RESEARCH ARTICLE

Fatigue Driving Detection via Multi-Head Transformer with Adaptive Weighted Loss

Ling Huang[1], **Shifeng Li**[1], **Yaxin Man**[1], **Xiaoyan Wang**[1,*], **Xiu Tang**[1] **and Rendong Ji**[1]

[1] Faculty of Electronic Information Engineering, Huaiyin Institute of Technology, Huai'an 223003, China

Abstract

Fatigue driving is widely recognized as one of the major factors contributing to traffic accidents, posing not only a serious threat to road safety but also potential risks to drivers' health and public security. With the rapid development of modern transportation, how to efficiently and accurately detect and warn against driver fatigue has become a critical issue in the field of intelligent transportation. To effectively address this issue, this paper proposes a novel fatigue driving detection method based on a Multi-Head Transformer with Adaptive Weighted Loss. In the proposed framework, the YOLOv8 model is first employed to efficiently and accurately locate key facial regions of the driver from real-time video streams, ensuring both high-speed processing and robustness. Subsequently, a Multi-Head Transformer model is introduced to capture the temporal dependencies and feature correlations among facial landmarks, enabling a more comprehensive characterization of fatigue-related behaviors such as eye closure, blinking, and yawning. In addition, an Adaptive Weighted Loss function is designed to dynamically balance the contributions of multiple fatigue features during training, effectively alleviating the class imbalance problem and enhancing the model's generalization capability. Experimental results demonstrate that the proposed method maintains stable and superior detection accuracy even under complex long-duration driving conditions. Compared with traditional approaches, the system improves fatigue detection accuracy by 7.2%, achieving 95.5%, while also satisfying real-time requirements. In summary, this study presents an intelligent, adaptive, and efficient fatigue driving detection framework that provides a reliable theoretical foundation for intelligent warning systems and holds significant value for enhancing road traffic safety.

Keywords: fatigue driving detection, facial keypoints, transformer, adaptive weighted loss.

1 Introduction

Fatigued driving significantly contributes to road accidents. According to relevant studies, traffic accidents caused by driver fatigue in the United States and Europe account for more than 20% of all accidents [1]. In China, the proportion of drivers who drove while fatigued reached 50%, and the accident rate caused by fatigue was eight times higher than that of drunk driving, and it accounted for 40% of major and especially serious traffic accidents [2, 3]. Fatigue

Academic Editor:
Jianlei Kong

Submitted: 25 September 2025
Accepted: 09 November 2025
Published: 05 March 2026

Vol. 3, **No.** 1, 2026.
10.62762/TIS.2025.633754

***Corresponding author:**
✉ Xiaoyan Wang
wxygxy@163.com

Citation
Huang, L., Li, S., Man, Y., Wang, X., Tang, X., & Ji, R. (2026). Fatigue Driving Detection via Multi-Head Transformer with Adaptive Weighted Loss. *ICCK Transactions on Intelligent Systematics*, 3(1), 55–69.

detection methods can be broadly categorized into three types: those based on physiological signals [4–7], vehicle dynamics information [8–10], and computer vision techniques [11]. Physiological signal-based detection offers high accuracy but relies on multiple sensor devices. Drivers are usually required to wear EEG caps or electrode patches, which introduces a certain degree of intrusiveness. Although non-contact approaches, such as embedding sensors in the steering wheel or seat, have been introduced to improve comfort, they still suffer from high equipment costs, complex maintenance, and sensitivity to individual physiological differences. Vehicle dynamics-based detection analyzes features such as steering wheel angle, acceleration fluctuations, and lane deviation. While it has the advantages of being non-intrusive, low-cost, and easy to integrate, it is highly susceptible to noise from factors such as weather, road conditions, traffic interference, and driving habits, which can lead to misjudgments and unstable accuracy.

Computer vision–based detection methods can directly exploit drivers' facial features to achieve high-accuracy, and non-intrusive recognition, making them a prominent focus of current research. In recent years, researchers have proposed novel fatigue driving detection strategies, such as multimodal fusion approaches. Cao et al. [12] constructed a multimodal feature fusion model by combining electroencephalogram (EEG), electrocardiogram (ECG), and facial features, achieving an accuracy of 94.87% under normal lighting conditions. Li et al. [13] proposed a lightweight joint framework integrating CNN and Vision Transformer features, which significantly reduced parameter complexity and achieved a real-time inference rate of 159.13 FPS; however, its detection accuracy remained unstable, dropping to as low as 93.54%. Chen et al. [14] compared the performance of cross-entropy, mean squared error (MSE), and D-InfoNCE loss functions in fatigue and distraction recognition, demonstrating that adaptive loss functions can effectively alleviate the issue of sample imbalance and enhance the model's ability to identify fatigue and distraction behaviors. These studies indicate that emerging detection methods outperform traditional CNN architectures in driving behavior analysis, as the self-attention mechanism in Transformers can effectively model temporal features and global dependencies. However, under complex conditions such as varying illumination and facial occlusion, these methods still face challenges in maintaining recognition accuracy and real-time performance. Therefore, this study constructs a model that integrates a multi-head Transformer with an adaptive weighted loss under long-duration day and night driving scenarios, aiming to improve both accuracy and robustness while ensuring real-time performance. The proposed approach provides a more stable and real-time fatigue detection solution, offering practical insights for the design of advanced driver assistance systems (ADAS) and contributing to the prevention of traffic accidents, casualties, and psychological trauma caused by driver fatigue.

This study proposes a highly accurate and robust fatigue driving detection method. YOLOv8 is employed for facial image detection and localization, while a multi-head Transformer is utilized to perform in-depth analysis of facial key points. The driver's 3D head pose is estimated using a Perspective-n-Point (PnP) algorithm, and estimation errors are minimized to improve fatigue detection performance. In addition, an adaptive weighted loss function is introduced, allowing the system to adapt to various driving environments and further enhancing detection accuracy and robustness. These innovations enable the proposed fatigue detection method to achieve significant advantages in terms of real-time performance, accuracy, and stability, providing strong support for driver safety warnings.

Subsequent sections are arranged as follows: Section 2 reviews existing studies, summarizing prevalent fatigue detection methods and their implementation in real-world scenarios. Section 3 presents the proposed framework, offering an in-depth analysis of the multi-head Transformer model and deriving a weighted loss function for fatigue level classification. Section 4 evaluates the experimental findings from multiple datasets and includes ablation experiments. Section 5 addresses the study's conclusions and potential avenues for future investigation.

2 Related Work

Machine vision-based fatigue driving detection collects facial images of the driver using cameras and utilizes computer vision and machine learning techniques to extract key features for assessing fatigue levels. This approach has been widely applied in the automotive industry because of its advantages of high accuracy, real-time monitoring, and non-contact detection. However, it also faces challenges, such as the need to improve recognition accuracy and stability under complex lighting conditions and diverse facial

