Vol. 20 (2026): Proceedings of 2026 International Conference on Materials Science, Energy Engineering and Automation Control (ICMSEAC 2026)
E-ISSN: 3105-0522 P-ISSN: 3105-0514
DOI: https://doi.org/10.70267/mseac.20264155

A Driver Fatigue Detection Mechanism Based on a Lightweight RGB-IR Fusion Algorithm

Weixuan Shao*

School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China

**Corresponding author: Weixuan Shao.*

Abstract

To address the degradation of feature representation in a single RGB modality under low-light conditions, and the lack of texture and semantic information in a single IR modality, this paper proposes a lightweight fatigue detection method based on RGB-IR dual-modal fusion. This method realizes cross-modal information fusion at the feature layer by building a dual-stream feature extraction framework, and introduces a feature alignment mechanism to improve the semantic consistency of RGB and IR; At the same time, it combines lightweight network design and unidirectional feature aggregation strategy to reduce the complexity of model calculation, and improve the real-time performance of the system through asynchronous dual-stream inference mechanism. The experimental results show that the dual-flow model based on the dual lightweight strategy shows better detection robustness under extremely dark light conditions: mAP@0.5 is upgraded from 0.314 of the single-stream RGB model to 0.547, mAP@0.5:0.95 increased from 0.200 to 0.359; Compared with the dual-flow RGB IR baseline model, the number of parameters and FLOPs decreased by 37.45% and 24.57% respectively. Under the asynchronous inference mechanism, the effective frame rate of the system reaches 72.09 FPS. It can be seen that the dual lightweight strategy achieves a better balance between detection performance, model complexity and real-time performance, providing a feasible solution for in-vehicle driver fatigue detection in low-light environments.

Keywords

driver fatigue detection, YOLO, RGB-IR, cross-modal fusion, intelligent transportation systems

1. Introduction

With the development of Intelligent Transportation Systems (ITS) and autonomous driving technology, driver fatigue has become an important factor affecting road traffic safety. Fatigue can lead to reduced attention and delayed reactions, which significantly increases the risk of traffic accidents. Therefore, real-time and reliable detection of driver fatigue status in complex driving environments has become an important research direction of Driver Monitoring System (DMS) [1, 2]. The existing driver fatigue detection methods mainly include methods based on physiological signals, vehicle behavioral features and computer vision. Methods based on physiological signals usually monitor the driver's status through EEG, ECG and other sensors. Although the detection accuracy is high, there are problems such as highly intrusive and inconvenient to wear [3, 4]. The method based on vehicle behavior realizes fatigue judgment by analyzing the dynamic information

of the steering wheel angle, lane offset and other vehicles, but it is susceptible to the influence of the environment and driving habits, and its robustness is limited [5]. In contrast, the computer vision-based method can stably obtain the driver's facial status information through the camera and realize end-to-end feature learning in combination with the deep learning model, so it has gradually become the mainstream research direction [6, 7]. However, most of the existing research focuses on the modeling of a single RGB image. The ability of the model to extract the tiny features of the eyes and mouth in low-light driving scenarios is often affected by light degradation [8]. Compared with RGB images, IR images stably retain the key contour information of the driver's face in low light, which can complement RGB images. Therefore, the introduction of RGB-IR cross-modal fusion mechanism is expected to improve the robustness of fatigue detection in low light conditions [9, 10]. However, cross-modal fusion will additionally introduce problems such as modal semantic inconsistency, high complexity of fusion computing and insufficient real-time deployment ability, limiting its application in in-vehicle embedded scenarios [11, 12].

In view of the problems of RGB feature degradation, large cross-modal fusion computing overhead and insufficient real-time deployment capacity in driver fatigue detection under low light conditions. This paper proposes a driving fatigue detection method based on lightweight RGB-IR cross-modal fusion, and conducts research on cross-modal feature coordination, network lightweight design and real-time reasoning mechanism. The overall framework adopts RGB-IR dual-stream architecture to carry out feature extraction and multi-scale semantic modeling for different modalities respectively, and complete cross-modal information fusion at the feature layer; At the same time, the cross-modal feature alignment mechanism is introduced to alleviate the problem of semantic shift between different modals. In terms of network structure design, combined with local Backbone lightweight and one-way feature aggregation structure, the dual-flow detection network is lightweight and optimized to reduce redundant computing overhead; In addition, in view of the data synchronous waiting problem in the process of dual-modal reasoning, an asynchronous dual-flow reasoning mechanism is designed to improve the real-time processing ability of the system and the adaptability of on-board deployment.

The main contributions of this article are as follows:

(1) An RGB-IR dual-flow feature fusion framework is proposed to enhance the expression ability of fatigue features under low light conditions through cross-modal feature interaction and semantic alignment mechanism, and improve the robustness of detection in low-light environments.

(2) Design a lightweight network structure and unidirectional feature aggregation strategy, while reducing the redundancy of cross-modal feature fusion computing, while taking into account the dual-mode feature expression ability and model complexity, and realize on-board-side-friendly lightweight deployment.

(3) The asynchronous dual-stream reasoning mechanism is proposed to reduce the delay expense caused by cross-modal synchronous waiting by decoupling the dual-modal data reading and model reasoning process, thus improving the real-time performance of the system.

In order to facilitate reproduction and experimental verification, the relevant code of this article has been open sourced to GitHub: https://github.com/swx0721/Lightweight-RGBIR-Fatigue-Detection.

2. Related Technologies

In the driving fatigue detection method based on computer vision, the YOLO single-stage object detection algorithm is widely used [13]. In order to meet the scenario of limited computing power on the vehicle, and introducing methods such as depthwise separable convolution, Ghost module and lightweight CSP structure, researchers reduce the complexity of model calculation by improving the backbone network and feature extraction module. However, because the lightweight structure will compress the number of channels, reduce the interaction of cross-channel features and reduce the extraction ability of deep semantics, the model's feature representation of small areas such as eyes and mouth is limited [14]. By introducing the channel and spatial attention mechanism or multi-scale feature fusion strategy, enhance the model's feature representation of the local area of the driver's face to alleviate the problem of visual degradation under low light conditions [15, 16]. However, this kind of method still mainly relies on a single RGB visual modality, which is prone to information loss and feature instability in extreme low-light scenarios. Therefore, it is difficult to meet the robust needs of driving fatigue detection in low-light environments by relying only on single-mode RGB information. The

introduction of IR modes that can stably characterize the target contour information under low light has become an important research direction at present.

