ISSN 2616-5775 Vol. 8, Issue 9: 17-25, DOI: 10.25236/AJCIS.2025.080903

Motorcycle Helmet Detection Model Based on Improved YOLOv11n

Tianhang Mu[a,*], Lei Ding[b]

School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an, Shaanxi, China
[a]lding@sust.edu.cn, [b]231612151@sust.edu.cn
**Corresponding author*

Abstract: *To address the performance degradation in current motorcycle helmet detection models caused by factors such as large variations in helmet scale, frequent occlusions, and complex backgrounds, this study proposes a motorcycle helmet detection model based on an improved YOLOv11n, which has undergone key improvements in the following areas: First, the C3k2 modules in the original model are partially replaced with C3k2-SCConv modules, where SCConv helps reduce redundant information and enhances the model's feature extraction capability in complex environments. Second, the iAFF module was introduced to replace the conventional concat operation for feature fusion, effectively leveraging detailed information from shallow layers and semantic information from deep layers, thereby improving the detection performance for small objects. Third, a MultiSEAM module is incorporated into the neck of the model to mitigate information loss caused by occlusion by learning the relationship between occluded and non-occluded regions, which helps reduce missed and false detections owing to occlusion. Finally, ADown modules were used to replace certain convolutional layers, reducing both the number of parameters and computational cost, thereby improving the detection speed. Experimental results demonstrate that the proposed model achieves a 3.4 percentage point improvement in mAP@0.5 compared to the baseline model, while maintaining a competitive detection speed, and overall outperforms existing mainstream object detection models in terms of comprehensive performance.*

Keywords: *Helmet Detection; YOLOv11n; Small Object; Occlusion Target Detection*

1. Introduction

Motorcycles, one of the most widely used means of transportation, significantly improve travel efficiency but also pose a higher risk of traffic accidents. Compared to drivers of other vehicles, motorcyclists are far more likely to suffer traumatic brain injuries in accidents. Wearing a standard-compliant safety helmet is one of the most effective measures for reducing the risk of such injuries. However, even after the nationwide implementation of the "One Helmet, One Belt" safety campaign for over a year, a large number of motorcycle riders and passengers still lack awareness of helmet-wearing [1].

To reduce the incidence of severe injuries in motorcycle-related traffic accidents, helmet-wearing compliance is typically monitored manually by law enforcement officers. However, this approach is time-consuming, labor-intensive, and prone to missed detections, making it challenging to ensure effective safety supervision.

With the rapid advancement of artificial intelligence, computer vision, and deep learning, researchers have proposed various object detection frameworks, including Fast R-CNN[2], SSD[3], RT-DETR[4], and the YOLO series. Leveraging object detection algorithms to monitor motorcycle helmet usage in real-world traffic scenarios can significantly reduce the human labor costs. Moreover, issuing penalties based on the detection results can enhance public awareness of helmet usage. Consequently, various motorcycle helmet detection methods have been developed based on different object detection frameworks.

For instance, Xie et al. enhanced the YOLOv5 model by integrating the Efficient Channel Attention (ECA-Net) mechanism to improve detection performance. They also introduced a Bi-FPN bidirectional feature pyramid to balance multi-scale features and adopted the Alpha-CIoU loss function to improve

