

DANGEROUS DRIVING BEHAVIOR DETECTION IN-CAR BASED ON IMPROVED YOLOV11

SHUO WANG¹, NGHIA THI MAI², MD ABDUS SAMAD KAMAL¹
IWANORI MURAKAMI¹ AND KOU YAMADA¹

¹Graduate School of Science and Technology
Gunma University

1-5-1 Tenjincho, Kiryu 376-8515, Japan
wgsu2025@gmail.com; { maskamal; murakami; yamada }@gunma-u.ac.jp

²Faculty of Electronics Engineering 1
Posts and Telecommunications Institute of Technology
122 Hoang Quoc Viet Road, Cau Giay District, Hanoi 122300, Vietnam
nghiamt@ptit.edu.vn

Received November 2025; revised February 2026

ABSTRACT. *In recent years, intelligent driving technologies have developed rapidly, leading to an increasing demand for driving safety. However, existing algorithms generally suffer from high computational complexity and slow inference speed. To enhance the real-time detection capability of dangerous driving behaviors inside vehicles, we propose a novel algorithm based on YOLOv11, termed YOLOv11-SGCB. The proposed model integrates Ghost Convolution for lightweight feature extraction, SCConv for spatial-channel reconstruction, and CBAM for attention-guided feature refinement. Experimental results demonstrate that YOLOv11-SGCB achieves a mean Average Precision (mAP) of 93.8%, representing a 1.9% improvement over the original YOLOv11, while inference speed is increased by 23%, significantly enhancing real-time performance. Moreover, the GFLOPs are reduced to 69.1% of the original model, enabling more efficient deployment on resource-constrained mobile devices.*

Keywords: Artificial intelligence, Deep learning, Image recognition, Safe driving, YOLOv11

1. Introduction. Road traffic accidents cause a high number of deaths and injuries around the world, creating a serious global problem. More than one million people die in traffic accidents every year [1]. In 2018, there are approximately 1.35 million deaths caused by road traffic accidents. Road traffic accidents are the eighth leading cause of death worldwide, accounting for about 2.37% of the 56.9 million deaths that occur each year [2]. The factors that contribute to road traffic crashes generally fall into three categories: 1) environmental, 2) vehicle-related, and 3) human. Among them, human factors are by far the most significant. Most traffic accidents are caused by drivers engaging in activities that are not related to driving [3]. Obviously, dangerous driving behavior is one of the main causes of traffic accidents and puts both drivers and others at serious risk.

Therefore, in the field of intelligent driving, conducting research on real-time monitoring and leveraging deep learning technologies to detect dangerous driving behaviors is of great significance. These researches aim to enhance driving safety and reduce the incidence of traffic accidents. In practice, to maintain road safety, transportation authorities have deployed substantial resources – such as assigning personnel to complex traffic areas or installing surveillance cameras – to monitor driver attentiveness. However, external

monitoring methods have inherent limitations. These include high operational costs and the presence of blind spots, which hinder real-time effectiveness and fail to address the problem at its core. As a result, developing systems capable of monitoring drivers' behavior from within the vehicle in real time is crucial. Such approaches can reduce accident risk, ensure driving safety, and contribute to the overall reduction of traffic accidents [4, 5, 6, 7].

With in-depth research on artificial intelligence theory and deep learning technology, object detection has become an important research direction in the field of computer vision. It has achieved many significant developments in recent years and has been widely applied in intelligent monitoring systems, autonomous driving, industrial production, quality inspection, and other areas. Object detection is a testing model in deep learning. The object detection model can recognize objects in images, locate them and identify them [8, 9, 10, 11, 12, 13]. Related researches mainly use deep learning algorithms to identify various behaviors of people in-car as a basis for determining whether or not there are safety hazards. With the rapid development of deep learning technology, significant achievements have been made in mature applications such as intelligent driving, providing new ideas for solving the problem of real-time monitoring of driver behavior. By using deep learning algorithms to identify various behaviors of passengers in vehicles, it is possible to determine whether or not there are safety hazards [14, 15, 16, 17]. Recently, Lu et al. proposed the model which is called Deformable and Dilated Faster R-CNN (DD-RCNN), which utilizes deformable convolution and attention mechanisms to achieve accurate behavior recognition in situations, where drivers are distracted [18]. Ritheesh et al. proposed a YOLO-based module for driver fatigue detection to reduce the incidence of accidents caused by fatigued drivers [19]. Li et al. enhanced it with the addition of Spatial and Channel Reconstruction Convolution (SCConv) to efficiently capture multiscale features based on spatial-channel information [20]. Wang et al. proposed an enhanced safety detection algorithm in-car based on the YOLOv11 architecture [21], in which lightweight GhostConv modules were incorporated to reduce computational cost and improve real-time detection efficiency. Although previous studies have made significant contributions to the detection of dangerous driving behaviors, several limitations remain. First, many existing models are computationally intensive and lack sufficient real-time performance, making them unsuitable for in-car deployment. In practical applications, lightweight, efficient, and accurate algorithms are essential. Therefore, this study focuses on reducing computational complexity and improving real-time detection through model lightweighting. Second, traditional feature extraction mechanisms show poor adaptability to complex environments, such as varying lighting conditions, occlusions, and different camera angles, leading to limited generalization capability. To address this issue, we expanded the dataset by incorporating diverse weather conditions and dangerous driving behaviors, thereby enhancing the model's adaptability to real-world scenarios. To overcome the above challenges, this study proposes an improved detection algorithm based on YOLOv11, which integrates a lightweight attention-enhanced convolutional module. The proposed architecture maintains high detection speed while achieving an effective balance between multi-feature fusion and recognition accuracy, improving the overall performance of in-car real-time monitoring systems.