The existing RGB-IR detection research mainly adopts cross-modal information interaction methods such as input-level fusion, feature-level fusion and decision-level fusion to realize the collaborative feature expression between RGB and IR modes. Input-level fusion usually uses channel concatenation to directly input RGB and IR images into the detection network. Although the structure is simple and the computational complexity is low, due to the lack of modal difference modeling ability, it is easy to introduce noise interference in the low-level feature stage. Decision-level fusion completes RGB and IR modal reasoning respectively, and weighted voting or bounding-box fusion is carried out at the output layer, which can reduce intermodal information interference to a certain extent. However, due to the lack of cross-modal feature interaction, it is difficult to make full use of the complementary information between dual modalities [17]. In contrast, feature-level fusion extracts RGB and IR features respectively through dual-stream networks, and conducts cross-modal information interaction in the back layer or Neck stage in the Backbone, which has become the mainstream scheme in the current RGB-IR detection task [17]. Existing studies usually use feature concatenation, attention weighting and Cross-Attention to achieve modal fusion. Among them, the method based on the attention mechanism can dynamically adjust the weights of RGB and IR modality according to the ambient light, so as to improve the ability to express features in low-light scenarios [18]; The fusion method based on Cross-Attention can further enhance the semantic alignment ability between modes and show excellent performance in terms of detection accuracy. However, complex cross-modal interaction structures often significantly increase the number of model parameters and computing overhead, which not only weakens the real-time performance of the system, but also makes it more difficult to deploy the dual-stream detection framework in on-board embedded scenarios.

In addition to the computational overhead brought about by the model structure itself, the RGB-IR dual-stream detection framework based on feature-level fusion will also face additional cross-modal synchronization delay problems in the actual reasoning process. Since RGB images and IR images usually need to be collected, transmitted and preprocessed independently, the data loading and synchronization mechanism will introduce additional waiting overhead; At the same time, cross-modal feature interaction and serial processing logic will further amplify the end-to-end total delay. In response to this problem, researchers use the BiSwift framework to build a global bandwidth control and multi-stream scheduling mechanism to dynamically allocate the computing and transmission resources of different video streams, so as to improve the overall throughput performance while ensuring fairness [19]; The pipeline scheduling method based on deep reinforcement learning improves GPU resource utilization and reduces end-to-end latency through fine-grained task splitting and execution order adjustment [20]. However, these methods are mainly aimed at general multi-video streams or single reasoning tasks, and have not been optimized for the sampling frequency inconsistency and time alignment deviation that may occur in RGB-IR dual- modality in-vehicle scenarios. Cross-modal synchronous waiting is still an important factor that restricts the real-time performance of the dual-stream detection system.

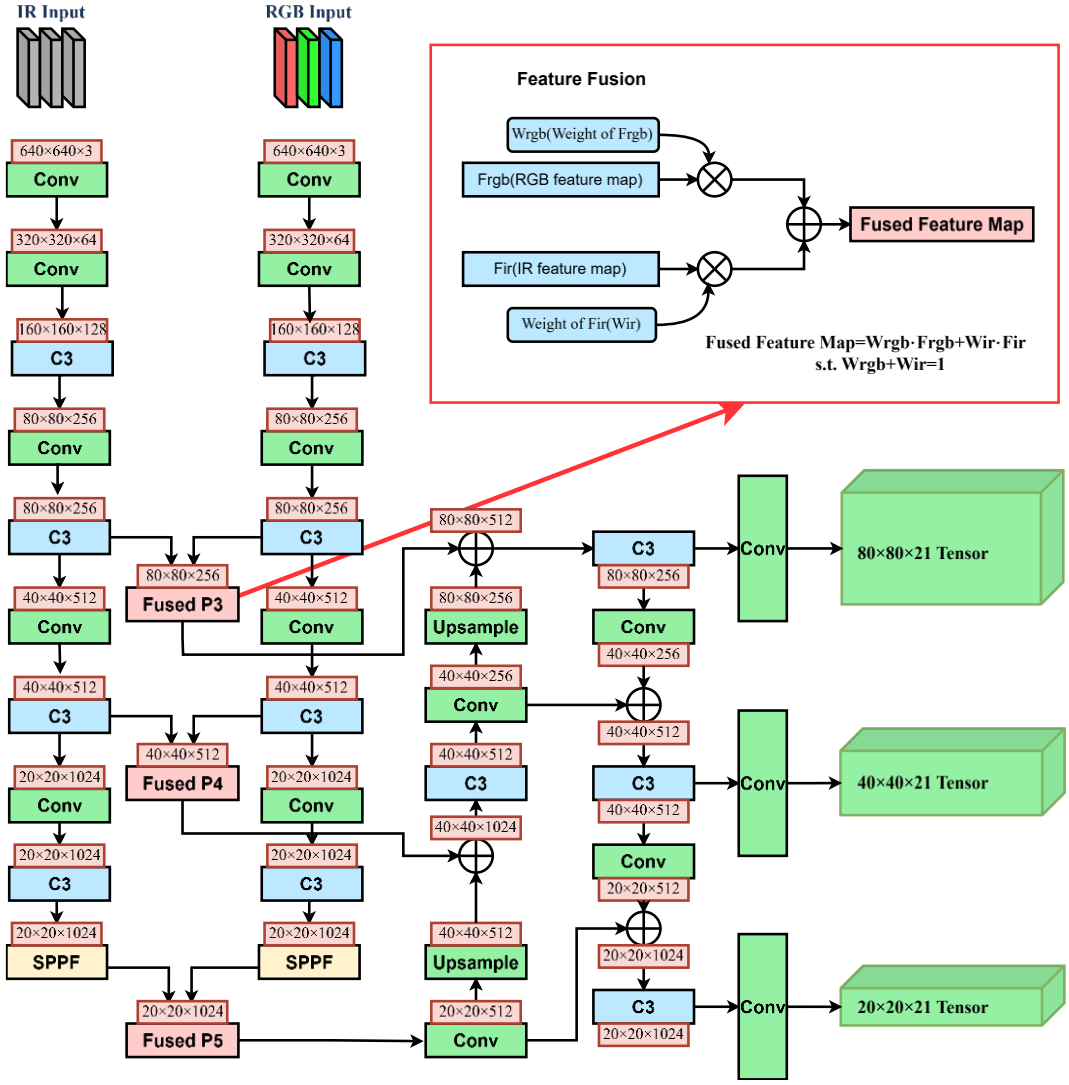
Therefore, how to realize lightweight cross-modal feature coordination and efficient real-time reasoning while ensuring the robustness of low-light detection is still a key issue in the current driving fatigue detection research. Although the existing cross-modal integration method based on attention mechanism or Cross-Attention can improve the ability of feature interaction. However, it is usually accompanied by high computational complexity, and it is difficult to meet the real-time deployment needs of the on-board end; Although simple feature concatenation has low computational overhead, it lacks an effective semantic alignment mechanism, and it is difficult to give full play to the complementary advantages between RGB and IR. In this regard, this paper adopts a linear weighted fusion strategy to reduce the computational complexity of cross-modal fusion, and introduces a cross-modal feature alignment mechanism to enhance the consistency of RGB and IR in semantic space. And combined with one-way feature aggregation and local backbone lightweighting to improve the efficiency of cross-modal information utilization, so as to achieve model lightweight deployment and efficient real-time reasoning performance on the basis of ensuring low-light robustness.

