Journal of Computers Vol. 37 No. 5, October 2026, pp. 357-372
https://doi.org/10.63367/19911599202610370502 2

Research on the Fitness for Duty of Public Transport Drivers Based on Micro-expression Recognition

Ying-Si Zhao[1*], Gulipiyan Ainiwan[1], and Wen-Hao Ye[2]

[1] School of Economics and Management, Beijing Jiaotong University,
Beijing 100044, China
{yszhao, 22120708}@bjtu.edu.cn

[2] School of Electronic Information Engineering, Beijing Jiaotong University,
Beijing 100044, China
22211110@bjtu.edu.cn

Received 10 July 2026; Revised 26 July 2026; Accepted 10 September 2026

Abstract. The fitness for duty of bus drivers is one of the critical factors ensuring public transportation safety. Many scholars domestically and internationally have conducted relevant research in this area. However, studies on bus drivers' fitness for duty based on micro-expression recognition remain relatively limited. This paper proposes the Y8n-V Face model, which integrates the Yolov8n model and the VGG16 model to achieve micro-expression recognition for public transport drivers. Compared with traditional models, the proposed framework demonstrates improvements in recall, precision, and F1-Score.

Keywords: fitness for duty, public transport drivers, micro-expression recognition, long-term driver monitoring, early risk intervention

1 Introduction

As a fundamental component of urban mobility systems, public transportation relies critically on the sustained fitness and operational performance of its drivers. Ensuring driver fitness is not only essential for maintaining high standards of operational safety but also directly influences service reliability, passenger satisfaction, and overall network efficiency. Investigating the multifaceted factors that affect driver fitness—and modeling their dynamic interactions—offers significant potential to enhance proactive safety management, optimize resource allocation, and support evidence-based operational planning.

Unlike static occupational assessments, the fitness for duty of public transport drivers is a highly dynamic construct, influenced by a confluence of factors including circadian rhythms, cumulative fatigue across consecutive shifts, unpredictable traffic events, and even emotional fluctuations triggered by passenger interactions. Traditional periodical health check-ups or post-accident evaluations are inherently inadequate to capture these real-time fluctuations. Therefore, transitioning from a static, retrospective evaluation paradigm to a dynamic, prospective, and continuous monitoring framework is not merely a technological upgrade, but a fundamental epistemological shift in transportation safety management. This study responds to this imperative by proposing a non-intrusive, vision-based system capable of providing second-by-second state inference, thereby bridging the critical gap between sporadic assessments and the continuous vigilance required for safe operations.

Recent advances in artificial intelligence, particularly deep learning architectures such as convolutional neural networks (CNNs), have revolutionized several fields related to visual analysis and behavioral inference. Models including Faster R-CNN [1], R-FCN [2], SSD [3], YOLO [4], VGGNet [5], and SENet [6] have set benchmarks in object detection, recognition, and segmentation, demonstrating robust capability in extracting meaningful patterns from complex visual data. Building upon these developments, this study introduces an integrated framework that employs micro-expression recognition as a non-invasive, real-time modality for assessing the psychological and physiological states of public transport drivers. Micro-expressions—brief, involuntary facial movements that reveal underlying affective and cognitive states—serve as a viable data source for inferring driver fatigue, stress, distraction, and emotional load.

* Corresponding Author

The proposed approach extends beyond real-time monitoring by incorporating temporal modeling and predictive analytics. By leveraging historical records of driver states alongside continuous micro-expression data, this research applies sequence-aware deep learning models to forecast near-term fitness degradation and potential risk states. Such predictive capacity enables transit operators to implement timely interventions—such as adjusted scheduling, mandated breaks, or targeted support—before impairment leads to safety-critical events.

This research contributes to both academic discourse and practical applications in transportation safety and human factors engineering. It provides public transport operators and regulatory agencies with a scalable, data-driven framework for driver fitness assessment, supporting decisions related to shift rotation, fatigue management, and individualized driver support programs. Ultimately, the work aims to reduce the incidence of fatigue-related accidents, enhance road safety, and improve the well-being of drivers through intelligent, adaptive monitoring.