In order to address these issues, this study reconstructs a dedicated experimental dataset for dangerous driving behavior detection and proposes an enhanced detection algorithm (YOLOv11-SGCB). To improve real-time performance, lightweight Ghost Convolution (GhostConv) is integrated to reduce redundancy and accelerate feature extraction. Furthermore, SCConv is introduced to strengthen spatial-channel feature representation,

while the Convolutional Block Attention Module (CBAM) is incorporated to refine important features through sequential channel and spatial attention. Together, these improvements lead to a more efficient and robust detection model. To further enhance the model's performance, we modify the placement of the GhostConv module within the network architecture. Subsequently, we adapt SConv to replace certain conventional convolutional layers, thereby improving feature representation and detection accuracy. In addition, we modify the position of CBAM to enhance the model's ability to identify occluded objects and optimize its position in the network. Following these modifications, we conduct a series of experiments to evaluate their effectiveness.

This paper is organized as follows. In Section 2, we introduce the structure and characteristics of YOLOv11, and propose methods to improve its performance. In Section 3, we conduct experiments and analyze the results. Section 4 summarizes this paper and makes plans for future work.

2. Improvement of YOLOv11.

2.1. Instructions for YOLOv11. Target detection algorithms [22, 23, 24] are usually divided into two-stage. One is target detection and the other is single-stage target detection. Representatives of two-stage target detection algorithms include Fast Region-based Convolutional Neural Networks (R-CNN) [25], Faster R-CNN [26], and so on. The two-stage method usually has higher accuracy, but due to the need for two-stage calculations, it is slow and then cannot meet real-time detection scenarios. Compared with two-stage target detection, single-stage target detection simplifies the entire detection process and directly solves the target detection task as an end-to-end problem, greatly improving the detection speed. YOLO, the representative of the single-stage target detection algorithm, predicts the position and category of the object through regression, and can directly output the detection result after a single detection [27]. YOLOv11 is a version of Ultralytics YOLO real-time object detector series, redefining what is possible with state-of-the-art accuracy, speed, and efficiency. Building on the impressive evolution of its previous version YOLOs, YOLOv11 makes significant improvements to its architecture and training methodology, making it a versatile choice for a wide range of computer vision tasks. Figure 1 shows a schematic diagram of the network algorithm of the YOLOv11 algorithm. The network consists of three main parts: the backbone for feature extraction, the neck for multi-scale feature fusion, and the decoupled detection head for classification and localization. The backbone adopts a lightweight C2f-based structure with SPPF to enhance feature representation, while the FPN-PAN neck strengthens information flow between different feature scales. Finally, the decoupled head predicts object categories and bounding boxes. The core architecture of YOLOv11 primarily consists of three innovative modules: C3K2 block, the SPPF module, and the C2PSA block. Here, C3K2 is an enhanced version based on the Cross Stage Partial (CSP) block, it optimizes complex feature extraction through different kernel sizes and channel separation, and improves the model's adaptability to objects of different scales. SPPF is an optimized version of the Spatial Pyramid Pooling (SPP) module, which enhances small object detection capabilities through multiscale feature fusion, especially suitable for complex scenes. C2PSA combines channel and spatial information, introduces a multi-head attention mechanism, optimizes feature maps, and improves the detection accuracy of occluded objects. These improvements significantly enhance model performance while maintaining high inference speed.