3. Research on a Fatigue Detection Mechanism Based on Lightweight RGB-IR Fusion

3.1 Construction of the RGB-IR Dual-Stream Fusion Framework

The lightweight RGB-IR fatigue detection framework constructed in this chapter adopts a dual-stream parallel structure. Using RGB and IR images as inputs, multi-scale semantic modeling is completed through independent feature extraction branches respectively, and then cross-modal fusion is performed at the feature layer, and the final detection results are output through the lightweight multi-scale aggregation module.

Figure 1: Overall Framework of the Lightweight RGB-IR Dual-Stream Fusion-Based Fatigue Detection Method



As shown in Figure 1, the RGB and IR branches maintain structural consistency in their feature extraction components; however, their parameters are updated independently, ensuring that each modality can learn its own distinct feature distribution and form complementary representations during the fusion stage. Finally, the fused multi-scale features are fed into the detection head to perform the classification and prediction of the driver's fatigue state, as well as object localization.

In the feature fusion stage, the features of the corresponding RGB and IR layers are linearly weighted and fused to achieve adaptive integration of cross-modal information. As shown in Formula (1), the fusion formula is defined as:

$$F_l = w^{RGB} \cdot F_l^{RGB} + w^{IR} \cdot F_l^{IR} \tag{1}$$

Here, w^{RGB} and w^{IR} are learnable fusion weights that satisfy normalization constraints, enabling the model to dynamically adjust the contribution ratio of the two modes according to different lighting conditions.

Thereby enhancing RGB texture information under normal lighting conditions, and strengthening IR structural representation capabilities under low-light conditions.

To further mitigate the semantic shift between the RGB and IR feature spaces, a cross-modal feature alignment loss is employed to impose consistency constraints on the high-level semantic representations of the two modalities. The alignment loss is defined as shown in Equation (2):

$$L_{align} = \frac{1}{L}\sum_{l=1}^{L}\left(1 - \frac{F_l^{RGB} \cdot F_l^{IR}}{\|F_l^{RGB}\|\|F_l^{IR}\|}\right) \tag{2}$$

Here, L denotes the number of feature layers involved in the constraint, and $\|\cdot\|$ denotes the vector norm. The fractional term represents the cosine similarity between the RGB and IR features at the l-th layer; by converting feature similarity into a distance metric, this loss function constrains the directional consistency of features from different modalities within the semantic space. At the same time, it avoids the scale sensitivity issues inherent in direct reliance on Euclidean distance, thereby enabling the stable alignment of features from different modalities within a normalized space.

Ultimately, the detection loss and alignment loss collectively constitute the overall optimization objective, as shown in Equation (3):

$$L = L_{box} + L_{obj} + L_{cls} + \lambda L_{align} \tag{3}$$

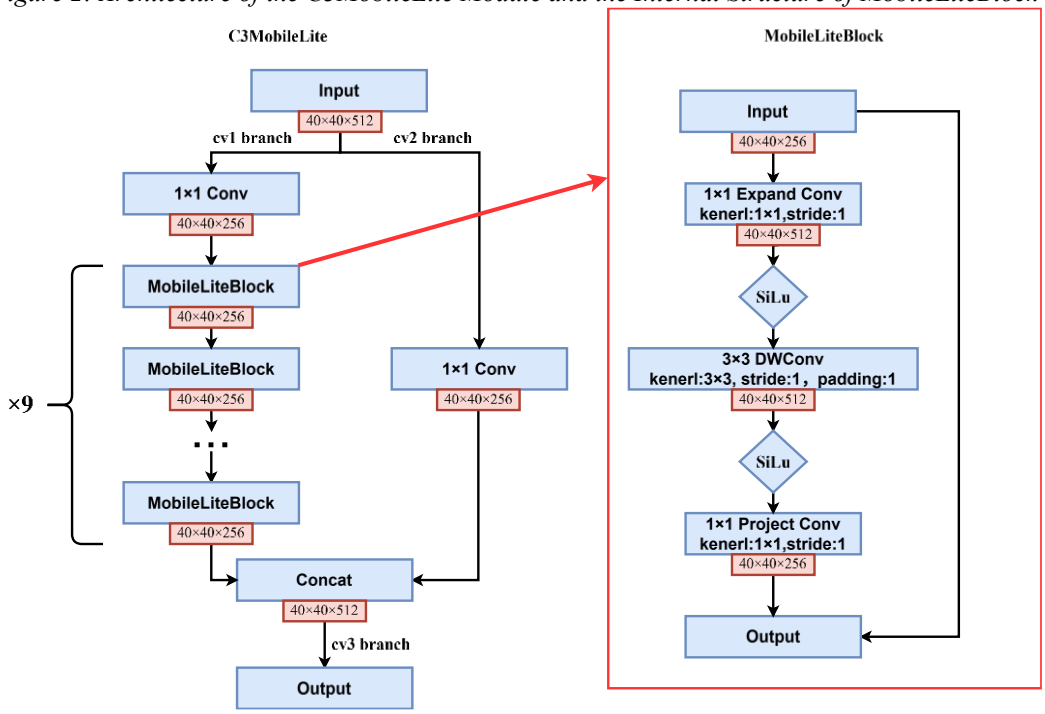
Specifically, λ is used to control the strength of the alignment constraint, balancing cross-modal consistency learning while preserving the primary objective of the detection task.

3.2 Lightweight Optimization Strategies for the RGB-IR Fusion Model

3.2.1 Local Lightweight Strategy Based on MobileLiteBlock

In the backbone network, this paper adopts a local modular lightweight optimization design. The C3MobileLite lightweight structure is only introduced in the middle feature block of the dual-stream feature extraction branch to replace the original standard C3 module with high computing overhead, while maintaining the overall expression ability of the backbone network as much as possible. Figure 2 shows the overall structure of C3MobileLite and the local details of its internal MobileLiteBlock: The input feature is first diverted into cv1 branch and cv2 branch, both of which undergo 1×1 convolution for channel compression, and then cv1 branch further stacks MobileLiteBlock to complete the lightweight feature transformation. Cv2 branch is used to retain the original feature information; The two-way features are merged and output by 1×1 convolution after Concat. The right side of Figure 2 also shows the MobileLiteBlock structure adopted by the cv1 branch in C3MobileLite to illustrate the specific calculation process of the lightweight unit.