ISSN 2616-5775 Vol. 8, Issue 9: 17-25, DOI: 10.25236/AJCIS.2025.080903

the localization accuracy. Although the modified model showed marked improvements in detecting small objects, it still exhibited high miss rates in the case of object occlusion[5]. Yuan et al. replaced the backbone of YOLOv8s with VanillaNet, incorporated the CARAFE module for upsampling, added an extra detection layer for tiny objects, and introduced the MPDIoU loss function. These enhancements significantly improved both the accuracy and speed of helmet detection, although there remains room for improvement under extreme conditions[6]. Yang et al. improved YOLOv8n by incorporating SPDConv and C2f-CGblock modules and replacing standard convolutions in the head with group convolutions. These changes enhanced the model's ability to detect low-resolution images collected from real road scenarios and reduced the computational cost. However, detection errors still occur under high-glare lighting conditions[7]. Zhou et al. adopted a progressive feature pyramid network to enhance detection performance in complex scenes and proposed the PCAHead and HelmetIoU loss function to optimise model understanding and data processing capabilities. Although these improvements increased the computational efficiency and accuracy, the model continued to suffer from missed and false detections for distant small targets[8]. Zhou et al. combined the strengths of MAFPN and BiFPN to propose the BIMAFPN structure, which was integrated into YOLOv10n to improve performance in complex traffic environments. They also replaced the traditional CIoU loss with the Inner-Wise MPDIoU loss to enhance accuracy and convergence speed, and introduced the LSCD detection head to reduce the number of parameters while improving performance. However, the adaptability of the model to diverse weather conditions remains limited[9].

Although the aforementioned improvements have contributed to motorcycle helmet detection, several challenges remain: missed and false detections of small targets persist; occlusions among multiple objects in dense traffic reduce detection performance; and the limited consideration of environmental conditions, along with the constraints on model size, computation, and inference speed, make it difficult to meet the requirements of real-time and edge deployment.

To balance the detection accuracy, parameter size, and computational efficiency, this was selects YOLOv11n as the baseline model[10]. As the latest version in the YOLO series, YOLOv11 inherits the strengths of its predecessors while offering excellent performance and a lightweight design that facilitates deployment. Based on YOLOv11n, this study proposes an improved real-time motorcycle helmet detection model to address the aforementioned issues. Specifically, the C3k2 modules were partially replaced with C3k2-SCConv modules to enhance feature extraction under complex backgrounds. The iAFF module was introduced to replace the simple concatenation of shallow and deep features, thereby improving the model's capability to detect small objects. To address occlusion-induced errors, the MultiSEAM module was added to the neck of the model to compensate for information loss. Finally, the ADown module replaces several convolutional layers to reduce both the parameter count and computational cost, enabling real-time detection and deployment on edge devices.

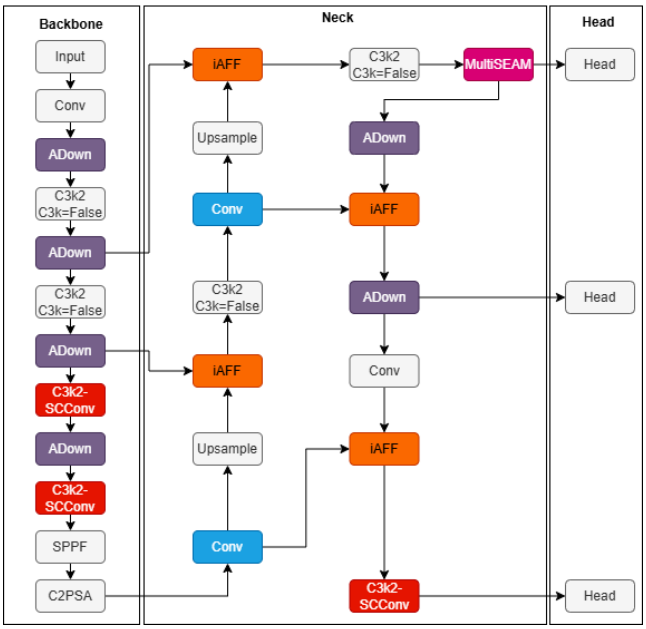

Figure 1: Improved YOLOv11 network structure.

ISSN 2616-5775 Vol. 8, Issue 9: 17-25, DOI: 10.25236/AJCIS.2025.080903

2. Improved YOLOv11 model

Although YOLOv11 demonstrated strong detection performance on conventional datasets, it did not perform well when applied to the task of motorcycle helmet detection. To address the typical challenges associated with this task, this study introduces targeted modifications to the backbone and neck of YOLOv11n. The resulting model architecture is illustrated in the following figure.

