



Helmet Detection Algorithm of Electric Bicycle Riders Based on YOLOv5 with CBAM Attention Mechanism Integration

Si-Yue Fu¹, Dong Wei^{1,2(✉)}, and Liu-Ying Zhou¹

¹ School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China

weidong@bucea.edu.cn

² Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing Municipal Science and Technology Commission, Beijing 100044, China

Abstract. Object detection algorithms can assist in detecting the helmet-wearing status of electric bicycle riders, thereby saving regulatory manpower costs. However, there is currently a lack of standardized and publicly available datasets. Additionally, the basic YOLOv5s object detection algorithm, due to its limited feature extraction capabilities, may lead to numerous instances of both false negative and false positive. To enhance the model's focus on critical information within the feature maps, this paper introduces the CBAM attention mechanism module into the Backbone section of YOLOv5s. This module sequentially infers attention maps from the input feature maps along both channel and spatial dimensions independently, and then multiplies these attention maps with the input feature maps to achieve adaptive feature optimization. This paper have established self-built dataset for experimental research, and the results indicate that compared to the original YOLOv5s model, the proposed method has improved the model's overall mAP score by 1.89%.

Keywords: YOLOv5 · Object detection · Attention mechanism · CBAM

1 Introduction

In recent years, with the rapid development of economy and the expansion of road traffic, electric bicycles have become one of the main means of transportation for mass travel. However, while people enjoy the convenience of electric bicycles, they often have serious safety accidents because they do not wear helmets. According to statistics, about 80% of the deaths of motorcycle and electric bicycle drivers are caused by craniocerebral injury, and the correct wearing of safety helmets can greatly reduce the risk of death in traffic accidents, which plays an important role in protecting the life safety of the masses. Therefore, since April 2020, the Ministry of Public Security has launched a nationwide “One helmet and one belt” safety protection action to correct unsafe behaviors such as motorcycle and electric bicycle riders not wearing safety helmets. The current supervision method is still based on the daily patrol of traffic police, and the workload

is large. In this paper, the object detection algorithm is used to realize the automatic identification of helmet wearing of electric bicycle drivers and passengers, so as to save labor costs and improve urban management efficiency.

Currently, object detection based on deep learning can be categorized into two main types: “two-stage detection” and “one-stage detection” [1]. The “two-stage detection” algorithms perform object detection in two steps. Firstly, they generate candidate object regions, and then combine these regions with CNN networks to extract features and perform regression-based classification. The “two-stage detection” category includes representative algorithms such as the RCNN series [2–4], SPPNet [5], and FPN [6]. These algorithms exhibit high detection accuracy but require significant computational resources. On the other hand, the “one-stage detection” algorithms employ a single neural network to simultaneously generate candidate regions, classify objects, and localize their positions, eliminating the need for explicit region proposal generation. This end-to-end training approach simplifies the complex processes of two-stage methods, significantly improving the detection speed. However, it often leads to a decrease in localization accuracy. Representative algorithms in this category include the YOLO series [7–10] and SSD series [11–13].

Following the release of YOLOv5 in 2020, Ultralytics Corporation introduced the latest version, YOLOv8, in January 2023. In comparison to YOLOv5, the YOLOv8 model employs a more complex network architecture, which allows it to achieve higher accuracy. However, it requires more training data and computational resources to reach optimal performance. On the other hand, YOLOv5 offers a comprehensive advantage of a lightweight model, speed, efficiency, and high accuracy. It has garnered widespread attention and applications across various domains. Considering that most traffic monitoring scenarios involve the use of low-power embedded devices and low-resolution cameras, which demand lightweight models, this paper chooses YOLOv5 as the foundational model for implementing object detection tasks. So far, numerous researchers have chosen YOLOv5 as a baseline network and made targeted improvements to achieve more accurate detection results. Wang J et al. [14] proposed an improved YOLOv5 network for real-time multi-scale traffic sign detection. They replaced the original feature pyramid network in YOLOv5 with AF-FPN, which improved the detection performance of YOLOv5 on multi-scale objects while ensuring real-time detection. Qi D et al. [15] developed a face detector called YOLO5Face based on YOLOv5, and they designed detectors of different sizes based on the deployment capabilities of different platforms. Qi J et al. [16] introduced a method to incorporate the SE attention mechanism module into the YOLOv5 model to enhance the extraction of crucial features for more accurate identification of tomato virus diseases. Wu et al. [17] replaced the original Bottleneck structure in the YOLOv5 model with the Ghost Bottleneck composed of Ghost modules, resulting in a novel neural network model called Yolov5-Ghost. This model was used for vehicle detection in virtual environments and effectively reduced computational complexity.

