

Article

Research on Distracted and Fatigue-Related Driving Behavior Detection Based on YOLOv12-LAD

Xiyao Liu ^{1,2}, Zhiwei Guan ^{1,3}, Qiang Chen ^{1,4,*} and Yi Ren ^{1,5}

¹ School of Automotive and Transportation, Tianjin University of Technology and Education, Tianjin 300222, China; 0311221005@tute.edu.cn (X.L.); guanzhiwei@tsguas.edu.cn (Z.G.); ry@hzpt.edu.cn (Y.R.)

² Department of Automotive and Transportation, Baotou Vocational & Technical College, Baotou 014030, China

³ School of Automobile and Rail Transportation, Tianjin Sino-German University of Applied Sciences, Tianjin 300350, China

⁴ Qinghai Provincial Key Laboratory of Plateau Vehicle Electrification and Intelligentization Technology, Xining 810003, China

⁵ School of Automotive Engineering, Hangzhou Polytechnic, Hangzhou 311402, China

* Correspondence: chen@tute.edu.cn

Abstract

Distracted and fatigue-related driving behaviors are major causes of road traffic accidents, creating an urgent need for reliable driver monitoring systems. Vision-based detection methods have garnered widespread attention due to their low cost of deployment and practical applicability. However, existing lightweight models often suffer from limited global contextual perception and insufficient preservation of fine details. Motivated by these challenges, this study introduces an improved distracted and fatigue-related driving behavior detection model, YOLOv12-LAD, built on the YOLOv12 architecture. The proposed framework integrates a Large Separable Kernel Attention module (LSKA) to enhance global contextual perception, an Adaptive Downsampling module (ADown) to mitigate information loss during feature compression, and a Dynamic Sampling module (DySample) to enable content-adaptive feature reconstruction and improve multi-scale behavior representation. Experimental results show that YOLOv12-LAD achieved 97.5% precision, 96.3% recall, and 98.4% mAP@50 with only 2.5 million parameters, 6.2 GFLOPs, and an inference speed of 249 FPS. Ablation studies, comparisons with representative models, cross-dataset evaluation, and real-vehicle tests further verify the effectiveness and robustness of the proposed method. The proposed method demonstrates strong performance while maintaining computational efficiency, making it suitable for real-time vision-based driver monitoring applications.

Keywords: distracted driving detection; fatigue-related driving behavior detection; vision-based driver monitoring; YOLOv12; object detection



Academic Editor: Silvia Liberata Ullo

Received: 31 March 2026

Revised: 21 April 2026

Accepted: 22 April 2026

Published: 26 April 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Road traffic accidents remain a major public safety issue worldwide. Data from the World Health Organization (WHO) indicate that nearly 1.19 million fatalities occur annually due to traffic crashes, and an additional 20–50 million people sustain non-fatal injuries, many of which result in permanent disability [1]. Road traffic accidents not only result in substantial loss of life but also impose considerable economic burdens on national development, as related losses account for approximately 1–3% of gross domestic

product (GDP) and can reach up to 6% in some countries [1]. These statistics clearly indicate that strengthening road safety interventions and reducing the incidence of traffic accidents are urgent priorities worldwide. With the ongoing development of intelligent transportation and advanced driver assistance systems (ADAS), driver safety increasingly depends on reliable in-vehicle perception and monitoring systems, making accurate driver-state detection essential to modern vehicle safety.

Previous studies have indicated that driver distraction is a major contributor to road traffic accidents. The issue has drawn growing attention over the last several years, mainly due to the expanding adoption of mobile devices and integrated cabin entertainment technologies. Consequently, distracted driving has become an important research topic in both academic research and engineering applications [2]. Driver distraction refers to a condition in which the driver's attention shifts away from driving and toward other activities unrelated to vehicle operation, thereby impairing driving perception and control capabilities [3]. This condition can readily lead to traffic accidents, such as collisions and rollovers, and may cause fatal consequences in severe cases [4]. Notably, fatigue driving is often regarded as a covert form of driver distraction associated with physiological attentional impairment and may pose an even greater safety risk. Under such conditions, drivers may still operate the vehicle physically, while their attention is no longer focused on the driving task. Therefore, effective fatigue detection requires accurate recognition of subtle local behavioral features, particularly those related to the eyes, mouth, face, and head posture.

Existing approaches for driver distraction and fatigue detection are generally grouped into four types: methods utilizing driver physiological signals, vehicle motion information, visual data, and multimodal fusion data [5]. Among these approaches, physiological-signal-based methods typically require specialized wearable devices to capture signals that reflect changes in driver attention. Early studies focused on wearable devices for monitoring driver galvanic skin response, blood oxygen saturation, and electroencephalogram (EEG) signals. Subsequently, to improve user comfort, Fan et al. [6] employed frontal EEG signals and developed an ensemble learning framework based on time-series data for detecting distracted and fatigue-related driving behaviors. In addition, Mewborne et al. [7] tracked drivers' hand positions using Bluetooth signals and achieved an accuracy of more than 95% without relying on visual data. However, such methods are more suitable for privacy-sensitive scenarios, and wearable devices may interfere with normal driving behavior. Moreover, their high cost limits large-scale deployment and practical application.

Detection methods based on vehicle motion data rely on information obtained from on-board sensors. Sun et al. [8] integrated multiple parameters, including steering wheel angle and yaw rate, used wavelet packet decomposition for feature extraction, and employed a Bi-LSTM network to recognize driver states. By analyzing inter-vehicle distance fluctuations and car-following behavior, Ucar et al. [9,10] identified distracted driving and achieved an accuracy of approximately 65%. Although the method itself is relatively simple, it still demonstrates promising potential for real-world deployment. Nakano and Chakraborty [11] collected vehicle control data using a driving simulator, including steering wheel angle and braking frequency, to construct personalized driving models. They employed convolutional neural networks (CNNs) for distracted driving detection. However, this approach can only infer driver states indirectly.

Detection methods based on driving image information primarily employ visual sensors, including cameras and infrared devices, to capture drivers' behaviors and identify distraction and fatigue-related states by analyzing facial features, head poses, and behavioral actions. Among these approaches, facial-feature-based methods use in-vehicle cameras to capture drivers' facial images and extract dynamic features from regions such

as the eyes and mouth for distracted driving detection. Zhu et al. [12] integrated YOLOv3 with Kalman filtering to improve the stability of facial detection. However, relying solely on facial expression detection still carries the risk of being inconsistent with actual behavioral states. Head-based detection methods primarily infer gaze shifts by estimating head rotation angles. Ansari et al. [13] introduced modified linear units into a Bi-LSTM to process time-series data of head angular acceleration, thereby identifying complex head movements. However, this method requires training on a large dataset, and data annotation also poses certain challenges. In addition, building high-quality datasets involves not only annotation costs but also challenges such as class imbalance. To address data imbalance, Li et al. [14] presented the SelectAug method, which selectively augments specific data to enhance feature learning for minority classes. This effectively mitigates the impact of imbalanced data distributions and improves the robustness of visual detection models. In contrast, behavior-based detection approaches focus more directly on observable driver actions. These methods detect distraction by identifying physical actions unrelated to driving tasks, such as using a mobile phone or drinking water. Wagner et al. [15] developed a deep learning-based classification framework capable of effectively recognizing driver distraction activities, such as mobile phone use and eating while driving.

Multimodal fusion-based detection methods integrate visual, physiological, and vehicular data to construct a comprehensive perception framework. Zhou et al. [16] presented a multimodal detection approach that combined physiological signals with eye-tracking information to identify both cognitive and visual distraction in conditionally automated driving environments. Furthermore, Vijeikis et al. [17] introduced a multitask fusion architecture that used multi-view visual inputs to analyze driver attention levels. Smartphone-based monitoring approaches have also attracted increasing attention. For example, Papatheocharous [18] proposed a portable, cost-effective smartphone-based distraction detection system. Although multimodal detection methods can enhance detection capability and reliability, adding additional sensors increases system complexity, thereby limiting practical deployment.

Overall, these studies indicate that driver monitoring is evolving from simple behavior recognition toward a more comprehensive assessment of distraction and fatigue under complex driving conditions. Recent research has further highlighted that practical driver monitoring systems are moving toward the tighter integration of on-board perception, driver state assessment, and real-time safety support in intelligent vehicle environments [17–19]. In particular, fatigue detection, as an important branch of driver monitoring, places greater demands on the recognition of subtle local cues, especially those related to the eyes, mouth, and head posture [5,12,13,20]. This trend underscores the need for detection frameworks that can preserve fine-grained behavioral details while maintaining robust performance in real-world in-vehicle environments.

Currently, most studies still utilize visual information from driving scenarios, as this approach provides a more direct representation of driver behavior at a lower deployment cost. Early studies predominantly adopted classical CNN architectures, such as VGGNet, ResNet, and MobileNet. For instance, Tran et al. [21] used the VGG-16 network and achieved 96.7% accuracy with a dual-camera setup. However, the computational speed of this method was limited to only 8 frames per second (FPS). In recent years, the YOLO family of detectors has gained widespread attention due to its effective trade-off between detection efficiency and accuracy. A lightweight detection framework, StarDL-YOLO, was developed by Shen et al. [22] based on the YOLOv8 architecture. This framework adopts structural optimization techniques to reduce redundant parameters while maintaining strong detection capability. Similarly, Li et al. [23] presented the YOLOv8s-FPNE model, which maintains strong detection accuracy while controlling computational overhead

through the combination of efficient feature learning and localized convolution strategies. In addition, Du et al. [24] developed a lightweight YOLO-LBS framework to enhance robustness in complex in-vehicle environments. These studies collectively indicate that improving architectural efficiency while preserving discriminative capability has become a central focus of research in driver behavior detection.

To improve detection accuracy, Zhang et al. [25] incorporated an attention mechanism into CenterNet to enhance detection capability for small-scale behaviors. In addition, Gong et al. [26] combined OpenPose-based key point extraction with an improved weighted KNN classifier and achieved a recognition rate of 94.04% on the State Farm dataset. Zhao et al. [27] proposed a cross-layer feature aggregation strategy for multi-scale feature learning and achieved a classification accuracy of 95.56% on the State Farm dataset. However, most of these improvements are built on medium- or small-kernel convolutions and local attention mechanisms, which inherently provide only a limited receptive field. Consequently, modeling long-range dependencies in global contextual information remains challenging. As a result, such models may exhibit insufficient discriminative ability in scenarios that require holistic posture perception, such as when drivers turn to communicate with rear-seat passengers.

In real-world deployment, model lightweighting and real-time performance optimization have emerged as key concerns. Researchers have adopted techniques such as depthwise separable convolution, channel pruning, and neural architecture search to reduce model complexity and meet the computational constraints of in-vehicle edge devices. Sahoo et al. [28] achieved high classification accuracy on a Raspberry Pi using SqueezeNet. Similarly, Elshamy et al. [29] proposed the P-YOLOv8 model with a size of only 2.84 MB, demonstrating strong deployment potential. In addition, Salakapuri et al. [30] proposed an integrated deep learning framework combining CNN and YOLO architectures, thereby enabling simultaneous detection of driver distraction and road objects to support Advanced Driver Assistance Systems (ADAS). These studies further indicate that lightweight design and real-time deployment have become an important research direction for practical driver monitoring systems. However, existing lightweight strategies still tend to compromise feature representation quality, mainly because traditional downsampling operations, such as pooling or strided convolutions, often cause the loss of spatial details. This makes it difficult to accurately recognize fine-grained driving actions. Moreover, distracted and fatigue-related driving behaviors exhibit multi-scale patterns, ranging from large-scale posture changes to subtle finger movements. This places higher demands on multi-scale feature representation. Therefore, achieving an effective balance between computational efficiency, global contextual perception, and fine-grained feature representation remains a key challenge in real-world deployment. In addition to computational constraints, data privacy and security are also important considerations in practical deployment. Shang et al. [31] proposed the FedBiKD framework, which improves accuracy and stability while alleviating data heterogeneity under privacy-preserving conditions.