Based on these improvements, we propose a model specifically designed for the detection of motorcycle helmets. The detailed architecture is shown in Figure 1, and the following sections provide an in-depth explanation of the introduced modules and their modifications.

2.1 C3k2-SCConv module

In real-world scenarios, motorcycle riders are often surrounded by various sources of interference such as traffic, pedestrians, buildings, and trees. Helmet targets are typically small and may share colors with the background, making accurate identification challenging. Conventional standard convolutions apply the same processing across all spatial locations and channels, and lack the ability to dynamically distinguish between critical helmet features and background noise. As a result, the model struggles to focus on useful information amidst overwhelming redundancy.

SCConv[11] enhances feature extraction efficiency through a unique dual reconstruction mechanism that operates across both the spatial and channel dimensions. It primarily consists of two core components: a Spatial Reconstruction Unit (SRU) and a Channel Reconstruction Unit (CRU). By jointly leveraging these two units, the SCConv effectively suppresses redundant information and enables the model to focus on salient features. The architecture of the SCConv is illustrated in Figure 2.

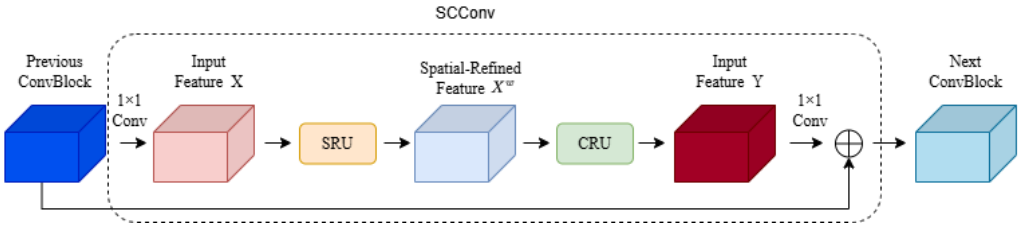

Figure 2: SCConv structure.

The C3k2 module serves as a fundamental building block of the YOLOv11 backbone and is responsible for feature extraction and downsampling. To further enhance the model's ability to extract features from complex backgrounds, we integrated SCConv into the C3k2 module, creating a new convolutional module named C3k2-SCConv. The C3k2-SCConv module replaces all C3k2 modules (C3k2=True) throughout the network.The structure of C3k2-SCConv is shown in Figure 3.

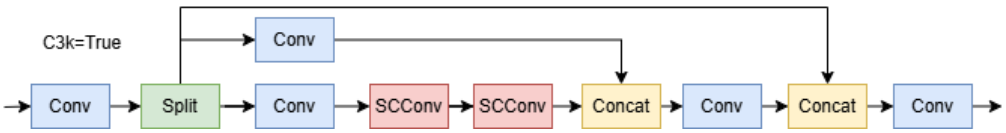

Figure 3: C3k2-SCConv structure.

2.2 iAFF module

Although shallow features retain the high-resolution details of small helmets, they lack high-level semantic information. In contrast, deep features contain rich semantic representations but often lose the spatial details of small objects owing to downsampling. Fusing shallow and deep features enables the integration of complementary information captured at different network depths, thereby reducing false and missed detections in small-object detection. In YOLOv11, feature fusion across different layers is performed by concatenation. However, this approach implicitly assumes equal importance across features and simply merges them without considering their semantic hierarchy, resolution sensitivity or saliency differences. This indiscriminate fusion may result in critical spatial details in shallow features being overwhelmed by strong semantic information in deep features, leading to blurred boundaries and texture loss, thereby reducing detection accuracy.

The iAFF module[12] uses the MS-CAM attention module and iterative optimization mechanism to

ISSN 2616-5775 Vol. 8, Issue 9: 17-25, DOI: 10.25236/AJCIS.2025.080903

ensure that the fused features contain advanced semantic information while also considering local spatial details, thereby improving the model's ability to detect small targets.The iAFF and MS-CAM structures are shown in Figure 4.