The YOLOv5 series of object detection networks consists of four versions: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. Among them, YOLOv5s is the smallest and lightest version, while the other three versions progressively increase the width and depth of the YOLOv5s model to achieve better detection accuracy. The lightweight

nature and lower computational requirements of YOLOv5s make it more suitable for embedded devices and low-resolution cameras. Zhou et al. [18] conducted training and testing of YOLOv5 models with different parameters for construction site safety helmet detection. The experimental results showed that YOLOv5s achieved an average detection speed of 110 FPS, meeting the requirements for real-time detection. Although the YOLOv5s series models have been applied in various domains, they require the construction of corresponding datasets specific to the detection tasks to enable the models to extract general features of the target objects and accomplish the object detection task. Furthermore, the basal YOLOv5s network suffers from insufficient feature extraction capabilities due to the wastage of computational resources in redundant background regions. Consequently, it exhibits varying degrees of missed detections and false detections in different detection tasks.

For addressing this issue, incorporating attention mechanisms at appropriate positions within object detection models can help enhance the model's feature extraction capabilities. In the current field of computer vision, there are two common types of attention mechanisms: channel attention and spatial attention. Channel attention primarily focuses on determining what important feature information is, while spatial attention emphasizes where the feature information is located. In this paper, a hybrid-domain attention mechanism known as CBAM [19] is employed. CBAM sequentially applies both channel and spatial attention modules to enhance the adaptive feature optimization capabilities of convolutional neural networks across various channels and pixel positions. As a result, CBAM achieves superior performance improvements compared to single-dimensional attention mechanisms such as SE [20] and ECA [21].

To achieve automated recognition of whether electric bicycle riders are wearing safety helmets, this paper independently constructed a dataset and conducted experimental research based on it. Ultimately, the paper introduced CBAM attention mechanism into the backbone section of the YOLOv5s model. This enhancement resulted in a significant 1.89% increase in the model's overall mAP, effectively reducing instances of false positives and false negatives.

2 Collection and Processing of the Dataset

Due to the absence of publicly available datasets for helmet detection, this study independently creates a dataset of electric bicycle riders and their helmets through web scraping and on-site photography at street intersections. The collected dataset comprises a total of 2172 images, with 1145 images obtained through web scraping and 1027 images captured through on-site photography. The dataset includes samples with varying weather conditions, different shooting angles, and different levels of congestion, as illustrated in Fig. 1.

After analyzing the dataset, it was found that non-helmeted electric bicycle riders exhibit various head features, including different hair colors and styles, as well as the presence of various hats that bear resemblance to helmets, as shown in Fig. 2. In order to reduce false detections, this paper proposes the introduction of an attention mechanism to enhance the feature extraction capability of the model.

For the detection task presented in this paper, identifying the riders on electric bicycles is the algorithm's crucial first step. Given that the second step involves recognizing



(a) varying weather conditions

(b) different shooting angles



(c) different levels of congestion

Fig. 1. Partial visualizations of the dataset images.



Fig. 2. Typical examples of non-helmeted individuals.

the head region, which occupies a smaller area compared to the entire rider target and contains limited feature information, it is essential to treat it as a distinct category for separate recognition. Therefore, the detection categories are divided into two classes: “NMV” (non-motorized vehicle) and “Helmet”. Additionally, other head features that do not correspond to wearing a safety helmet are considered as background negative samples during training. After determining the category criteria, an online annotation tool called Make Sense was utilized to annotate the detection objects in the dataset, generating YOLO-formatted txt files. The content of the txt files is illustrated in Table 1, where each row represents five pieces of information for a detection box: the category

index, normalized coordinates of the center point (X, Y), and the normalized width and height of the bounding box.

Table 1. The content meaning of the annotated file in YOLO format.

Class	Centre_X	Centre_Y	Width	Height
0	0.508446	0.516470	0.618243	0.841216
1	0.453125	0.175042	0.098818	0.158361

After the annotation task was finished, the images along with their corresponding annotation files were randomly split into training, validation, and testing sets in a ratio of 8:1:1.