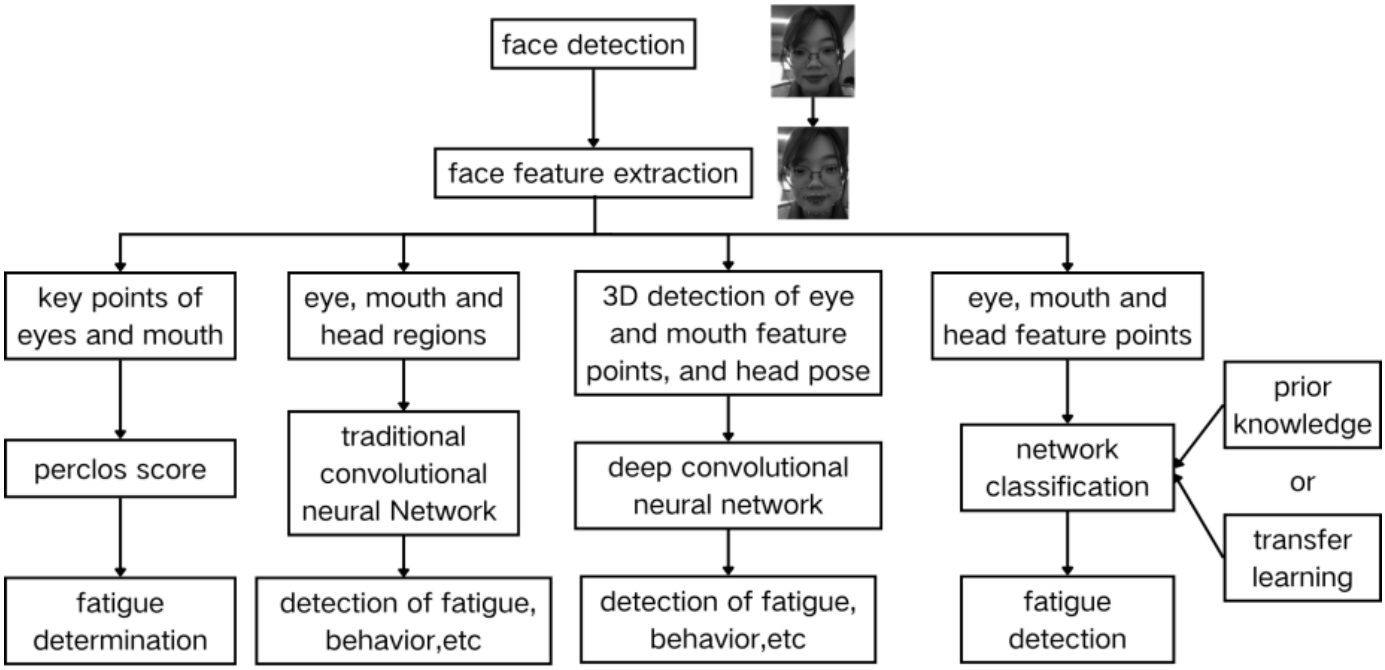

Figure 1. Common methods for detecting driver fatigue.

features, as well as the need to reduce false positives and missed detections.

Fatigue detection technology based on vision mainly relies on extracting features from videos or images to determine the driver's fatigue state. Currently, there are four common implementation methods, as shown in Figure 1. Method 1 utilizes facial keypoint tracking technology to accurately locate and extract key points and then determines fatigue based on threshold values. In 2022, Zhu et al. [15] proposed a multi-threaded optimized Dlib for face detection, which determines fatigue levels based on blinking frequency, eye closure duration (PERCLOS), and yawning frequency. However, this method requires further optimization to handle subtle facial changes and complex lighting conditions. Method 2 analyzes facial features by detecting characteristic expressions and then uses convolutional neural networks (CNN) to recognize feature states. Gu et al. [16] proposed a multi-task hierarchical CNN for facial state detection, followed by multi-scale pooling (MSP-Net) for parameter processing. However, this method suffers from poor generalization ability, leading to significant errors in scenarios involving different angles and facial expressions, making it difficult to accurately recognize fatigue states. Method 3 employs 3D facial keypoint detection or head-pose estimation algorithms to determine head movement patterns. Ran et al. [17] proposed a deep hyperparameterized convolution and attention grid-based 3D feature extraction method, while Li et al. [18] used ConvNet and 3D face reconstruction techniques for detection. However, when the head angle is too large, the tracking performance of the 3D keypoints degrades, and the method lacks consideration for cumulative driving fatigue. Method 4 uses pretrained model parameters for network classification, followed by deep learning models for fatigue state recognition. The deep learning models include CNN and long short-term memory (LSTM) networks proposed by Liu et al. [19], residual network models proposed by Tao et al. [20], and attention networks proposed by Peng et al. [21]. Other models, such as gated recurrent units (GRUs) based on driver eye features and vehicle operation data [22], also exist. However, these methods often struggle to effectively capture long-term dependencies, exhibit insufficient parallel processing capabilities, and have weak robustness in handling noise and missing data, leading to inaccurate detection results [23].

Compared with the above several methods, the work in this study can be regarded as a deep learning framework for multimodal learning, aiming to simultaneously handle object detection and feature analysis tasks in images. This study innovatively

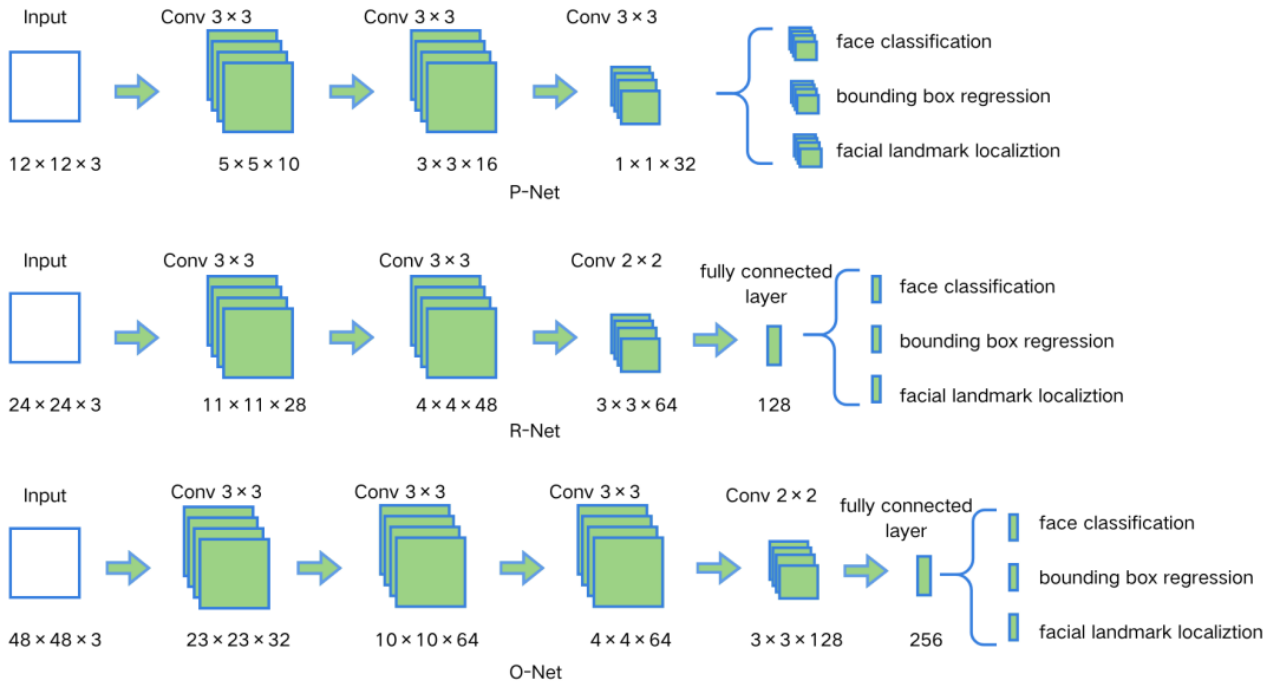

Figure 2. Schematic of MTCNN network structure.

combines the object detection capability of YOLOv8 with the feature representation learning of the Transformer and introduces a multilevel adaptive learning strategy in the model design. This allows the model to leverage its powerful sequence modeling ability to detect driver facial fatigue signs while extracting richer representation information. Furthermore, it can maintain high accuracy and generalization ability when variations in lighting, angles, and facial expressions are encountered.

3 Fatigue Driving Detection Based on the Combination of Transformer and YOLOv8

The installation of cameras in the driver's cabin results in the driver occupying a large number of pixels in an image. To address this, this study employs a method specifically designed for face detection and keypoint localization: the Multi-task Cascaded Convolutional Neural Network (MTCNN). The network structure of the MTCNN consists of a Proposal Network (P-Net), a Refinement Network (R-Net), and an Output Network (O-Net) [24]. P-Net was employed to efficiently propose a large set of face candidate regions. The R-Net filters and refines these candidate boxes and performs bounding box regression. The O-Net further refines the results processed by the P-Net and R-Net, outputting more accurate face classification results, bounding box regression results, and facial keypoint coordinates, achieving precise face detection for subsequent analysis and processing. The cascading structure of MTCNN is shown in Figure 2.

