SSB-YOLO: A vehicle object detection algorithm based on improved YOLOv8

Mingda Wang

20211204048@chnu.edu.cn

Huaibei Normal University

Luyuan Ren

Huaibei Normal University

Research Article

Keywords: Vehicle detection, YOLO, Shuffle attention, BWIoU

Posted Date: December 14th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3743453/v1

Additional Declarations: No competing interests reported.

SSB-YOLO: A vehicle object detection algorithm based on improved YOLOv8

Mingda Wang[1*], Luyuan Ren[2]

[1*]Network Engineering, Huaibei Normal University, Lieshan Distict, Huaibei, 235000, Anhui, China.

[2]Visual Communication Design, Huaibei Normal University, Lieshan Distict, Huaibei, 235000, Anhui, China.

*Corresponding author(s). E-mail(s): 20211204048@chnu.edu.cn;
Contributing authors: 20211204048@chnu.edu.cn; 20210606027@chnu.edu.cn;

Abstract

In the field of computer vision, vehicle object detection has been a topic of significant and complex interest. With the rise of intelligent transportation systems and autonomous driving technology, the importance of vehicle object detection continues to be highlighted. Given the current issues of low precision, high miss rate, and poor robustness in existing algorithms, this study introduces an improved vehicle detection algorithm, SSB-YOLO, based on the YOLOv8 model. The SSB-YOLO algorithm integrates the Shuffle Attention mechanism to filter out unimportant factors and enhance model performance; it also incorporates the spatial and channel reconstruction convolution mechanism to reduce spatial and channel redundancy between features in convolutional neural networks. Furthermore, a new and better algorithm based on Wise-IoU optimization is proposed, which yields superior bounding box regression performance throughout the training period. The model demonstrated improved detection accuracy and reduced computational cost. The experimental results indicate that, compared to the YOLOv8n model, SSB-YOLO achieves a 1.6% increase in mAP@50. This approach outperforms other object detection algorithms, enhancing the overall system's robustness and accuracy and thereby providing higher precision in the field of vehicle detection.

Keywords: Vehicle detection, YOLO, Shuffle attention, BWIoU

1 Introduction

In the field of computer vision, vehicle target detection has been an important and complex issue. With the rise of intelligent transportation systems and autonomous driving technology, the significance of vehicle target detection continues to grow. An accurate and efficient vehicle target detection system is paramount in achieving intelligent transportation and vehicle autonomous driving. [1]Additionally, vehicle target detection places strict constraints on computational resources, making it a major challenge to maintain high accuracy while minimizing computational resources in complex road environments. [2]

Currently, there are two main types of detection models in the target detection field: dual-stage algorithms, such as Faster-Rcnn[3], which have better accuracy but slower speed; and single-stage target detection algorithms, such as the YOLO series**Error! Reference source not found.**, SSD[4]and RetinaNet[6], which have faster speeds but slightly lower accuracy. Despite the high accuracy of dual-stage algorithms, the two-stage processing of images and slow processing speed make them unsuitable for vehicle detection. On the other hand, single-stage algorithms extract features only once for detection, leading to faster processing but at the cost of reduced accuracy.

To address the shortcomings of target detection in vehicle recognition, ZHANG et al.[7] proposed the YOLOv7-RAR algorithm, which involves restructuring the YOLOv7 backbone network with the introduction of the Res3Unit structure. They also incorporated the ACmix attention mechanism to reduce interference from other targets and added the RFLA module at the connection of the detection head and feature fusion area to enhance the network model's receptive field. WANG et al.[8] proposed the YOLOv5-NAM algorithm, which added the NAM attention module and proposed methods for tracking small target vehicles, embedding the feature extraction process into the joint training of the prediction head. Moreover, Farid et al. [9] enhanced the accuracy of vehicle detection by modifying YOLO weights and utilizing transfer learning. Nitika et al.[10] employed a region-based convolutional neural network to detect moving vehicles both during the day and at night, optimizing the detection performance under different weather conditions. ZHAO et al.[11] utilized an attention mechanism to suppress interference features in images through both channel and spatial dimensions while also modifying the network structure to enhance effective features. YUAN et al. [12] introduced a high-performance bounding box proposal matching module and a keypoint selection strategy to compress collective perception messages and address the data fusion problem for multiple vehicles. These algorithms have to some extent improved the performance of target detection algorithms, but challenges related to robustness, real-time performance, and precision in complex road scenarios remain. Vehicle target detection systems need to be able to handle various complexities, such as weather, lighting, and road conditions, while achieving real-time and efficient target detection.