The subsequent sections of this paper are organized as follows. In Section 2, we review related work on fitness for duty and micro-expression recognition. In Section 3, The Y8n-V Face model is introduced. Section 4 presents the experimental design, results, and corresponding analysis and conclusions. Finally, Section 5 outlines future directions for improvement and research.

2 Related Work

Current research on the fitness for duty of public transport drivers links driver states with various physiological and psychological conditions to explore methods for assessing fitness for duty based on these factors. In the field of facial expression recognition, state-of-the-art models have achieved high recognition accuracy through deep learning while maintaining rapid processing speeds via GPU acceleration. This section provides a concise overview of studies on factors affecting fitness for duty and micro-expression recognition models.

2.1 Factors Affecting Public Transport Drivers' Fitness for Duty

Early investigations into the factor model of public transport drivers' fitness for duty focused on psychological, physiological, and behavioral characteristics. For example, Chizuru pointed out that physiological indicators such as electroencephalogram (EEG), electrocardiogram (ECG), and respiratory rate can be used to evaluate the operational state of railway drivers, offering a non-invasive and real-time technical approach for monitoring driver conditions [7]. Martina's research found a positive correlation between the physical activity levels of bus drivers and their work ability, indicating that regular physical activity serves as a key protective factor in maintaining and enhancing the occupational capacity of professional bus drivers [8]. Kyoung et al. found that a history of traffic accidents is significantly associated with professional pride, job satisfaction, and risky driving behaviors. Through an empirical model, they demonstrated that improving drivers' work attitudes and mental health serves as a key mechanism for breaking the vicious cycle of "accidents – negative mindset – increased risk-taking behavior," thereby enhancing long-term road safety [9].

Recent studies have shown that psychological disposition, physical health, driving behavior, and demographic characteristics collectively influence a driver's fitness-for-duty status. For instance, the paper by Ye et al. contributes a significant theoretical model of psychological attributes to human resource management and safety management in the high-speed railway sector. It emphasizes the critical role of this model in practical applications such as personnel selection, evaluation, and training [10]. Meanwhile, Wang et al. employed the entropy weight method combined with the TOPSIS model to objectively and accurately evaluate bus drivers' driving behaviors, deriving comprehensive scores for drivers based on objective weighted indicators [11]. Furthermore, Liu et al.'s study demonstrates that key demographic variables—including age, years of driving experience, and monthly income—significantly influence professional mental health outcomes. Their findings indicate that these factors are closely linked to psychological well-being in high-stress occupations, with implications for both individual support and workplace mental health strategies [12]. Additionally, research by Zhixin Li et al. confirmed that integrating multidisciplinary perspectives (such as ergonomics, psychology, management, and transportation engineering) has deepened the understanding of the coupling mechanism between occupational health and safety for bus drivers [13].

Understanding the mechanistic pathways through which psychological and physiological states translate into impaired driving performance is essential for justifying the focus on facial expression analysis. Fatigue,

for instance, manifests through progressive declines in psychomotor vigilance, prolonged reaction times, and compromised executive functions—particularly working memory and inhibitory control, which are critical for hazard perception and emergency response. Concomitantly, elevated stress levels and negative emotional states (e.g., anger, anxiety, frustration) have been shown to narrow the attentional field, increase risk-taking propensity, and impair situational awareness, collectively elevating the probability of unsafe driving behaviors such as tailgating, abrupt lane changes, and signal violations. Critically, these internal states are not only manifested through facial muscle activations—including drooping eyelids, reduced blink rates, perioral tension, and brow furrowing—but also exhibit temporal accumulation patterns over the course of a work shift. This bidirectional relationship between internal states and facial manifestations provides the physiological and psychological foundation for our proposed monitoring approach: by continuously observing subtle facial dynamics, the system infers the latent cognitive and affective states that govern driving performance, thereby enabling timely intervention before these states compromise operational safety.

A review of existing research indicates that current studies on driver suitability factors remain limited in scope, and the integration of psychological and physiological indicators continues to present challenges. Most models overlook implicit factors such as personality traits and cognitive patterns, and their strong dependence on specific industries or scenarios restricts generalizability and scalability. Moreover, existing research predominantly remains at the theoretical or selection-and-training stages, lacking long-term tracking, dynamic early warning mechanisms, and systematic modeling. Consequently, effective analysis and intervention regarding drivers' suitability status remain difficult to achieve.