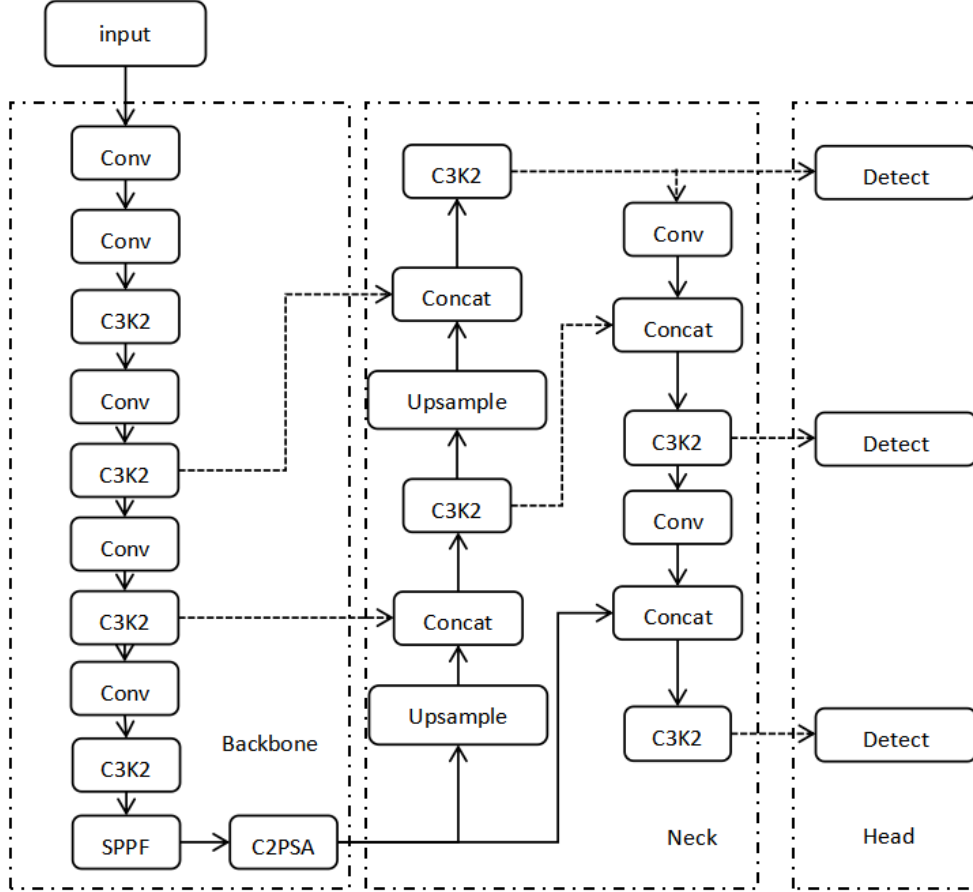


FIGURE 1. YOLOv11 model architecture diagram

2.2. GhostConv. To enhance the real-time performance of the proposed detection model, the GhostConv [28] was introduced to reduce computational complexity while maintaining feature representation capability. Traditional convolution operations involve high computational costs, which limit their deployment in real-time in-car environments. GhostConv generates additional feature maps through inexpensive linear operations instead of standard convolutions, significantly reducing the number of parameters and Floating-Point Operations (FLOPs). By integrating this module into the YOLOv11 architecture, the model achieves faster inference speed and improved efficiency without compromising detection accuracy. GhostConv is a convolution operation method aimed at solving the problems of excessive parameter count and computational complexity in traditional convolutional neural networks, especially suitable for resource limited devices. An illustration of the convolutional layer and the Ghost module is shown in Figure 2. All GhostConv layers first apply standard convolutions with a kernel size of 1×1 , stride = 1, and groups = 1, while keeping the output feature map size unchanged. Subsequently, the resulting feature maps are processed sequentially by cheap linear convolutions (depth-wise convolutions) with a kernel size of 5×5 , stride = 1, and groups = 7, using the SiLU activation function. A channel shuffle operation is then employed to generate “ghost” features. The final (10th) layer is the SPPF module, which performs pooling operations with four kernel sizes (1×1 , 2×2 , 3×3 , and 6×6) on the input feature maps. The SPPF layer captures objects at multiple scales while preserving high-level semantic information. This multi-scale pooling strategy enables accurate understanding and recognition of targets with varying sizes, where Φ represents the cheap operation. Conventional convolution to

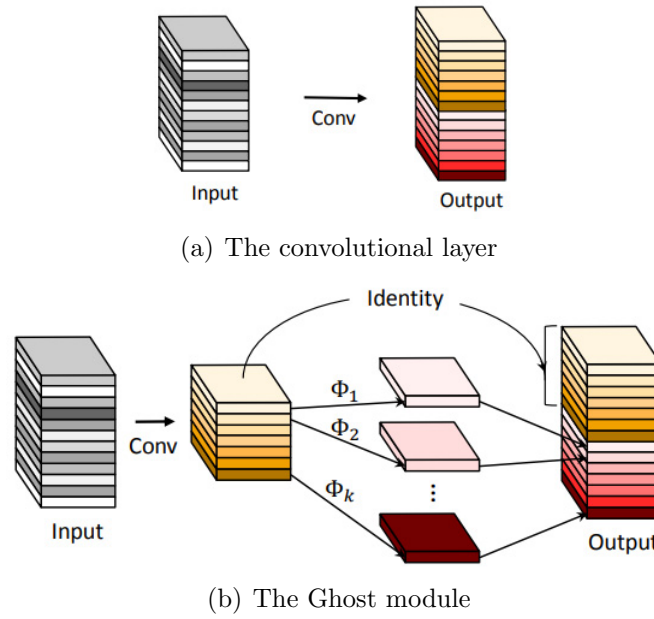


FIGURE 2. An illustration of the convolutional layer and the Ghost module

get intrinsic feature maps is used. Then we adopt the feature map of each channel to generate the Ghost feature map using the operation. Finally, the intrinsic feature map obtained in the first step and the Ghost feature map obtained in the second step are concatenated (identity connection) to obtain the final result Output. GhostConv allows the model to simultaneously consider both virtual and actual inputs with zero padding, thereby enhancing the robustness of the model and its robustness to feature distributions. The Ghost module generates more feature maps through efficient operations. Based on a set of intrinsic feature maps, we apply a series of low-cost linear transformations to generating many ghost feature maps that can fully reveal the information behind the intrinsic features. These operations are computationally efficient and can generate a large number of feature maps without increasing computational complexity. The intrinsic feature map and the generated ghost feature map are connected to obtain the final output feature map, which not only retains the key information of the original feature map, but also increases the diversity of features through the ghost feature map.

2.3. Spatial and channel reconstruction convolution. To compensate for the slight loss in accuracy caused by the lightweight design, the SConv module was introduced. SConv enhances the model's feature discrimination capability by reconstructing spatial and channel information, enabling more effective extraction of key features. By integrating SConv into the YOLOv11 algorithm, the proposed model achieves improved detection accuracy while maintaining real-time performance in in-car environments.

SConv is a structure designed to improve the ability to express characteristics of Convolutional Neural Networks (CNNs). By introducing reconstruction mechanisms in two dimensions, space and channels, it enhances the receptive field and adaptability of features, thereby improving the performance of the model in tasks such as object detection and image classification [29].

The SConv module is designed to improve the efficiency and effectiveness of convolutional neural networks by introducing two core components: the Spatial Reconstruction Unit (SRU) and the Channel Reconstruction Unit (CRU). The SRU enhances spatial feature representations, while the CRU focuses on refining channel-wise information. By

simultaneously addressing spatial and channel redundancies, SCConv enables the generation of more compact and informative features. The outputs from the SRU and CRU are integrated into the original residual connection and passed to the subsequent convolutional block. SCConv can be directly used as a substitute for standard convolution layers, significantly reducing computational costs while maintaining – or even enhancing – overall model performance. This architecture enables CNNs to produce more discriminative features by explicitly leveraging richer contextual information. Notably, the self-calibrated design of SCConv is simple and general-purpose, allowing it to augment standard convolutional layers without introducing additional parameters or computational complexity. The architecture of SCConv is integrated with SRU and CRU. SCConv illustrated in Figure 3 is composed of two sequentially arranged components: SRU and CRU. Specifically, given the intermediate input features X within a bottleneck residual block, we first apply the SRU to obtaining spatially refined features X^w . These are then processed by the CRU to produce the channel-refined output Y . The SCConv module effectively leverages both spatial and channel-wise redundancies in the feature maps. It can be seamlessly integrated into existing CNN architectures to reduce redundancy in intermediate representations and enhance overall feature expressiveness.