Figure 2: Architecture of the C3MobileLite Module and the Internal Structure of MobileLiteBlock



MobileLiteBlock employs a lightweight architecture consisting of "expansion—depthwise separable convolution—projection." Let the input features be x; the computation process can be summarized by Equation (4):

$$y = f_{proj}(f_{dw}(f_{exp}(x))) \tag{4}$$

Among these, f_{exp} is the channel expansion layer, f_{dw} is the depthwise separable convolution layer, and f_{proj} is the channel projection layer. As shown in Equation (5), when the number of input and output channels is identical and the stride is 1, the module employs a residual connection:

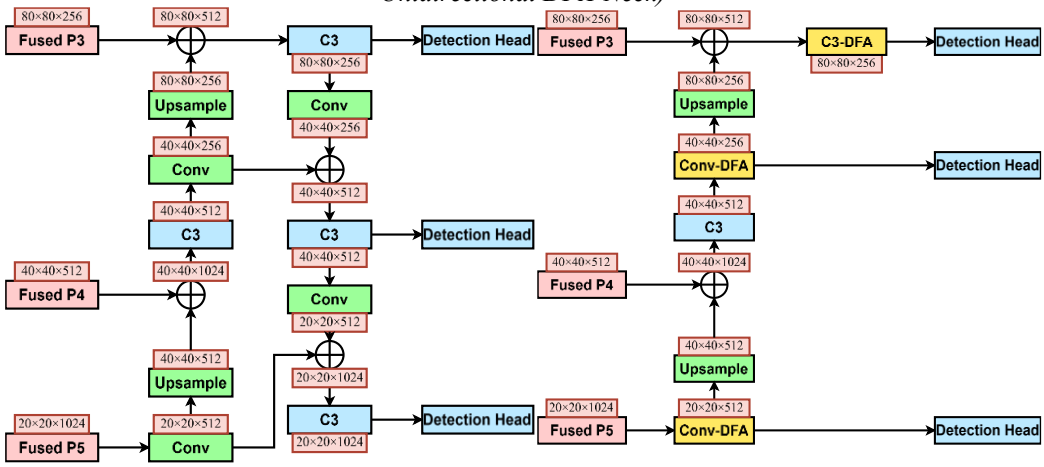
$$Output = Input + y \tag{5}$$

Both the channel expansion layer and the depthwise separable convolution layer employ the SiLU activation function to enhance the nonlinear expressive power of features. The channel projection layer maintains a linear mapping to ensure numerical spatial consistency between the output and the input, thereby preventing semantic drift when added via residual connections.

3.2.2 Lightweight Strategy Based on Unidirectional Feature Aggregation

The Neck part of the original YOLO adopts a two-way multi-scale feature fusion structure that combines FPN (Feature Pyramid Network) and PAN (Path Aggregation Network) to enhance the expression ability of objects at different scales. The two-way path will introduce cross-scale redundant information transmission and repeated convolution calculation. In view of the characteristics of target scale concentration and focus area fixation in driver fatigue detection, this paper proposes a one-way feature aggregation mechanism DFA (Directional Feature Aggregation), which reduces redundant information transmission while retaining the expression of low, medium and high-level features, thus reducing the complexity of calculation and improving stability.

Figure 3: Comparison of Lightweight Neck Architectures (Left: Original Bidirectional FPN+PAN; Right: Unidirectional DFA-Neck)



As shown in Figure 3, unlike the low, medium, and high-level feature maps in the left image after bidirectional multi-scale fusion, these are different. The figure on the right transmits information via a unidirectional path, thereby aggregating deep semantic information with shallow positional information, and directly inputs it into the detection head for fatigue state recognition. Let the fusion input for the l-th layer be F_l; then the unidirectional aggregation process can be expressed by Equation (6):

$$F_l^{out} = \phi\big(\text{Concat}(\text{Upsample}(F_{l+1}), F_l)\big) \tag{6}$$

Here, $\phi(\cdot)$ denotes the convolution transform. Compared to the traditional bidirectional PAN, this architecture reduces redundant cross-scale information flow while halving the number of channels in the intermediate and high-level feature maps; this effectively lowers computational complexity while preserving the representation of critical semantic information.

3.3 Asynchronous Decoupled Inference Strategy

In order to further improve the real-time performance of the model in the in-vehicle embedded environment, this paper designs an asynchronous dual-stream inference strategy based on the producer-consumer mechanism. The proposed strategy decouples RGB–IR data acquisition from model inference, and reduces the synchronous waiting delay caused by dual- modality serial processing through dual-thread parallel reading and main thread reasoning.

Algorithm 1 gives an asynchronous RGB-IR dual-stream reasoning process. The system establishes independent buffer queues for RGB and IR video streams respectively, and continuously completes dual-modal video frame acquisition and caching through two independent reading threads. As a consumer, the main reasoning thread obtains the set of candidate frames in the sliding time window from the RGB queue and the IR queue respectively, and completes the RGB-IR frame-level alignment based on the minimum time difference criterion. Subsequently, the aligned dual-modal image is entered into the fusion detection model after quantification and normalization processing, and finally the detection results are generated through Non-Maximum Suppression (NMS).

Algorithm 1 Asynchronous RGB–IR Dual-Stream Inference

Input: RGB video stream V_{RGB}, IR video stream V_{IR}, RGB–IR detection model M
Output: Fatigue detection result sequence Y

Process:
1: $Q_{RGB}, Q_{IR} \leftarrow$ build_buffer(V_{RGB}, V_{IR})
2: launch(*RGBReader*)
3: launch(*IRReader*)
4: while streams active do
5: $C_{RGB} \leftarrow \text{window}(Q_{RGB}, W)$
6: $C_{IR} \leftarrow \text{window}(Q_{IR}, W)$
7: $(R^*, I^*) \leftarrow \arg\min_{R \in C_{RGB}, I \in C_{IR}} |t_R - t_I|$

8: $x_{RGB} \leftarrow \text{preprocess}(R^*)$
9: $x_{IR} \leftarrow \text{preprocess}(I^*)$
10: pred $\leftarrow M(x_{RGB}, x_{IR})$
11: $y \leftarrow \text{NMS}(pred)$
12: $Y \leftarrow Y \cup y$
13: end while
14: stop(*RGBReader*)
15: stop(*IRReader*)
16: return Y

Subsequently, the system employs a frame-level alignment mechanism based on a sliding time window to synchronize and match the dual video streams. For the asynchronous RGB and IR buffer queues, the system constructs a set of candidate RGB frames C_{RGB} and a set of candidate IR frames C_{IR} within a specific time window W. It then searches these candidate sets to identify the RGB–IR frame pair with the minimal timestamp difference, designating it as the final alignment result. Let the acquisition times for the RGB and IR frames be t_{RGB} and t_{IR}, respectively; then the frame-level alignment objective can be expressed as Equation (7):

$$\min \mid t_{\text{RGB}} - t_{\text{IR}} \mid \tag{7}$$

This mechanism is capable of effectively mitigating timing offsets caused by sampling frequency fluctuations in dual-modal systems under asynchronous reading conditions, thereby enhancing the stability and robustness of cross-modal feature fusion.

