):

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

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

where, TP represents the number of true positive predictions, which are cases where the model correctly predicts a positive sample as positive. FP represents the number of false positive predictions, which are cases where the model incorrectly predicts a negative sample as positive. FN represents the number of false negative predictions, which are cases where the model incorrectly predicts a positive sample as negative. Precision is the ratio of the number of true positive predictions to the total number of positive predictions, while Recall represents the proportion of correctly identified positive samples relative to the total number of positive samples.

PR curve illustrates the relationship between precision and recall. Average precision (AP) is the area under the PR curve. A higher AP indicates a more accurate model. Its definition is shown in Eq. (8):

$$AP = \int_0^1 P(R)dR, \tag{8}$$

mAP is the average of the individual class AP and is defined as shown in Eq. (9):

$$mAP = \frac{\sum_{n=0}^{N} AP_n}{N}. \tag{9}$$

This article use P, R and mAP as evaluation metrics to assess the detection performance of the model.

4.2. Analysis of results

4.2.1. Model comparison analysis

The comparison of detection results on the same dataset for Ghost-YOLOv5s and YOLOv5s after training is shown in Table 2. Ghost-YOLOv5s, in comparison to YOLOv5s, exhibits a reduction in overall model size by 32.8 % and a decrease in parameter count by 33.1 %, while the model's accuracy and recall rates remain largely unchanged. Through an analysis of the Ghost Bottleneck module, it was found that this module obtains a portion of feature maps through a small number of standard convolutions and then generates more feature maps through a simple linear operation. In contrast to the working principle of regular convolutions, this operation weakens the network's feature extraction capabilities, resulting in a decrease in detection accuracy and causing a drop of 0.013 in mAP.

Table 2. Comparison of YOLOv5s and Ghost-YOLOv5s

Model	mAP	Precision	Recall	Parameters (M)	Model size (M)
YOLOv5s	0.945	0.951	0.942	7.27	14.3
Ghost-YOLOv5s	0.932	0.949	0.941	4.86	9.6

After making initial improvements to YOLOv5s by incorporating the Ghost Bottleneck module, the network was further enhanced by introducing an attention mechanism known as the CBAM module, resulting in the model referred to as CBAM-YOLOv5s. Table 3 presents a

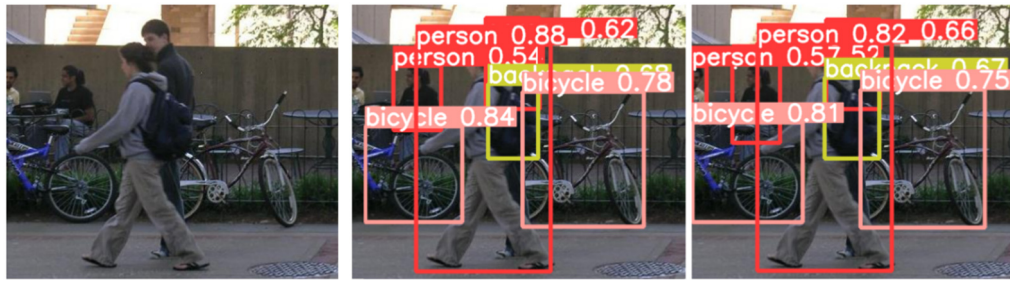
comparison of detection results for YOLOv5s, Ghost-YOLOv5s, and CBAM-YOLOv5s before and after the inclusion of the attention mechanism CBAM module. Compared to the original YOLOv5s, CBAM-YOLOv5s exhibits an increase of 1.1 percentage points in mAP, a reduction of 34.8 % in parameter count, and a 34.2 % reduction in model size. When compared to Ghost-YOLOv5s, CBAM-YOLOv5s achieves a 0.3 % increase in accuracy, a 2.4 percentage point improvement in mAP, and a further reduction in model size.

Table 3. Comparison of YOLOv5s, Ghost-YOLOv5s and CBAM-YOLOv5s

Model	mAP	Precision	Recall	Parameters (M)	Model size (M)
YOLOv5s	0.945	0.951	0.942	7.27	14.3
Ghost-YOLOv5s	0.932	0.949	0.941	4.86	9.6
CBAM-YOLOv5s	0.956	0.952	0.953	4.74	9.4

4.2.2. Comparison of results

Fig. 8 illustrates a comparison of the detection results between YOLOv5s and CBAM-YOLOv5s for the same road pedestrian images. It is clear that the improved algorithm with the addition of CBAM has an obvious advantage over the unimproved YOLOv5 algorithm in pedestrian detection, the YOLOv5 algorithm with the addition of CBAM is not disturbed by similar objects, and can effectively filter out the background interference in pedestrian detection, reducing the false detection rate and improving the detection accuracy.



a) Original image b) YOLOv5s algorithm detection c) CBAM-YOLOv5s algorithm detection

Fig. 8. Comparison of YOLOv5s and CBAM-YOLOv5s test results

5. Conclusions

This paper introduces a pedestrian detection algorithm based on lightweight improvements to YOLOv5. Firstly, the Ghost Bottleneck module is employed to reduce parameter count and eliminate redundant features, resulting in a more lightweight model. Then, an attention mechanism called the CBAM module is added to enhance the algorithm's feature extraction capabilities and detection accuracy. The experimental results show that the improved algorithm is significantly better than the traditional YOLOv5 algorithm in terms of target detection accuracy and leakage rate, and it can improve the detection speed and be easily deployed in edge devices while ensuring detection accuracy.