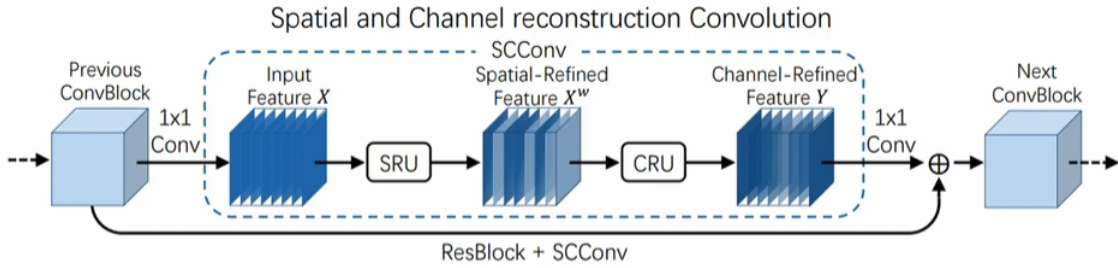


FIGURE 3. The exact position of our SCConv module within a ResBlock

2.4. CBAM. To further enhance feature representation, the Convolutional Block Attention Module (CBAM) [30] was integrated into the proposed model. CBAM sequentially applies channel attention and spatial attention to refining feature maps, allowing the network to focus on the most informative regions and suppress irrelevant background noise. By adaptively weighting both channel-wise and spatial information, CBAM strengthens the model’s ability to capture key features related to dangerous driving behaviors. This integration effectively improves detection accuracy without significantly increasing computational complexity, making it suitable for real-time in-car applications. CBAM is a lightweight attention mechanism module designed to enhance the feature representation capability of CNNs. It effectively focuses on the target area by jointly modeling channel attention and spatial attention, thereby improving the performance of the network. The overview of CBAM is summarized in Figure 4. Here, the Channel Attention Module

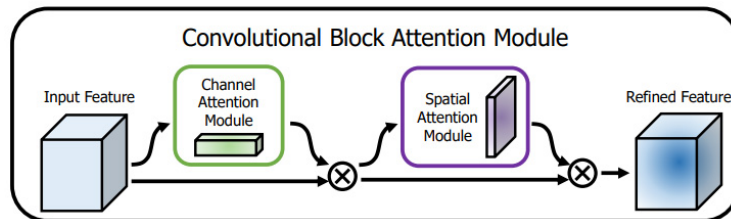


FIGURE 4. The overview of CBAM

(CAM) generates channel attention maps by exploiting the inter-channel relationships among features. Due to each channel of the feature map being treated as a feature detector, channel attention will be focused on important parts. The structure of CAM is shown in Figure 5. Spatial Attention Module (SAM) generates a spatial attention map by utilizing the inter-spatial relationship of features. Different from the channel attention, SAM focuses on informative parts. A spatial attention map is generated by leveraging the inter-spatial relationships of features. Unlike channel attention, which emphasizes what is important, spatial attention focuses on where the informative regions are, making it complementary to channel attention. Applying pooling operations along the channel axis is effective in highlighting these informative spatial regions. The structure of SAM is shown in Figure 6.

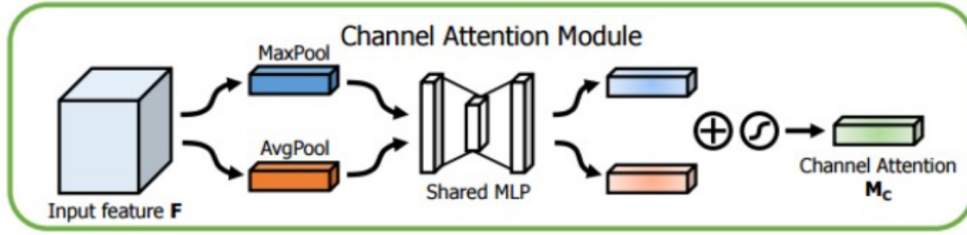


FIGURE 5. The structure of CAM

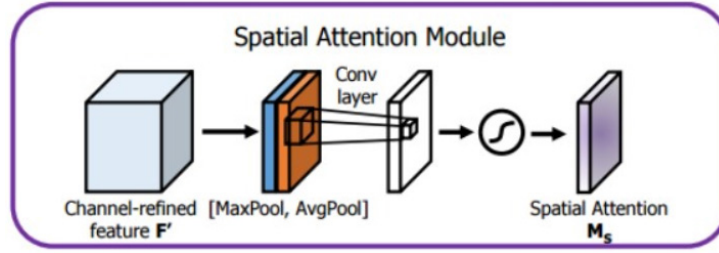


FIGURE 6. The structure of SAM

2.5. Adaptive threshold focal ATFL loss. To improve the model's detection performance on hard samples and address the imbalance between positive and negative samples, the Adaptive Threshold Focal Loss (ATFL) is employed. A new adaptive threshold focal ATFL loss function decouples the target and the background, and utilizes the adaptive mechanism to adjust the loss weight to force the model to allocate more attention to target features [31]. This builds upon the Focal Loss (FL) by introducing a dynamic threshold mechanism to better focus on hard samples during training.

FL is designed to handle class imbalance. The formula for FL is written by

$$\text{FL}(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t), \quad (1)$$

where p_t denotes the predicted probability for the true class, α_t is a balancing factor to weigh positive and negative samples differently and $(1 - p_t)^\gamma$ is a modulation term that emphasizes hard samples ($\gamma > 0$) by reducing the loss contribution from well-classified examples.

ATFL extends focal loss by introducing a dynamic threshold that adapts during training. This mechanism provides a more flexible way to differentiate between easy and hard samples as the model evolves. The general form of ATFL can be expressed as

$$\text{ATFL}(p_t, \tau) = -\alpha_t(1 - \hat{p}_t)^\gamma \log(p_t), \quad (2)$$

where $\hat{p}_t = \max(0, p_t - \tau)$ is the adjusted probability, with τ being the dynamic threshold and τ is a threshold that is adjusted as training progresses, typically determined using a distribution of predictions or statistical measurements.