Nevertheless, current driver monitoring methods still face challenges in jointly capturing global contextual information, preserving subtle local behavioral cues, and recognizing multi-scale behavior patterns under computational constraints. To address these issues, an enhanced detection framework, YOLOv12-LAD, based on the YOLOv12 architecture, is proposed. This framework addresses the limitations of existing distracted and fatigue-related driving detection methods, particularly those related to global contextual perception and fine-grained feature retention. It incorporates the Large Separable Kernel Attention (LSKA), Adaptive Downsampling (ADown), and Dynamic Sampling (DySample) modules to enhance feature representation under these constraints. Experimental evaluation shows that

YOLOv12-LAD maintains strong detection accuracy while preserving high computational efficiency. The main contributions of this study are summarized as follows:

- (1) A YOLOv12-based detection framework, YOLOv12-LAD, is developed for distracted and fatigue-related driving detection to improve the balance between detection accuracy and computational efficiency.
- (2) The LSKA, ADown, and DySample modules are integrated into the YOLOv12 architecture to enhance global contextual perception, reduce information loss during down-sampling, and improve adaptive feature reconstruction, respectively, thereby strengthening fine-grained detail preservation and multi-scale behavior representation.
- (3) Extensive experiments, including ablation studies, comparisons with representative models, cross-dataset evaluation, and real-vehicle tests, are conducted to comprehensively validate the effectiveness, robustness, and practical applicability of the proposed framework.

2. Methods

2.1. Architecture of YOLOv12

YOLOv12, introduced in 2025, adopts a single-stage detection pipeline and achieves efficient inference performance. The framework mainly includes three components: the backbone, neck, and detection head, as illustrated in Figure 1.

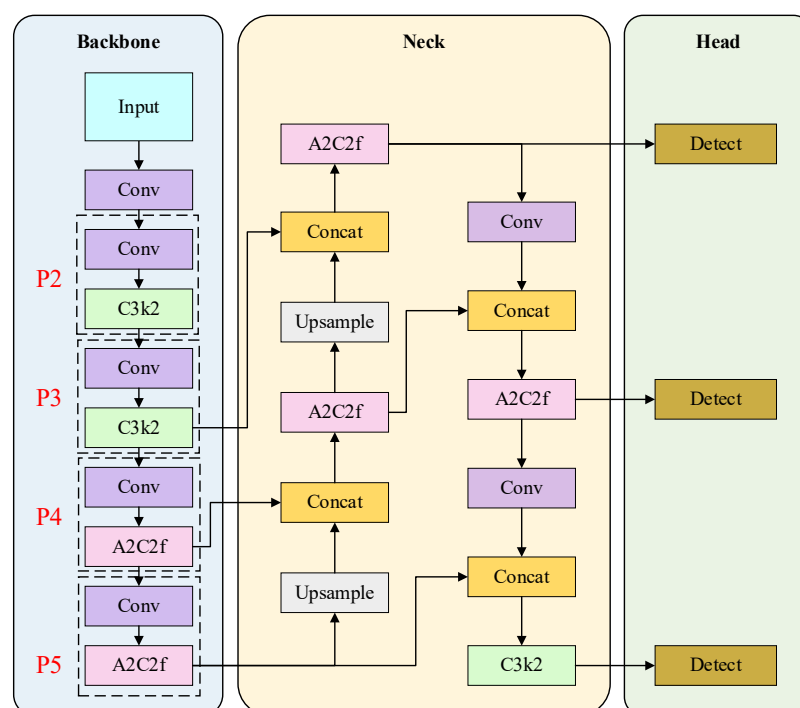


Figure 1. Network architecture of the baseline YOLOv12 model.

YOLOv12 is a real-time detection framework that incorporates attention mechanisms. Its performance gains mainly stem from the A2 module (Area Attention), the R-ELAN (Residual Efficient Layer Aggregation Network) structure, and several architectural refinements.

The A2 module incorporates the Flash Attention mechanism to reduce memory access overhead during attention computation, thereby enabling richer contextual modeling while maintaining low inference latency. In addition, improved cross-scale feature interaction further enhances detection performance for objects of different sizes.

YOLOv12 adopts the R-ELAN architecture, an improved version of ELAN [32]. Compared with ELAN, R-ELAN introduces residual shortcut connections that add the input directly to the output with a default scaling coefficient of 0.01. This design helps stabilize gradient propagation while maintaining computational efficiency. R-ELAN also uses 7×7 large-kernel separable convolutions to enhance spatial perception without significantly increasing computational cost. Compared with RT-DETR, YOLOv10, and YOLOv11, YOLOv12 provides a better balance between detection accuracy and inference efficiency, particularly in resource-limited environments. Its region-based attention mechanism also improves object recognition in heterogeneous backgrounds [33], which is beneficial in complex in-vehicle driving scenarios.

2.2. Architecture of the Proposed YOLOv12-LAD Model

Figure 2 illustrates the architecture of the proposed YOLOv12-LAD model for distracted and fatigue-related driving behavior detection. Although YOLOv12 provides a strong real-time detection baseline, distracted and fatigue-related behavior driving recognition still requires further improvement in global contextual modeling, the preservation of subtle local cues, and the adaptive representation of behavior patterns across different scales. First, the LSKA module is introduced to broaden the receptive field and strengthen global contextual perception through long-range dependency modeling. Second, the AD-own module is incorporated to alleviate information loss during downsampling while reducing computational overhead, thereby preserving fine-grained feature details during feature compression. Finally, the lightweight dynamic sampling module DySample is employed to improve feature representation across different scales, particularly for small-object detection. Together, these three modules jointly enhance global contextual perception, fine-grained feature preservation, and multi-scale behavior recognition, all of which are critical to accurate driver monitoring in complex in-vehicle scenarios.

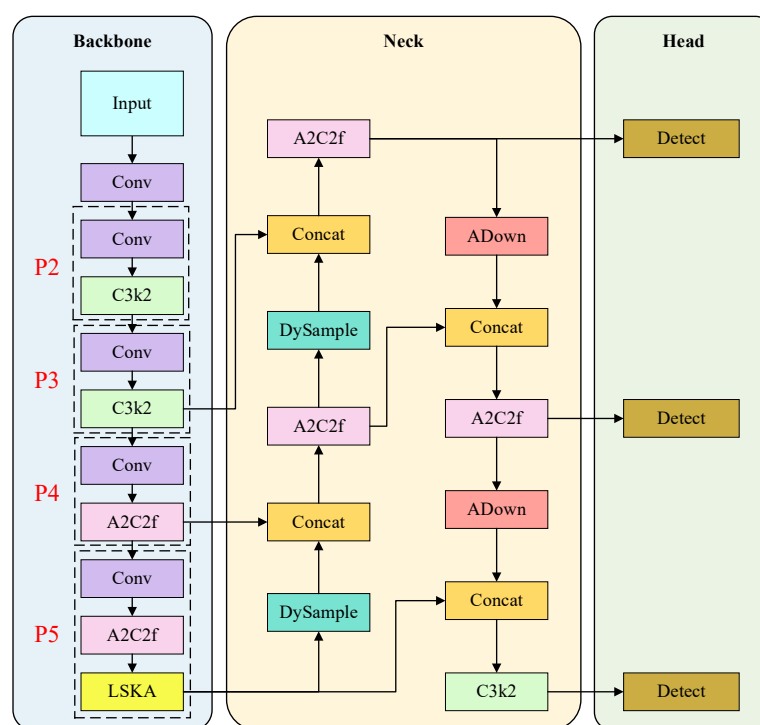


Figure 2. Network architecture of the proposed YOLOv12-LAD model.

2.2.1. LSKA Module

Effective object detection requires sufficient receptive field coverage and robust modeling of long-range spatial dependencies. While larger convolution kernels, such as 7×7 or 9×9 , are effective for expanding the receptive field, they introduce significantly higher computational overhead due to the quadratic growth of computational complexity with respect to kernel size. This increases the parameter count and memory usage, making efficient deployment more difficult in deep networks. The YOLOv12-LAD backbone, therefore, incorporates the LSKA mechanism [34] to enlarge the receptive field while maintaining computational efficiency. The structure of the LSKA module is shown in Figure 3.

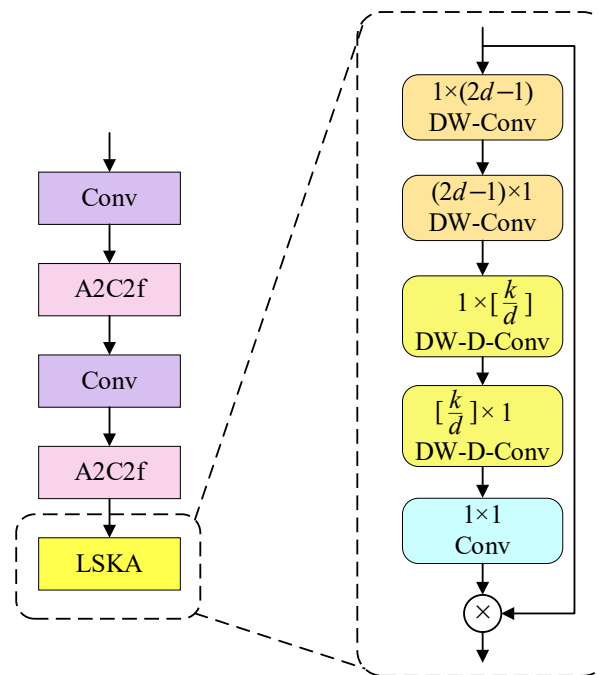


Figure 3. Structural architecture of the LSKA module.

The LSKA module is based on the Large Kernel Attention (LKA) mechanism [35]. One-dimensional convolution kernels are first applied in a spatially separable manner to produce intermediate feature maps, which are then used to generate attention maps. The resulting attention maps are fused with the original feature maps to enhance feature representation. Compared with conventional two-dimensional convolutions, cascaded one-dimensional kernels allow the network to capture spatial information over a larger receptive field at lower computational cost. This design enhances the modeling of long-range spatial dependencies while maintaining effective feature extraction. For driver monitoring tasks, this enables the network to better capture long-range spatial dependencies among the head, hands, and upper body, thereby strengthening the global perception of driver posture and overall behavioral context.

The LSKA mechanism can be formulated as follows:

$$\bar{Z}^C = \sum_{H,W} W_{(2d-1) \times 1}^C * \left(\sum_{H,W} W_{1 \times (2d-1)}^C * F^C \right) \quad (1)$$

$$Z^C = \sum_{H,W} W_{\lfloor \frac{k}{d} \rfloor \times 1}^C * \left(\sum_{H,W} W_{1 \times \lfloor \frac{k}{d} \rfloor}^C * \bar{Z}^C \right) \quad (2)$$

$$A^C = W_{1 \times 1} * Z^C \quad (3)$$

$$\bar{F}^C = A^C \odot F^C \quad (4)$$

In the above equations, Z^C denotes the depth-wise convolution output; H and W represent the spatial dimensions of the feature map; “ $*$ ” denotes the convolution operator; F^C denotes the original input feature map; k indicates the kernel size; A^C denotes the attention feature map; $\lfloor \cdot \rfloor$ denotes the floor operation; and d denotes the dilation rate; $\odot$ represents the Hadamard product; C indexes the channel dimension; and $\bar{F}^C$ denotes the enhanced output feature map.