This study focuses on long-term tracking and analysis of suitability status to enhance practical feasibility. Subsequent research will explore the integration of multiple influencing factors, among other directions.

2.2 The Model of Micro-expression Recognition

Prior to the ascendancy of deep learning, face-expression recognition (FER) predominantly relied upon handcrafted feature extraction methodologies that leveraged domain-specific knowledge from psychology and computer vision. These approaches can be broadly categorized into three families: (1) appearance-based methods, which extract spatiotemporal features directly from image sequences—prominent examples include Local Binary Patterns from Three Orthogonal Planes (LBP-TOP), which encodes texture variations across spatial and temporal dimensions; Histograms of Oriented Gradients (HOG) for capturing edge and shape information; and Optical Flow algorithms for estimating motion vectors between consecutive frames. (2) geometry-based methods, which track the displacement of fiducial facial landmarks (e.g., eye corners, mouth corners, brow positions) and construct feature vectors representing the dynamic deformation patterns of facial structures. (3) hybrid approaches that combine appearance and geometric features to exploit their complementary strengths. While these methods have demonstrated reasonable performance on controlled datasets, they suffer from inherent limitations including (a) vulnerability to illumination variations and partial occlusions; (b) sensitivity to landmark localization errors, which cascade into significant feature distortion; and (c) limited representational capacity to capture the subtle, transient, and spatially localized muscle activations characteristic of genuine micro-expressions. These deficiencies motivated the transition to deep learning architectures, which offer end-to-end feature learning, hierarchical abstraction, and superior robustness to input variations—advantages that have progressively rendered traditional approaches obsolete in contemporary FER research.

Recently, numerous methods have been proposed to improve the accuracy and response speed of FER. Advances in FER technologies have contributed to the refinement of facial recognition systems for public transport drivers. For example, Fan et al. proposed GLA-FERNet, a lightweight automatic neural architecture search method based on gradient optimization for facial expression recognition in resource-constrained environments. The method efficiently designs networks that balance accuracy and computational cost, outperforming manual and reinforcement learning-based approaches in efficiency while maintaining competitive accuracy on benchmark datasets [14]. Similarly, Wang et al. developed a multimodal micro-expression recognition method that integrates visual features with textual semantics. By leveraging text as prior knowledge to guide attention toward discriminative facial regions, the approach effectively addresses the challenges of brief and subtle micro-expressions. Experiments show it outperforms visual-only models across multiple benchmarks, demonstrating the value of cross-modal fusion for accurate micro-expression analysis [15].

Current applications of facial expression recognition technology in driver fitness-for-duty monitoring still exhibit significant limitations, primarily reflected in insufficient applicability and generalizability in long-term real-world driving scenarios. Existing research predominantly focuses on short-term static expression recognition

in laboratory settings, lacking systematic modeling of domain-specific behavioral states in complex driving environments (such as micro-expression variations across fatigue levels or facial dynamic features under stress conditions). Moreover, these studies often fail to adequately account for the challenging conditions of actual in-vehicle environments, including variable lighting, changes in driving posture, and facial occlusions (e.g., glasses, masks), all of which compromise recognition stability.

Furthermore, existing approaches demonstrate notable shortcomings in multimodal data fusion, as they do not effectively integrate complementary information such as eye movements, physiological signals (e.g., heart rate variability), and vehicle operation parameters to enhance the robustness of state assessment. More critically, most studies lack a longitudinal continuous analysis framework for driver states, making it difficult to capture the dynamic patterns of state evolution and thus failing to support long-term, continuous fitness-for-duty evaluation.

These limitations collectively hinder the practical implementation of facial expression recognition technology in real-world driving monitoring scenarios, necessitating breakthroughs through interdisciplinary methodological innovation and scenario-specific data modeling.