From the perspective of latency modeling, under the synchronous inference mode, RGB and IR data must sequentially undergo reading, pairing, and inference; therefore, the end-to-end latency can be approximately expressed by Equation (8):

$$t_{\text{sync}} = t_{\text{RGB}} + t_{\text{IR}} + t_{\text{inf}} \tag{8}$$

Among these, t_{RGB} and t_{IR} represent the latency for reading and preprocessing the RGB and IR data, respectively, while t_{inf} represents the latency for model inference and post-processing. In contrast, in the asynchronous dual-stream inference mode, the bimodal reading process can be executed in parallel; therefore, the overall latency can be expressed by Equation (9):

$$t_{\text{async}} = \max(t_{\text{RGB}}, t_{\text{IR}}) + t_{\text{inf}} \tag{9}$$

The asynchronous decoupling mechanism ensures that the system's total latency is primarily determined by the greater of the two stream-reading times and the model inference time, thereby enhancing the system's overall real-time performance.

4. Experimental Analysis

4.1 Dataset and Parameter Settings

This paper adopts DMD (Driver Monitoring Dataset) as the experimental data source. The data set contains RGB and IR dual- modality synchronous video sequences, which provides good data support for cross-modal fusion research under low light conditions. Select 14 of the subjects as experimental subjects, and construct the data set according to the principle of "dividing by person", that is, the data of the same subject will not appear in the training set, verification set and test set at the same time, so as to avoid data leakage and improve the generalization ability of the model. In terms of data division, the subjects are divided into training set (Train), verification set (Val) and test set (Test) according to the ratio of 8:3:3. The test set only contains the data of subjects who did not participate in the training, which is used to evaluate the generalization ability of the model in cross-individual scenarios. In order to further verify the robustness of the model under low light conditions, two subsets of low light and extremely low light are built on the basis of the test set, and different degrees of light degradation are simulated through gamma transformation, brightness adjustment, contrast compression, noise and motion blurring. The degradation parameters are shown in Table 1:

Table 1: Illumination Degradation Settings for the Test Set

Degradation Operation	Low-Light Condition	Extremely Low-Light Condition
Gamma Correction	1.8	2.8
Brightness Factor	0.6	0.35
Contrast	0.85	0.6
Gaussian Noise	5	15
Motion Blur	No	Yes

The experiments were conducted in a unified hardware environment, as shown in Table 2.

Table 2: Training and Optimization Parameter Settings

Parameter	Setting
Input Size	640 × 640
Optimizer	SGD
Momentum	0.937
Weight Decay	0.0005
Initial Learning Rate (lr0)	0.01
Learning Rate Schedule	Linear Decay
Final Learning Rate Factor (lrf)	0.01
Batch size	8
Number of Training Epochs	25

4.2 Evaluation Metrics

(1) Mean Average Precision (mAP)

This paper employs Mean Average Precision (mAP) to comprehensively evaluate the classification and localization performance in driver fatigue detection, with a particular focus on two evaluation metrics: mAP@0.5 and mAP@0.5:0.95. mAP@0.5 represents the average detection precision calculated at an IoU threshold of 0.5, and is primarily used to evaluate a model's object detection capabilities. mAP@0.5:0.95 represents a comprehensive metric obtained by averaging the AP values across IoU thresholds ranging from 0.5 to 0.95, with a step size of 0.05; its calculation is formulated as shown in Equation (10):

$$mAP_{0.5:0.95} = \frac{1}{10}\sum_{k=0}^{9} A\,P_{0.5+0.05k} \tag{10}$$

mAP@0.5:0.95 imposes stricter requirements on the localization accuracy of target bounding boxes than mAP@0.5, thereby providing a more comprehensive reflection of the model's overall performance regarding bounding box regression and detection stability.

(2) Model Complexity

To assess the computational complexity and lightweight nature of the model, this paper introduces the number of parameters and Floating Point Operations (FLOPs) as evaluation metrics. Specifically, the number of parameters represents the total scale of learnable parameters within the model, while FLOPs measure the computational cost required for a single forward pass.

At the overall network level, the model's FLOPs represent the cumulative sum of the computational operations across all convolutional layers, as shown in Equation (11):

$$FLOPs_{total} = \sum_{l=1}^{L} FLOP\,s_l \tag{11}$$

The computational cost of a single convolutional layer is typically approximated using the following form, as shown in Equation (12):

$$FLOPs = H_{out} \cdot W_{out} \cdot C_{out} \cdot (C_{in} \cdot K^2) \tag{12}$$

Here, H_{out} and W_{out} denote the height and width of the output feature map, C_{in} and C_{out} denote the number of input and output channels, respectively, and K denotes the convolution kernel size.

(3) Inference Speed

Inference speed measures a model's real-time processing capability during actual deployment. This paper employs the effective frame rate (Frames Per Second, FPS) as the real-time performance evaluation metric,

defined as $FPS = \frac{N}{T}$. Here, N denotes the total number of frames processed, and T denotes the corresponding total inference time.

4.3 Results Analysis

4.3.1 Analysis of Model Complexity Results

The model complexity experiments primarily compare different model architectures across two dimensions—parameter count and floating-point operations—to assess the impact of lightweighting strategies on computational overhead. The experimental results are presented in Table 3.

Table 3: Comparison of Model Complexity

Model	Parameters (M)	FLOPs (G)	Reduction in Parameters After Lightweight Optimization	Reduction in FLOPs After Lightweight Optimization
Single-Stream RGB	7.24	8.31	/	/
Dual-Stream RGB-IR Baseline	14.74	16.66	/	/
Local Backbone Lightweight Optimization	14.17	15.75	3.86%	5.46%
Unidirectional Feature Aggregation	9.80	13.46	33.51%	19.21%
Dual-Strategy Lightweight Optimization	9.23	12.56	37.45%	24.57%

Compared to the single-stream RGB modality, the parameter count of the dual-stream RGB-IR baseline model increased from 7.24M to 14.74M, and the FLOPs rose from 8.31G to 16.66G—increases of 103.59% and 100.48%, respectively—indicating that dual-modality parallel modeling introduces significant additional computational overhead.