2.2.2. ADown Module

To mitigate the information loss inherent in conventional downsampling methods, ADown [36] is introduced as an adaptive downsampling module. ADown adopts an adaptive downsampling strategy that integrates the complementary advantages of average pooling and max pooling and applies differentiated processing across feature channels to preserve informative structural details. As shown in Figure 4, the ADown module employs a multi-branch downsampling architecture that integrates smooth structural information with locally salient features.

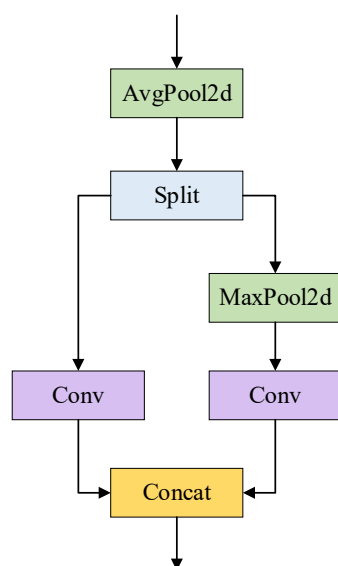


Figure 4. Structural architecture of the ADown module.

Specifically, given an input tensor $X \in \mathbb{R}^{B \times C \times H \times W}$, average pooling first reduces the spatial resolution while preserving the overall structural characteristics of the feature representation. Here, B refers to the batch size, C denotes the channel dimension, and H and W represent the spatial dimensions of the feature map. This operation is formulated in Equation (5):

$$X_{\text{avg}} = \text{AvgPool2d}(X) \quad (5)$$

where X_{avg} denotes the output after average pooling. Subsequently, X_{avg} is evenly split along the channel dimension, as shown in Equation (6):

$$X_1, X_2 = \text{Split}(X_{\text{avg}}, 2) \quad (6)$$

Here, the function $\text{Split}(X_{\text{avg}}, 2)$ separates the feature map into two tensors along the channel axis, X_1 and X_2 , where X_1 and $X_2 \in \mathbb{R}^{B \times \frac{C}{2} \times H' \times W'}$. This channel partition reduces computational complexity. A dual-branch processing strategy is then applied. For X_1 , a

3×3 convolution is applied to extract relatively smooth features and suppress background noise. The computation is given in Equation (7):

$$X_{1'} = \text{Conv}(X_1) \quad (7)$$

For the other branch X_2 , max pooling first enhances edge and texture features, followed by a 1×1 convolution to further extract salient features related to distracted driving behaviors, as shown in Equation (8):

$$X_{2'} = \text{Conv}(\text{MaxPool2d}(X_2)) \quad (8)$$

Finally, the outputs of the two branches are concatenated along the channel dimension to generate a more informative feature representation:

$$X_{\text{out}} = \text{Concat}(X_{1'}, X_{2'}) \quad (9)$$

where $X_{\text{out}} \in \mathbb{R}^{B \times C \times H' \times W'}$ represents the resulting feature map produced by the ADown module. Through this design, computational cost is reduced, and inference efficiency is improved while maintaining effective feature representations. This design allows the model to preserve critical fine-grained details, such as subtle hand movements, facial cues, and behavior-related object features, even after feature compression, thereby improving fine-grained feature retention in distracted driving recognition.

2.2.3. DySample Module

In YOLOv12-based distracted driving image analysis, traditional nearest-neighbor interpolation has several limitations. During image resizing, pixel values are determined from only a single neighboring pixel, without considering information from surrounding regions. This may produce aliasing artifacts, such as jagged edges or mosaic effects, which can blur object boundaries and reduce detection accuracy. Furthermore, conventional interpolation methods lack adaptive mechanisms for local feature optimization, making it difficult to retain fine-grained visual details in complex scenarios. To overcome this issue, the framework incorporates the DySample module [37]. DySample is a lightweight dynamic sampling module based on a content-aware mechanism. By directly learning the coordinates of sampling points from the input feature maps, DySample enables adaptive sampling. The core procedure comprises three stages that enable the model to restore spatial details while maintaining computational efficiency.

In the first stage, the DySample module uses a sampling offset generator to estimate adaptive sampling coordinates. Specifically, a linear projection layer predicts offset values for each spatial position in the input feature map. As shown in Figure 5, the DySample module uses the `grid_sample` operation to sample features from the input feature map X according to the generated sampling coordinates. The resulting output feature map X' has a size of $C \times sH \times sW$, and this operation can be written as:

$$X' = \text{grid_sample}(X, S) \quad (10)$$

where $S \in \mathbb{R}^{2g \times sH \times sW}$, $2g$ encodes the sampling coordinate offsets in the x- and y-directions, and s indicates the sampling scale factor.

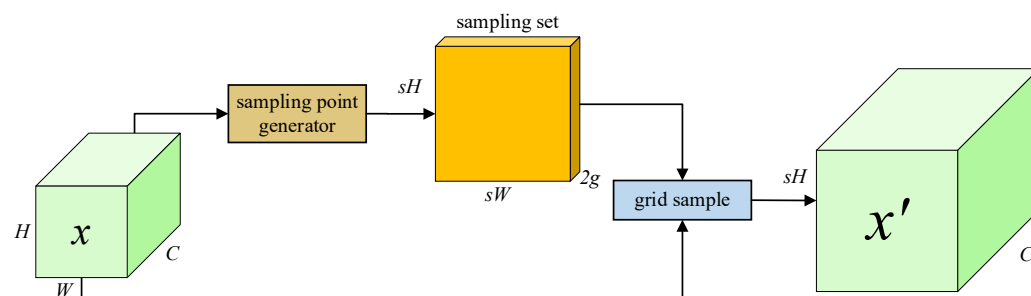


Figure 5. Architectural design of the DySample dynamic oversampling module.

In the second stage, the module incorporates a stability-enhancing offset constraint mechanism to improve training convergence during the early optimization phase and to restrict the magnitude of sampling offsets.

Specifically, DySample adopts a bilinear initialization strategy combined with a static scaling factor of 0.25 to regulate the offset distribution. As illustrated in Figure 6, the feature map X is fed into a linear projection module to estimate the offset tensor o . This tensor can be computed as follows:

$$o = 0.25 \cdot \text{Linear}(X) \quad (11)$$

where the scaling factor constrains the offset magnitude.

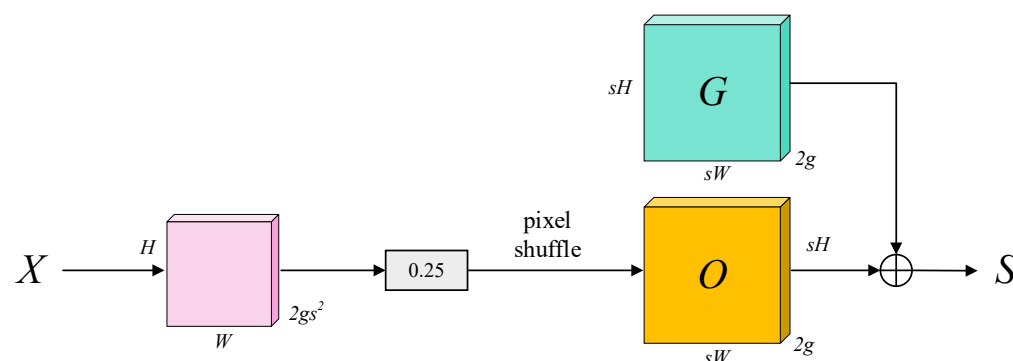


Figure 6. Sampling point generator with an introduced static sampling factor.

Subsequently, the predicted offsets are reshaped via a pixel reorganization operation into a spatial format of $sH \times sW \times 2g$. At the same time, a regular sampling grid G , corresponding to the standard sampling positions without offsets, is maintained. By adding the reshaped offset tensor to the regular grid, the final dynamic sampling set S is obtained. With this design, the sampling behavior during the early stage of training approximates bilinear interpolation. This helps preserve the spatial continuity of the feature maps while avoiding excessive dispersion of sampling points.

In the third stage, the dynamic sampling mechanism exploits its content-aware capability. Compared with nearest-neighbor interpolation, DySample adaptively adjusts sampling locations according to the semantic characteristics of the input feature map. The output pixel values are therefore determined by features in their surrounding neighborhoods. This design improves the reconstruction of small-object boundaries and helps separate target contours from background interference. By adaptively reconstructing spatial details at different resolutions, DySample enhances the recognition of both large-scale driver postures and small behavior-related objects, thereby improving multi-scale behavior recognition.

3. Experiments

3.1. Dataset

Publicly available image datasets for distracted driving behaviors are still limited. To establish a more comprehensive benchmark for unsafe driving behavior recognition, this study utilized a publicly available dataset hosted on GitHub, the FIRC Dataset [38], proposed by Fu, S and Li, S. The dataset is released under the FIRC License v1.0, which permits non-commercial academic research, citation, and the limited display of dataset samples with proper attribution.

The dataset includes common distracted and fatigue-related driving behaviors. Distracted driving behaviors are divided into ten categories: drinking, doing hair and makeup, operating the radio, reaching behind, safe driving, calling left, calling right, talking to a passenger, texting left, and texting right, as shown in Figure 7.

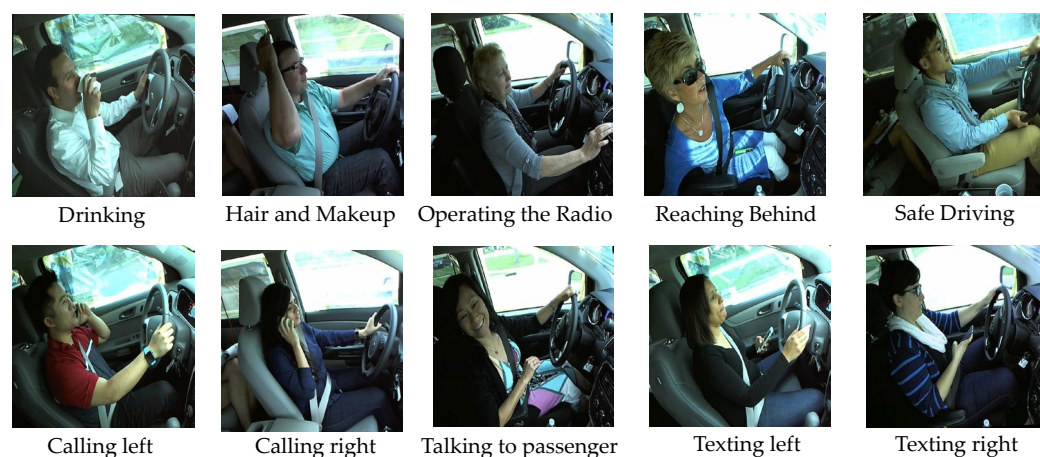


Figure 7. Distracted driving behavior categories in the FIRC Dataset.