3 YOLOv5s Algorithm with Integrated Attention Mechanism.

3.1 CBAM Attention Mechanism

The CBAM attention mechanism [19], introduced by Sanghyun Woo et al. in 2018, consists of two key modules: the Channel Attention Module (CAM) and the Spatial Attention Module (SAM). It is a prominent example of a hybrid-domain attention mechanism.

The CAM submodule generates channel-domain attention masks by analyzing the relationships among channels in the feature map. Its purpose is to capture the differences in importance among channels in the feature map. Since each channel of the feature map is considered as a feature detector, CAM focuses attention on the question of what important feature information exists in the image. The structure of CAM is illustrated in Fig. 3.

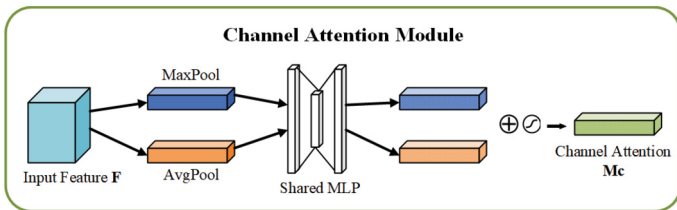


Fig. 3. Diagram of Channel Attention Module.

The CAM module first performs the global maximum pooling and the global average pooling operations as *MaxPool* and *AvgPool* on the input original feature map F to aggregate spatial information of the feature map. This results in two spatial context descriptors, F_{max}^c and F_{avg}^c . These descriptors are further processed by a shared network *MLP* to obtain two sets of weight vectors. The weight vectors are element-wise summed and then passed through a Sigmoid function σ to merge and output a set

of one-dimensional channel attention weights $\mathbf{M}_c(\mathbf{F})$. The computation process is as follows:

$$\begin{aligned} \mathbf{M}_c(\mathbf{F}) &= \sigma(MLP(MaxPool(\mathbf{F})) + MLP(AvgPool(\mathbf{F}))) \\ &= \sigma\left(W_1(W_0(\mathbf{F}_{Max}^c)) + W_1(W_0(\mathbf{F}_{Avg}^c))\right). \end{aligned} \quad (1)$$

The shared network MLP consists of a multi-layer perceptron with a single hidden layer. It has parameter vectors $W_0 \in \mathbb{R}^{\frac{C}{r} * C}$ and $W_1 \in \mathbb{R}^{C * \frac{C}{r}}$, where r is the reduction ratio. Finally, the weight vector $\mathbf{M}_c(\mathbf{F})$ is multiplied element-wise with the original feature map \mathbf{F} to obtain the new feature map \mathbf{F}' . The computation process is as follows:

$$\mathbf{F}' = \mathbf{M}_c(\mathbf{F}) \otimes \mathbf{F}. \quad (2)$$

The SAM submodule generates spatial-domain attention masks by analyzing the spatial relationships among features. Its purpose is to capture the differences in importance among features in the spatial dimension of the feature map. In contrast to CAM, which focuses on what feature information is important, SAM focuses on where the feature information is located. SAM and CAM complement each other in this regard. The structure of SAM is illustrated in Fig. 4.

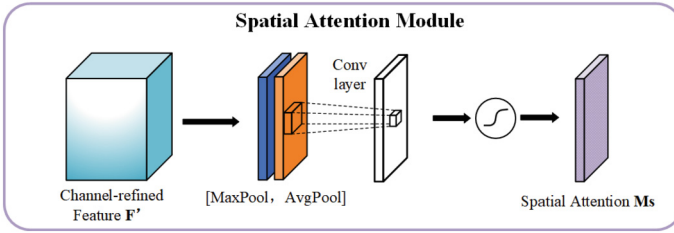


Fig. 4. Diagram of Spatial Attention Module.