This study focuses more on the overall performance enhancement of the system rather than solely on individual improvements. Moreover, this research introduces unique methodological considerations aimed at comprehensively improving the overall system's robustness and accuracy. This paper aims to improve the YOLOv8 model by (1) introducing attention mechanisms, (2) improving convolutional structures, and (3) optimizing loss functions to enhance the system's adaptability and accuracy in complex scenarios. Through empirical evidence, this study demonstrates the feasibility and effectiveness of the new algorithm, providing a more advanced and practical solution for vehicle target detection in the fields of intelligent transportation and autonomous driving. This not only holds significant academic importance but also plays a proactive role in advancing industry applications.

2 Methods

2.1 YOLOv8 network structure

The latest algorithm in the YOLO series, YOLOv8, is an optimized and upgraded version based on the previous generation and has improved performance and accuracy. In contrast to dual-stage models such as Faster-Rcnn, the YOLO algorithm ensures faster detection while maintaining a certain level of accuracy.

YOLOv8 comprises three main components: the backbone, neck, and head, as shown in Figure 1. The backbone network consists of CBS, C2f, and SPPF. Compared to YOLOv5[13], YOLOv8 eliminates the top-down upsampling phase of the PAN[14]-FPN [15]in the neck layer, replacing the C3 module with the more gradient-enriched C2f. The head network adopts the mainstream decoupled head structure, separating detection from classification and replacing the anchor-based concept with Anchor-free, reducing model computations and improving convergence speed and effectiveness.

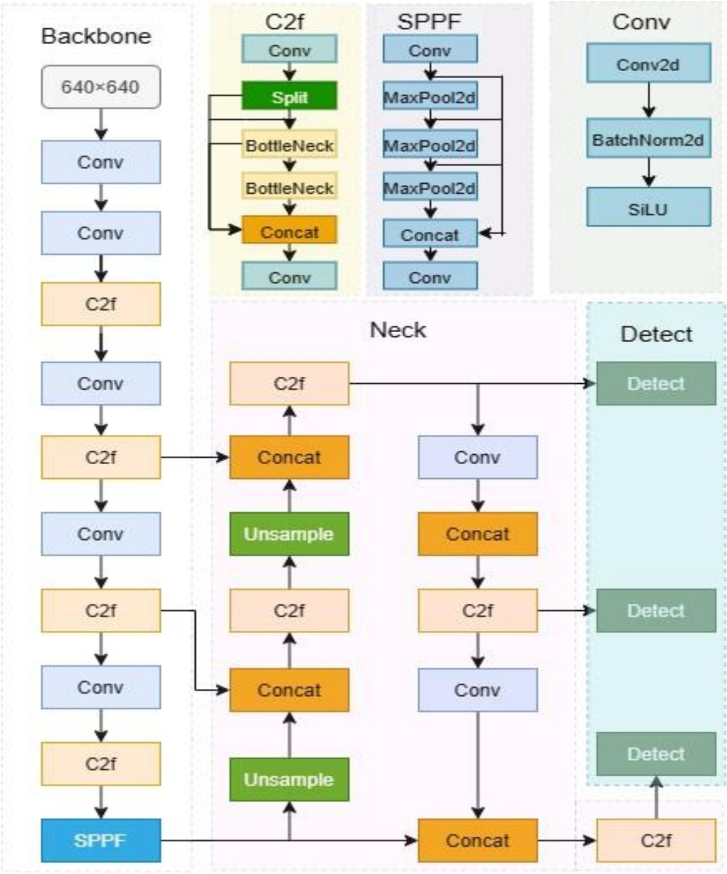

Fig. 1. YOLOv8n

2.2 Improved YOLOv8 model

2.2.1 BWIoU

The regression loss function for bounding boxes is crucial in the field of object detection. YOLOv8 adopts the CIoU [16]as the loss function, which continuously improves the accuracy of the predicted bounding boxes. However, the CIoU metric does not consider the balance between easy and hard samples in the dataset, leading to slow convergence and low efficiency of the network. In contrast, the WIoU [17]utilizes a nonmonotonic aggregation mechanism to construct dynamic gradient gain factors, providing a clear gain allocation strategy, as shown in Equations (1)–(3).