Through the lightweighting of the local Backbone, the model's parameter count and FLOPs were reduced to 14.17M and 15.75G, respectively—representing a decrease of only 3.86% and 5.46% compared to the dual-stream baseline. This result demonstrates that simply introducing lightweight replacements in the backbone network offers limited computational compression. The reason is that the Backbone merely reduces local convolutional complexity, while the overall dual-stream architecture continues to account for the dominant computational overhead; consequently, it is difficult to significantly reduce system-level complexity. In contrast, the unidirectional feature aggregation strategy reduced the number of parameters and FLOPs to 9.80M and 13.46G, respectively, representing reductions of 33.51% and 19.21%. This indicates that the cross-scale bidirectional reflow structure in the Neck stage still suffers from significant computational redundancy. The adoption of a unidirectional feature aggregation design can effectively reduce the computational overhead associated with redundant feature fusion. The dual-strategy lightweight optimization approach further compresses the parameter count and FLOPs to 9.23M and 12.56G, representing reductions of 37.45% and 24.57%, respectively, compared to the dual-stream baseline. This demonstrates that the combined application of these two lightweighting strategies enables a significant reduction in overall computational complexity while preserving the dual-modality structure.

4.3.2 Analysis of Illumination Robustness Results

To validate the detection stability of the proposed method under varying lighting conditions, comparative experiments were conducted in three scenarios—normal lighting, low light, and extremely dim light—the results of which are presented in Tables 4 and 5, respectively.

Table 4: Detection Performance Under Different Illumination Conditions (mAP@0.5)

Model	Normal Lighting	Low-Light Condition	Extremely Low-Light Condition
Single-Stream RGB	0.865	0.827	0.314
Dual-Stream RGB-IR Baseline	0.826	0.800	0.378
Local Backbone Lightweight Optimization	0.825	0.774	0.226
Unidirectional Feature Aggregation	0.775	0.775	0.513
Dual-Strategy Lightweight Optimization	0.832	0.754	0.547

Table 5: Detection Performance Under Different Illumination Conditions (mAP@0.5:0.95)

Model	Normal Lighting	Low-Light Condition	Extremely Low-Light Condition
Single-Stream RGB	0.569	0.546	0.200
Dual-Stream RGB-IR Baseline	0.563	0.535	0.217
Local Backbone Lightweight Optimization	0.561	0.529	0.105
Unidirectional Feature Aggregation	0.525	0.533	0.317
Dual-Strategy Lightweight Optimization	0.575	0.531	0.359

Under normal and low-light conditions, the mAP@0.5 of the single-stream RGB model is 0.865 and 0.827, respectively, slightly higher than the 0.826 and 0.800 of the dual-stream RGB-IR baseline model. This indicates that in well-illuminated or mildly degraded scenes, the RGB modality is already capable of providing sufficiently rich textural and semantic information. Under extremely low lighting conditions, the mAP@0.5 of the single-stream RGB model drops to 0.314, a decrease of approximately 63.7% compared to normal lighting conditions; The dual-stream RGB-IR baseline model still achieves a score of 0.378—an improvement of approximately 20.4% over the single-stream RGB model—demonstrating that the IR modality plays a significant compensatory role under extreme low-light conditions.

The locally lightweighted backbone model maintains a score of 0.825 under normal lighting conditions; however, under extremely dark lighting, it drops to 0.226—a decline of approximately 40.2% compared to the dual-stream baseline. This indicates that applying localized lightweighting solely to the backbone network significantly impairs the capability to extract mid-to-high-level semantic features. In contrast, the unidirectional feature aggregation strategy performs slightly below the dual-stream baseline under both normal and low-light conditions; however, Under extremely low-light conditions, it reaches a score of 0.513—representing an improvement of approximately 35.7% compared to the dual-stream baseline. It demonstrates that reducing cross-scale bidirectional redundant feedback facilitates the stable transmission of high-level semantics to low-level features, thereby effectively suppressing noise propagation and enhancing robustness under low-light conditions. Furthermore, the dual-strategy lightweight optimization approach achieved scores of 0.832, 0.754, and 0.547 under normal, low-light, and extremely dark lighting conditions, respectively. In extremely low-light scenarios, performance improves by approximately 6.6% compared to unidirectional feature aggregation, and by approximately 44.7% compared to the dual-stream baseline. These results indicate that, building upon the effective suppression of redundant cross-scale backflow achieved through unidirectional feature aggregation, the introduction of moderate backbone lightweighting does not compromise critical semantic representations. On the contrary, by reducing redundant convolutional responses, it renders the cross-modal fusion features more focused, thereby further enhancing the model's robustness.

Based on the mAP@0.5:0.95 results, this trend is further amplified. In extremely dark scenes, Single-stream RGB achieves a score of only 0.200; the Dual-stream baseline improves to 0.217—an 8.5% increase over Single-stream RGB. Unidirectional feature aggregation was further improved to 0.317, representing a 46.1% increase compared to the dual-stream baseline. The dual-strategy lightweight approach achieved a score of 0.359, marking a further 13.2% improvement over the unidirectional feature aggregation method. Therefore, it can be observed that as lighting conditions progressively diminish, the compensatory effect of the IR mode gradually intensifies. In the context of lightweight optimization methods, unidirectional feature aggregation determines the lower bound of robustness; conversely, lightweighting the backbone—when its capacity is reasonably controlled—can actually enhance the efficiency of feature aggregation. The synergy between these two elements enables the achievement of optimal illumination robustness.

4.3.3 Analysis of Real-Time Performance Results

To validate the real-time performance of the proposed lightweight structure and asynchronous inference mechanism, this chapter evaluates the single-stream RGB model, the dual-stream RGB-IR baseline model, and various lightweight dual-stream models; the results are presented in Table 6.