The SAM submodule takes the feature map \mathbf{F}' generated by the previous module as input. It applies $MaxPool$ and $AvgPool$ operations along the channel axis to pool all channels at the same pixel position, aggregating the channel information of the feature map. This process results in two two-dimensional feature maps, F'^s_{Max} and F'^s_{Avg} . .. These two feature maps are then concatenated and convolved with a $7 * 7$ filter f^{7*7} . The resulting output is passed through a Sigmoid function σ to generate a two-dimensional spatial attention feature map $\mathbf{M}_s(\mathbf{F}')$. The computation process is as follows:

$$\begin{aligned} \mathbf{M}_s(\mathbf{F}') &= \sigma(f^{7*7}([MaxPool(\mathbf{F}'); AvgPool(\mathbf{F}')])) \\ &= \sigma\left(f^{7*7}[F'^s_{Max}; F'^s_{Avg}]\right). \end{aligned} \quad (3)$$

Finally, the module performs element-wise multiplication between the spatial attention feature map $\mathbf{M}_s(\mathbf{F}')$ and the input feature map \mathbf{F}' , resulting in the final feature map \mathbf{F}'' enhanced by the attention mechanism in both dimensions. The computation process is as follows:

$$\mathbf{F}'' = \mathbf{M}_s(\mathbf{F}') \otimes \mathbf{F}'. \quad (4)$$

The overall structure of the CBAM attention mechanism is illustrated in Fig. 5. The original feature map sequentially passes through the CAM and SAM modules, enabling the adaptive refinement of key feature information in both the channel and spatial domains.

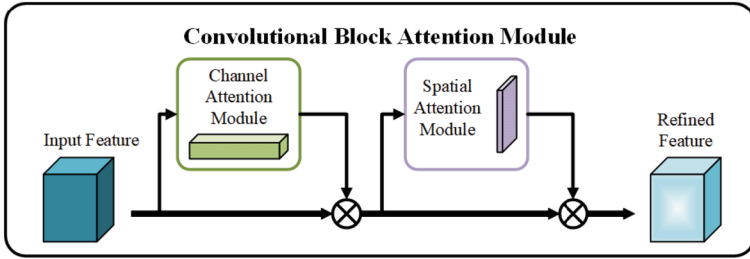


Fig. 5. The overview of CBAM.

From the above computation process, it can be observed that both the CAM and SAM sub-modules utilize *MaxPool* and *AvgPool* pooling methods in parallel. This parallel utilization allows for the preservation of a greater amount of feature information compared to using only one of these pooling methods, thereby enhancing the feature extraction capability. Furthermore, the CBAM attention mechanism adopts a sequential structure design, where the CAM sub-module is applied before the SAM sub-module. This design has been validated through comparative experiments in literature [19], conducted on the benchmark architecture of ResNet50, demonstrating superior application effectiveness compared to the parallel utilization of the two sub-modules or a sequential structure with the order reversed.

3.2 The improved YOLOv5s + CBAM model

In this paper, an additional layer of CBAM attention mechanism is integrated into the backbone section of YOLOv5s, aiming to maximize the enhancement of the model's feature extraction capability. The improved YOLOv5s + CBAM model structure, as shown in Fig. 3, is primarily divided into four components: the input end (Input), feature extraction network (Backbone), feature fusion network (Neck), and prediction end (Head). The attention mechanism module is introduced in the backbone section, as indicated by the red text in Fig. 6.

The Input component is responsible for receiving image data as input to the algorithm. It mainly includes the Mosaic data augmentation method, adaptive anchor box calculation, and adaptive image scaling. The Mosaic data augmentation method randomly scales multiple images and then concatenates them into a single image, effectively enriching the dataset and improving the performance of small object detection as well as the robustness of the model. The adaptive anchor box calculation helps the model automatically set the initial anchor box size when the dataset changes. Adaptive image scaling processes the input image size to meet the required input dimensions of

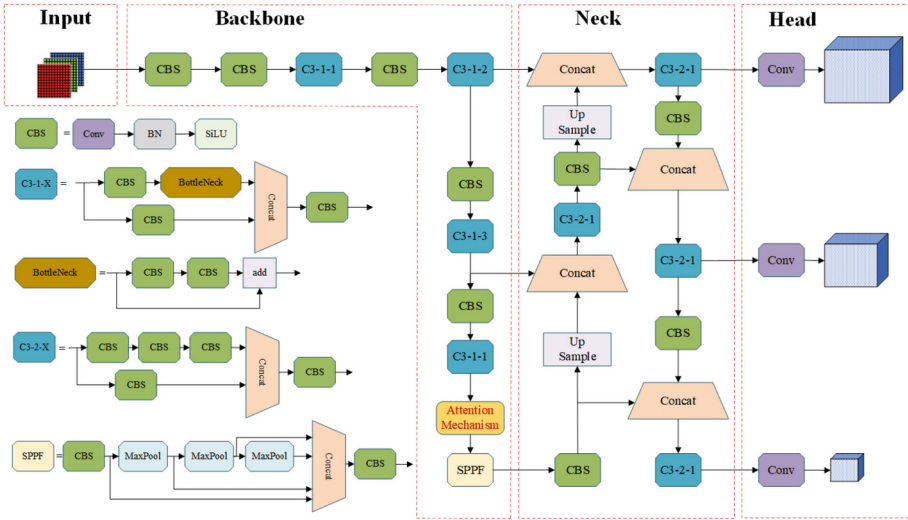


Fig. 6. YOLOv5s + CBAM network structure.