$$\mathcal{L}_{WIoU} = \mathcal{R}_{WIoU} \times \mathcal{L}_{IoU} \tag{1}$$

$$\mathcal{R}_{WIoU} = \exp\left(\frac{\left(x - x_{gt}\right)^2 + \left(y - y_{gt}\right)^2}{\left({W_g}^2 + {H_g}^2\right)^*}\right) \tag{2}$$

$$\mathcal{L}_{IoU} = 1 - IoU \tag{3}$$

Specifically, $\mathcal{R}_{WIoU}$ represents the penalty term of the WIoU, $\mathcal{L}_{IoU}$ is the IoU loss function, x and y refer to the coordinates of the center point of the predicted box, and x_{gt} and y_{gt}represent the coordinates of the predicted box for the ground truth box. Additionally, W_g and H_g denote the width and height, respectively, of the minimum rectangle formed by the predicted box and the ground truth box, while the IoU represents the intersection over union of the predicted box and the ground truth box.

Conversely, blindly enhancing the boundary box regression on low-quality samples in the training data diminishes the model's generalization performance. A good loss function should attenuate the penalty on geometrical factors when the overlap between the predicted box and the target box is high, enabling the model to achieve better generalization. Building on the improvements in the WIoU, this model introduces the BWIoU, as demonstrated in Equation (4).

$$\mathcal{L}_{BWIoU} = \mathcal{R}_{WIoU} \times \mathcal{L}_{IoU} + \frac{3}{5}\left(\frac{\left(x - x_{gt}\right)^2 + \left(y - y_{gt}\right)^2}{{W_g}^2 + {H_g}^2}\right) \tag{4}$$

The BWIoU algorithm dynamically adjusts the bounding box regression: In the early stage of training, when the model's IoU is low, the IoU between the predicted candidate box and the actual object annotation box should be improved. In the later stage of training, when the IoU is higher, the model automatically prioritizes the regression of the center point and aspect ratio of the candidate box, thus optimizing the model's performance.

2.2.2 SCConv

The complex backgrounds of vehicle images and numerous interfering factors weaken the adaptability and generalization performance of traditional convolution methods. This paper introduces spatial and channel reconstructive convolution (SCConv) [18]to reduce spatial and channel redundancy among features in convolutional neural networks, compressing the model and improving its performance.

As shown in Figure 2, the SCConv module consists mainly of a spatial reconstructive unit (SRU) and a channel reconstructive unit (CRU). The SRU reduces spatial redundancy in input features through analysis and reconstruction methods, while the CRU employs a segmentation and fusion strategy to reduce channel redundancy, effectively enhancing the model's robustness.

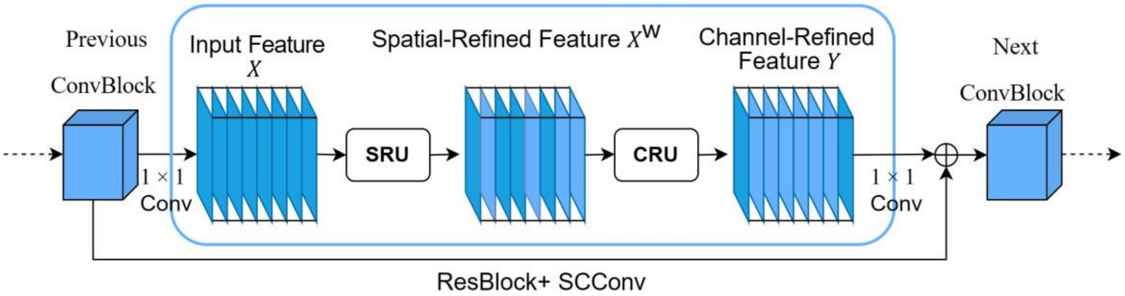

Fig. 2. Spatial and Channel reconstruction Convolution

The SCConv is integrated within the bottleneck of the c2f model and replaces the last three layers of the neck module, reducing spatial and channel redundancy. By optimizing the feature extraction process, resource consumption is minimized, and network performance is enhanced.

2.2.3 Shuffle Attention

In vehicle detection, the use of multilayer convolutional processing on images leads to inefficient training resources being allocated to nonvehicle images, resulting in poor training efficiency. To address this issue, the model incorporates the Shuffle Attention (SA) mechanism[19], which is placed before the SPPF layer to filter out irrelevant factors and reduce the complexity of the model. The SA module structure is illustrated in Figure 3.