In this study, the facial recognition component is designed to enhance adaptability to complex in-vehicle environments and improve medium- to long-term deployability under real driving conditions. Additionally, a dataset specifically tailored for drivers in operational scenarios will be developed to improve recognition capability during driving states. In contrast to the aforementioned state-of-the-art approaches, the distinctive positioning of this study is twofold. First, the methods reviewed above are predominantly designed for general-purpose facial expression recognition tasks, with their architectural designs and optimization objectives failing to adequately account for the specific environmental constraints prevalent inside bus cabins—including drastic illumination fluctuations, frequent head rotations of drivers, and facial occlusions caused by safety masks or eyeglasses. Our proposed framework explicitly addresses these challenges through architectural innovations including Mosaic data augmentation, adaptive anchor computation, and PAN-FPN-based multi-scale feature fusion, thereby achieving scenario-specific adaptability. Second, existing approaches predominantly concentrate on single-frame recognition accuracy, largely neglecting the temporal evolution patterns of driver states. By embedding micro-expression recognition within a long-term continuous monitoring paradigm, our framework captures the dynamic trajectories of fatigue accumulation and stress progression through sequential analysis, thus filling a critical gap in the longitudinal assessment dimension—an aspect conspicuously absent in current literature.

3 System Model

This study proposes a long-term monitoring framework for assessing bus drivers' fitness for duty using facial recognition technology. The framework integrates the YOLOv8n detection model with the VGG-Face feature extractor to construct the Y8n-V Face system, which enables robust micro-expression recognition in complex in-vehicle environments. The system employs adaptive preprocessing techniques—including Mosaic data augmentation, adaptive anchor box computation, and adaptive grayscale padding—to enhance detection stability under variable lighting, occlusion, and posture changes. The extracted facial expression features are analyzed temporally to continuously evaluate drivers' psychological and physiological states. Experimental results demonstrate that the proposed framework achieves reliable recognition performance while maintaining practical deployability, offering a viable approach for proactive safety management in public transportation.

3.1 The Face Detection Model

The face detection model presented in this study is implemented within the PyTorch deep learning framework, employing a computationally efficient and structurally streamlined architecture. As depicted in Fig. 1, its design adheres to a modular and hierarchical organization, consisting of four fundamental and functionally differentiated components: the Input module, Backbone network, Neck structure, and Head output layer. The complete algorithmic logic governing the detection pipeline is further formalized through the pseudocode illustrated in Table 1, providing a clear, stepwise representation of its operational flow.

At the Input processing stage, a suite of three advanced methodologies is systematically integrated to significantly bolster the model's robustness and operational adaptability: Mosaic data augmentation, adaptive anchor box computation, and adaptive grayscale padding. The Mosaic augmentation technique synthesizes a

composite training image by algorithmically stitching four distinct images after subjecting them to stochastic geometric transformations—including random scaling, cropping, and spatial reconfiguration. This process enriches the global contextual information available during training and particularly enhances the model's capability to detect objects of small spatial dimensions. From a theoretical perspective, the efficacy of Mosaic augmentation can be formalized through the lens of data distribution coverage. Let D denote the original training distribution over a single image space X, and D′ represent the augmented distribution after applying random stitching operations. The mixing process generates a new sample x′ by combining four independently sampled images $x_1, x_2, x_3, x_4 \sim D$ with a random cropping mask M, formulated as:

$$x^{'} = \sum_{k=1}^{4} M_k \odot \mathrm{Aug}_k\left(x_k\right), \text{ where} \sum_{k=1}^{4} M_k = 1. \quad (1)$$

Here, $M_k \in \{0,1\}^{H \times W}$ is a binary mask partitioning the composite canvas, $\odot$ denotes element-wise multiplication, and $\mathrm{Aug}_k(\cdot)$ represents stochastic geometric transformations (e.g., scaling, rotation, and color jitter) applied independently to each component. This augmentation strategy induces a substantial expansion in the effective training distribution, effectively increasing the support of the empirical risk minimization objective. Specifically, the expected risk over the augmented distribution $E_{x' \sim D'}[L(f(x'),y]$ provides a tighter upper bound on the true population risk compared to the empirical risk over D alone, as it forces the model $f(\cdot)$ to be invariant to a broader set of spatial configurations and object co-occurrence patterns. Consequently, the model's generalization error bound improves, particularly for detecting small-scale facial instances that are frequently obscured or truncated in single-image views. This theoretical advantage directly translates to enhanced robustness in complex in-vehicle environments where partial occlusions and multi-scale object presentations are prevalent.