The MS-CAM module performs global and local feature processing on the input linear fusion features through two branches. The main difference between them lies in the application of global average pooling. The branch that applies global average pooling can capture global feature information, whereas the branch that does not apply global average pooling retains local feature information. After processing through these two branches, their outputs are combined and passed through a sigmoid function to generate weights. These weights are then applied to the original features, and through an iterative optimization design, the initial fusion results are subjected to another weighted fusion, addressing the bottleneck issues caused by the initial fusion. When fusing small helmet features, the weights generated by the MS-CAM module favor shallow-layer features to preserve helmet detail information while minimizing the loss of deep-layer semantic information, enabling the model to achieve better detection performance when detecting small helmet targets.

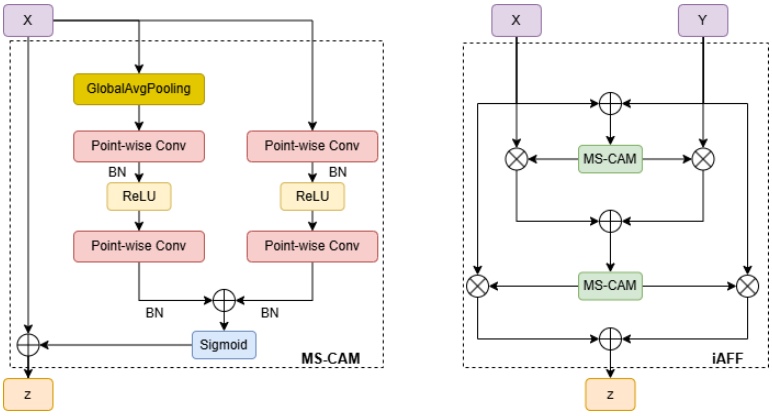

Figure 4: MS-CAM and iAFF structure.

2.3 MultiSEAM module

During motorcycle operation, vehicles, pedestrians, cyclists, electric bicycles, and obstacles on both sides of the road may partially obstruct the motorcycle, riders, or passengers. YOLOv11's fixed receptive field-based convolutional operations cannot focus on the local visible features of these occluded targets, leading the model to either over-rely on global features and misdetect, or fail to detect due to key features being occluded. To address the misdetection and false negatives caused by occlusion, the MultiSEAM module was introduced at the neck region.

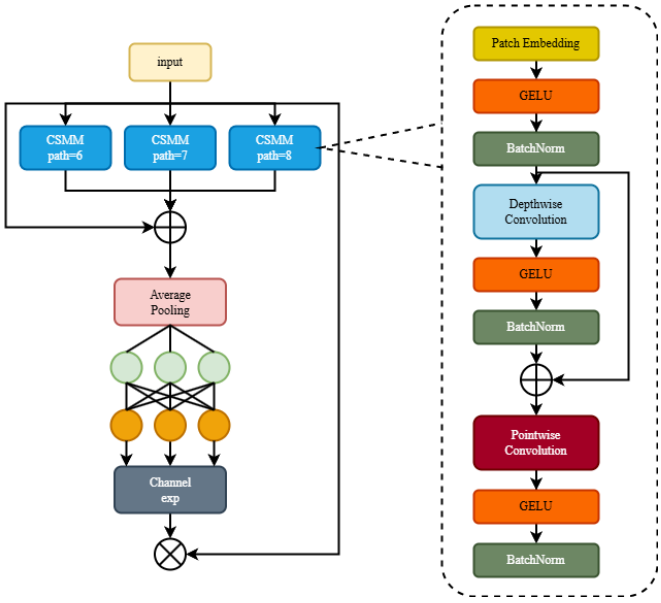

Figure 5: MultiSEAM structure.