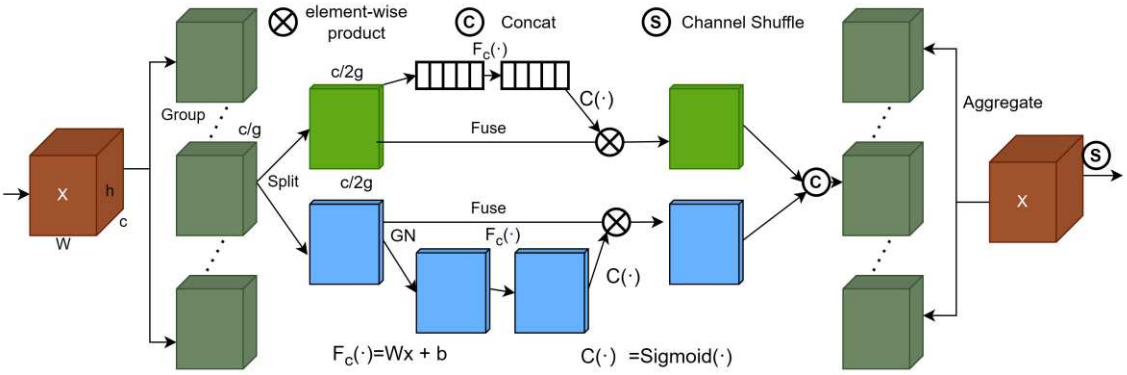

Fig. 3. Shuffle Attention

It inherits the design concept of the SGE attention mechanism[20], dividing the channel dimension into multiple subfeatures and utilizing the shuffling units to aggregate all subfeatures by taking into account their spatial and channel dependencies. Furthermore, it introduces the parallel use of two types of attention mechanisms—spatial and channel—via random channel partitioning and finally performs random mixing on all SA units to obtain the final output feature map, thereby effectively combining the two types of attention mechanisms.

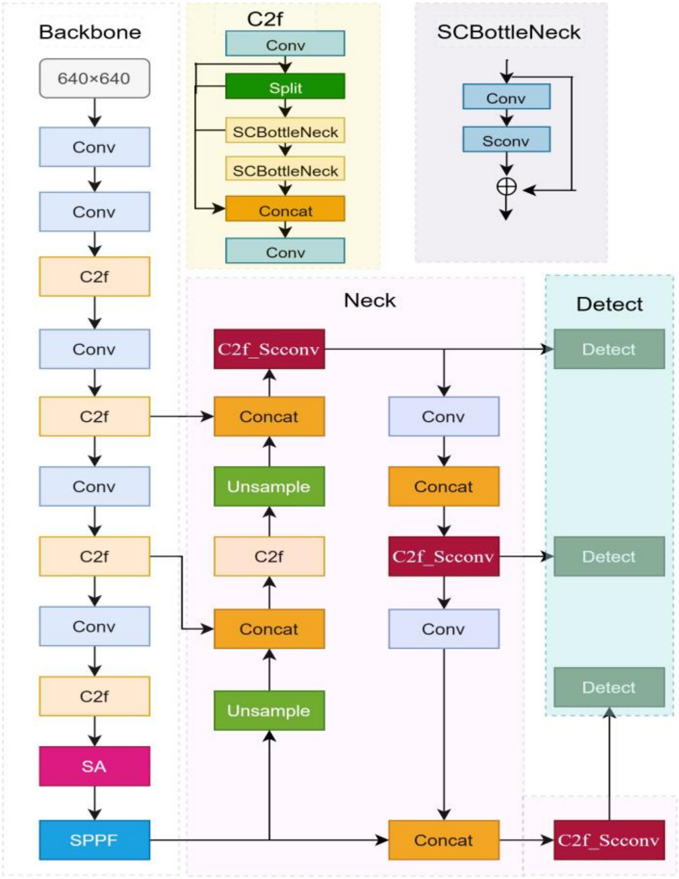

Fig. 4. SSB-YOLO

2.3 Datasets

The experiment tested the performance of the model using the Pascal VOC dataset. A total of 4610 images were extracted from the training and validation sets of Pascal VOC 2007 and 2012 for five vehicle categories: car, bus, train, bicycle, and motorbike. The images were randomly divided into training, validation, and testing sets at a ratio of 7:1:2, with 3227 images in the training set, 461 in the validation set, and 922 in the testing set.