Directly using the eye and mouth feature point positions to detect the driver's fatigue state after feature point detection may result in errors. This research introduces a fatigue detection framework that integrates Transformer with YOLOv8. It utilizes deep learning-based facial region and feature point detection for drivers. Based on the detection results, the states of the eyes and mouth were further analyzed. The Transformer model is then used to learn the dynamic changes in these features in the time series and accurately predict the driver's fatigue state. Finally, by combining the driver's current state information with historical state information, the driver's fatigue level was estimated using the transformer model. The fatigue driving detection process is shown in Figure 3.

3.1 Face State Detection Based on YOLOv8

In fatigue driving detection, deep learning -based face detection technology is critical. YOLOv8 has demonstrated a powerful performance in object detection and feature extraction. With the YOLOv8 model, facial regions, including the eyes and mouth, can be detected quickly and accurately, and the key

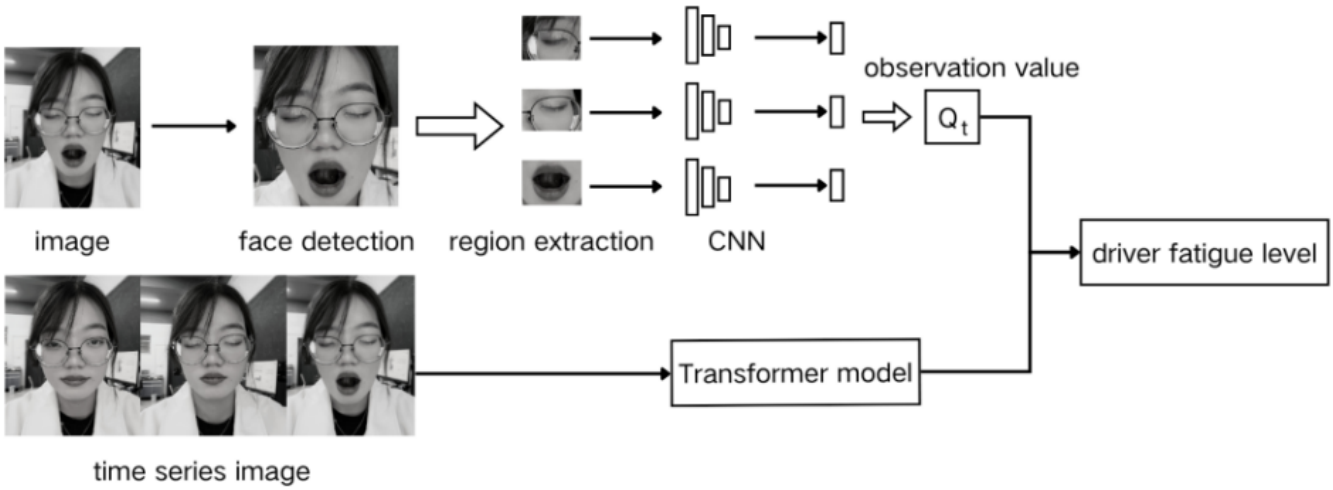

Figure 3. Flowchart of the fatigue driving detection.

point coordinates of these areas can be extracted. Based on the detected eye and mouth regions, further identification is performed, and the corresponding characteristic information is extracted.. YOLOv8 can effectively handle face detection under different lighting conditions and adapt to head angle deviations during driving, ensuring the system's stability and accuracy in complex environments.

3.1.1 Convolutional Neural Network Layers

Face state recognition is a key step in determining the driver's fatigue state, where eye closure and yawning are important indicators of fatigue. This study used the YOLOv8 model for face region detection and state recognition, with custom classification labels "open" and "closed" during data annotation. When processing the model output, the "open" class was mapped to 0, and the "closed" class is mapped to 1, achieving binarization for the binary classification output.

The convolutional neural network of YOLOv8 accomplishes state recognition through multiple modules working collaboratively. The input layer generates multilayer feature maps that fit the network size. The feature extraction layer extracts structural and textural cues from the image, batch normalization helps stabilize the training process, and the activation layer introduces non-linearity to enhance the model's ability to distinguish subtle differences, such as open eyes, closed eyes, and yawning. The feature pyramid network processes features at different scales, ensuring the accurate recognition of target features in complex backgrounds. The SE module focuses on key regions while ignoring irrelevant backgrounds, and the anchor box mechanism helps to locate the eye and mouth regions. The output layer ultimately generates predictions for the eye and mouth states, enabling the real-time monitoring of the driver's fatigue state. The corresponding convolutional neural network structure is shown in Figure 4.

3.1.2 Head Pose Localization

Head pose localization is an important component for assessing the driver's state. When a driver is fatigued, they may frequently nod or maintain a downward head position for extended periods. In contrast, when a driver is distracted, such as by using a phone or drinking water, the head tilt angle may increase, potentially posing a safety risk. As shown in Figure 5, this study selected five feature points from the standard facial model as three-dimensional references. The PnP algorithm maps key points, extracted from the 2D image by YOLOv8, onto a predefined three-dimensional facial model. Subsequently, the rotation matrix and translation matrix of the head are calculated, allowing the determination of the head's rotation angles in three-dimensional space.

Head rotations along the X, Y, and Z axes were expressed in terms of the pitch, yaw, and roll Euler angles to provide a more intuitive representation of head pose. These angle values effectively reflect the driver's attention state and are used as input-embedded features for the transformer model. By combining a multitask learning strategy, the accuracy of head-pose classification was further optimized. Guided by the 3D face model, the detection

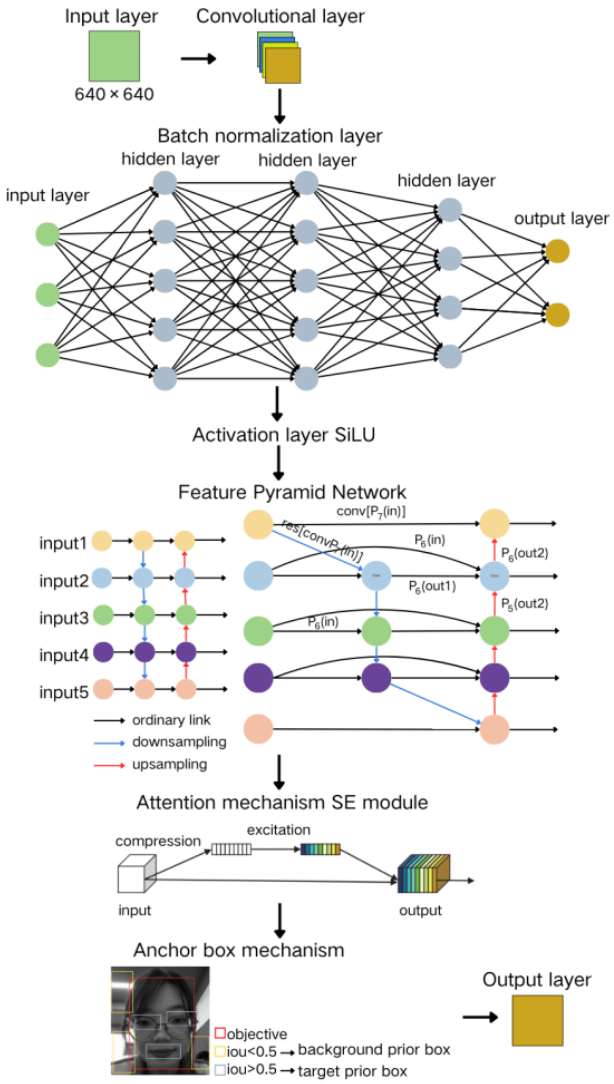

Figure 4. Convolutional neural network structure of the YOLOv8 model.