ISSN 2616-5775 Vol. 8, Issue 9: 17-25, DOI: 10.25236/AJCIS.2025.080903

The MultiSEAM [13] module is designed to compensate for the information loss caused by occlusion by learning the relationships between the occluded and unoccluded regions. It begins by capturing both local and global information through three CSMM modules with different patch sizes(3, 5, and 7 in this study). After the CSMM modules, average pooling was applied to downsample the output. Subsequently, a two-layer fully connected network was employed to further integrate information across channels, enhancing inter-channel relationships while suppressing irrelevant signals. This strengthens the model's ability to handle fine-grained features in both the occluded and unoccluded regions. The outputs from the fully connected layers were then passed through an exponential transformation, extending their value range from [0,1] to [1,e]. This exponential normalization provides monotonic mapping that facilitates the effective integration of features from both the occluded and visible regions. Finally, the resulting values are used as attention weights and multiplied by the original features, enabling the model to effectively address occlusion-related challenges. The structure of the module is illustrated in Figure 5.

The MultiSEAM module enhances the detection performance of the model in occluded scenarios by integrating spatial attention with feature enhancement mechanisms. This allows the model to focus more effectively on the visible regions of the target, thereby optimizing the overall feature representation and improving the detection accuracy in the presence of occlusions.

2.4 ADown module

To enhance the real-time performance of motorcycle-helmet detection, several convolutional layers in the backbone and neck of the YOLOv11 network were replaced with the ADown module from YOLOv9[14]. The ADown module reduces model complexity and improves computational efficiency, enabling fast and accurate target detection, even in resource-constrained environments. Its structure is shown in Figure 6.

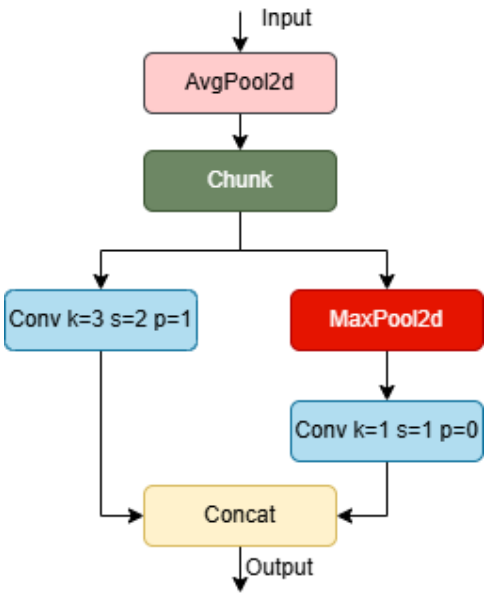

Figure 6: ADown structure.

This design cleverly combines two different feature extraction methods, convolution and pooling, to retain diverse and rich spatial information while significantly reducing the number of parameters and computational complexity. This is achieved because convolution operates on only half of the input channels, pooling operations have no parameters and low computational overhead, and 1×1 convolution is used to efficiently integrate channel information. Through this asymmetric branch structure, the ADown module effectively reduces model complexity and improves inference speed, making it highly suitable for achieving efficient and stable feature downsampling on resource-constrained edge devices.

ISSN 2616-5775 Vol. 8, Issue 9: 17-25, DOI: 10.25236/AJCIS.2025.080903

3. Experimental results and analysis

3.1 Dataset

This study utilized the "Osf HELMET" dataset [15] and the dataset provided by the 2024 AI City Challenge Track 5 [16]. To ensure dataset completeness, samples were selected from various traffic conditions, including sunny, rainy, nighttime, and foggy weather, as well as low-traffic and high-traffic road segments, totalling 5,489 images. Three categories were annotated using the LabelImg tool, as illustrated in Figure 7: green bounding boxes represent the 'motorperson' class, red boxes represent 'Helmet', and blue boxes represent 'NoHelmet.'

Figure 7: Data annotation method.