Fatigue-related behaviors are categorized into four types: eye closure, eye opening with gaze deviating from the driving direction, drowsiness, and yawning, as shown in Figure 8.



Figure 8. Fatigue-related driving behavior categories in the FIRC Dataset.

The two sets of behaviors were merged and reorganized by the dataset authors (e.g., unifying texting left and right into a single “texting” category) to establish the twelve driving behavior categories of the FIRC Dataset, as shown in Figure 9. The images were captured using an RGB camera installed on the passenger side of the front seat, covering real in-vehicle driving scenarios. All images were annotated using the Labelling tool, with bounding boxes defined for both distracted and fatigue-related behaviors.

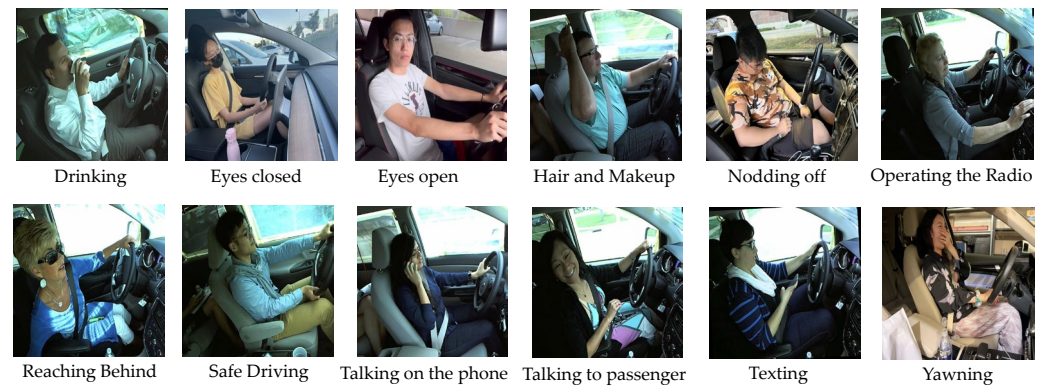


Figure 9. Combined twelve driving behavior categories used in this study.

The combined dataset consists of 22,581 images, which are divided into training, validation, and test subsets following an 8:1:1 ratio. Figure 10 illustrates the class-wise data distribution and provides detailed statistics of the annotated bounding boxes during training.

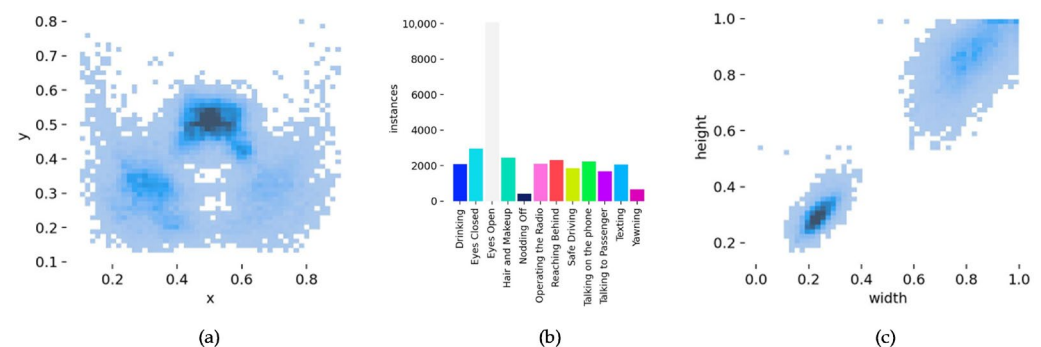


Figure 10. Statistical analysis of annotation distribution in the distracted driving dataset. (a) Spatial distribution of bounding box center coordinates. (b) Sample distribution across the twelve distracted driving behavior categories. (c) Width and height distribution of annotated bounding boxes. The different colors in (b) are used only to distinguish different behavior categories.

The eye-opening category appears in multiple behaviors, leading to a higher number of samples compared to other categories. Although this imbalance may introduce a potential risk of prediction bias, subsequent category-level performance analysis and heatmap visualization indicate that no obvious bias is observed in the current setting.

3.2. Experimental Setup and Metrics

All experiments were conducted on a workstation with Windows 11, an Intel Core i5-13600KF CPU, an NVIDIA RTX 4080 GPU, 64 GB RAM, CUDA 12.2, PyTorch 2.2.2, Python 3.9, and Ultralytics package version 8.3.108. The models were trained for 200 epochs with a batch size of 32 and an initial learning rate of 0.01. The input image size was set to 640×640 pixels. The optimizer used was SGD with a momentum of 0.9 and a weight decay of 0.0005. The learning rate was gradually reduced during training.

Online data augmentation was applied to enhance model robustness. The augmentations followed the default Ultralytics YOLO pipeline, including random horizontal flipping, scaling, translation, HSV color adjustment, Mosaic, and MixUp.

To assess detection performance, four standard metrics are used: Precision (P), Recall (R), $F1$ -score, and mean Average Precision (mAP) [39,40], computed as:

$$P = \frac{TP}{TP + FP} \quad (12)$$

$$R = \frac{TP}{TP + FN} \quad (13)$$

$$F1 = \frac{2 \times P \times R}{P + R} \quad (14)$$

$$mAP = \frac{\sum_{q=1}^Q AP(q)}{Q} \quad (15)$$

where TP , FP , and FN represent true positive, false positive, and false negative predictions, respectively. $F1$ is the harmonic mean of P and R , and mAP is averaged across all categories.

Model efficiency was evaluated using four indicators: parameter count, floating-point operations (FLOPs), model size, and inference speed (FPS, Frames Per Second). Parameters reflect network scale, FLOPs measure computational cost, model size indicates storage requirements, and FPS represents real-time inference capability.

3.3. Experimental Results and Analysis

3.3.1. Model Performance Analysis

In this study, YOLOv12 was adopted as the baseline for comparison. Experiments were conducted on the constructed distracted driving dataset using both YOLOv12 and the proposed YOLOv12-LAD. Quantitative results are summarized in Table 1.

Table 1. Performance comparison of YOLOv12 and YOLOv12-LAD on the FIRC dataset.

Model	P (%)	R (%)	F1 (%)	mAP@50 (%)	Params (M)	GFLOPs	FPS (f/s)
YOLOv12	96.2	95.6	95.9	97.7	2.6	6.5	251
YOLOv12-LAD	97.5	96.3	96.9	98.4	2.5	6.2	249

Compared to the baseline, YOLOv12-LAD improves precision, recall, F1-score, and mAP@50 by 1.3%, 0.7%, 1.0%, and 0.7%, respectively. In addition, the parameter count is reduced by 0.1 M, and computational complexity decreases by 0.3 GFLOPs. The FPS remains at 249 f/s, which is close to the baseline and still meets real-time requirements. Although the absolute gains are modest, the baseline already exhibits strong performance; under these conditions, the proposed model achieves balanced improvements across multiple metrics, demonstrating the effectiveness of the LSKA, ADown, and DySample modules.

The proposed model also demonstrates a better balance between false positive and false negative errors. Specifically, LSKA enlarges the receptive field through large-kernel convolutions, enabling more discriminative feature learning. DySample enhances multi-scale feature representation, particularly for small targets such as hands and mobile phones, thereby contributing to the observed precision improvement. Despite using large kernels, LSKA maintains computational efficiency through kernel decomposition. In addition, ADown further reduces computational complexity while preserving feature representation capability, leading to reductions of 0.3 GFLOPs and 0.1 M parameters.

Figure 11 further compares the overall detection performance of the two models using precision–recall (P–R) curves at an IoU threshold of 0.5. In general, a larger enclosed area indicates better overall performance across the precision–recall trade-off. The P–R curves show that YOLOv12-LAD achieves better overall detection performance than the baseline model.

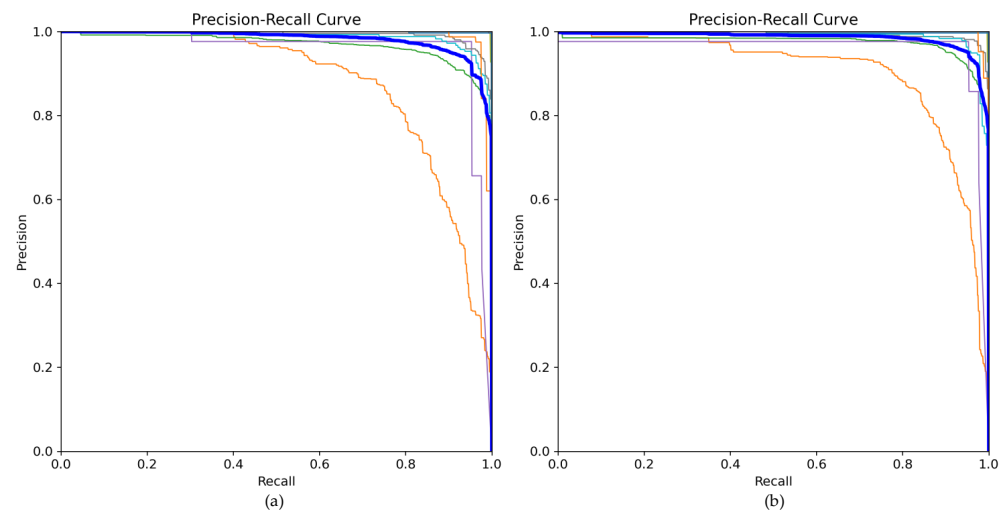


Figure 11. P–R curves of the baseline YOLOv12 and the proposed YOLOv12-LAD models. (a) Baseline YOLOv12; (b) Proposed YOLOv12-LAD. The different colored curves represent different behavior categories, and the thick blue curve represents the overall performance across all categories.

Beyond the overall performance advantage, it is also important to examine whether the larger number of ‘Eyes Open’ samples introduces category-level prediction bias. To this end, Table 2 compares the AP for each behavior category and the overall mAP@50 achieved by both the baseline and proposed models. The “Images” column denotes the number of samples in the test set for each category, provided here to facilitate category-level performance comparison.

Table 2. Per-category AP results in the test set.

Class	Images	(YOLOv12) AP (%)	(YOLOv12-LAD) AP (%)
Drinking	266	99.5	99.5
Eyes Closed	374	89.5	92.4
Eyes Open	1236	96.5	98.0
Hair and Makeup	267	99.4	99.5
Nodding Off	43	95.6	96.8
Operating the Radio	260	99.2	99.5
Reaching Behind	279	99.2	99.5
Safe Driving	266	98.9	99.0
Talking on the phone	281	99.0	99.5
Talking to Passenger	191	98.2	98.5
Texting	290	98.6	99.3
Yawning	80	98.8	99.3
All (mAP@50)	2259	97.7	98.4

Table 2 shows that, although the ‘Eyes Open’ category contains substantially more samples than other behavior categories, the AP values for the remaining categories are still competitive. The larger number of ‘Eyes Open’ samples does not introduce obvious category-level prediction bias in the current experimental setting.

Figure 12 compares feature heatmaps from YOLOv12 and YOLOv12-LAD and demonstrates that YOLOv12-LAD captures more precise and complete discriminative features.