2.4 Device

The experiment was conducted on a Windows 10 operating system using Python 3.11.5, CUDA 11.7, and PyTorch 2.0.0. The training was performed on an Nvidia GeForce RTX 3090 using YOLOv8n as the base model, with an image input size of 640x640, a batch size of 16, 2 threads, an initial learning rate of 0.01, and training for 300 epochs using the SGD optimization function.

3 Results

3.1 Evaluation metrics

This experiment uses Precision, Recall, and the mean Average Precision (mAP) as evaluation metrics. Referring to equations (5)~(7),

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{5}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{6}$$

$$\mathrm{mAP} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{AP_i} \tag{7}$$

The evaluation of the model's performance involves several key elements, as represented by the following equations. True positive predictions (TP) indicate instances where both the predicted and true values are positive. Conversely, false positive predictions (FP) occur when the predicted value is positive but the true value is negative. On the other hand, false negative predictions (FN) arise when the predicted value is negative, yet the true value is positive. Moreover, the area under the Precision-Recall curve is denoted as $\mathrm{AP_i}$ for each detection target class, and the average mAP is calculated based on multiple classes. Additionally, the model's advantages are emphasized through the consideration of parameters such as size, computational workload, and model file size as evaluation criteria.

3.2 Ablation experiment

To evaluate the performance of each module, we conducted experiments using the YOLOv8n base model and vehicle images extracted from the Pascal VOC dataset. We designed a series of comparative ablation experiments and used the precision (P), recall (R), mAP@0.5, and GFLOPS as quantitative evaluation metrics. The experimental results are presented in Table 1.

Table 1. Ablation experiment

Method	*BWIou*	*SA*	*SCConv*	*P(%)*	*R(%)*	*mAP@50(%)*	*GFLOPs(G)*
1				*90.0*	*76.3*	*85.9*	*16.4*
2	√			*89.9*	*77.0*	*86.6*	*17.8*
3		√		*90.5*	*76.0*	*85.9*	*16.4*

4			√	*87.2*	*76.4*	*85.5*	*17.2*
5	√		√	*91.4*	*76.5*	*86.9*	*16.0*
6	√	√	√	*89.9*	*77.1*	*87.5*	*15.8*

The optimal results of the ablation experiments were selected for analysis. In Table 1, when comparing Method4 and Method5, the model's introduction of BWIou led to a 1.7% increase in mAP@50 and an approximately 1.9% increase in precision, while the GFLOPs decreased by 1.2G. When comparing Method2 and Method5, the model's introduction of SCConv led to an approximately 2.8% increase in precision, an approximately 0.7% increase in mAP@50, and an approximately 0.7% increase in recall. When comparing Method1 and Method3, the model's introduction of the SA model resulted in virtually unchanged recall and an approximately 0.3% increase in precision. The data show that the SSB-YOLO model, based on the improved YOLOv8, exhibits enhanced performance and accuracy.

3.3 Model Comparison

To verify the detection performance of the SSB-YOLO algorithm model, quantitative indicators such as parameters, GFLOPS, mAP@0.5, and model_size were used. A quantitative analysis was conducted to compare the results of Faster R-CNN, SSD, YOLOv3, YOLOv4, YOLOv5, and YOLOv8 with those of mainstream algorithms for vehicle detection on the PASCAL VOC dataset. The comparative numerical results are shown in Table 2.

Table 2. Comparison of SSB-YOLO with other models

Method	Parameters(10^{-6})	mAP@50(%)	GFLOPs(G)
Faster-RCNN	136.8	76.4	369.8
SSD	24.1	84.0	61.2
YOLOvXs	8.9	84.1	26.8
YOLOv5n	2.5	86.4	14.4
YOLOv6n	4.2	85.1	23.8
YOLOv8n	3.0	85.9	16.4
SSB-YOLO	2.9	87.5	15.8

After analyzing Table 2 and comparing SSB-YOLO with the Faster R-CNN and SSD algorithms, it is evident that Faster R-CNN uses ResNet50 as the backbone network, while SSD uses VGG. SSB-YOLO shows a significant improvement in GFLOPs and average precision across all categories, with a noticeable reduction in computational complexity. Compared to YOLOvXs, YOLOv5n, YOLOv6n, and YOLOv8n, SSB-YOLO exhibited improvements in average precision of 3.4%, 1.1%, 2.4%, and 1.6%, respectively. Comparative experiments show that SSB-YOLO outperforms mainstream detection models and the original models in terms of performance and accuracy. It is more suitable for the deployment and application of vehicle target detection models, demonstrating superior overall performance compared to other algorithms.