the network, such as by compressing or adding gray borders. This allows YOLOv5s to adapt to different scenes and image resolutions.

The Backbone component adopts the CSPDarknet53 structure, which mainly consists of the CBS (Convolution BatchNormalization SiLU) module, C3 (CSP Bottleneck with Three Convolutions) module, and SPPF (Spatial Pyramid Pooling-Fast) module, as depicted on the left side of Fig. 3. Serving as the backbone network of YOLOv5, the Backbone component is responsible for extracting generic features of the targets through convolutional operations and transforming them into high-level semantic features for subsequent object detection. An attention mechanism network layer is added at the end of the Backbone, enabling the model to enhance its focus on key parts of the feature map through a feature extraction process with a global view.

The Neck component has been incorporated into the backbone network and prediction layers since YOLOv3 to fuse the features extracted by the backbone network. The multi-scale prediction module within the Neck network layer enables the detection of objects at three different scales simultaneously. In YOLOv5, the Neck network layer adopts a combined structure of FPN (Feature Pyramid Network) [22] and PAN (Path Aggregation Network) [23], leveraging their complementary roles. The FPN layer, which conveys strong semantic features top-down, is combined with the PAN layer, which conveys strong localization features bottom-up. This aggregation process integrates high-level semantic information and low-level positional information to produce three fused and effective feature output layers, which are then passed to the prediction end.

Lastly, the Head component generates anchor boxes of different sizes based on the feature maps outputted by the feature extraction stage. It then utilizes non-maximum suppression (NMS) [24] to remove redundant bounding boxes and generate the positional information of the target bounding boxes as well as their class probabilities, enabling object detection and classification tasks.

4 Experimental Study and Result Analysis

4.1 Experimental Environment

The experimental platform used for training and testing in this paper employed Windows 10 as the operating system. The hardware specifications included 32 GB of memory, an Intel(R) Core(TM) i7-9700K CPU, and an NVIDIA GeForce RTX 3060 GPU with 16 GB of VRAM. The experiments were conducted using the PyTorch deep learning framework, with the IDE environment being PyCharm 2019. The CUDA version used was 11.6, and the programming language employed was Python 3.9.7.

4.2 Evaluation Metrics

In this paper, the performance of the model is evaluated using the following metrics: Precision, Recall, AP (Average Precision) for each class, and mAP (mean Average Precision) overall.

TP (True Positive) refers to the cases where positive samples are correctly detected. FP (False Positive) refers to the cases where negative samples are incorrectly detected as positive, including both localization errors and classification errors. FN (False Negative) refers to the cases where positive samples are incorrectly detected as negative, also known as missed detections.

Precision is the ratio of TP to the total number of predicted boxes. It measures the false detection rate of the model.

$$Precision = \frac{TP}{TP + FP}. \quad (5)$$

Recall is the ratio of TP to the total number of positive samples. It measures the missed detection rate of the model.

$$Recall = \frac{TP}{TP + FN}. \quad (6)$$

AP is the average precision at different recall levels. It is calculated as the area under the Precision-Recall curve and reflects the accuracy of predictions for each class.

$$AP = \int_0^1 Precision(Recall)d(Recall). \quad (7)$$

mAP is the average of AP values across all classes. It provides an overall measure of the model's accuracy.

$$mAP = \frac{AP_{NMV} + AP_{Helmet}}{2}. \quad (8)$$

Table 2. Performance comparison of different models

Model	mAP@0.5(%)	Parameters	GFLOPs	FPS
YOLOv3-SPP	90.75	61,502,815	154.6	35
YOLOv5s	90.50	7,015,519	15.8	55
YOLOv8n	88.80	3,006,038	8.1	188
YOLOv8s	90.93	11,126,358	28.4	125

4.3 Experimental Results and Analysis

In the selection of the algorithm’s base model, this paper considered four models: YOLOv3-SPP, YOLOv5s, YOLOv8n, and YOLOv8s. Performance tests were conducted on each of these models using the same dataset. During the training phase, the number of iterations (epochs) was set to 100, and the batch size was consistent at 16. The performance results obtained are summarized in Table 2.