Beyond diversifying the training distribution, this strategy promotes batch-level contextual learning, thereby improving overall training efficacy and convergence stability. In conventional detection frameworks, anchor box parameters (dimensions and aspect ratios) are typically pre-defined through heuristic or empirical analysis of dataset characteristics. In marked contrast, the proposed adaptive anchor computation module employs unsupervised clustering algorithms, notably K-means, to dynamically infer anchor configurations that are statistically optimal with respect to the shape and scale distribution of target objects (faces) within the specific training dataset. This data-driven, self-adapting mechanism yields more precise initial bounding box priors, expedites the gradient-based optimization process during training, and demonstrably elevates the recall metric—the model's ability to identify all relevant facial instances.

This clustering process can be formally articulated as follows. Let $G = \{g_1, g_2, \ldots, g_N\}$ denote the set of all ground-truth bounding boxes in the training dataset, where each bounding box is represented by its width and height as $g_j = (w_j, h_j)$. The objective of the adaptive anchor generation algorithm is to partition these bounding boxes into K clusters $C = \{C_1, C_2, \ldots, C_k\}$ such that the within-cluster sum of squared distances is minimized:

$$\arg\min_{C} \sum_{k=1}^{K} \sum_{g \in C_k} \| g - \mu_k \|^2. \quad (2)$$

Where $\mu_k = (w_k, h_k)$ denotes the centroid of the k-th cluster, representing the dimensions of a corresponding anchor box. In contrast to conventional K-means that adopts Euclidean distance, this study employs the inverse of the Intersection over Union (IoU) as the distance metric to better reflect the localization accuracy inherent in detection tasks:

$$d(g,\mu) = 1 - IoU(g,\mu) = 1 - \frac{g \cap \mu}{g \cup \mu}. \quad (3)$$

By minimizing the aforementioned objective function, the clustering algorithm adaptively identifies the most representative distribution of anchor shapes and scales within the dataset, thereby minimizing the prior bias between the initialized anchors and the actual ground-truth distribution. This directly facilitates convergence efficiency during model training and theoretically elevates the upper bound of the recall metric.

Furthermore, during the necessary image resizing operation to standardize input dimensions, padding is applied to preserve the original aspect ratio and avoid spatial distortion. While conventional padding often employs a uniform, constant color value (e.g., black or white), the implemented adaptive grayscale padding

technique computes a locally coherent fill color. This is achieved by analyzing the mean pixel intensity along the perimeter of the original image content. The resultant padding seamlessly blends with the inherent tonal characteristics of the input, thereby minimizing the introduction of spurious visual artifacts or hard edges that could adversely affect subsequent feature extraction and representation learning in the downstream network layers.

By adopting a holistic optimization paradigm that strategically unites data-aware anchor design, intrinsically motivated augmentation strategies, and intelligent adaptive preprocessing, the proposed face detection model attains multiple synergistic benefits. These include accelerated training convergence, markedly improved generalization performance across varied and challenging real-world operational environments (encompassing diverse poses, occlusion levels, and illumination conditions), and ultimately, superior overall accuracy in the core task of face detection. This comprehensive and integrated architectural design not only directly tackles prevalent obstacles in practical deployment but also provides a robust, extensible, and high-fidelity visual front-end. As such, it establishes a critical and scalable computational foundation for the subsequent, more nuanced stage of micro-expression recognition and analysis within the overarching Y8n-V Face framework proposed in this work.