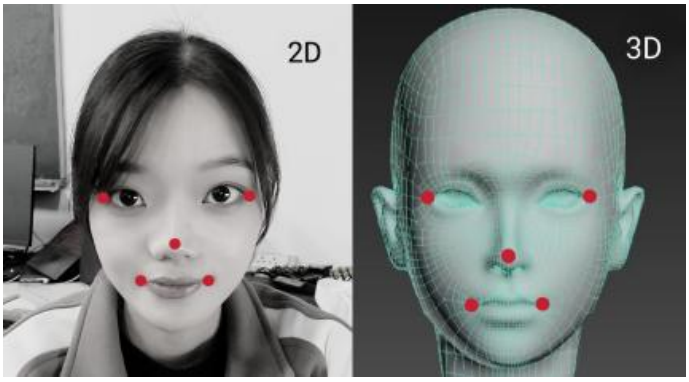

Figure 5. 2D and 3D images of the five key facial points.

results became more stable and intuitive, as shown in Figure 6, achieving an accurate prediction of the head position and motion trends.

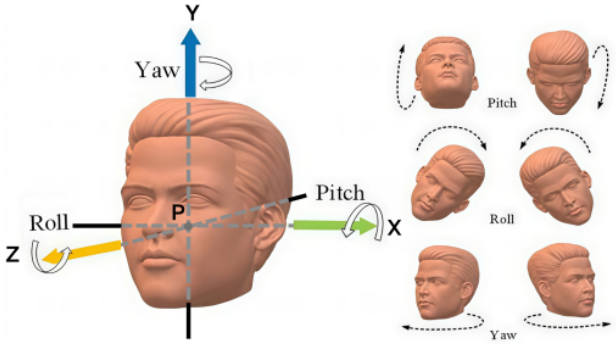

Figure 6. Head pose rotation angles.

Assume point P is located at (X, Y, Z) in the 3D world frame, at (u, v) in the image, and at (x_c, y_c, z_c) in camera coordinates. The f_x and f_y represent the camera's focal lengths in the x and y directions, with units in pixels. The head rotation is encoded by matrix R and the spatial displacement by vector T, with s acting as a scale factor to normalize 3D coordinates onto the 2D image plane. A homogeneous transformation matrix is used to relate the world coordinates to the image coordinates, as given in equation (1).

$$s\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} R & T \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{1}$$

Given the five facial feature points, ten image coordinates, and five 3D coordinates as known quantities, the task is to solve for the three rotation parameters and three translation parameters. This can be achieved by using the least squares method to fit and obtain the best angle estimates. In this study, the PnP algorithm was used to solve this problem, minimizing the projection error to optimize the accuracy of the solution. Figure 7 shows the head pose angle values displayed on the detected image, providing a more intuitive representation of the driver's head pose.

3.2 Fatigue State Recognition Based on Transformer

The YOLOv8 model was used to detect facial keypoints as feature inputs, and the Multi-Head Transformer was employed to learn features and fatigue behavior. However, eye, mouth, and other visible features are still calculated to assist the Transformer in performing temporal analysis and learning the dynamic changes in the driver's fatigue behavior.

Table 1. Fatigue parameter index level.

Indicator Level	PERCLOS (%)	Blink Frequency (times/min)	Yawn Frequency (times/min)	Head Nod Frequency (times/min)
Normal	0-20	>20	0	5
Mild Fatigue	20-40	0.7751	1-2	5-10
Moderate Fatigue	40-60	5-10	2-3	10-15
Severe Fatigue	60	5	3	15

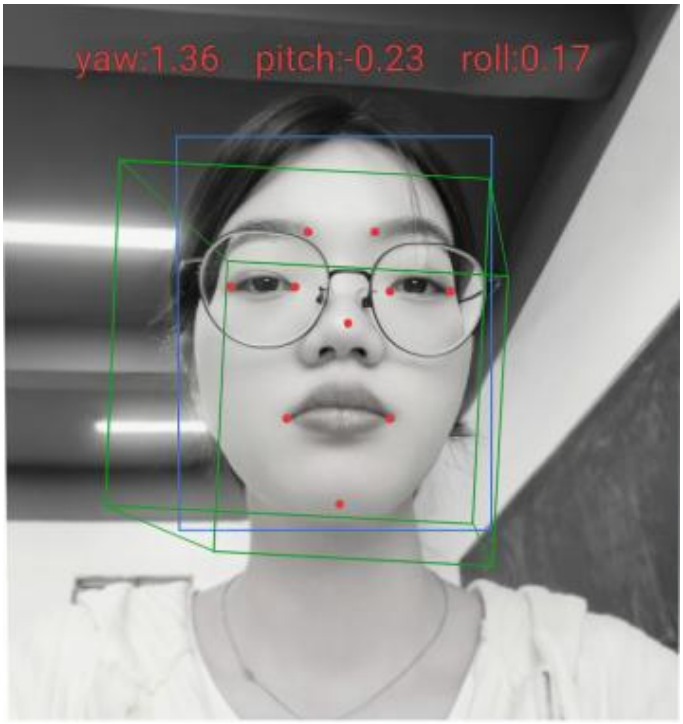

Figure 7. Head pose detection image.

$$\mathrm{EAR} = \frac{\|p_2 - p_6\| + \|p_3 - p_5\|}{2\|p_1 - p_4\|}, \tag{2}$$

In equation (2), EAR represents the Eye Aspect Ratio, and p denotes the eye keypoint coordinates. When the EAR is below the threshold (e.g., 0.25), it is classified as a closed-eye state, as shown in equation (2). Because head rotation affects the calculation of the EAR, the keypoint positions are corrected through head pose estimation to reduce the impact of rotation. The Mouth Aspect Ratio (MAR) is judged in the same way as the EAR. Using a 5-second time window and 200 frames per cycle, if the EAR dropped below the set threshold, it was classified as a closed-eye frame, and the roll was incremented by 1, with the mouth state accounting for a weight of 0.2. When the PERCLOS model score exceeds 0.3, the driver is classified as being in a fatigued state, as shown in equation (3).

$$\mathrm{PERCLOS} = \frac{\mathrm{Roll}_{\mathrm{eye}}}{\mathrm{Roll}} + \frac{\mathrm{Roll}_{\mathrm{mouth}}}{\mathrm{Roll}} \times 0.2, \tag{3}$$

3.2.1 *Head Pose Localization*

The Transformer, built upon a self-attention architecture, has become a core framework in areas like natural language processing and computer vision, demonstrating strong capability in modeling sequential information. It maps input data to a high-dimensional space, capturing the relationships between different time steps, and thus effectively extracts features.

In this study, the driver's fatigue state was determined by multiple factors, including eye closure time (PERCLOS), average blink count (Neye), yawning frequency (Nyawn), and head-down frequency (Nhead). Based on these parameters, the model classifies the driver's fatigue state into four levels: normal, mild, moderate, and severe fatigue, as shown in Table 1.

The core components of the Transformer model can be represented as $T = \{E, P, Q, K, V, H\}$, where the defined parameters Q, K, and V are the query, key, and value matrices, respectively, and the other meanings are as follows.

Input feature embedding (E): Let the input feature sequence be $X = [x_1, x_2, ..., x_T]$, where T is the number of time steps and d is the feature dimension at each time step. Here, x_t represents the feature at the t-th time step (such as N_{yawn}, N_{eye}, and N_{head}).

$$E = \{e_1, e_2, ..., e_T\}, \quad e_i = f(x_i)W^E, \tag{4}$$

In equation (4), $f(x_i)$ is the feature extraction function and $W^E \in \mathbb{R}^{d \times d_{\mathrm{model}}}$ is the linear transformation matrix that embeds the raw features into the model dimension d_{model}.

Positional Encoding (P): By adding temporal positional information, the model can recognize

different time steps and represent the input sequence as

$$z_i = e_i + p_i. \tag{5}$$

Specifically, e_i denotes the feature embedding at time step i, and p_i represents the positional encoding, and z_i is the final input representation at time step i.