Table 6: Comparison of Real-Time Performance Between Synchronous and Asynchronous Inference Across Different Models

Model	Sync FPS	Async FPS	Sync Latency(ms)	Async Latency(ms)
Single-Stream RGB	94.99	95.34	10.5	10.5
Dual-Stream RGB-IR Baseline	18.71	61.49	53.4	16.3
Local Backbone Lightweight Optimization	19.11	61.91	52.3	16.2

Unidirectional Feature Aggregation	19.23	70.36	52.0	14.2
Dual-Strategy Lightweight Optimization	20.34	72.09	49.2	13.9

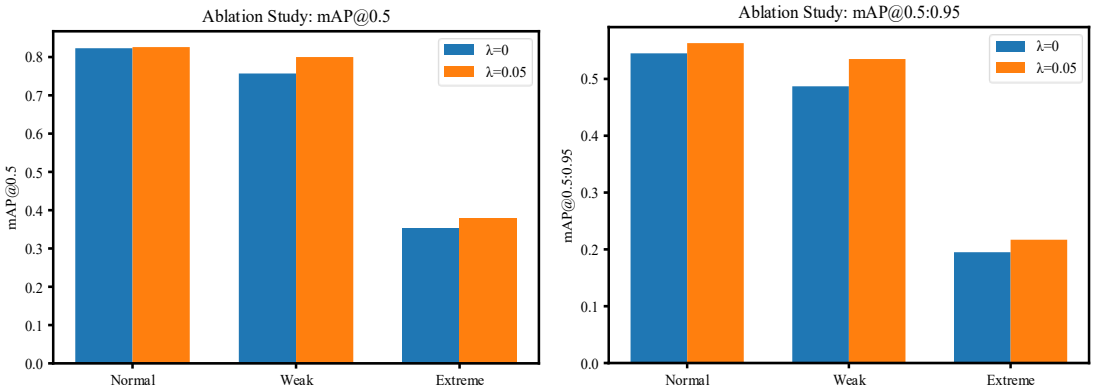
The single-stream RGB model achieved 94.99 FPS in synchronous mode; however, after switching to asynchronous mode, performance increased only marginally to 95.34 FPS, indicating that its system bottleneck does not primarily lie in data synchronization. Since the model computation itself is already lightweight, the benefits yielded by the asynchronous mechanism are limited. Compared to the single-stream RGB configuration, the dual-stream RGB-IR baseline model exhibits a frame rate of only 18.71 FPS in synchronous mode—a decrease of approximately 80.3% relative to the single-stream RGB setup. Furthermore, its average latency reaches as high as 53.4 ms, representing an increase of approximately 408.6% over the single-stream RGB baseline. These results indicate that dual-modal parallel processing and synchronization wait times significantly degrade the system's throughput. Following asynchronous inference, the FPS increased to 61.49—an improvement of approximately 228.6% compared to the synchronous mode—while latency dropped to 16.3 ms, a reduction of roughly 69.5% relative to the synchronous mode. This indicates that the primary performance bottleneck of the dual-stream system does not stem solely from the inference computation itself, but rather from the synchronization waiting and serial pairing overhead between the RGB and IR streams.

In the lightweight dual-stream model, the lightweight local backbone and unidirectional feature aggregation components achieved frame rates of 19.11 FPS and 19.23 FPS, respectively, under synchronous mode—representing improvements of approximately 2.1% and 2.8% compared to the dual-stream baseline. This indicates that structural compression contributes to real-time performance to some extent, but the improvement is limited. True real-time performance gains are still derived from asynchronous strategies: under the asynchronous mode, localized Backbone lightweighting achieved 61.91 FPS—an improvement of approximately 0.7% over the dual-stream baseline—while unidirectional feature aggregation reached 70.36 FPS, representing a gain of about 14.4%; the dual-strategy lightweight optimization approach further boosted performance to 72.09 FPS, an increase of approximately 17.3%. In particular, both unidirectional feature aggregation and dual-strategy lightweight optimization exceeded 70 FPS under an asynchronous mechanism; this indicates that, once the overhead of synchronous waiting is eliminated, the performance boost to overall throughput provided by the lightweight architecture is further amplified. The corresponding latencies decreased to 16.2 ms, 14.2 ms, and 13.9 ms, representing reductions of approximately 69.7%, 73.4%, and 73.9%, respectively, compared to the dual-stream baseline. Overall, the real-time bottlenecks in dual-stream systems primarily stem from data synchronization and cross-modal serial processing; consequently, asynchronous inference serves as the key to enhancing system-level real-time performance. Furthermore, lightweight architectures build upon this foundation by further reducing the inference burden; the combination of these two approaches enables an optimal balance between detection performance and deployment efficiency.

4.4 Ablation Study of the Alignment Loss

To verify the effectiveness of the proposed cross-modal alignment loss in RGB-IR fusion detection, this chapter presents ablation experiments on a dual-stream RGB-IR baseline model. A comparative analysis was conducted to examine the variations in the model's detection performance under different lighting conditions, specifically when the alignment loss weights were set to $\lambda=0$ and $\lambda=0.05$. The experimental results are presented in Figure 4.

Figure 4: Results of the Alignment Loss Ablation Study Under Different Illumination Conditions (Left: mAP@0.5; Right: mAP@0.5:0.95)



As shown in the left figure of Figure 4, the introduction of cross-modal alignment loss brings performance improvement under different lighting conditions. On the mAP@0.5 index, the RGB-IR baseline model (λ = 0) achieved 0.823 in normal light, decreased to 0.757 in low light, and further reduced to 0.354 in extremely dark light; After introducing the alignment loss and setting the weight λ=0.05, the model increased to 0.826 under normal lighting conditions, an increase of about 0.4%; It increased to 0.800 under low light conditions, an increase of about 5.7%; it increased to 0.378 under extremely dark light conditions, an increase of about 6.8%. It shows that the cross-modal alignment mechanism can enhance the consistency between RGB and IR features, among which the improvement is more obvious in low-light and extremely dark scenarios, indicating that the mechanism can effectively alleviate the modality shift caused by light degradation.

Under the mAP@0.5:0.95 indicator shown in the figure on the right, the trend is further amplified. The RGB-IR baseline model (λ = 0) achieved 0.545 under normal lighting conditions, decreased to 0.487 under low light conditions, and further decreased to 0.195 Under extremely low-light conditions. After introducing the alignment loss, the model was upgraded to 0.563 under normal lighting conditions, an increase of about 3.3%; It is increased to 0.535 under low light conditions, with an increase of about 9.9%; It is increased to 0.217 under extremely dark light conditions, an increase of about 11.3%. It shows that the alignment loss not only improves the target detection ability, but also has a more significant improvement effect on the positioning accuracy under the high IoU threshold.

Judging from the overall trend, cross-modal alignment loss brings performance improvement under different lighting conditions. Among them, the improvement is more obvious in low-light and extremely dark scenes. By enhancing the consistency of RGB and IR features, modal shift can be alleviated, thus improving the stability and discrimination ability of fusion features. The method achieves the overall performance improvement under low light conditions without increasing additional reasoning overhead.