This annotation scheme offers several advantages: the 'motorperson' class treats the rider and motorcycle as a single entity, effectively filtering out interference from cyclists and pedestrians, while the 'Helmet' and 'NoHelmet' classes indicate the helmet-wearing status. The dataset was split into 4,873 images for training and 616 images for testing.

3.2 Experimental environment

The experiments were conducted on a platform running the Ubuntu 20.04 operating system, equipped with a 12-core Intel(R) Xeon(R) Platinum 8352V CPU @ 2.10GHz, 90GB RAM, and an NVIDIA vGPU with 32GB memory. The model was initialized with the pretrained weights of YOLOv11n trained on the COCO dataset. The training parameters were set as follows: number of epochs = 150, batch size = 32, and number of worker processes = 0. The input images were resized to 640 × 640 pixels. The AdamW optimizer was employed to optimize the learning rate with an initial learning rate of 0.01.

3.3 Comparative experiment

To evaluate the performance of the proposed algorithm, comparative experiments were conducted against several mainstream object detection algorithms, including YOLOv5s, YOLOv7-tiny, YOLOv8n, RT-detr, and NanoDet, as well as classic methods such as Faster R-CNN and SSD. For all algorithms used in the experiments, the input images were standardized to a resolution of 640 × 640 pixels. Each model was initialized with pretrained weights on public datasets and trained for 150 epochs. The evaluation metrics included the number of parameters, computational complexity (FLOPs), precision, recall, and mean average precision at IoU threshold 0.5 (mAP@0.5) to comprehensively assess model performance.

Table 1: Algorithm performance comparison.

method	Params/M	FLOPs/G	Precision/%	Recall/%	mAP0.5/%
Faster-RCNN	136.2	360.0	76.6	74.3	78.8
SSD	25.4	131.7	74.6	73.4	75.2
RT-detr	15.5	37.4	84.9	79.0	84.2
nanodet	2.44	2.97	-	70.2	83.5
yolov5s	7.02	15.8	85.7	81.0	85.9
yolov7-tiny	6.01	13.0	83.5	77.5	84.8
yolov8n	3.01	8.1	82	79.5	85.5
yolov11n	2.58	6.3	82.9	79.6	84.8
ours	2.69	5.3	88.9	82.5	88.2

ISSN 2616-5775 Vol. 8, Issue 9: 17-25, DOI: 10.25236/AJCIS.2025.080903

As shown in the experimental results in Table 1, all compared algorithms outperform the classical Faster R-CNN and SSD methods across various metrics. Regarding the mAP@0.5 metric, the proposed model achieved the highest score among the nine evaluated models, surpassing YOLOv5s, YOLOv7-tiny, YOLOv8n, and YOLOv11n by 2.3%, 3.4%, 2.7%, and 3.4%, respectively. Compared with RT-detr and NanoDet, the proposed model improved by 4.0% and 4.7%, respectively. In terms of parameter count and computational cost, the model was only surpassed by NanoDet. Although the improved model increases parameters by 4.2% compared with the original YOLOv11n, it achieves the highest detection accuracy, demonstrating significant advantages over the baseline and other models under limited computational resources.

3.4 Ablation experiment

To evaluate the effectiveness of each module incorporated into the YOLOv11n model, a series of ablation experiments was conducted with all improvements integrated into the network architecture. The results are shown in Table 2. Here, a denotes the C3k2-SCConv module, b the iAFF module, and c the MultiSEAM module.

As shown in Table 2, the baseline YOLOv11n achieves an mAP@0.5 of 0.848 with 2.58 million parameters and 6.3 GFLOPs. In experiment group A, the introduction of the C3k2-SCConv module enhanced the feature extraction capability of the model under complex backgrounds, resulting in a 0.8% increase in mAP@0.5. Group B employed the iAFF module for feature fusion, which improved small-object detection, yielding an AP@0.5 increase of 0.8% and 3.3% for the Helmet and NoHelmet classes, respectively, and a 1.4% overall mAP@0.5 improvement. Group C integrated the MultiSEAM module to enhance detection under occlusion, resulting in a 2.0% increase in mAP@0.5.