Self-Attention Mechanism: Each input feature z_i undergoes a linear transformation to obtain the query, key, and value matrices:

$$Q = ZW^Q, \quad K = ZW^K, \quad V = ZW^V. \tag{6}$$

where $W^Q, W^K, W^V \in \mathbb{R}^{d_{\text{model}} \times d_k}$ are weight matrices.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{7}$$

where Q is the query matrix that represents the information of the current frame. The key matrix K contains information from all historical frames. The value matrix V holds the feature values for the historical frames. The softmax function is used to compute the weight of each historical state.

As shown in equation (6), the multi-head attention mechanism H extends the self-attention mechanism into multiple parallel attention heads to capture fatigue patterns in different **subspaces**.

$$H = \text{Concat}(\text{head}_1, \text{head}_2, ..., \text{head}_h)W^o, \tag{8}$$

In equation (6), each head represents an independent computation of the self-attention mechanism, h denotes the number of attention heads, and W^o is the weight matrix for the linear transformation.

Feed Forward Network (FFN): After the multi-head attention mechanism, the output at each time step undergoes a nonlinear transformation through a feed-forward network. The FFN formula is shown in equation (7).

$$\text{FFN}(z) = \text{ReLU}(zW_1 + b_1)W_2 + b_2, \tag{9}$$

Output Classification (C): An additional output layer is added after the Transformer encoder, where the output sequence is represented as

$$\hat{Z} = \{Z_1, Z_2, ..., Z_T\}.$$

Through a linear layer and softmax activation function, the extracted features are mapped to a probability distribution of fatigue levels:

$$\hat{y} = \text{softmax}\left(Z_T W^c + b^c\right), \tag{10}$$

In equation (8), the classification layer's weight matrix is represented by W^c, and $\hat{y}$ shows the fatigue-levels probability distribution.

After training, the output value $\hat{y}$ can be used to determine the fatigue level, such as normal, mild, moderate,

3.2.2 *Parameter Estimation and Optimization of Transformer*

As shown in equation (9), the parameter estimation of the Transformer outputs the predicted probability distribution of fatigue levels through forward propagation. Then, based on the true label y and the model's predicted $\hat{y}$, the cross-entropy loss is calculated to optimize the classification task:

$$L(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij} \log \hat{y}_{ij}, \tag{11}$$

In the equation (9): N is the number of samples. C represents the number of fatigue level categories (normal, mild, moderate, severe fatigue). y_{ij} is the one-hot encoding of the true labels. $\hat{y}_{ij}$ is the predicted probability distribution of the model.

The implementation of backpropagation follows the chain rule, propagating the gradient of the loss function layer by layer to each parameter. An optimization algorithm (such as Adam or SGD) is then used to update the parameters.

The model parameters (such as W^E, W^O, W^K, W^V) are denoted as θ, the learning rate is η, and $\nabla_\theta L(\theta)$ represents the gradient under the current parameters. The model parameters θ are defined in equation (10):

$$\theta^{(t+1)} = \theta^{(t)} - \eta \cdot \nabla_\theta L(\theta^{(t)}), \tag{12}$$

As shown in equation (11), L_{reg} regularization is applied to constrain the weights in order to prevent model overfitting and improve training stability.

Dropout randomly drops a portion of neurons during the training process to increase the model's robustness.

$$L_{reg} = L + \lambda \sum \|\theta\|^2, \tag{13}$$

During the parameter estimation process, the model's performance needs to be evaluated at the end of each epoch by calculating the loss and accuracy on the validation set to avoid overfitting. The model's prediction accuracy and generalization ability across different fatigue levels are assessed. Finally, based on the predicted distribution for each sample in the validation set, the class with the highest probability is selected as the fatigue level.

4 Experimental Results and Analysis

The experiments were conducted in a software environment with the Windows 11 operating system, and the hardware environment consisted of an AMD Ryzen 9 7950X processor running at 4.5GHz, an NVIDIA GeForce RTX 4090 graphics processor with 24GB of VRAM, and 32GB of RAM.

The dataset used for this study was composed of images and videos selected and preprocessed from the YawDD, Driver Face Detection Dataset, and a self-collected dataset. The self-collected video dataset was recorded under simulated daytime and nighttime driving conditions. It contains 3,200 annotated frames from five drivers, covering eye-open, eye-closed, yawning, and normal states. This dataset contains 16,000 images depicting closed eyes, yawning, using a phone, drinking water, and normal driving behavior with open eyes.

The dataset was partitioned by random selection into 13,500 training samples, 1,400 validation samples, and 1,100 testing samples. The dataset labels are divided into four fatigue levels: normal, mild, moderate, and severe fatigue. When evaluating the driver's fatigue state, head pose serves as an important reference indicator, especially evident in behaviors such as head-down or head-turning. In this study, facial keypoint extraction combined with a head pose estimation algorithm was applied to each frame to assist in fatigue state assessment. Figure 8 presents the head pose estimation results of several drivers from the YawDD dataset under different fatigue levels, where behaviors such as head-down and frequent head-turning become more apparent under severe fatigue, confirming the strong correlation between head pose and fatigue-related behaviors.

The loss function is a key component in evaluating model performance, as it measures the difference between the model's predicted output and the actual target. As shown in Figure 9, the box loss, classification loss, and objective loss for both the training and validation sets show a decreasing trend and eventually stabilize, indicating that the model's performance improves gradually during training and enhances its ability to adapt to and learn from the data, while maintaining a high recall rate and stable precision, with the balance between these two metrics influencing the model's practicality and accuracy; furthermore, the average detection accuracy at different IoU thresholds also demonstrates the model's excellent target detection performance.

In evaluating fatigue driving detection models, the F1 score is a crucial metric combining precision and recall, with values closer to 1 indicating a better trade-off between the two. The F1 scores for "Yawn" and "Close" are relatively high, especially in the mid-confidence threshold range, indicating that the model performs well in detecting these behaviors. However, the F1 score for "noYawn" is relatively low, possibly due to the lack of distinct visual features, making it harder for the model to distinguish. As shown in Figure 10, the F1 score reaches its peak at approximately 0.217 confidence threshold, with a maximum value of 0.55. This threshold balances false positives and false negatives, allowing the model to achieve relatively optimal performance.

By evaluating the F1 score and confidence threshold, the target detection results are obtained and then fused with the global features extracted by the Transformer model to calculate the probability of the driver being in different fatigue states. A higher state probability leads to a higher chance of the current instance being classified accordingly. Specifically, the F1 score offers a combined assessment of precision and recall for the detection model, while confidence quantifies the trustworthiness of its outputs. The combination of these two metrics provides a comprehensive quantification of detection performance and lays a solid foundation for subsequent fatigue state analysis.

To validate the robustness of the proposed Multi-Head Transformer with Adaptive Weighted Loss framework, integrated with YOLOv8 for facial region detection, experiments were conducted under varying illumination conditions (daylight, dusk, and nighttime) and occlusion scenarios (sunglasses, hand blocking, and steering wheel interference). The

Figure 8. Head pose estimation results of drivers under different fatigue levels.