The lightweight model lacks robustness to facial recognition in complex environments and lighting conditions, making it difficult to accurately detect targets in some real-world situations. The next step will be to further optimize the network structure and improve the loss function to improve the robustness, and to study the problems and solutions encountered after the model is deployed at the edge so that the model can be refined and improved in real-world pedestrian detection applications.

Acknowledgements

The authors have not disclosed any funding.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Author contributions

Yunfeng Jin: conceptualization, methodology, software, investigation, formal analysis, writing – original draft. Zhizhan Lu: data curation. RuiLi Wang: visualization. Chao Liang: conceptualization, funding acquisition, resources, supervision, writing - review and editing.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] T. Liu, J. Cheng, M. Yang, X. Du, X. Luo, and L. Zhang, "Pedestrian detection method based on self-learning," in *2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pp. 2161–2165, Dec. 2019, https://doi.org/10.1109/iaeac47372.2019.8997629

[2] Luo Y. et al., "An overview of deep learning based pedestrian detection algorithm," (in Chinese), *Journal of Image and Graphics*, Vol. 27, No. 14, pp. 2094–2111, 2022, https://doi.org/10.11834/jig.200831

[3] Hong Cheng, Nanning Zheng, and Junjie Qin, "Pedestrian detection using sparse Gabor filter and support vector machine," in *IEEE Proceedings. Intelligent Vehicles Symposium, 2005.*, pp. 583–587, 2005, https://doi.org/10.1109/ivs.2005.1505166

[4] N. Dalal, "Histograms of oriented gradients for human detection.," in *IEEE Computer Society*, 2005, https://doi.org/10.1109/cvpr.2005.177

[5] M. Saeidi and A. Ahmadi, "Deep learning based on CNN for pedestrian detection: an overview and analysis," in *2018 9th International Symposium on Telecommunications (IST)*, pp. 108–112, Dec. 2018, https://doi.org/10.1109/istel.2018.8661043

[6] S. Zhang, X. Yang, Y. Liu, and C. Xu, "Asymmetric multi-stage CNNs for small-scale pedestrian detection," *Neurocomputing*, Vol. 409, pp. 12–26, Oct. 2020, https://doi.org/10.1016/j.neucom.2020.05.019

[7] H. Xu, M. Guo, N. Nedjah, J. Zhang, and P. Li, "Vehicle and pedestrian detection algorithm based on lightweight YOLOv3-promote and semi-precision acceleration," *IEEE Transactions on Intelligent Transportation Systems*, Vol. 23, No. 10, pp. 19760–19771, Oct. 2022, https://doi.org/10.1109/tits.2021.3137253

[8] G. Li, J. Yang, and Z. Kang, "Pedestrian detection algorithm based on improved YOLOv3_tiny," (in Chinese), *Proceedings of 2021 Chinese Intelligent Automation Conference*, Vol. 42, No. 14, pp. 98–106, 2022, https://doi.org/10.1007/978-981-16-6372-7_12

[9] L. Li, M. Liu, L. Sun, Y. Li, and N. Li, "ET-YOLOv5s: toward deep identification of students' in-class behaviors," *IEEE Access*, Vol. 10, pp. 44200–44211, 2022, https://doi.org/10.1109/access.2022.3169586

[10] S. Li, Y. Li, Y. Li, M. Li, and X. Xu, "YOLO-FIRI: improved YOLOv5 for infrared image object detection," *IEEE Access*, Vol. 9, pp. 141861–141875, 2021, https://doi.org/10.1109/access.2021.3120870

[11] J. Chu, Z. Guo, and L. Leng, "Object detection based on multi-layer convolution feature fusion and online hard example mining," *IEEE Access*, Vol. 6, pp. 19959–19967, 2018, https://doi.org/10.1109/access.2018.2815149

[12] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, "DetNet: design backbone for object detection," in *Computer Vision – ECCV 2018*, pp. 339–354, 2018, https://doi.org/10.1007/978-3-030-01240-3_21

[13] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944, Jul. 2017, https://doi.org/10.1109/cvpr.2017.106

[14] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: a metric and a loss for bounding box regression," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 658–666, Jun. 2019, https://doi.org/10.1109/cvpr.2019.00075

[15] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: more features from cheap operations," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1577–1586, Jun. 2020, https://doi.org/10.1109/cvpr42600.2020.00165

[16] C. Yang, C. Zhang, X. Yang, and Y. Li, "Performance study of CBAM attention mechanism in convolutional neural networks at different depths," in *2023 IEEE 18th Conference on Industrial Electronics and Applications (ICIEA)*, pp. 1373–1377, Aug. 2023, https://doi.org/10.1109/iciea58696.2023.10241832

Yunfeng Jin received bachelor's degree in Railway Traffic Signal and Control from the School of Electrical Engineering at Jiangsu Normal University, Xuzhou, China, in 2020. Now he pursuing a master's degree in Transportation and Traffic Engineering at Beihua University's School of Civil Engineering and Transportation, under the guidance of Professor Chao Liang. His current research interests include Deep Learning, Image processing and Computer Vision.

Zhizhan Lu is currently a Master student at Department of Vehicle and Civil Engineering of Beihua University, Jilin City, China. His research interests include digital image processing and intelligent transportation.

Ruili Wang is currently a Master student at Department of Vehicle and Civil Engineering of Beihua University, Jilin City, China. His research interests include digital image processing and environment awareness for intelligently connected vehicles.

Chao Liang obtained master's degree and currently works as Associate Professor at Department of Vehicle and Civil Engineering of Beihua University, Jilin City, China. His domain of research includes electrical control of vehicles, intelligent transportation, etc.