Experiments D and E combined the advantages of groups B and C, respectively, with group A, further improving the detection accuracy; mAP@0.5 increased by 1.4% and 1.9% over group A, respectively. Group F combined the improvements of groups B and C, achieving a 1.6% mAP@0.5 increase over that of group B. Finally, group G incorporated all three improvements simultaneously, resulting in a 3.8% mAP@0.5 increase compared with the baseline model.

Table 2: Ablation Experiment.

model	modules			AP@0.5			mAP@0.5	Params/M	FLOPs/G	FPS
	a	b	c	motorperson	Helmet	NoHelmet				
v11n				0.966	0.888	0.689	0.848	2.58	6.3	119
A	√			0.965	0.892	0.711	0.856	2.46	6.2	88
B		√		0.967	0.896	0.722	0.862	2.59	6.3	89
C			√	0.965	0.898	0.741	0.868	2.95	6.6	95
D	√	√		0.969	0.901	0.739	0.870	2.46	6.2	72
E	√		√	0.971	0.903	0.752	0.875	2.81	6.5	78
F		√	√	0.972	0.901	0.763	0.879	2.95	6.6	79
G	√	√	√	0.974	0.905	0.778	0.886	2.82	6.5	62
H	G+ADown			0.971	0.904	0.772	0.882	2.69	5.3	71

Considering the real-time requirements of motorcycle-helmet detection, the frames-per-second (FPS) metric was evaluated for each model during the ablation experiments. As shown in Table 2, although Group G achieved the highest mAP@0.5, its FPS was only 62 fps. To improve the inference speed, Group H replaced some standard convolutional layers in the model with the ADown module. The results of Group H demonstrate that, while the mAP@0.5 decreased slightly by 0.4% compared with Group G, the parameter count and computational cost were reduced by 4.6% and 15.9%, respectively, and the FPS improved by 14.5%.

3.5 Visualization and Analysis of Detection Results

To compare the detection performance of the improved algorithm with that of the baseline YOLOv11n model, both models were used to perform inference on the test set. Representative images were selected from the prediction results for visual analysis (Figure 8). Subfigure (A) shows the original image, (B) shows the detection results of YOLOv11n, and (C) presents the results of the improved model. Dark blue bounding boxes indicate the ‘motorperson’ class, light blue represents ‘Helmet’, and white represents ‘NoHelmet.’

ISSN 2616-5775 Vol. 8, Issue 9: 17-25, DOI: 10.25236/AJCIS.2025.080903

Figure 8: Visualization of Detection Performance Before and After Model Improvements.

From column (a), it can be observed that YOLOv11n fails to detect motorcycles in the distant background, which is attributed to its insufficient capability in detecting small objects. In contrast, the improved model successfully and accurately identified these distant targets. Columns (b) and (d) show that YOLOv11n exhibits both missed detections and false positives under nighttime and foggy conditions, indicating its vulnerability in complex environments. However, the improved model maintained a robust performance under these challenging scenarios. In column (c), the rear passenger without a helmet is occluded by the driver, causing YOLOv11n to miss the detection. The improved model correctly detected the occluded target, validating its superior detection capability under occlusion.

In summary, the proposed model effectively addresses the missed and false detection issues encountered by the baseline model when handling small or occluded objects and maintains a reliable performance under complex environmental conditions.

4. Conclusion

To meet the demand for accurate motorcycle-helmet detection under various complex environmental conditions, we introduced the following improvements to the YOLOv11n model: first, parts of the original C3k2 modules were replaced with C3k2-SCConv modules to enhance the model's feature extraction capability, enabling better adaptation to diverse scenes. Second, an iAFF module was employed for feature fusion, improving the model's ability to detect small objects. Third, the MultiSEAM module was integrated into the neck of the network to address the challenges caused by occlusion and target overlap, which often result in missed detections. Finally, several standard convolutional layers were replaced with ADown modules to reduce the parameter count and computational cost while increasing the detection speed, enabling the model to meet the requirements of edge deployment and real-time detection.