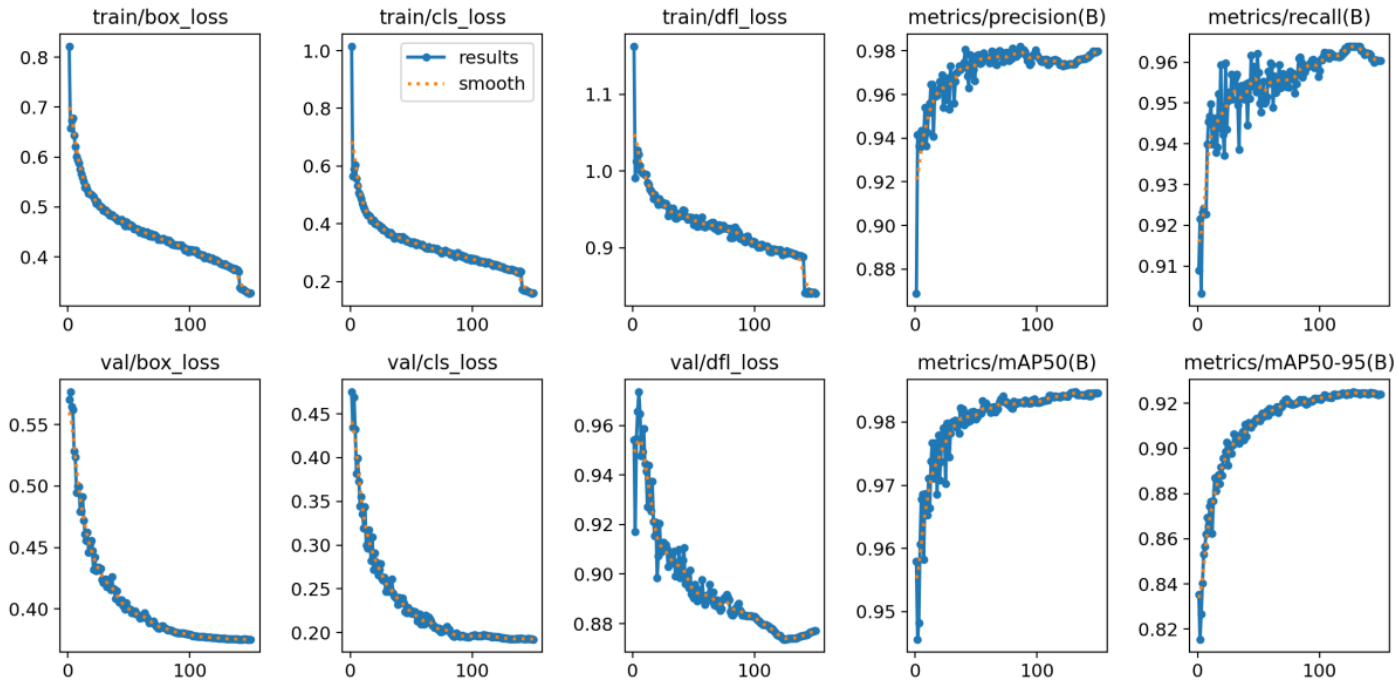

Figure 9. Performance evaluation of the feature training process.

results show that the model maintained detection accuracies of 95.8%, 92.1%, and 90.3% across the three illumination levels, exhibiting only a slight performance degradation compared with the standard scenario. In the occlusion tests, the attention mechanism of the Multi-Head Transformer effectively captured visible facial cues, achieving an average F1-score of 0.85, which demonstrates strong robustness and adaptability to real-world environments. As illustrated in Figure 11, the confidence–F1 score distributions remain stable under different illumination and Occlusion conditions, indicating that the proposed framework possesses strong adaptability to brightness variation and partial facial occlusion. These findings confirm that the integration of YOLOv8 and the Multi-Head Transformer with Adaptive Weighted Loss effectively mitigates the influence of lighting and occlusion, outperforming baseline CNN-based models whose accuracy typically drops by more than 10% under similar conditions.

This study employs a dataset containing a large number of driver facial images with annotated fatigue states serving as the foundation for algorithm training and validation, with performance analysis and comparative studies confirming the method's

Table 2. Comparison of different algorithms for driver fatigue detection.

Reference	Algorithm Model	Detection Features	Modal Type	Contact	Accuracy
Alameen [25]	3DCNN, LSTM	Eyes, Mouth	Vision	No	93.00%
Debsi [26]	ME-YOLOv8	Eyes, Mouth	Vision	No	92.80%
Azmi [27]	Vision Transformer	Eyes, Mouth,Head	Vision	No	93.07%
Xu [28]	Swin Transformer	Eyes, Mouth,Head	Vision	No	95.39%
Ours	**YOLOv8, Multi-Head Transformer**	**Eyes, Mouth, Head**	Vision	**No**	**95.50%**

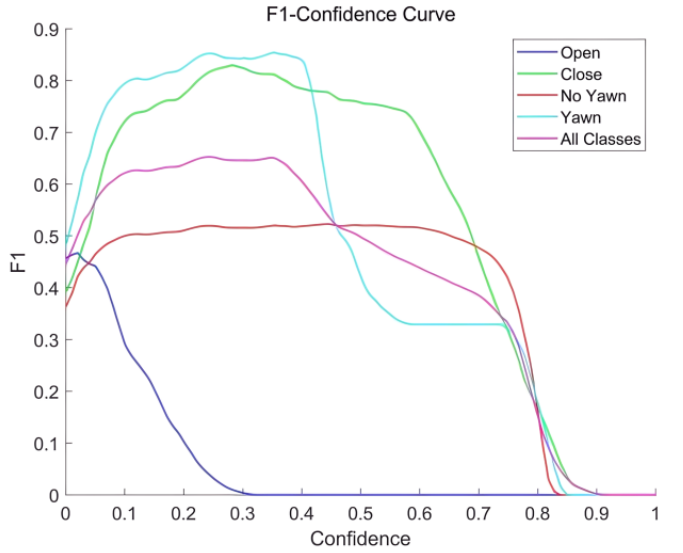

Figure 10. Relationship curve between F1 score of driving behavior and confidence threshold.

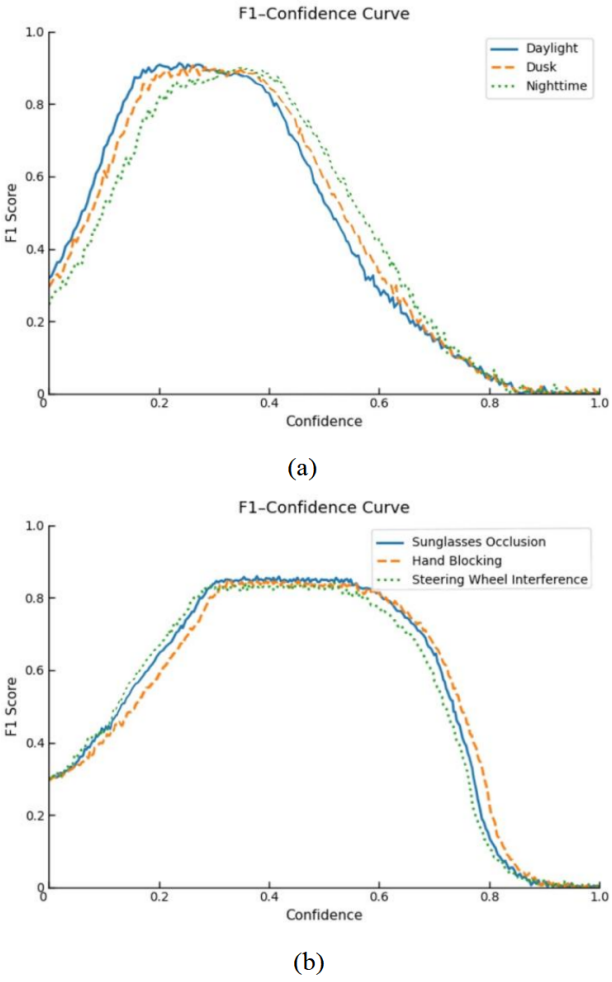

Figure 11. F1-Confidence Curves under (a) Different Illumination and (b) Occlusion Conditions.

effectiveness and robustness.

The proposed Multi-Head Transformer with Adaptive Weighted Loss integrated with YOLOv8 achieves an accuracy of 95.5% in driver fatigue detection, marking a 7.2% improvement over the traditional YOLOv8 model, which attains 88.3% accuracy. This enhancement primarily results from the multi-head Transformer's ability to extract and analyze fatigue-related temporal and spatial dependencies, enabling the model to capture subtle behavioral cues such as gradual eye closure, yawning, and micro facial movements that often precede fatigue. Furthermore, the Adaptive Weighted Loss function dynamically adjusts the contribution of different fatigue indicators during training, improving robustness under varying lighting conditions and driver postures.

In terms of efficiency, the proposed network achieves a real-time detection speed of approximately 60 frames

Table 3. Analysis of the contribution of each module to the model's overall performance.