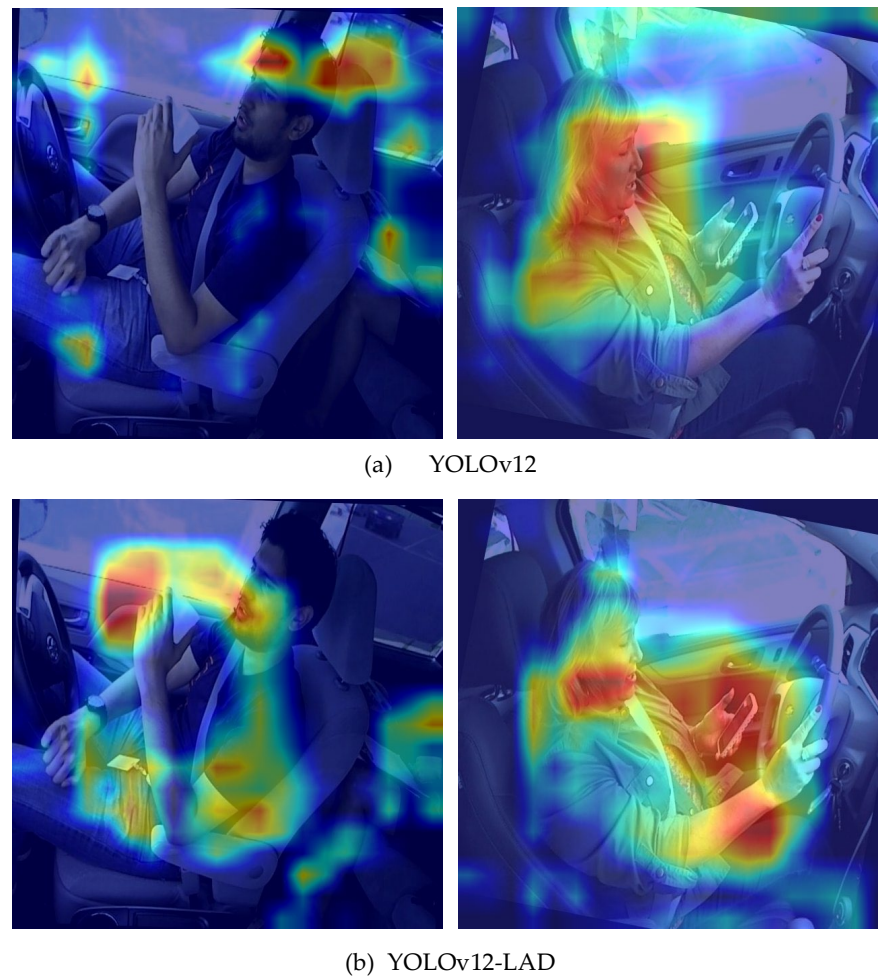


Figure 12. Comparison of feature heatmaps generated by the baseline YOLOv12 and the proposed YOLOv12-LAD models. (a) YOLOv12; (b) YOLOv12-LAD. In the heatmaps, warmer colors indicate stronger feature responses or higher attention intensity, whereas cooler colors indicate weaker responses.

The two heatmaps illustrate distracted driving behaviors, including drinking and mobile phone use. Although both behaviors involve the eyes-open state, the heatmap visualizations show that YOLOv12-LAD focuses on key objects (e.g., cups and mobile phones) and critical body regions (e.g., the hands and mouth), rather than solely focusing on the eye region. This provides visual evidence that the larger sample size in the ‘Eyes Open’ category does not result in obvious attention bias in the current setting.

3.3.2. Additional Analysis on the Effect of Class Imbalance

To further examine the effect of class imbalance, an additional experiment was conducted. The ‘Eyes Open’ samples in the training set were randomly reduced to approximately 20% of their original number. YOLOv12-LAD was retrained under the same settings. The resulting model was then evaluated using the same test set to ensure a fair comparison. Figure 13 shows a comparison of distributions of samples in the training-set before and after the adjustment. Table 3 presents the AP results for each category on the same test set under the original and adjusted training set distributions.

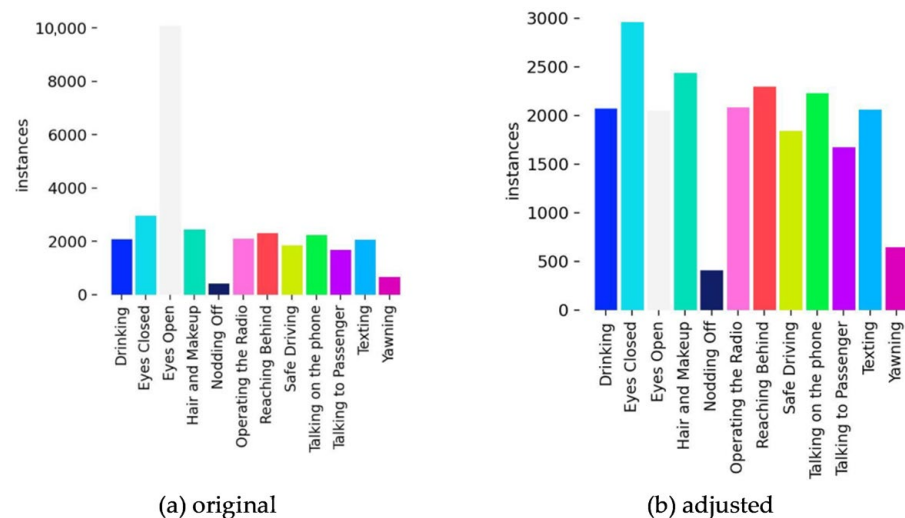


Figure 13. Comparison of training-set sample distributions across the twelve distracted and fatigue-related driving behavior categories: (a) original distribution; (b) adjusted distribution after reducing the “Eyes Open” samples for the additional class-imbalance experiment.

Table 3. Effect of reducing the ‘Eyes Open’ training samples on per-category AP performance of YOLOv12-LAD.

Class	Original AP (%)	Adjusted AP (%)	Difference
Drinking	99.5	99.5	0.0
Eyes Closed	92.4	92.5	+0.1
Eyes Open	98.0	22.9	−75.1
Hair and Makeup	99.5	99.5	0.0
Nodding Off	96.8	95.3	−1.5
Operating the Radio	99.5	99.5	0.0
Reaching Behind	99.5	99.5	0.0
Safe Driving	99.0	99.4	+0.4
Talking on the phone	99.5	99.5	0.0
Talking to Passenger	98.5	98.7	+0.2
Texting	99.3	99.2	−0.1
Yawning	99.3	99.3	0.0
All (mAP@50)	98.4	92.1	−6.3

As shown in Table 3, the AP values of most categories related to distraction and fatigue remain largely stable after the ‘Eyes Open’ samples are reduced in the training set. However, the AP of the ‘Eyes Open’ category itself drops sharply from 98.0% to 22.9%, and the overall mAP@50 decreases from 98.4% to 92.1%. These results suggest that the original class imbalance does not cause obvious prediction bias in the remaining categories within the current task setting. However, substantially reducing the ‘Eyes Open’ samples severely weakens recognition of this category and lowers the overall detection performance. This further indicates that a larger number of ‘Eyes Open’ samples mainly benefits recognition of this category itself, rather than producing obvious bias towards other categories.

3.3.3. Ablation Studies

Eight ablation experiments were conducted to evaluate the contributions of the LSKA, ADown and DySample modules. Table 4 summarizes the results of these experiments. Specifically, Experiment 1 corresponded to the YOLOv12 baseline model without any additional modules. Experiments 2–4 involved introducing the LSKA, ADown, and DySample modules individually, in order to examine their respective contributions. Experiments 5–7

evaluated combinations of two modules to analyse their collaborative effects. Experiment 8 involved the YOLOv12-LAD model with all three modules integrated together.

Table 4. Ablation experiment results.

Experiment	LSKA	ADown	DySample	P (%)	R (%)	mAP@50 (%)	Params (M)	Model Size (MB)	GFLOPs	FPS (f/s)
1				96.2	95.6	97.7	2.6	5.5	6.5	251
2	✓			96.7	96.3	97.9	2.6	5.7	6.4	238
3		✓		96.5	96	97.8	2.4	5.2	6.2	257
4			✓	96.8	96.5	98	2.6	5.5	6.3	255
5	✓	✓		96.3	96.5	98.2	2.5	5.4	6.2	245
6	✓		✓	96.8	95.9	98.2	2.6	5.7	6.4	245
7		✓	✓	96.9	95.9	98.1	2.4	5.3	6.2	259
8	✓	✓	✓	97.5	96.3	98.4	2.5	5.4	6.2	249

After incorporating the LSKA mechanism into the backbone network, the precision, recall, and mAP@50 increased by 0.5%, 0.7%, and 0.2%, respectively. Meanwhile, the additional attention computation slightly reduces FPS to 238 f/s, but real-time performance remains sufficient. The LSKA mechanism primarily enhances global contextual perception by enlarging the receptive field and strengthening the modeling of long-range dependencies among driver-related features. Replacing the original stride-2 convolution layers in the neck with ADown increases precision, recall, and mAP@50 by 0.3%, 0.4%, and 0.1%, respectively, while reducing the number of parameters, the model size, and the GFLOPs by 0.2 million, 0.3 MB, and 0.2 GFLOPs, respectively. Accordingly, FPS increases to 257 f/s, demonstrating improved efficiency. ADown's lightweight downsampling design reduces downsampling-induced information loss and lowers computational overhead. This preserves fine-grained features during compression, particularly for subtle hand, facial, and object-related cues. Introducing the dynamic sampling module, DySample, improves precision, recall, and mAP@50 by 0.6%, 0.9%, and 0.3%, respectively, while keeping model size unchanged and maintaining comparable inference speed. FPS remains at 255 f/s, indicating that DySample achieves performance improvements with negligible extra computation. By learning content-aware sampling locations, DySample enhances adaptive feature reconstruction across different resolutions. This improves multi-scale behavior recognition for both large driver postures and small behavior-related objects, such as hand gestures and mobile phones.

In Experiment 5, the LSKA and ADown modules were incorporated, resulting in improvements of 0.2% in recall and 0.3% in mAP@50 compared to the LSKA-only configuration, while reducing model size by 0.3 MB and GFLOPs by 0.2 GFLOPs. The FPS reaches 245 f/s, confirming that ADown effectively alleviates the computational overhead introduced by LSKA. These results indicate a better balance between global perception and fine-grained feature preservation. However, scale-adaptive behavior recognition remains limited compared to the full model. In contrast, Experiment 6 explores the combination of LSKA and DySample, which provides a slight improvement in mAP@50 over the single-module setting, while precision remains comparable and recall decreases slightly. Although this combination enhances global contextual perception and multi-scale feature reconstruction, the absence of ADown still limits the preservation of subtle local details during downsampling, which may affect the stability of fine-grained behavior detection. The FPS remains at 245 f/s. Experiment 7 demonstrates that combining ADown with DySample improves precision by 0.4% and mAP@50 by 0.3% compared to the ADown-only model. This combination also maintains low computational cost. Benefiting from the high efficiency of both modules, the FPS increases to 259 f/s, the highest among all ablation

settings. The results show that the design effectively balances detection accuracy and computational efficiency. However, without LSKA, the model still exhibits relatively limited global contextual perception of driver posture and overall behavioral context compared to configurations that include LSKA.

Among all ablation configurations, Experiment 8, which combines LSKA, ADown, and DySample, demonstrates the best overall performance. Compared to the baseline, the precision, recall, and mAP@50 improve by 1.3%, 0.7%, and 0.7%, respectively. Additionally, parameter count, model size, and FLOPs are reduced. The FPS stabilizes at 249 f/s, achieving an optimal balance between accuracy and real-time performance. These results indicate that the joint design enhances detection performance without increasing computational cost, as the three modules complement each other in global perception, fine-grained feature preservation, and multi-scale behavior recognition.