Experimental results show that the proposed model outperforms existing mainstream object detection models in terms of detection accuracy. Notably, the improved model achieves this performance with fewer parameters and lower computational complexity than most counterparts, making it particularly advantageous under limited computational resources and increasing its potential for deployment on-edge devices. A comparison of the detection results before and after the improvements in different environments further validated the superiority of the proposed model.

References

[1] Qin L., Zheng W., Ning P., et al. Awareness and implementation effectiveness of China's "One Helmet, One Belt" safety campaign among the public[J]. Chinese Journal of Public Health, 2023, 39(09): 1197-1200.

[2] Ross Girshick. Fast R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision. Santiago: NJ: IEEE, 2015:1440-1448.

[3] Wei Liu, Dragomir Anguelov, Dumitru Erhan, et al. SSD: Singleshot multibox detector[C]// Proceedings of European Conference on Computer Vision.Cham. Amsterdam: Springer, 2016: 21-37.

[4] Mingxing Tan, Ruoming Pang, Quoc V Le. Efficientdt:Scalable and efficient object detection[C]//

ISSN 2616-5775 Vol. 8, Issue 9: 17-25, DOI: 10.25236/AJCIS.2025.080903

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: NJ: IEEE, 2020: 10781-10790.
[5] Xie P., Cui J., Zhao M. A helmet-wearing detection algorithm for electric bicycle riders based on an improved YOLOv5[J]. Computer Science,2023,50(S1):420-425.
[6] Yuan Y., Tang, W. Helmet-wearing detection for electric bicycle riders based on an improved YOLOv8s model [J]. Journal of Hubei Minzu University (Natural Science Edition), 2024, 42(03): 355-367+367.(in Chinese)
[7] Yang J., Hu P., Dai, J. Helmet-wearing detection algorithm for electric bicycle riders based on YOLOv8-scG neural network [J/OL]. Journal of Chongqing Technology and Business University (Natural Science Edition),2024-07-15.
[8] Zhou S., Peng Z., Zhang H., et al. Helmet-YOLO: A high-precision road safety helmet detection algorithm [J]. Computer Engineering and Applications,2025,61(2):135-144.
[9] Zhou X., Wang K., Zhou X., et al. An improved YOLOv10n-based helmet-wearing detection algorithm for electric bicycles[J]. Electronic Measurement Technology,2025,48(05):40-49.
[10] Khanam R, Hussain M. Yolov11: An Overview of the Key Architectural Enhancements[EB/OL]. https://www.arxiv.org/abs/2410.17725. 2024-10-23.
[11] Jiafeng Li, Ying Wen, Lianghua He.SCConv: Spatial and Channel Reconstruction Convolution for Feature Redundancy[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: NJ: IEEE, 2023:6153-6162.
[12] Yimian Dai, Fabian Gieseke, Stefan Oehmcke,et al.Attentional Feature Fusion[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2021: 3560-3569.
[13] Ziping Yu, Hongbo Huang, Weijun Chen, et al.YOLO-FaceV2: A scale and occlusion aware face detector[J].Pattern Recognition, 2024,vol 155:110714.
[14] Wang CY, Yeh IH, Liao HYM. YOLOv9: Learning what you want to learn using programmable gradient information[C]//Proceedings of the 18th European Conference on ComputerVision. Milan: Springer, 2024. 1-21.
[15] Hanhe Lin, Jeremiah D. Deng, Deike Albers, et al.Helmet Use Detection of Tracked Motorcycles Using CNN-Based Multi-Task Learning[J].IEEE Access,2020,vol 8:162073-162084.
[16] Shuo Wang, David C. Anastasiu, Zheng Tang, et al. The 8th AI City Challenge[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: NJ: IEEE, 2024: 7261-7272.