Model Configuration	Transformer	Adaptive Weighted Loss	Accuracy	F1-score
YOLOv8 (Baseline)	×	×	88.30%	0.47
YOLOv8 + Transformer	✓	×	93.60%	0.52
YOLOv8 + Transformer + Adaptive Weighted Loss	✓	✓	95.50%	**0.55**

per second on an NVIDIA GeForce RTX 4090 GPU with 24GB of VRAM. By incorporating the Transformer's parallel attention mechanism, redundant feature computations are significantly reduced, which enhances model efficiency and lays the foundation for lightweight, real-time fatigue detection in embedded or in-vehicle systems. This study also compares the proposed method with several fatigue detection models representing the latest state-of-the-art (SOTA) achievements in the field of vision-based driver fatigue detection. The 3D-CNN + LSTM model employs a traditional spatio-temporal hybrid framework capable of capturing both spatial image features and temporal dependencies simultaneously. The ME-YOLOv8 model is a lightweight, multi-feature-enhanced real-time detection network that reflects the current SOTA level of real-time visual detection. The Vision Transformer is a typical Transformer-based visual model that learns global contextual dependencies through self-attention mechanisms, while the Swin Transformer adopts a hierarchical shifted-window attention structure and is considered one of the most powerful spatio-temporal visual Transformer frameworks to date. All compared models are vision-based and non-contact, consistent with the experimental scenario of this study, ensuring fair comparison under the same data modality. By integrating traditional spatio-temporal modeling methods with the latest Transformer architectures, these benchmark models establish a comprehensive evaluation framework, demonstrating that the proposed YOLOv8 combined with the Multi-Head Transformer maintains high accuracy, adaptability, and robustness even under complex driving conditions. The detailed comparison results are shown in Table 2.

To quantify the contribution of each module in the proposed framework, an ablation study was conducted by selectively removing the Multi-Head Transformer and the Adaptive Weighted Loss components from the full architecture. As shown in Table 3, the baseline YOLOv8 model achieved an accuracy of 88.3% and an F1-score of 0.47. Incorporating the Multi-Head Transformer improved temporal feature modeling, increasing accuracy to 93.6% and F1-score to 0.52. When the Adaptive Weighted Loss was additionally applied, the model achieved the best performance with an accuracy of 95.5% and an F1-score of 0.55.

These results clearly demonstrate that the Transformer significantly enhances temporal dependency learning, while the Adaptive Weighted Loss function further balances inter-class contributions and strengthens robustness under varying driving conditions. The ablation study confirms that each designed component plays a distinct and complementary role in enhancing the overall detection performance.

5 Conclusion

Driver fatigue is a dynamic process transitioning from wakefulness to fatigue. This study proposes a driver fatigue detection method that integrates a Transformer with YOLOv8. To overcome the limitations of conventional approaches in modeling temporal dependencies, YOLOv8 is employed to efficiently detect facial regions, including the eyes and mouth, while the Transformer's self-attention mechanism is utilized to capture temporal dependencies and improve discriminative capability. The evaluation results indicate that the method attains recognition rates of 92.3% for mild fatigue, 89.7% for moderate fatigue, and 87.5% for severe fatigue. The average per-frame inference time is approximately 20 ms, meeting real-time processing requirements. The self-attention mechanism exhibits strong capabilities in global feature modeling, effectively capturing complex relationships among facial features. This is particularly beneficial for detecting subtle expressions such as blinking and yawning, thereby mitigating the limitations of single-frame detection inherent to YOLOv8. Furthermore, the Transformer enables robust temporal modeling for recognizing fatigue-related behavioral transitions. Moreover, Transformer exhibits powerful multi-modal integration capabilities, allowing fatigue detection to incorporate not only visual information but also driving-related indicators such as steering behavior and vehicle dynamics, which further enhances reliability. In challenging conditions, such as varying lighting or complex environments, other detection methods may struggle to extract accurate features. However, the combination of Transformer and YOLOv8 can effectively weight features at different time points, maintaining robust recognition performance and enhancing system accuracy and reliability, ultimately providing more reliable technological support for driving safety.

Nevertheless, challenges such as vehicle vibrations and lighting variations may still affect detection stability. Greater emphasis should be placed on reinforcing robustness and interference immunity in complex scenarios in future developments. Additionally, integrating physiological signals and other multi-modal data should be explored to further

enhance detection accuracy, providing a more reliable technological foundation for intelligent driving systems.

These findings support the JRS mission by advancing the scientific basis for early fatigue detection and offering practical tools that can be integrated into intelligent vehicles to prevent fatigue-related crashes and mitigate their consequences on public health.

Data Availability Statement

The datasets used in this study are publicly available at https://ieee-dataport.org/open-access/yawdd-yawning-detection-dataset, and https://www.nexdata.ai/datasets/computervision/1588. The self-collected dataset consists of driving video recordings from five drivers, all of whom provided informed consent. Due to privacy protection concerns, this dataset cannot be publicly shared.

Funding

This work was supported in part by the Natural Science Research Project of Higher Education Institutions in Jiangsu Province of China under Grant 24KJA510002; in part by the Postgraduate Research & Practice Innovation Program of Jiangsu Province under Grant SJCX25_2194; in part by the Postgraduate Science and Technology Innovation Program of Huaiyin Institute of Technology under Grant HGYK202516.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

This work does not contain any studies with human participants or animals performed by any of the authors that would require ethical approval. Although a self-collected dataset was used, all participants provided written informed consent, and all data were anonymized to ensure privacy.

References

[1] Zhang, G., Yau, K. K., Zhang, X., & Li, Y. (2016). Traffic accidents involving fatigue driving and their extent of casualties. *Accident Analysis & Prevention, 87*, 34-42. [CrossRef]

[2] Zhang, H., Ni, D., Ding, N., Sun, Y., Zhang, Q., & Li, X. (2023). Structural analysis of driver fatigue behavior: A systematic review. *Transportation Research Interdisciplinary Perspectives, 21*, 100865. [CrossRef]

[3] Meng, F., Li, S., Cao, L., Li, M., Peng, Q., Wang, C., & Zhang, W. (2015). Driving fatigue in professional drivers: a survey of truck and taxi drivers. *Traffic injury prevention, 16*(5), 474-483. [CrossRef]

[4] Jiao, Y., Zhang, C., Chen, X., Fu, L., Jiang, C., & Wen, C. (2023). Driver fatigue detection using measures of heart rate variability and electrodermal activity. *IEEE Transactions on Intelligent Transportation Systems, 25*(6), 5510-5524. [CrossRef]

[5] Wang, L., Song, F., Zhou, T. H., Hao, J., & Ryu, K. H. (2023). EEG and ECG-based multi-sensor fusion computing for real-time fatigue driving recognition based on feedback mechanism. *Sensors, 23*(20), 8386. [CrossRef]

[6] Shi, J., & Wang, K. (2023). Fatigue driving detection method based on Time-Space-Frequency features of multimodal signals. *Biomedical Signal Processing and Control, 84*, 104744. [CrossRef]

[7] Du, G., Wang, H., Su, K., Wang, X., Teng, S., & Liu, P. X. (2022). Non-interference driving fatigue detection system based on intelligent steering wheel. *IEEE Transactions on Instrumentation and Measurement, 71*, 1-11. [CrossRef]

[8] Sukumar, N., & Sumathi, P. (2024). An improved lane detection and lane departure warning framework for adas. *IEEE Transactions on Consumer Electronics, 70*(2), 4793-4803. [CrossRef]

[9] Bonfati, L. V., Mendes Junior, J. J., Siqueira, H. V., & Stevan Jr, S. L. (2022). Correlation analysis of in-vehicle sensors data and driver signals in identifying driving and driver behaviors. *Sensors, 23*(1), 263. [CrossRef]

[10] Li, R., Chen, Y. V., & Zhang, L. (2021). A method for fatigue detection based on Driver's steering wheel grip. *International Journal of Industrial Ergonomics, 82*, 103083. [CrossRef]

[11] Xiao, W., Liu, H., Ma, Z., Chen, W., & Hou, J. (2024). FPIRST: fatigue driving recognition method based on feature parameter images and a residual Swin Transformer. *Sensors, 24*(2). [CrossRef]