To further evaluate the effectiveness of each individual improvement strategy presented in Table 4, Gradient-weighted Class Activation Mapping (Grad-CAM) was employed to generate attention heatmaps for all evaluated models. Figure 14 illustrates these heatmaps extracted from the detection layer, where brighter regions correspond to stronger attention, and red-to-yellow regions denote the network's primary areas of interest.

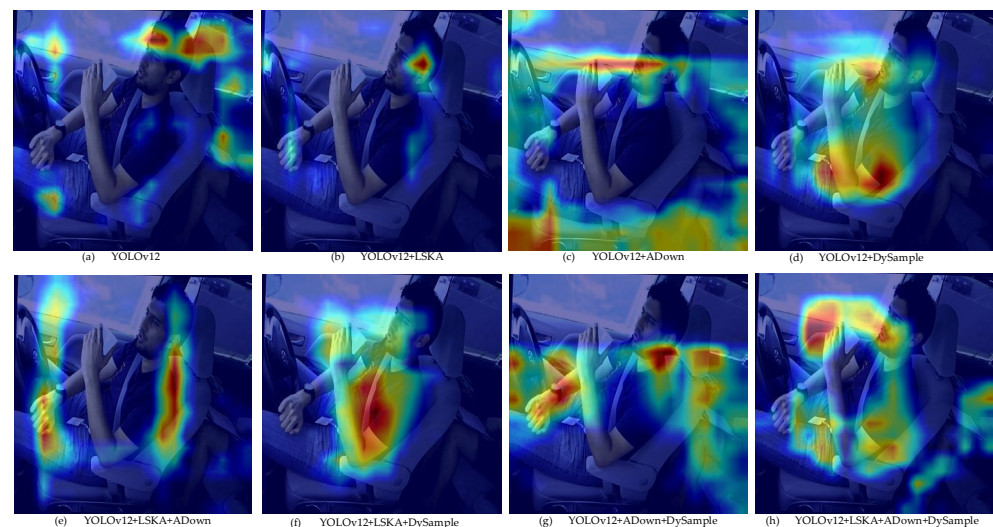


Figure 14. Attention heatmap visualization for driver drinking distraction detection: (a) YOLOv12; (b) YOLOv12 + LSKA; (c) YOLOv12 + ADown; (d) YOLOv12 + DySample; (e) YOLOv12 + LSKA + ADown; (f) YOLOv12 + LSKA + DySample; (g) YOLOv12 + ADown + DySample; (h) YOLOv12 + LSKA + ADown + DySample. Brighter red-to-yellow regions indicate stronger attention and the network's primary areas of interest.

The baseline YOLOv12 mainly focuses on the driver's head and several irrelevant peripheral regions. As a result, the hand–cup interaction area, which is essential for recognizing drinking behavior, is not captured consistently. Considerable background noise is observed, suggesting a limited ability to distinguish behavior-related regions from irrelevant surroundings.

Each module improves the attention pattern from a different perspective. After LSKA is introduced, the activation becomes more concentrated around the driver's head and upper body, while background interference is noticeably reduced. This indicates a stronger global perception of the overall posture and behavioral context. However, the attention to fine-grained local cues, such as the hand and the cup, remains relatively limited. After ADown is introduced, the attention range expands to cover a larger portion of the driver's body and the action-related region. This suggests that more structural details are preserved during downsampling. Nevertheless, the activation around key drinking-related cues

is still insufficiently concentrated, and some redundant responses remain in non-critical regions. In contrast, DySample produces clearer activation for small behavior-related targets, especially the hand and the cup, demonstrating its advantage in reconstructing multi-scale detail. However, without sufficiently strong global contextual guidance, some background interference persists.

The dual-module combinations show better complementarity. However, their performance remains inferior to that of the full model. The LSKA + ADown model achieves more stable global attention and broader coverage of the driver's action region. Still, the activation of small behavior-related targets remains relatively weak. The LSKA + DySample model enhances both global saliency and local target activation, with clearer responses in the head, hand, and cup region. Nevertheless, the continuity of local structural details is still not as good as that of the full model. The ADown + DySample model improves coverage of action-related regions and enhances activation of local details. Yet, the overall attention distribution remains less stable because global contextual perception remains insufficient.

In contrast, the full YOLOv12-LAD model exhibits the most balanced and discriminative attention distribution. It not only suppresses irrelevant background responses but also maintains stable attention to the driver's upper body. At the same time, the hand and cup regions are accurately highlighted. These results indicate that LSKA mainly improves global perception, ADown helps preserve fine-grained structural information during down-sampling, and DySample enhances multi-scale behavior recognition through adaptive detail reconstruction. Their complementary integration enables the model to achieve more reliable behavior localization than any single- or dual-module configuration.

3.3.4. Comparison with Representative Detection Methods

Table 5 compares the proposed method with several representative driver distraction detection models, including RetinaNet, TOOD, RT-DETR, RT-DETRv2, YOLOv8, YOLOv9, YOLOv10, and YOLOv11.

Table 5. Overall performance of representative detectors.

Detector	P (%)	R (%)	F1 (%)	mAP@50 (%)	Params (M)	GFLOPs	FPS (f/s)
RetinaNet	93.7	94.1	93.9	95.9	36.6	83.5	41
TOOD	95.1	95.3	95.2	97.1	32.0	79.1	43
RT-DETR [41]	96.6	95.4	96.0	97.4	20.0	57.0	69
RT-DETRv2 [42]	97.1	94.5	95.8	97.6	20.0	57.0	70
YOLOv8	96.2	96.0	96.1	97.8	3.0	8.1	287
YOLOv9	96.3	95.8	96.0	97.9	1.9	7.6	194
YOLOv10	95.2	96.2	95.7	97.6	2.3	6.5	213
YOLOv11	96.2	95.5	95.8	97.7	2.6	6.3	245
YOLOv12-LAD	97.5	96.3	96.9	98.4	2.5	6.2	249

Among the compared methods, RetinaNet achieves the lowest precision, recall, and mAP@50. Its model complexity is relatively high, with 36.6 M parameters and 83.5 GFLOPs. This increases the computational burden and leads to higher inference latency. Accordingly, its FPS is only 41 f/s, which does not meet real-time requirements. These characteristics make it less suitable for real-time driver monitoring scenarios. TOOD achieves higher detection accuracy than RetinaNet. However, it still requires 32.0 M parameters and 79.1 GFLOPs. The FPS of TOOD is only 43 f/s, indicating that its real-time performance is still limited by high computational cost.

In addition to CNN-based detectors, representative real-time Transformer-based models, including RT-DETR and RT-DETRv2, were also evaluated. RT-DETR achieves com-

petitive detection performance, with an mAP@50 of 97.4%. However, it requires 20.0 M parameters and 57.0 GFLOPs. RT-DETRv2 further improves detection accuracy to an mAP@50 of 97.6% while maintaining a similar model scale. Compared with earlier CNN-based detectors such as RetinaNet and TOOD, RT-DETR achieves a noticeable reduction in parameters and FLOPs. However, the Transformer encoder–decoder architecture still introduces additional attention-related computation. Therefore, the computational overhead remains significant. As a result, the FPS values of RT-DETR and RT-DETRv2 are only 69 f/s and 70 f/s, respectively, which are much lower than those of lightweight YOLO-series detectors.

By comparison, YOLO-series models generally maintain a good balance between computational cost and detection accuracy. YOLOv8 improves precision to 96.2% and achieves an mAP@50 of 97.8%, with 3.0 M parameters and 8.1 GFLOPs. Benefiting from its lightweight design, its FPS reaches 287 f/s. Within the series, YOLOv9 uses the fewest parameters (1.9 M) while still attaining an F1-score of 96.0% and an mAP@50 of 97.9%. Its FPS is 194 f/s, indicating good real-time performance. Additionally, YOLOv10 and YOLOv11 focus on reducing computational demand. Their GFLOPs are reduced to 6.5 and 6.3, respectively, while maintaining similar detection accuracy. Their FPS values reach 213 f/s and 245 f/s, further confirming the efficiency advantage of the YOLO series.

Although these models achieve competitive results, they do not simultaneously provide both superior detection accuracy and computational efficiency. In contrast, the proposed YOLOv12-LAD achieves the best overall performance among the compared detectors. Specifically, YOLOv12-LAD achieves 97.5% precision, 96.3% recall, 96.9% F1-score, and 98.4% mAP@50, with only 2.5 M parameters and 6.2 GFLOPs. Its FPS reaches 249 f/s. This value is higher than that of YOLOv9, YOLOv10, and YOLOv11, and is only slightly lower than YOLOv8. This demonstrates a more favorable balance between accuracy and real-time performance.

As shown in Table 5, YOLOv12-LAD achieves the most favorable trade-off between accuracy and complexity among RetinaNet, TOOD, RT-DETR, and advanced YOLO detectors, including YOLOv8, YOLOv9, YOLOv10, and YOLOv11. Many existing YOLO-based detection methods rely on lightweight convolutions, single attention mechanisms, or simple multi-scale fusion. These designs struggle to simultaneously maintain global perception, feature stability, and fine-grained feature representation. This limitation is particularly evident in driver monitoring tasks, where reliable recognition requires the simultaneous understanding of global driver posture, subtle hand–face interactions, and behavior-related objects with large-scale variation.

The superior performance of YOLOv12-LAD stems from the complementary roles of its three integrated modules. LSKA expands the receptive field and improves the modeling of long-range spatial dependencies, thereby improving the global perception of driver posture and overall behavior. ADown reduces information loss during feature downsampling, thereby preserving fine-grained structural details of the hands, face, and objects related to distracted behaviors. DySample strengthens adaptive multi-scale feature reconstruction. As a result, it improves recognition of both large-scale posture features and small behavior-related objects, such as mobile phones, cups, and other hand-held items. Through the complementary integration of these modules, YOLOv12-LAD outperforms the other detectors evaluated in this study in terms of accuracy, robustness, and efficiency.

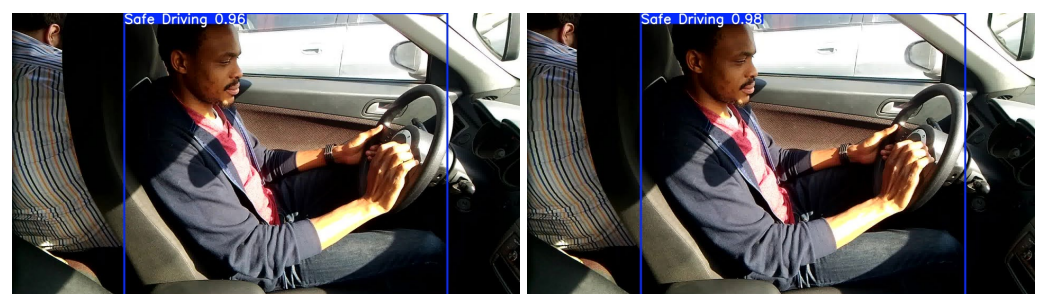
3.3.5. Cross-Dataset Evaluation

Experiments were conducted on two publicly available datasets, the AUC (American University in Cairo Distracted Driver Dataset) [43,44] and the 3MDAD (Multiview, Multimodal and Multispectral Driver Action Dataset). Table 6 summarizes the performance of

YOLOv12-LAD and the baseline YOLOv12 model under identical experimental settings. Qualitative detection results are shown in Figure 15.