3. Experiments and Results.

3.1. Experimental environment. The experimental environment used in this paper is summarized as follows: Hardware configuration is that the operating system is Windows 10, 64-bit, and CPU is AMD Ryzen 5 5600X. We use NVIDIA GeForce RTX 3060 as the GPU uses. The learning algorithm uses PyTorch 2.5.1, the programming language is Python-3.12.8 and cuda-12.4. The entire training process is divided into 400 epochs, and the batch size is 32.

3.2. Dataset. The dataset used in this article has 6,000 photos of driving behavior in-car, of which 5,000 are from the Internet [32] and 1,000 are self-made datasets, which are from vehicle monitoring. Among them, 4,000 are used as training sets, 1,000 as verification sets, and 1,000 as test sets. Image is shown in Figure 7.



FIGURE 7. Image acquisition

3.3. Evaluation indicators. When evaluating YOLO object detection models, several standard metrics are used to assess both accuracy and efficiency. Precision (P) measures the ratio of correctly predicted positives, Recall (R) reflects the proportion of actual positives correctly detected, and mean Average Precision (mAP) provides an overall accuracy measure. In terms of efficiency, GFLOPs indicate computational complexity, while FPS represents real-time processing capability. Together, these metrics offer a comprehensive evaluation algorithm for comparing different YOLO variants. The calculation formulas for these evaluation metrics are as follows:

$$P = \frac{TP}{TP + FP}, \quad (3)$$

$$R = \frac{TP}{TP + FN}, \quad (4)$$

$$mAP = \frac{\sum_{j=1}^m AP_j}{m} \quad (5)$$

and

$$FPS = \frac{N_{\text{frames}}}{T_{\text{total}}}. \quad (6)$$

Here,

- 1) True Positive (TP): the number of samples correctly detected as targets.
- 2) False Negative (FN): the number of real targets that were not detected.

- 3) False Positive (FP): refers to the number of non-target samples that are incorrectly predicted as targets by the model.
- 4) Average Precision (AP): represents the area under the Precision-Recall (P-R) curve for a single class, reflecting the model's detection accuracy for that specific category.
- 5) GFLOPs indicate the computational complexity of a model or algorithm, whereas the number of parameters reflects the model's size. In general, lower parameter counts and GFLOPs correspond to reduced computational demands, making deployment on end devices easier and lowering hardware requirements.
- 6) Frames Per Second (FPS): measures how many frames the model can process each second. The larger the FPS value, the faster the model detection speed and the better the real-time performance. N_{frames} is total number of frames processed; T_{total} is total time (seconds).

3.4. Proposed method (YOLOv11-SGCB). The proposed YOLOv11-SGCB model replaces the conventional convolutional modules in the backbone with GhostConv modules to significantly reduce the model's parameters and overall computational burden. Meanwhile, the neck network retains traditional convolutional structures to enhance multi-scale feature fusion and maintain robust feature representation. This hybrid design strategy effectively balances accuracy and efficiency, enabling the model to sustain high detection performance while improving computational efficiency and practical deployability. As a result, the model becomes more lightweight and well-suited for deployment on resource-constrained platforms, such as edge devices and embedded systems. In addition to the lightweight GhostConv modules, the YOLOv11-SGCB model integrates the SCConv and the Convolutional Block Attention Module (CBAM) to further enhance feature extraction and representation. The SCConv module reconstructs spatial and channel information, thereby improving the network's ability to capture discriminative features and compensating for the slight accuracy loss caused by model lightweighting. Meanwhile, the CBAM module adaptively applies channel and spatial attention, allowing the model to focus on the most informative regions while suppressing background noise. The combination of these two mechanisms significantly strengthens multi-dimensional feature learning and boosts detection accuracy without introducing substantial computational overhead, making the model more effective and robust for real-time in-car dangerous driving behavior detection.

Then the improved YOLOv11 network is summarized in Figure 8.

3.5. Result analysis. In this subsection, we present and analyze the experimental results obtained from YOLOv5s, YOLOv8n, YOLOv11, YOLOv11-GCB, and the proposed YOLOv11-SGCB. A comprehensive comparison of these models is provided in Table 1.

Based on the experimental results presented in Table 1, the YOLOv11-SGCB achieves the best overall performance among all evaluated models. Compared with the original YOLOv11, YOLOv11-SGCB shows clear improvements across multiple metrics: Precision increases from 93.4% to 95.3%, Recall improves from 92.2% to 92.4%, and mAP rises from 92.9% to 93.8%. These gains demonstrate that the integration of GhostConv, SCConv, and CBAM significantly enhances feature representation while maintaining a lightweight architecture. Furthermore, relative to YOLOv11-GCB, the YOLOv11-SGCB achieves additional improvements, with Precision increasing from 94.1% to 95.3% and mAP increasing from 93.6% to 93.8%, confirming the effectiveness of introducing SCConv for spatial-channel reconstruction. Although YOLOv11-SGCB exhibits a slightly lower FPS than YOLOv11-GCB, it still maintains high real-time performance and far exceeds traditional models such as Faster R-CNN.