[12] Cao, S., Feng, P., Kang, W., Chen, Z., & Wang, B. (2025). Optimized driver fatigue detection method using multimodal neural networks. *Scientific Reports, 15*(1), 12240. [CrossRef]

[13] Li, Z., Zhao, X., Wu, F., Chen, D., & Wang, C. (2024). A lightweight and efficient distracted driver detection model fusing convolutional neural network and vision transformer. *IEEE Transactions on Intelligent Transportation Systems*. [CrossRef]

[14] Chen, J., Zhang, Q., Chen, J., Wang, J., Fang, Z., Liu, Y., & Yin, G. (2024). A Driving Risk Assessment Framework Considering Driver's Fatigue State and Distraction Behavior. *IEEE Transactions on Intelligent Transportation Systems*. [CrossRef]

[15] Zhu, T., Zhang, C., Wu, T., Ouyang, Z., Li, H., Na, X., ... & Li, W. (2022). Research on a real-time driver fatigue detection algorithm based on facial video sequences. *Applied Sciences, 12*(4), 2224. [CrossRef]

[16] Gu, W. H., Zhu, Y., Chen, X. D., He, L. F., & Zheng, B. B. (2018). Hierarchical CNN-based real-time fatigue detection system by visual-based technologies using MSP model. *IET Image Processing, 12*(12), 2319-2329. [CrossRef]

[17] Ran, X., He, S., & Li, R. (2023). Research on fatigued-driving detection method by integrating lightweight yolov5s and facial 3d keypoints. *Sensors, 23*(19), 8267. [CrossRef]

[18] Li, Y., Sun, B., Wu, T., & Wang, Y. (2016, September). Face detection with end-to-end integration of a convnet and a 3d model. In *European Conference on Computer Vision* (pp. 420-436). Cham: Springer International Publishing. [CrossRef]

[19] Liu, M. Z., Xu, X., Hu, J., & Jiang, Q. N. (2022). Real time detection of driver fatigue based on CNN-LSTM. *IET Image Processing, 16*(2), 576-595. [CrossRef]

[20] Tao, S., Li, Y., Huang, Y., & Lan, X. (2021, March). Face detection algorithm based on deep residual network. In *Journal of Physics: Conference Series* (Vol. 1802, No. 3, p. 032142). IOP Publishing. [CrossRef]

[21] Peng, B., Zhang, Y., Wang, M., Chen, J., & Gao, D. (2023). TA-MFFNet: Multi-feature fusion network for EEG analysis and driving fatigue detection based on time domain network and attention network. *Computational Biology and Chemistry, 104*, 107863. [CrossRef]

[22] Li, D., Zhang, X., Liu, X., Ma, Z., & Zhang, B. (2023). Driver fatigue detection based on comprehensive facial features and gated recurrent unit. *Journal of Real-Time Image Processing, 20*(2), 19. [CrossRef]

[23] Hassan, O. F., Ibrahim, A. F., Gomaa, A., Makhlouf, M. A., & Hafiz, B. (2025). Real-time driver drowsiness detection using transformer architectures: a novel deep learning approach. *Scientific Reports, 15*(1), 17493. [CrossRef]

[24] Khan, S. S., Sengupta, D., Ghosh, A., & Chaudhuri, A. (2024). MTCNN++: A CNN-based face detection algorithm inspired by MTCNN. *The Visual Computer, 40*(2), 899-917. [CrossRef]

[25] Alameen, S. A., & Alhothali, A. M. (2023). A Lightweight Driver Drowsiness Detection System Using 3DCNN With LSTM. *Computer Systems Science & Engineering, 44*(1). [CrossRef]

[26] Debsi, A., Ling, G., Al-Mahbashi, M., Al-Soswa, M., & Abdullah, A. (2024). Driver distraction and fatigue detection in images using ME-YOLOv8 algorithm. *IET Intelligent Transport Systems, 18*(10), 1910-1930. [CrossRef]

[27] Azmi, M. M. B. M., & Zaman, F. H. K. (2024, May). Driver drowsiness detection using vision transformer. In *2024 IEEE 14th Symposium on Computer Applications & Industrial Electronics (ISCAIE)* (pp. 329-336). IEEE. [CrossRef]

[28] Xu, M., Zhan, A., Wu, C., & Wang, Z. (2025). A Novel Driver Fatigue Detection Method Based on Dual-Stream Swin-Transformer. *IEICE Transactions on Information and Systems, 2024EDL8094*. [CrossRef]

Ling Huang M.Eng., Huaiyin Institute of Technology, has published one SCI (Q4) paper and one EI-indexed paper. She has received several awards, including the First Prize in the 16th Lanqiao Cup EDA Design and Development Provincial Competition, the Second Prize in the 20th China Graduate Electronic Design Competition (Provincial Level), the First Prize in the Jiangsu Provincial Graduate Mathematical Modeling Scientific Innovation Competition, and the Third Prize in the 19th China Graduate Electronic Design Competition (Provincial Level). (Email: 19816096061@163.com)

Shifeng Li M.Eng., Huaiyin Institute of Technology, has participated in one scientific research project and was awarded the Second-Class Scholarship of the university twice in 2023. He also received the Second Prize in the 16th Lanqiao Cup EDA Design and Development Provincial Competition, the Second Prize in the 19th China Graduate Electronic Design Competition (Provincial Level), and the Second Prize in the 20th China Graduate Electronic Design Competition (Provincial Level). (Email: 17736736790@163.com)

Yaxin Man M.Eng., Huaiyin Institute of Technology, holds a Master's degree in Engineering. His academic training focuses on transportation, traffic information and control, and related fields, with research interests in deep learning-based image processing and intelligent transportation systems. During his postgraduate studies, he actively participated in scientific research projects and academic activities, demonstrating solid professional knowledge and practical skills. (Email: 15866676685@163.com)

Xiaoyan Wang Ph.D. in Engineering from Nanjing University of Aeronautics and Astronautics, is an Professor at Huaiyin Institute of Technology. She also serves as a Jiangsu Provincial Science and Technology Associate, an Excellent Graduation Design Team Instructor of Jiangsu Province, and a leading talent of the "533 Talent Program" in Huai'an. She has long been engaged in research on photoelectric detection, with a focus on spectral detection and analysis, as well as signal and

information processing. She has presided over multiple research projects funded by the National Natural Science Foundation of China, the Ministry of Science and Technology's Spark Program, Jiangsu Province Industry–University–Research Cooperation, and major university projects, and has participated in more than 10 provincial or ministerial-level projects. She has published over 20 papers, been granted 2 invention patents and 8 software copyrights, and contributed to one textbook and one monograph. (Email: wxygxy@163.com)

Xiu Tang M.Eng., Huaiyin Institute of Technology. She participated in a scientific research project, undertaking key technical research and experimental work, demonstrating solid research capabilities. In addition, she has been granted three invention patents, showcasing her outstanding innovation ability. She also achieved the Third Prize in the 15th Lanqiao Cup Provincial Competition and the Third Prize in the 16th Lanqiao Cup EDA Design and Development Provincial Competition, reflecting her strong professional competence. (Email: 2474855223@qq.com)

Rendong Ji received his Ph.D. in Engineering from Nanjing University of Aeronautics and Astronautics and is currently a Professor at Huaiyin Institute of Technology. He also serves as a member of the Optical Testing Professional Committee of the Chinese Optical Society and actively participates in various academic activities and professional organizations related to optical engineering and measurement technologies. In recognition of his academic contributions and professional achievements, he has been selected as a leading talent in Huai'an City's "533 Talent Program." He has presided over more than 10 national and provincial-level research projects and has also undertaken over 20 industry-funded research projects in collaboration with enterprises. His research achievements have resulted in the publication of more than 20 papers indexed by SCI and EI, and he has been granted 6 invention patents. In addition, he has led a teaching reform project funded by the Ministry of Education, contributing to the improvement of engineering education and curriculum development. He has also published one academic monograph and one textbook in his research field. (Email: jrdgxy@163.com)