Table 6. Cross-dataset evaluation results on the AUC and 3MDAD datasets.

Evaluation Metrics	AUC		3MDAD	
	YOLOv12	YOLOv12-LAD	YOLOv12	YOLOv12-LAD
P (%)	98.8	99.9	88.2	92.2
R (%)	98.8	99.5	87.6	92.8
mAP@50 (%)	98.5	99.5	88.5	92.5
mAP@50-95 (%)	98.0	99.1	71.5	78.4
Params (M)	2.6	2.5	2.6	2.5
GFLOPs	6.3	6.2	6.2	6.2



YOLOv12

YOLOv12-LAD

(a) AUC Dataset



YOLOv12

YOLOv12-LAD

(b) 3MDAD Dataset

Figure 15. Cross-dataset detection results of YOLOv12-LAD on the AUC and 3MDAD datasets. (a) AUC dataset; (b) 3MDAD dataset.

Overall, YOLOv12-LAD consistently outperforms the baseline on both datasets, achieving higher precision, recall, and mAP@50 while requiring fewer computational resources. The AUC dataset provides higher image resolution (1920×1080 pixels), a single camera viewpoint, daytime-only samples, and fewer behavior categories. These characteristics reduce variation in viewpoint and illumination and provide clearer visual details. As a result, the model can capture behavior-related cues more effectively, leading to high precision and recall across all distraction and fatigue categories.

By contrast, the 3MDAD dataset presents several challenges that collectively contribute to a noticeable performance drop (mAP@50: 92.5% on 3MDAD vs. 99.5% on AUC). Specifically, 3MDAD provides lower-resolution images (640×480 pixels), includes 50 drivers and 16 distraction categories, uses two camera views (left-front and side views), and contains both daytime and nighttime samples. In addition, about 150 images are provided for each behavior category of each driver. These factors increase the difficulty of detection by reducing the availability of fine-grained visual cues, especially for subtle fatigue-related behaviors such as eye closure, nodding off, and yawning, as well as small-scale distraction actions involving phones or other small in-cabin objects. As a result, both false positives and false negatives become more likely.

Although YOLOv12-LAD still performs well on 3MDAD, these dataset-specific factors make subtle behaviors more difficult to recognize and partly explain the performance gap relative to AUC. These results indicate that the proposed model maintains relatively stable performance across different public datasets. However, its performance is still affected by differences in image quality, viewpoint, and scene complexity.

3.3.6. Real-World Driving Evaluation

An on-road evaluation was conducted in November 2025 at Baotou Vocational & Technical College in Qingshan District, Baotou, Inner Mongolia. Eight drivers participated, and all trials were conducted using the same vehicle model, the Geely Galaxy L7. Data on distracted and fatigue-related driving behaviors were collected across multiple sessions.

After data preprocessing, a real-world driving behavior dataset containing 2200 samples was constructed. To demonstrate the model's detection performance in real-world scenarios, Figures 16–18 present detection results for several typical driving behaviors under real-world road conditions.

As shown in Figures 16–18, both YOLOv12 and the proposed YOLOv12-LAD identify a range of distracted and fatigue-related driving behaviors in real-world scenes. These behaviors include representative cases such as drinking, eye closure, texting, and yawning. Overall, these results demonstrate that the proposed YOLOv12-LAD framework is applicable to real-world in-vehicle behavior monitoring. For visually salient behaviors involving clear hand–object interactions or relatively large posture changes, such as drinking, phone use, and texting, the detections are generally stable and have relatively high confidence. These results suggest that the model is more reliable for behaviors with strong visual saliency and clear spatial cues in real-world driving scenes.

However, the real-world evaluation also indicates that some categories involving complex local actions, such as doing hair and makeup and operating the radio, remain more challenging in real-world in-cabin environments. In these cases, subtle local actions and complex cabin backgrounds may make detection more sensitive to localization deviations and background interference, thereby reducing detection stability. Similarly, fatigue-related behaviors such as eye closure, nodding off, and yawning rely more heavily on subtle facial and head-related cues. Their detection performance may therefore be more sensitive to variations in driver posture, camera viewpoint, illumination, and occlusion.



Figure 16. Detection results for the first four distracted driving behavior categories.



Figure 17. Detection results for the next four distracted driving behavior categories.



Figure 18. Detection results for the remaining four distracted driving behavior categories.

Therefore, the real-world results reported in this study should be interpreted within the scope of the current evaluation setting. Since the on-road experiments involved only eight drivers and a single vehicle model, the reported results mainly reflect the model's

performance under relatively constrained validation conditions rather than providing a comprehensive assessment across diverse real-world scenarios. More specifically, the present findings are more representative of performance on visually prominent distraction behaviors under conditions similar to the current experimental setting. These conditions include a fixed camera viewpoint, a single vehicle cabin layout, limited driver diversity, and relatively constrained environmental variations. In contrast, more subtle fatigue-related or locally complex behaviors may show larger performance fluctuations when the model is deployed in environments with different drivers, vehicle cabins, camera viewpoints, lighting conditions, occlusions, and road contexts.

4. Discussion

The experimental results offer several important insights into the effectiveness of the proposed framework for distracted and fatigue-related driving behavior detection. Distracted and fatigue-related driving behaviors typically involve multi-scale visual patterns, including both overall postural changes and subtle local movements. By expanding the receptive field, the LSKA module enhances global contextual perception. In parallel, ADown reduces information loss during downsampling and helps preserve fine-grained structural details, whereas DySample improves content-adaptive feature reconstruction across different resolutions, thereby strengthening multi-scale behavior representation.

Many representative driver monitoring methods based on improved YOLO models have been proposed in recent years. For instance, some improved YOLOv8-based schemes mainly focus on lightweight convolution and simple multi-scale fusion to reduce computational cost, while YOLOv10-based methods improve efficiency by removing redundant network branches and simplifying inference pipelines. Although these strategies improve model speed or accuracy to a certain extent, specialized designs for global perception and fine-grained feature preservation in driving scenarios remain limited. In contrast, the proposed integrated design of LSKA, ADown, and DySample helps overcome the limitations of single- or dual-module designs. This integration represents a key design contribution of the present study.

Ablation studies further confirm the complementary interaction among the three modules. Their combined design yields more balanced performance gains, indicating that the improvements primarily stem from stronger feature representation capabilities within the network, rather than from simple model scaling. Heatmap visualizations also show that the three-module configuration can capture more complete and accurate behavioral features. This level of feature completeness and accuracy is not observed in single- or dual-module combinations.

Cross-dataset evaluation further shows that YOLOv12-LAD maintains relatively stable performance gains across different public datasets, which suggests good adaptability to varying data distributions. The improvements are more evident under conditions with richer spatial information and more consistent imaging settings. This finding suggests that the expanded receptive field and content-aware sampling strategy enable the model to better exploit informative visual cues. At the same time, the model still retains relatively strong performance under more challenging conditions, demonstrating a certain degree of robustness to variations in image quality, camera viewpoint, cabin layout, and environmental factors. Through the coordinated effects of its integrated modules, YOLOv12-LAD achieves a better balance between detection precision and computational efficiency than mainstream YOLO-based detectors, making it more suitable for real-time in-vehicle monitoring tasks.

Real-world road test results also provide further evidence of the practical value of the proposed approach. The framework can reliably detect visually prominent distracted-driving behaviors while also capturing subtle fatigue-related movements. In complex

in-vehicle environments with multiple displays and ambient lighting, the joint modeling of global and local features supports more robust behavior detection. However, the current findings are more representative of performance on visually prominent distraction behaviors under conditions similar to those in the current experimental setting. In contrast, more subtle fatigue-related behaviors and locally complex actions may be more sensitive to variations in driver characteristics, vehicle cabin layouts, camera viewpoints, lighting conditions, and road environments. This may lead to greater performance fluctuations in practical applications.

Despite these encouraging findings, several limitations should be acknowledged when interpreting the present results and considering the broader deployment of the proposed method. First, existing public distracted driving datasets remain limited in scene diversity and do not fully capture the complexity of real driving environments, such as adverse weather, varying road contexts, and diverse in-vehicle layouts. This limitation may restrict the model's adaptability when it is deployed beyond the datasets used in this study. In addition, varying road contexts may affect model robustness by increasing background complexity, changing illumination conditions, and introducing image instability caused by vehicle motion. This limitation is also consistent with the cross-dataset evaluation results, which show that the performance of YOLOv12-LAD is influenced by factors such as image resolution, camera viewpoint, cabin layout, and environmental conditions. Although the model maintains relatively strong overall performance across different public datasets, subtle fatigue- and distraction-related behaviors remain more susceptible to interference on the more challenging 3MDAD dataset. Second, practical testing involved only a small number of drivers and a single vehicle model, covering a limited range of age groups, driving habits, and physical characteristics. This limitation may weaken the framework's generalizability, since driver posture, behavioral expression, and cabin appearance can vary considerably across different individuals and vehicle types. Finally, the current framework relies solely on monocular RGB images, which introduces inherent limitations in complex visual conditions. Such inputs are susceptible to ambient lighting changes, partial occlusions, and adverse weather, which may degrade feature representation and reduce detection stability and reliability in real-world applications. Accordingly, the performance reported in this study should be interpreted within the scope of the current datasets and validation settings, rather than as a fully generalizable estimate for all real-world driving environments.

To address these limitations and further improve the practical applicability of the proposed method, future research may focus on three directions. First, the dataset can be expanded to include more diverse driving scenes, thereby improving the model's adaptability to complex environments. Second, practical tests involving more drivers and multiple vehicle models can be conducted to further evaluate and enhance the model's generalizability. Third, multimodal perception strategies, such as infrared imaging and vehicle sensor data, can be explored to improve detection robustness under challenging real-world conditions.

5. Conclusions

This work proposes YOLOv12-LAD, an enhanced YOLOv12-based framework for distracted and fatigue-related driving behavior detection. By integrating the LSKA, ADown, and DySample modules, the proposed framework improves global contextual perception, fine-grained feature preservation, and multi-scale behavior representation.

Comprehensive experiments, including ablation studies, comparisons with representative models, cross-dataset evaluation, and real-world vehicle validation, demonstrate

that the proposed framework achieves higher detection accuracy while maintaining high computational efficiency.

The results indicate that YOLOv12-LAD provides a promising vision-based framework for driver monitoring in intelligent driving systems. Real-world evaluation further supports its practical applicability for distracted and fatigue-related driving behavior detection under the current experimental setting. However, further research is still needed to improve its applicability and generalizability in complex real-world scenarios. Future work will focus on expanding the dataset to cover more diverse driving scenes, conducting larger-scale real-vehicle validation across more drivers and vehicle models, and exploring multimodal perception strategies to further improve robustness under challenging driving conditions.