3.4 Elucidation of the Results

Figures 5 and 6 compare the actual detection differences between YOLOv8n and SSB-YOLO. In image number 003576, there is a bus, two cars, and a motorbike. The scene with four vehicles is complex and has considerable overlap. YOLOv8n failed to detect one car and misidentified the motorbike as a bicycle, resulting in detection errors. In contrast, SSB-YOLO accurately detects all the vehicles in this scene. In image 2008_004326, the six cars overlap. YOLOv8n detected only four cars, whereas SSB-YOLO successfully detected all six cars.

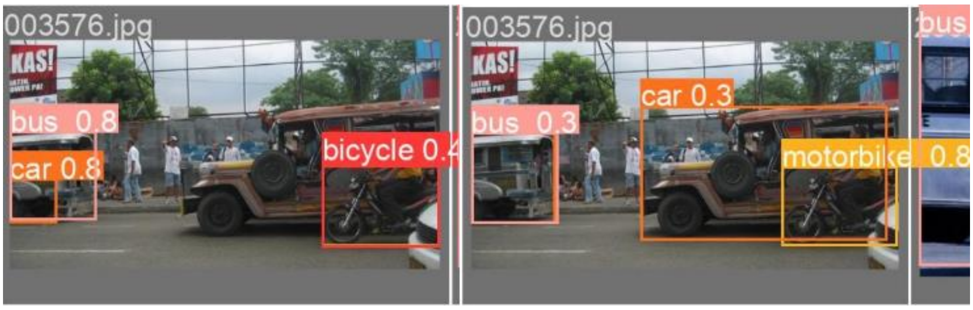

Fig. 5. YOLOv8n (left) and SSB-YOLO (right)

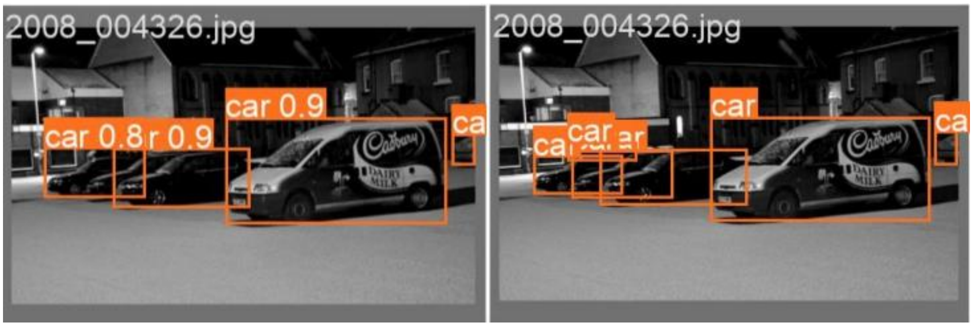

Fig. 6. YOLOv8n (left) and SSB-YOLO (right)

4 Discussion

To enhance the accuracy and performance of vehicle detection algorithms, this study proposes a new vehicle detection algorithm, SSB-YOLO, based on improvements to YOLOv8. SSB-YOLO incorporates the SA mechanism to filter out irrelevant factors, reducing model complexity. Additionally, the BWIoU algorithm, an improvement over the WIoU algorithm, is introduced to enhance model generalizability throughout the training period. Furthermore, the C2f module in the neck layer is enhanced to reduce computational resource consumption and improve model detection performance. Compared with YOLOv8n on the PASCAL VOC 2007 and PASCAL VOC 2012 vehicle datasets, SSB-YOLO shows a 1.6% improvement in mAP@50 while reducing GFLOPs by 0.6%. Consequently, in comparison with the original model, SSB-YOLO achieves better accuracy and reduced computational resource consumption.

Statements and Declarations

Competing interests

The authors did not receive support from any organization for the submitted work. The authors have no competing interests to declare that are relevant to the content of this article. in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Author Contributions

Conceptualization: Mingda Wang; Methodology: Mingda Wang; Formal analysis and investigation: Mingda Wang,Luyuan Ren; Writing - original draft preparation: Mingda Wang; Writing - review and editing: Mingda Wang; Resources: Mingda Wang; Supervision: Luyuan Ren

References

[1] X,Han., J,Chang., K,Wang.:Real-time object detection based on YOLO-v2 for tiny vehicle object.Procedia Computer Science.Volume 183.pp.61-72. (2021)

[2] Z, Wang., J, Zhan., C,Duan., X,Guan., P,Lu., K, Yang.:A Review of Vehicle Detection Techniques for Intelligent Vehicles/in IEEE Transactions on Neural Networks and Learning Systems.vol.34,no.8,pp.3811-3831.Aug.(2023) https://doi.org/10.1109/TNNLS.2021.3128968.