Based on the data from Table 2, it is evident that YOLOv3-SPP, YOLOv5s, and YOLOv8s models exhibit very similar detection accuracy, all of which surpass YOLOv8n. Furthermore, YOLOv5s stands out with less parameters and lower computational load compared to both YOLOv3-SPP and YOLOv8s. Additionally, YOLOv5s achieves a high frame rate of 55 fps, making it suitable for subsequent video dynamic detection requirements. Overall, when compared to the older version YOLOv3 and the newer version YOLOv8, YOLOv5 demonstrates its characteristics of being lightweight, fast, efficient, and accurate. Therefore, this paper chose to build upon the YOLOv5s base model for further improvements.

During the process of improving the base model, the experiments were conducted using a transfer learning approach, where the pre-trained weights file “yolov5s.pt” and hyperparameters file “hyp.scratch.yaml” were utilized. The training was performed on a custom dataset. During the training phase, the number of epochs was set to 100, and a batch size of 12 was used. The model was trained using the stochastic gradient descent (SGD) optimizer, with a learning rate of 0.01, momentum of 0.937, and weight decay of 0.0005. The Mosaic data augmentation method was employed to enhance the detection capability for small objects.

To validate the effectiveness of the proposed method, this paper compared the detection algorithms before and after improvement under the same experimental conditions and data augmentation strategies. The resulting detection accuracy is presented in Table 3.

According to the statistical data from Table 3, introducing the CBAM attention mechanism into the YOLOv5 network results in a 1.89% improvement in the overall mAP (mean Average Precision) score of the model. This experimental result confirms that incorporating the attention mechanism network helps strengthen the YOLOv5s model’s focus on key information within the feature maps, making the feature extraction process more efficient and ultimately enhancing the model’s overall mAP score. Regarding the improvements in algorithm precision and recall metrics, this paper conducted visual

Table 3. The detection accuracy comparison between YOLOv5s and YOLOv5s + CBAM

Algorithm	Classes	Precision(%)	Recall(%)	AP/mAP@0.5(%)
YOLOv5s	NMV	87.19	92.93	91.47
	Helmet	79.78	94.30	89.53
	all	82.50	93.58	90.50
YOLOv5s + CBAM	NMV	86.03	96.22	94.86
	Helmet	79.69	94.49	89.92
	all	82.94	95.40	92.39

comparisons of the detection algorithms before and after improvement on the same test dataset. Some of the detection result comparisons are illustrated in Fig. 7.

From Fig. 7, it can be observed that the improved YOLOv5s + CBAM algorithm mitigates some of the false positives or false negatives that were present in the original algorithm. In the first image, the electric bicycle rider on the far right is moving away from the camera. Due to the smaller and less distinct feature region at the rear of the electric bicycle, this target is prone to being missed by the detection algorithm. The pink helmet in the second image and the red helmet in the third image both represent the head features of electric bicycle passengers. Compared to the position of the driver, these two targets exhibit a slight offset within the entire NMV detection box. Moreover, since the training dataset contains relatively few samples of multiple individuals riding on an electric bicycle, these targets are more likely to be missed. In the fourth test image, the baseball cap worn by the electric bicycle rider in the middle can be misidentified as a helmet due to its similar shape. In response to these instances of false negatives and false positives, the object detection algorithm with the integrated attention mechanism enhances the model's feature extraction capabilities from the feature maps, thereby improving the model's detection accuracy. Consequently, the YOLOv5s + CBAM algorithm effectively mitigates the aforementioned errors.