Author Contributions: Conceptualization, X.L. and Z.G.; methodology, X.L. and Q.C.; software, X.L.; validation, X.L. and Y.R.; formal analysis, X.L. and Q.C.; investigation, X.L. and Y.R.; data curation, X.L. and Y.R.; writing—original draft preparation, X.L.; writing—review and editing, Z.G. and Q.C.; visualization, X.L.; supervision, Z.G. and Q.C.; project administration, Q.C.; funding acquisition, Q.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Tianjin Transportation Science and Technology Development Plan Project of the Tianjin Municipal Transportation Commission, grant number 2022-32; and by the Qinghai Provincial Key Laboratory of Plateau Vehicle Electrification and Intelligentization Technology of Qinghai Vocational and Technical University, grant number QZDSZ03-202504.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: FIRC Dataset can be found at <https://github.com/flsl/firc-dataset/tree/main> (accessed on 10 April 2026). The other data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. World Health Organization. *Global Status Report on Road Safety 2023*; World Health Organization: Geneva, Switzerland, 2023. Available online: <https://www.who.int/teams/social-determinants-of-health/safety-and-mobility/global-status-report-on-road-safety-2023> (accessed on 1 March 2026).
2. Javid, R.; Sadeghvaziri, E.; Jeihani, M. A driving simulator study to understand the impact of cell phone blocking apps on distraction. *Highlights Veh.* **2023**, *1*, 17–28. [CrossRef]
3. Dunn, N.; Dingus, T.; Socolich, S.; Hornsey, W. Investigating the impact of driving automation systems on distracted driving behaviors. *Accid. Anal. Prev.* **2021**, *156*, 106152. [CrossRef]
4. Singh, H.; Kathuria, A. Analyzing driver behavior under naturalistic driving conditions: A review. *Accid. Anal. Prev.* **2021**, *150*, 105908. [CrossRef]
5. Fu, S.; Yang, Z.; Ma, Y.; Li, Z.; Xu, L.; Zhou, H. Advancements in the Intelligent Detection of Driver Fatigue and Distraction: A Comprehensive Review. *Appl. Sci.* **2024**, *14*, 3016. [CrossRef]
6. Fan, C.; Peng, Y.; Peng, S.; Zhang, H.; Wu, Y.; Kwong, S. Detection of train driver fatigue and distraction based on forehead EEG: A time-series ensemble learning method. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 13559–13569. [CrossRef]
7. Mewborne, T.; Lee, Y.; Tan, S.; Yang, J. Distracted Driving Detection Utilizing Wearable-Based Bluetooth. In Proceedings of the 2022 IEEE 19th International Conference on Mobile Ad Hoc and Smart Systems (MASS), Denver, CO, USA, 19–23 October 2022; pp. 485–490.
8. Sun, Q.; Wang, C.; Guo, Y.; Yuan, W.; Fu, R. Research on a cognitive distraction recognition model for intelligent driving systems based on real vehicle experiments. *Sensors* **2020**, *20*, 4426. [CrossRef]
9. Ucar, S.; Oguchi, K. Demo: Distracted Driving Behavior Detection to Avoid Rear-End Collisions. In Proceedings of the 2021 IEEE Vehicular Networking Conference (VNC), Ulm, Germany, 10–12 November 2021; pp. 115–116.
10. Ucar, S.; Sisbot, E.; Muralidharan, H.; Oguchi, K. Distracted Driving Detection in Stop-and-Go Traffic. In Proceedings of the 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), Macau, China, 8–12 October 2022; pp. 1196–1201.

11. Nakano, K.; Chakraborty, B. Real-Time Distraction Detection from Driving Data Based Personal Driving Model Using Deep Learning. *Int. J. Intell. Transp. Syst. Res.* **2022**, *20*, 238–251. [\[CrossRef\]](#)
12. Zhu, F.; Chen, J.; Chen, J.; Yan, M.; Xiang, L. Driver fatigue detection based on improved YOLOv3. *Sci. Technol. Eng.* **2022**, *23*, 11730–11738.
13. Ansari, S.; Naghdy, F.; Du, H.; Pahnwar, Y. Driver mental fatigue detection based on head posture using new modified reLU-BiLSTM deep neural network. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 10957–10969. [\[CrossRef\]](#)
14. Li, Y.; Mi, W.; Ge, J.; Hu, J.; Li, H.; Zhang, D.; Li, T. SelectAug: A Data Augmentation Method for Distracted Driving Detection. In *Advances in Knowledge Discovery and Data Mining, PAKDD 2022, Part II*; Springer: Cham, Switzerland, 2022; Volume 13281, pp. 405–416.
15. Wagner, B.; Taffner, F.; Karaca, S.; Karge, L. Vision based detection of driver cell phone usage and food consumption. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 4257–4266. [\[CrossRef\]](#)
16. Zhou, Y.; Chen, Y.; Zhang, Y. Driver distraction detection in conditionally automated driving using multimodal physiological and ocular signals. *Electronics* **2025**, *14*, 3811. [\[CrossRef\]](#)
17. Vijeikis, R.; Dada, I.D.; Abayomi-Alli, A.A.; Raudonis, V. Enhancing driver monitoring systems based on novel multi-task fusion algorithm. *Sensors* **2025**, *25*, 6799. [\[CrossRef\]](#)
18. Papatheocharous, E.; Kaiser, C.; Moser, J.; Stocker, A. Monitoring distracted driving behaviours with smartphones: An extended systematic literature review. *Sensors* **2023**, *23*, 7505. [\[CrossRef\]](#)
19. Visconti, P.; Rausa, G.; Del-Valle-Soto, C.; Velázquez, R.; Cafagna, D.; De Fazio, R. Innovative Driver Monitoring Systems and On-Board-Vehicle Devices in a Smart-Road Scenario Based on the Internet of Vehicle Paradigm: A Literature and Commercial Solutions Overview. *Sensors* **2025**, *25*, 562. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Essahraoui, S.; Lamaakal, I.; El Hamly, I.; Maleh, Y.; Ouahbi, I.; El Makkaoui, K.; Filali Bouami, M.; Plawiak, P.; Alfarraj, O.; Abd El-Latif, A.A. Real-Time Driver Drowsiness Detection Using Facial Analysis and Machine Learning Techniques. *Sensors* **2025**, *25*, 812. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Tran, D.; Do, H.; Lu, J.; Sheng, W. Real-Time Detection of Distracted Driving Using Dual Cameras. In Proceedings of the 2020 IEEE/RIS International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 24 October 2020–24 January 2021; pp. 2014–2019.
22. Shen, Q.; Zhang, L.; Zhang, Y.; Li, Y.; Liu, S.; Xu, Y. Distracted driving behavior detection algorithm based on lightweight StarDL-YOLO. *Electronics* **2024**, *13*, 3216. [\[CrossRef\]](#)
23. Li, F.; Gu, S.; Lu, L.; Ren, B.; Zhang, L.; Wu, W. A lightweight and efficient approach for distracted driving detection based on YOLOv8. *Electronics* **2026**, *15*, 34. [\[CrossRef\]](#)
24. Du, Y.; Liu, X.; Yi, Y.; Wei, K. Optimizing road safety: Advancements in lightweight YOLOv8 models and GhostC2f design for real-time distracted driving detection. *Sensors* **2023**, *23*, 8844. [\[CrossRef\]](#)
25. Zhang, Q.; Zhu, Z.; Bai, Y.; Liao, G.; Liu, T. Distracted driving detection based on the improved CenterNet with attention mechanism. *Multimed. Tools Appl.* **2022**, *81*, 7993–8005. [\[CrossRef\]](#)
26. Gong, Y.; Shen, X. An algorithm for distracted driving recognition based on pose features and an improved KNN. *Electronics* **2024**, *13*, 1622. [\[CrossRef\]](#)
27. Zhao, D.; Li, H.; Fu, Z.; Dong, S.; Li, Z.; Zhong, Y.; Hou, J.; He, W. Distracted Driving Behavior Recognition Model for Electric Bus Based on Fusion of Cross-Layer Multi-scale Features and Attention Mechanism. *Int. J. Automot. Technol.* **2026**, *27*, 133–144. [\[CrossRef\]](#)
28. Sahoo, G.; Das, S.; Singh, P. A deep learning-based distracted driving detection solution implemented on embedded system. *Multimed. Tools Appl.* **2023**, *82*, 11697–11720. [\[CrossRef\]](#)
29. Elshamy, M.; Emara, H.; Shoaib, M.; Badawy, A. P-YOLOv8: Efficient and Accurate Real-Time Detection of Distracted Driving. In Proceedings of the 2024 IEEE High Performance Extreme Computing Conference (HPEC), Wakefield, MA, USA, 23–27 September 2024; pp. 1–6.
30. Salakapuri, R.; Navuri, N.K.; Vobbilineni, T.; Ravi, G.; Karmakonda, K.; Vardhan, K.A. Integrated deep learning framework for driver distraction detection and real-time road object recognition in advanced driver assistance systems. *Sci. Rep.* **2025**, *15*, 25125. [\[CrossRef\]](#)
31. Shang, E.; Liu, H.; Yang, Z.; Du, J.; Ge, Y. FedBiKD: Federated Bidirectional Knowledge Distillation for Distracted Driving Detection. *IEEE Internet Things J.* **2023**, *10*, 11643–11654. [\[CrossRef\]](#)
32. Wang, C.; Bochkovskiy, A.; Liao, H. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
33. Tian, Y.; Ye, Q.; Doermann, D. YOLOv12: Attention-centric real-time object detectors. *arXiv* **2025**, arXiv:2502.12524.
34. Lau, K.; Po, L.; Rehman, Y. Large separable kernel attention: Rethinking the large kernel attention design in CNN. *Expert Syst. Appl.* **2024**, *236*, 121352. [\[CrossRef\]](#)

35. Guo, M.; Lu, C.; Liu, Z.; Cheng, M.; Hu, S. Visual attention network. *Comput. Vis. Media* **2023**, *9*, 733–752. [[CrossRef](#)]
36. Cai, S.; Zhou, X.; Wei, L.; Mo, Y. ALW-YOLOV8n: A Lightweight underwater detector Enhanced by Attention mechanism, ADown Block and Wise-WIoU on YOLOv8n. *Res. Sq.* **2024**. [[CrossRef](#)]
37. Liu, W.; Lu, H.; Fu, H.; Cao, Z. Learning to upsample by learning to sample. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 6027–6037.
38. Fu, S.; Li, S. FIRC Dataset: Fine-Grained Industrial Image Recognition Classification Dataset. GitHub. 2023. Available online: <https://github.com/flsl/firc-dataset/tree/main> (accessed on 10 April 2026).
39. Wang, C.; Sun, W.; Wu, H.; Zhao, C.; Teng, G.; Yang, Y.; Du, P. A Low-Altitude Remote Sensing Inspection Method on Rural Living Environments Based on a Modified YOLOv5s-ViT. *Remote Sens.* **2022**, *14*, 4784. [[CrossRef](#)]
40. Liu, W.; Quijano, K.; Crawford, M. YOLOv5-Tassel: Detecting Tassels in RGB UAV Imagery with Improved YOLOv5 Based on Transfer Learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 8085–8094. [[CrossRef](#)]
41. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs beat YOLOs on real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024.
42. Lv, W.; Zhao, Y.; Chang, Q.; Huang, K.; Wang, G.; Liu, Y. RT-DETRv2: Improved baseline with bag-of-freebies for real-time detection transformer. *arXiv* **2024**, arXiv:2407.17140.
43. Eraqi, H.; Abouelnaga, Y.; Saad, M.; Moustafa, M. Driver distraction identification with an ensemble of convolutional neural networks. *J. Adv. Transp.* **2019**, *2019*, 4125865. [[CrossRef](#)]
44. Abouelnaga, Y.; Eraqi, H.; Moustafa, M. Real-time distracted driver posture classification. *arXiv* **2018**, arXiv:1706.09498. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.