[3] Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: Toward Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence.39.1137-1149. (2015)

[4] REDMON, J., DIVVALA, S., GIRSHICK, R., et al.:You Only Look Once: Unified, Real-time Object Detection.Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Las Vegas . IEEE.2016:779-788.(2016)

[5] Liu, W., Dragomir Anguelov, D. Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu and Alexander C. Berg. :SSD: Single Shot MultiBox Detector.European Conference on Computer Vision .(2015)

[6] LIN T Y., GOYAL ,P., GIRSHICK, R., et al. :Focal loss for Dense Object Detection. Proceedings of the IEEE International Conference on Computer Vision(ICCV). Venice : IEEE , 2017: 2980-2988.(2017)

[7] Zhang, Y., Sun, Y.,Wang, Z.,Jiang, Y.: YOLOv7-RAR for UrbanVehicle Detection. Sensors 2023, 23,1801. (2023)

[8] Wang, J., Dong, Y., Zhao, S., Zhang, Z. :A High-Precision Vehicle Detection and Tracking Method Based on the Attention Mechanism. Sensors 2023, 23, 724. (2023)

[9] Farid, A., Hussain, F., Khan,K., Shahzad, M., Khan, U.,Mahmood,Z.: A Fast and Accurate Real-TimeVehicle Detection Method UsingDeep Learning for UnconstrainedEnvironments. Appl. Sci. 2023, 13,3059. (2023)

[10] Arora, N., Kumar, Y., Karkra, R., et al. :Automatic vehicle detection system in different environment conditions using fast R-CNN. Multimed Tools Appl 81, 18715–18735 (2022). https://doi.org/10.1007/s11042-022-12347-8

[11] J, Zhao., et al.:Improved Vision-Based Vehicle Detection and Classification by Optimized YOLOv4, in IEEE Access, vol. 10, pp. 8590-8603. (2022) doi: 10.1109/ACCESS.2022.3143365.

[12] Y, Yuan., H,Cheng.,M, Sester.:Keypoints-Based Deep Feature Fusion for Cooperative Vehicle Detection of Autonomous Driving.in IEEE Robotics and Automation Letters. vol. 7. no. 2, pp. 3054-3061, April (2022). doi: 10.1109/LRA.2022.3143299.

[13] ZHOU, F., ZHAO ,H., NIE ,Z.: Safety helmet detection based on YOLOv5.2021 IEEE International conference on power electronics, computer applications. Shenyang: IEEE, 2021: 6-11.(2021)

[14] LIU,S.,QI,L., QIN,F., et al.: Path Aggregation Network for Instance Segmentation. In: IEEE Conf. Comput. Vis. Pattern Recognit. Salt Lake City.pp.8759-8768. (2018)

[15] LIN, Y., DOLLAR,P., GIRSHICK,R., et al.: Feature Pyramid Networks for Object Detection. In: IEEE Conf. Comput. Vis. Pattern Recognit. Honolulu.pp.936-944 .(2017)

[16] Zheng, Z., Wang, P., Ren, D., Liu, W., Ye, R., Hu, Q., Zuo, W.: Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance Segmentation. IEEE Transactions on Cybernetics, 52, 8574-8586. (2020)

[17] Tong, Z., Chen, Y., Xu, Z., Yu, R. :Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism. ArXiv.abs/2301.10051. (2023)

[18] Li,J.,Wen,Y.,He,L.: SCConv:Spatial and Channel Reconstruction Convolution for Feature Redundancy.In: IEEE Conf. Comput. Vis. Pattern Recognit.Vancouver.pp.6153–6162. (2023)

[19] Zhang, Q., Yang, Y. :SA-Net: Shuffle Attention for Deep Convolutional Neural Networks. In: IEEE Int. Conf. Acoust. Speech Signal Process.Toronto.pp.2235-2239. (2021)

[20] Li, X., Hu, X., Yang, J.: Spatial Groupwise Enhance: Improving Semantic Feature Learning in Convolutional Networks. ArXiv. abs/1905.09646. (2019).