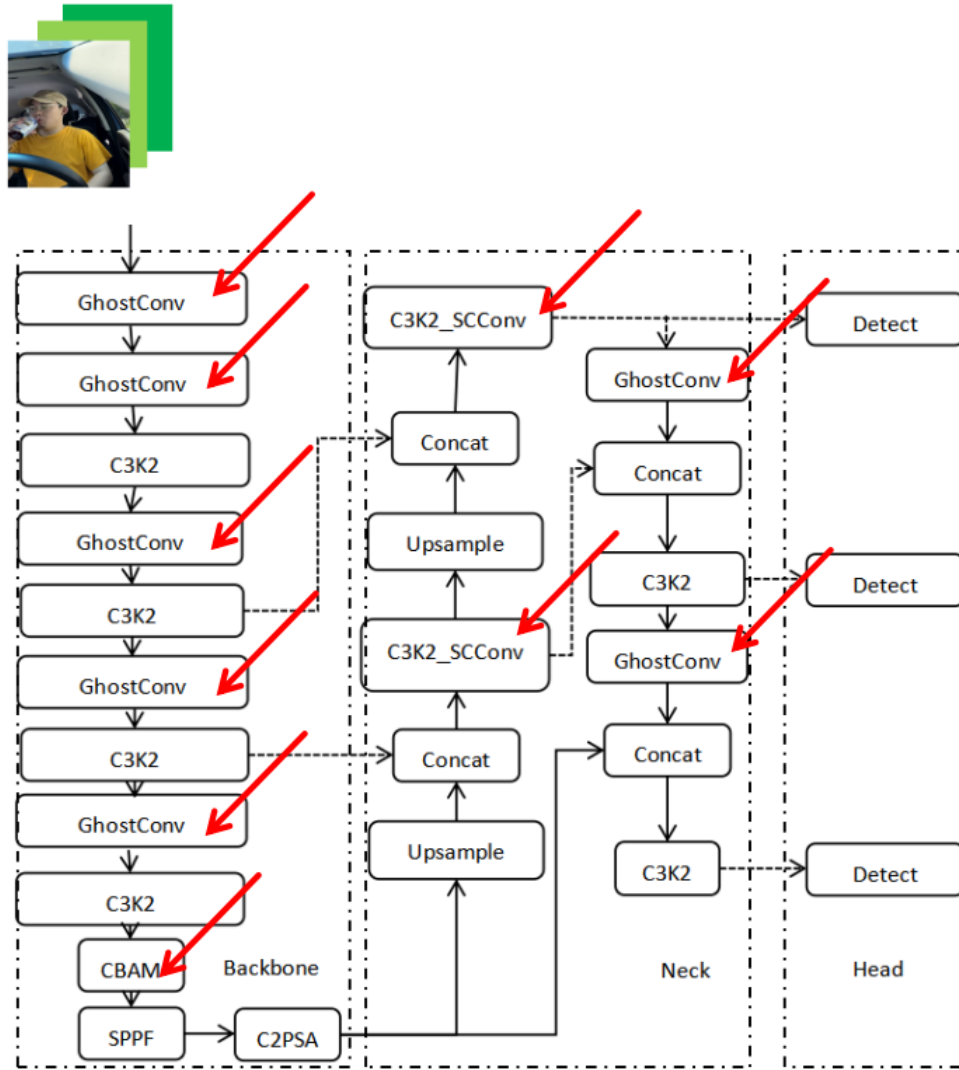


FIGURE 8. The improved YOLOv11 network (YOLOv11-SGCB)

TABLE 1. Comparison results of the proposed model and state-of-the-art models

Algorithm	GhostConv	CBAM	SCConv	P (%)	R (%)	mAP (%)	GFLOPs (MB)	Speed (FPS)
Faster R-CNN	×	×	×	85.9	91.8	89.2	256	21
YOLOv5s	×	×	×	90.5	88.4	90.4	23.7	37
YOLOv8n	×	×	×	91.2	87.4	91.7	9.8	93
YOLOv11	×	×	×	93.4	92.2	92.9	8.1	110
YOLOv11-GCB	○	○	×	94.1	91.8	93.6	5.2	140
YOLOv11-SGCB	○	○	○	95.3	92.4	93.8	5.6	136

Overall, the results indicate that the combined use of GhostConv, SCConv, and CBAM provides a more expressive and robust feature extraction capability, enabling YOLOv11-SGCB to achieve superior detection accuracy with only minimal additional computational cost. This verifies the model's practicality and advantages for real-time dangerous driving behavior detection from Table 1.

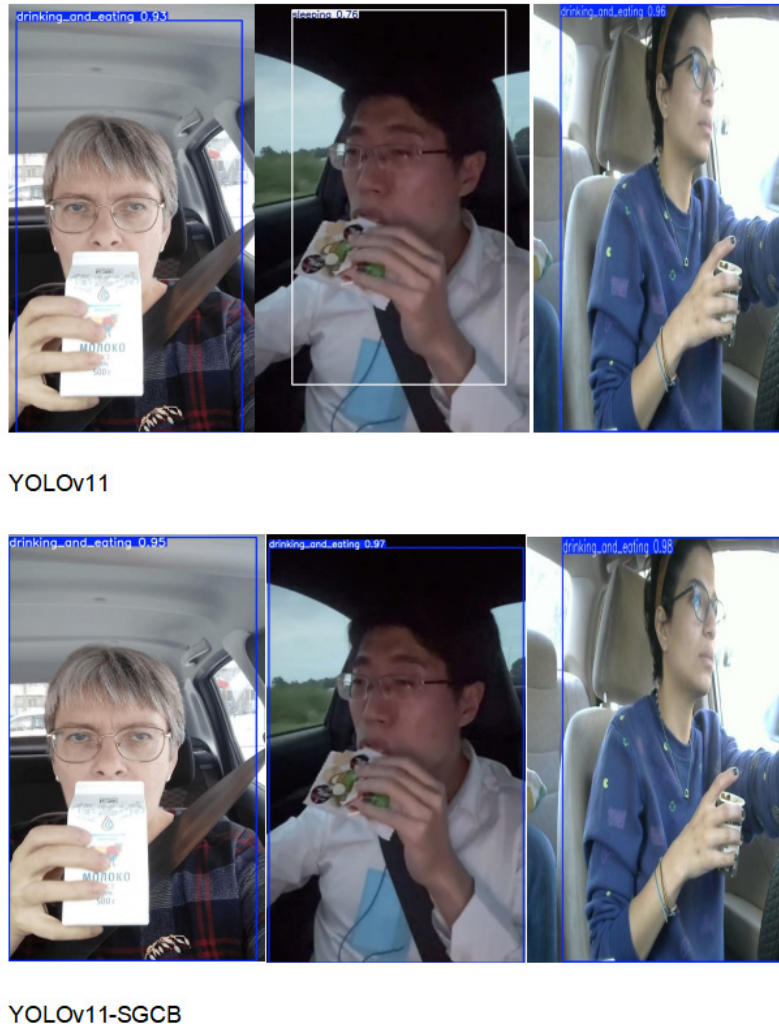


FIGURE 9. Comparison of model detection results

As shown in Figure 9, the detection accuracy of the improved YOLOv11-SGCB is slightly higher than that of YOLOv11. Figure 9 presents qualitative comparisons between the baseline YOLOv11 model and the proposed YOLOv11-SGCB under representative driving scenarios. As shown in the first row, YOLOv11 exhibits several issues, including inaccurate bounding box localization and misclassification, particularly in cases involving drinking and eating behaviors. In contrast, the proposed YOLOv11-SGCB model produces more precise bounding boxes and higher confidence scores, while correcting several erroneous detections observed in the baseline results.