4.5 Conclusions

This paper proposes a YOLOv5 detection method based on lightweight RGB-IR fusion, builds an RGB-IR dual- stream feature extraction and linear weighted fusion framework to achieve cross-modal information fusion, and uses cross-modal feature alignment loss to constrain the semantic consistency between different modes; Lightweight optimization introduces a dual lightweight strategy, that is, Neck-end one-way feature aggregation to reduce cross-scale redundancy backflow, and Backbone local MobileLiteBlock to compress redundancy calculations, So as to reduce the complexity of the model while ensuring the ability to express features. The experimental results show that the method achieves 0.547 mAP@0.5 and 0.359 mAP@0.5:0.95 Under extremely low-light conditions, and significantly reduces the parameter count and computational cost; Combined with the asynchronous dual-stream reasoning mechanism, the system achieves a real-time detection performance of 72.09 FPS. In summary, while significantly compressing the calculation volume of the model and improving the real-time, this method effectively enhances the robustness of detection in low light and realizes the coordinated optimization of accuracy and efficiency.

References

[1] Fu, B., Boutros, F., Lin, C.-T., & Damer, N. (2024). A survey on drowsiness detection: Modern applications and methods. *IEEE Transactions on Intelligent Vehicles, 9*(11), 7279–7300. https://doi.org/10.1109/TIV.2024.3395889

[2] Ayas, S., Donmez, B., & Tang, X. (2024). Drowsiness mitigation through driver state monitoring systems: A scoping review. *Human Factors, 66*(9), 2218–2243. https://doi.org/10.1177/00187208231208523

[3] Jiao, Y., Zhang, Y., Liu, W., & Jiao, Z. (2026). Detecting driver sleepiness from physiological indicators using a CNN-LSTM self-attention model. *IEEE Journal of Biomedical and Health Informatics, 30*(5), 4082–4095. https://doi.org/10.1109/JBHI.2025.3629974

[4] V. P. B., & Chinara, S. (2021). Automatic classification methods for detecting drowsiness using wavelet packet transform extracted time-domain features from single-channel EEG signal. *Journal of Neuroscience Methods, 347*, Article 108927. https://doi.org/10.1016/j.jneumeth.2020.108927

[5] Baccour, M. H., Driewer, F., Schäck, T., & Kasneci, E. (2022). Comparative analysis of vehicle-based and driver-based features for driver drowsiness monitoring by support vector machines. *IEEE Transactions on Intelligent Transportation Systems, 23*(12), 23164–23178. https://doi.org/10.1109/TITS.2022.3207965

[6] Ganguly, B., Dey, D., & Munshi, S. (2025). An attention deep learning framework-based drowsiness detection model for intelligent transportation systems. *IEEE Transactions on Intelligent Transportation Systems, 26*(4), 4517–4527. https://doi.org/10.1109/TITS.2025.3544138

[7] Khan, M. A., Malik, S. A., & Ullah, H. (2022). Intelligent driver drowsiness detection for traffic safety based on multi CNN deep model and facial subsampling. *IEEE Transactions on Intelligent Transportation Systems, 23*(10), 19743–19752. https://doi.org/10.1109/TITS.2021.3134222

[8] Aprilia, S. J. N., & Fitrianah, D. (2024). Drowsiness detection in low-light conditions using YOLO and CLAHE augmentation for real-time applications. In *Proceedings of the 2024 7th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)* (pp. 742–746). https://doi.org/10.1109/ISRITI64779.2024.10963648

[9] Kim, K.-J., Baek, J. W., Lim, K.-T., & Shin, M. (2021). Low-cost real-time driver drowsiness detection based on convergence of IR images and EEG signals. In *Proceedings of the 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)* (pp. 362–367). https://doi.org/10.1109/ICAIIC51459.2021.9415193

[10] Gowda, M. S., Talasila, V., Umar, S. A. R., & Alva, R. (2024). A multimodal approach to detect driver drowsiness. In *Proceedings of the 2024 5th International Conference on Circuits, Control, Communication and Computing (I4C)* (pp. 554–560). https://doi.org/10.1109/I4C62240.2024.10748433

[11] Wang, S., Sun, G., Dong, L., & Zheng, B. (2026). CARNet: Cross-attention guided feature reconstruction for RGB-T object detection. *Expert Systems with Applications, 298*, Article 129865. https://doi.org/10.1016/j.eswa.2025.129865

[12] Palo, P., Nayak, S., Gupta, K., & Uttarkabat, S. (2025). A three-stage multimodal approach to driver activity recognition with RGB and infrared video analysis. In *Proceedings of the 2025 IEEE International Conference on Imaging Systems and Techniques (IST)*. https://doi.org/10.1109/IST66504.2025.11268357

[13] Herath, D., Abeyrathne, C., & Jayaweera, P. (2025). Vision-based driver drowsiness monitoring: Comparative analysis of YOLOv5–v11 models. *arXiv*. https://arxiv.org/abs/2509.17498

[14] Dou, X., Wang, T., & Shao, S. (2023). A lightweight YOLOv5 model integrating GhostNet and attention mechanism. In *Proceedings of the 2023 4th International Conference on Computer Vision, Image and Deep Learning (CVIDL)* (pp. 348–352). https://doi.org/10.1109/CVIDL58838.2023.10166155

[15] Saxena, S., Angel, Khurana, M., Tiwari, S., & Shakya, H. K. (2026). Low-light driver drowsiness detection for real-time safety assistance using dual attention mechanisms in deep learning model. *Scientific Reports, 14*(1), Article 44442. https://doi.org/10.1038/s41598-026-44442-3

[16] Liang, X., Yao, W., Fang, X., & Zhang, C. (2024). FMIF: Facial multi-feature information fusion for driver fatigue detection. *Signal, Image and Video Processing, 19*, 121. https://doi.org/10.1007/s11760-024-03573-8

[17] Sun, Y., Meng, Y., Wang, Q., Tang, M., Shen, T., & Wang, Q. (2024). Visible and infrared image fusion for object detection: A survey. In *International Conference on Image, Vision and Intelligent Systems (ICIVIS)* (pp. 236–248). Springer. https://doi.org/10.1007/978-981-97-0855-0_24

[18] Bouraffa, Y., Benzaoui, A., & Bouridane, A. (2026). Feature-disentangling RGB-NIR fusion network for remote driver physiological measurement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 657–666). https://doi.org/10.1109/WACV57317.2026.00400

[19] Sun, L., Wang, W., Yuan, T., Mi, L., Dai, H., Liu, Y., & Fu, X. (2024). BiSwift: Bandwidth orchestrator for multi-stream video analytics on edge. In *Proceedings of IEEE INFOCOM* (pp. 1–10). https://doi.org/10.48550/arXiv.2312.15740

[20] Pereira, D., Ghosh, S., & Dey, S. (2024). DRL-based multi-stream scheduling of inference pipelines on edge devices. In *Proceedings of the IEEE International Conference on VLSI Design (VLSID)* (pp. 324–329). https://doi.org/10.1109/VLSID60093.2024.00060

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access