(a) Object detection results of YOLOv5s



(b) Object detection results of YOLOv5s+CBAM

Fig. 7. Object detection results of YOLOv5s and YOLOv5s + CBAM

5 Conclusion

Currently, the supervision of helmet usage among electric bicycle riders consumes a substantial amount of manpower. In order to save labor costs, this paper first constructs a self-made dataset and then selects YOLOv5s as the foundational model for object detection, further improving it to achieve automated detection. To solve the problem of insufficient feature extraction capability of YOLOv5s model, this paper proposes an improved algorithm of combined CBAM attention mechanism. The introduced CBAM attention mechanism uses CAM submodule and SAM submodule successively to enhance the adaptive feature optimization ability of convolutional neural network in different channels and pixel positions, thus improving the detection accuracy of the model. The experimental results show that the improved YOLOv5s + CBAM algorithm can increase the overall mAP index of the model by 1.89%, and significantly reduce instances of both false negative and false positive.

Due to the discrepancy in target sizes between electric bicycle riders and helmets, along with their inherent hierarchical relationship, the current improvement in this paper is primarily observed in the boosted AP for electric bicycle riders. In the next step, we plan to divide the detection task into two separate models. Initially, the first object detection model will be employed to detect electric bicycle riders and export them as input for the subsequent model. Subsequently, the second model will be utilized to conduct helmet detection on each individual electric bicycle rider target. This strategy will enable targeted performance optimization for each model. Moreover, we will persist in augmenting and refining the self-constructed dataset in terms of both its quantity and quality.

Acknowledgments. This work was supported by the High Level Innovation Team Construction Project of Beijing Municipal Universities (No. IDHT20190506), the Science and Technology Project of China Ministry of Housing and Urban-Rural Development (No. 2019-K-149) and the Pyramid Talent Training Project of Beijing University of Civil Engineering and Architecture (GJZJ20220803).

References

1. Zou, Z., Chen, K., Shi, Z., et al.: Object detection in 20 years: a survey. In: Proceedings of the IEEE (2023)
2. Girshick, R., Donahue, J., Darrell, T., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
3. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
4. Ren, S., He, K., Girshick, R., et al.: Faster r-cnn: towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **28** (2015)
5. He, K., Zhang, X., Ren, S., et al.: Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(9), 1904–1916 (2015)
6. Lin, T.Y., Dollár, P., Girshick, R., et al.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)

7. Redmon, J., Divvala, S., Girshick, R., et al.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
8. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271 (2017)
9. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint [arXiv:1804.02767](https://arxiv.org/abs/1804.02767) (2018)
10. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint [arXiv:2004.10934](https://arxiv.org/abs/2004.10934) (2020)
11. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
12. Li, Z., Zhou, F.: FSSD: feature fusion single shot multibox detector. arXiv preprint [arXiv:1712.00960](https://arxiv.org/abs/1712.00960) (2017)
13. Fu, C.Y., Liu, W., Ranga, A., et al.: Dssd: Deconvolutional single shot detector. arXiv preprint [arXiv:1701.06659](https://arxiv.org/abs/1701.06659) (2017)
14. Wang, J., Chen, Y., Dong, Z., et al.: Improved YOLOv5 network for real-time multi-scale traffic sign detection. *Neural Comput. Appl.* **35**(10), 7853–7865 (2023)
15. Qi, D., Tan, W., Yao, Q., et al.: YOLO5Face: why reinventing a face detector. In: European Conference on Computer Vision. Cham: Springer Nature Switzerland, pp. 228–244 (2022). https://doi.org/10.1007/978-3-031-25072-9_15
16. Qi, J., Liu, X., Liu, K., et al.: An improved YOLOv5 model based on visual attention mechanism: application to recognition of tomato virus disease. *Comput. Electron. Agric.* **194**, 106780 (2022)
17. Wu, T.H., Wang, T.W., Liu, Y.Q.: Real-time vehicle and distance detection based on improved yolo v5 network. In: 2021 3rd World Symposium on Artificial Intelligence (WSAI). IEEE, pp. 24–28 (2021)
18. Zhou, F., Zhao, H., Nie, Z.: Safety helmet detection based on YOLOv5. In: 2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA), pp. 6–11. IEEE (2021)
19. Woo, S., Park, J., Lee, J.Y., et al.: CBAM: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
20. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
21. Wang, Q., Wu, B., Zhu, P., et al.: ECA-Net: efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11534–11542 (2020)
22. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 936–944 (2017). <https://doi.org/10.1109/CVPR.2017.106>
23. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 8759–8768 (2018). <https://doi.org/10.1109/CVPR.2018.00913>
24. Neubeck, A., Van Gool, L.: Efficient non-maximum suppression. In: 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, pp. 850–855 (2006). <https://doi.org/10.1109/ICPR.2006.479>