The improved performance can be attributed to the enhanced feature representation capability introduced by the GhostConv, SCConv, and CBAM modules, which enable the model to better focus on salient regions related to dangerous behaviors. Overall, the qualitative results demonstrate that YOLOv11-SGCB achieves more robust and reliable detection performance, especially in challenging in-car scenarios.

4. Conclusion. In this study, we proposed YOLOv11-SGCB, an improved detection model for dangerous behaviors in-car. By integrating GhostConv, SCConv, and CBAM, and employing the ATFL loss, the model achieves enhanced feature extraction, higher detection accuracy, and faster inference compared with YOLOv11 and YOLOv11-GCB. Experimental results demonstrate its superior performance and robustness. Future work

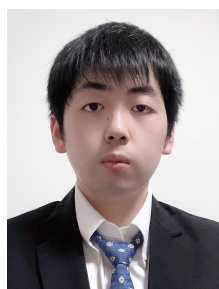
will focus on expanding the dataset with more diverse driving scenarios, incorporating additional dangerous behavior categories, and optimizing the model for real-world deployment, aiming to further improve accuracy, generalization, and real-time applicability.

REFERENCES

- [1] World Health Organization, Road traffic injuries, *WHO Newsroom – Fact Sheets*, <https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries>, 2023.
- [2] M. S. AL-Quraishi, S. S. Azhar Ali, M. AL-Qurishi, T. B. Tang and S. Elferik, Technologies for detecting and monitoring drivers' states: A systematic review, *Heliyon*, vol.10, no.20, 2024.
- [3] E. T. Petridou and M. Moustaki, Human factors in the causation of road traffic crashes, *European Journal of Epidemiology*, vol.16, no.9, 2000.
- [4] S. Jafarpour and V. Rahimi-Movaghar, Determinants of risky driving behavior: A narrative review, *Med. J. Islam. Repub. Iran*, 2014.
- [5] C. S. Dula and E. S. Geller, Risky, aggressive, or emotional driving: Addressing the need for consistent communication in research, *Journal of Safety Research*, vol.34, no.5, pp.559-566, 2004.
- [6] Y. Niu, Z. Li and Y. Fan, Analysis of truck drivers' unsafe driving behaviors using four machine learning methods, *International Journal of Industrial Ergonomics*, vol.86, 2021.
- [7] P. Mao, K. Zhang and D. Liang, Driver distraction behavior detection method based on deep learning, *IOP Conference Series Materials Science and Engineering*, vol.782, no.2, 2020.
- [8] L. Ali, F. Alnajjar, M. M. A. Parambil, M. I. Younes, Z. I. Abdelhalim and H. Aljassmi, Development of YOLOv5-based real-time smart monitoring system for increasing lab safety awareness in educational institutions, *Sensors*, vol.22, no.22, 2022.
- [9] X. Chen, D. L. Peng and Y. Gu, Real-time object detection for UAV images based on improved YOLOv5s, *Opto-Electron. Eng.*, vol.49, no.3, 2022.
- [10] X. Jia, Y. Tong, H. Qiao, M. Li, J. Tong and B. Liang, Fast and accurate object detector for autonomous driving based on improved YOLOv5, *Scientific Reports*, vol.13, 2023.
- [11] X. Jin, Z. Li and H. Yang, Pedestrian detection with YOLOv5 in autonomous driving scenario, *2021 5th CAA International Conference on Vehicular Control and Intelligence (CVCI)*, Tianjin, China, pp.1-5, 2021.
- [12] G. Zhang, K. K. Yau, X. Zhang and Y. Li, Traffic accidents involving fatigue driving and their extent of casualties, *Accident Anal. Prevention*, vol.87, pp.34-42, 2016.
- [13] S. Wang, N. T. Mai, K. Hashikura, M. A. S. Kamal, I. Murakami and K. Yamada, Research on detection algorithm for finding lost pets based on improved YOLOV5, *ICIC Express Letters*, vol.19, no.7, pp.757-766, 2025.
- [14] Y. He, S. Yu, S. Zhang, L. Lan and G. Li, Research on driver distracted driving behavior detection method based on target detection, *Journal of Highway and Transportation Research and Development*, vol.39, no.10, pp.153-161, DOI: 10.3969/j.issn.1002-0268.2022.10.019, 2022.
- [15] Y.-S. Poon, C. Kao, Y.-K. Wang, C.-C. Hsiao, M.-Y. Hung and Y.-C. Wang, Driver distracted behavior detection technology with YOLO-based deep learning networks, *2021 IEEE International Symposium on Product Compliance Engineering – Asia (ISPCE-ASIA)*, Taipei, Taiwan, pp.1-5, 2021.
- [16] E. Shen, Y. Hu, G. Chen, P. Lou and H. Hu, Deep convolutional neural network for real-time target detection in intelligent driving, *Journal of Automotive Safety and Energy Conservation*, vol.11, no.1, 2020.
- [17] H. Singh and A. Kathuria, Analyzing driver behavior under naturalistic driving conditions: A review, *Accident Anal. Prevention*, vol.150, 2021.
- [18] M. Lu, Y. Hu and X. Lu, Driver action recognition using deformable and dilated Faster R-CNN with optimized region proposals, *Int. J. Speech Technol.*, vol.50, no.4, pp.1100-1111, 2020.
- [19] Y. R. V. Ritheesh, S. Reddy and R. G. Rajan, Driver drowsiness detection and alert system using YOLO, *2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICES)*, Chennai, India, pp.1-6, 2022.
- [20] R. Li, C. Yu, X. Qin, X. An, J. Zhao and W. Chuai, YOLO-SGC: A dangerous driving behavior detection method with multiscale spatial-channel feature aggregation, *IEEE Sensors Journal*, vol.24, no.21, pp.36044-36056, 2024.
- [21] S. Wang, N. T. Mai, M. A. S. Kamal, I. Murakami and K. Yamada, Detection algorithm for dangerous driving behavior detection in the car based on YOLOv11, *ICIC Express Letters*, vol.20, no.6, pp.603-611, 2026.

- [22] W. Li, X. Feng, K. Zha, S. Li and H. Zhu, Summary of target detection algorithms, *Journal of Physics: Conference Series*, 2021.
- [23] D. Li, K. D. Wong, Y. H. Hu and A. M. Sayeed, Detection, classification, and tracking of targets, *IEEE Signal Processing Magazine*, vol.19, no.2, pp.17-29, 2002.
- [24] C. L. Philip Chen, H. Li, Y. Wei, T. Xia and Y. Y. Tang, A local contrast method for small infrared target detection, *IEEE Transactions on Geoscience and Remote Sensing*, vol.52, no.1, pp.574-581, 2014.
- [25] R. Girshick, Fast R-CNN, *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, pp.1440-1448, 2015.
- [26] S. Ren, K. He, R. Girshick and J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.39, no.6, pp.1137-1149, 2016.
- [27] W. Zhan, C. Sun, M. Wang, J. She, Y. Zhang, Z. Zhang and Y. Sun, An improved YOLOv5 real-time detection method for small objects captured by UAV, *Soft Comput.*, vol.26, pp.361-373, 2022.
- [28] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu and C. Xu, GhostNet: More features from cheap operations, *IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp.1580-1589, 2020.
- [29] J. Li, Y. Wen and L. He, SCConv: Spatial and channel reconstruction convolution for feature redundancy, *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, pp.6153-6162, 2023.
- [30] S. Woo, J. Park, J. Y. Lee and I. S. Kweon, CBAM: Convolutional Block Attention Module, *Proc. of the European Conference on Computer Vision (ECCV)*, Munich, Germany, pp.3-19, 2018.
- [31] B. Yang, X. Zhang, J. Zhang, J. Luo, M. Zhou and Y. Pi, EFLNet: Enhancing feature learning network for infrared small target detection, *IEEE Transactions on Geoscience and Remote Sensing*, vol.62, pp.1-11, 2024.
- [32] hshshahs22334, *[Dataset] Dangerous Driving Behavior Dataset 6000 + Images*, CSDN, https://blog.csdn.net/hshshahs22334/article/details/149157375?fromshare=blogdetail&sharetype=blogdetail&sharerId=149157375&sharerefer=PC&sharesource=hshshahs22334&sharefrom=from_link, 2025.

Author Biography



Shuo Wang received a B.E degree from Dalian Polytechnic University, China, in 2018 and a master degree from Gunma University, Japan, in 2023. He is currently pursuing his research as a Ph.D. at Gunma University. His research interests include visual recognition and automatic control.



Nghia Thi Mai received the B.S., M.S. and Dr. Eng. degrees from Gunma University, Japan, in 2009, 2011 and 2014, respectively. From 2014 to 2015, she was with the Human Resources Cultivation Center, Gunma University, as a research associate. From 2015 to 2021, she worked on research on damping control for automobiles at Exedy Co., Ltd. Since 2022, she has been working as a lecturer at the Faculty of Electronics Engineering 1, Posts and Telecommunications Institute of Technology (PTIT), Vietnam. In addition, she is currently working as a visiting associate professor and part-time lecturer at the Department of Electronics and Mechanical Engineering, Gunma University. Her research interest includes Smith predictor, internal model control and robotics.



Md Abdus Samad Kamal received the B.Sc. degree in Electrical and Electronic Engineering from Khulna University of Engineering and Technology (KUET), Bangladesh, in 1997; Master and Doctor degrees from Kyushu University from Graduate School of Information Science and Electrical Engineering, Japan, in 2003 and 2006, respectively. He was a post-doctoral fellow in Kyushu University till November 2006. He is currently an associate professor at Graduate School of Science and Technology, Gunma University, Japan. His current research interests are reinforcement learning, intelligent transportation systems and multiagent systems. He is a member of IEEE and SICE.



Iwanori Murakami received the B.E., M.E. and Dr. Eng. degrees from Gunma University, Japan, in 1992, 1994 and 1997, respectively. He is currently an associate professor at Graduate School of Science and Technology, Gunma University. His research interests include control problems in the mechanical fields and robotics.



Kou Yamada received his B.S. and M.S. degrees in Electrical and Information Engineering from Yamagata University, Japan, in 1987 and 1989, respectively; and the Dr. Eng. Degree from Osaka University, Japan, in 1997. He is currently a full-time professor at Graduate School of Science and Technology, Gunma University, Japan. His research interests include robust control, repetitive control, process control and control theory for inverse systems and infinite-dimensional systems. Prof. Yamada received the 2005 Yokoyama Award in Science and Technology, the 2005 Electrical Engineering/Electronics, Computer, Telecommunication and Information Technology International Conference (ECTI-CON2005) Best Paper Award, the Japanese Ergonomics Society Encouragement Award for Academic Paper in 2007, the 2008 Electrical Engineering/Electronics, Computer, Telecommunication and Information Technology International Conference (ECTI-CON2008) Best Paper Award and the 4th International Conference on Innovative Computing, Information and Control Best Paper Award in 2009, the 14th International Conference on Innovative Computing, Information and Control Best Paper Award in 2019, Outstanding Achievement Award from Kanto Branch of Japanese Society for Engineering Education in 2022 and JSME (The Japan Society of Mechanical Engineers) Education Award in 2023. He is a member of IEEE and SICE and a fellow of JSME.