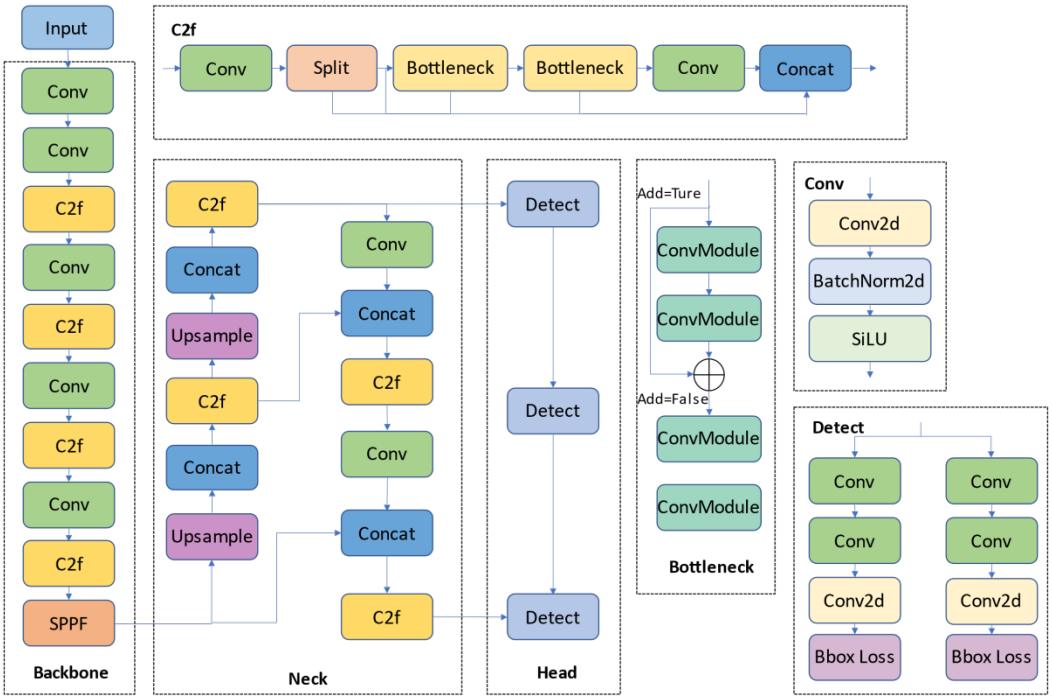


Fig. 1. The Architecture Flowchart of the Proposed Face Detection Model, illustrating the data flow from the Input module through the Backbone (CBS, C2f, SPPF), Neck (PAN-FPN), to the Decoupled Head for final detection

The Backbone network integrates three core architectural modules: the Conv-BatchNorm-SiLU (CBS) module, the C2f module engineered for enhanced gradient propagation, and the Spatial Pyramid Pooling-Fast (SPPF) module. The C2f module represents a structural refinement of the Cross Stage Partial (CSP) architecture, explicitly designed to improve gradient flow by incorporating initial base features from earlier layers with progressively transformed higher-level features through a residual-like feature fusion mechanism. This hybrid integration strategy yields a more diverse and representative feature set for downstream processing stages. Additionally, the channel dimensionality within the C2f module is dynamically configured based on the specific model variant scale, enabling adaptive representational capacity across different computational budgets. The SPPF module is architected to efficiently capture multi-scale contextual information through a series of sequentially applied MaxPooling operations with strategically varied kernel sizes and strides. This cascaded pooling mechanism allows for hierarchical aggregation of features across multiple receptive field sizes, thereby enriching the representational capacity within a unified feature map while maintaining computational efficiency.

The Neck module implements an enhanced Path Aggregation Network combined with a Feature Pyramid Network (PAN-FPN) architecture, which facilitates bidirectional and multi-scale feature fusion. Building upon the classical FPN framework that primarily propagates semantic information in a top-down manner, the PAN-FPN introduces an auxiliary bottom-up pathway dedicated to preserving and propagating fine-grained spatial details. This bidirectional design forms an integrated "dual-directional feature pyramid" that strengthens direct connections across different hierarchical levels of the network. By mitigating information degradation during feature transmission and explicitly reinforcing low-level spatial cues, this architecture substantially improves detection sensitivity, particularly for small-scale and occluded objects commonly encountered in real-world driving scenarios.

The Head module adopts a Decoupled-Head design, wherein the object classification task and bounding box regression task are processed by separate, specialized subnetworks. This structural separation allows each head to optimize its respective objective without mutual interference, leading to more precise and balanced performance. Concurrently, the detection paradigm shifts from an Anchor-Based mechanism to an Anchor-Free approach, eliminating the need for predefined anchor boxes and thereby reducing hyperparameter sensitivity and computational overhead. Finally, a Non-Maximum Suppression (NMS) post-processing stage is applied to filter redundant detections and ensure that each facial instance is represented by a single, high-confidence bounding box, thus enhancing the overall reliability and usability of the detection outputs for subsequent micro-expression analysis.

Table 1. Algorithmic pseudocode of the YOLO-based detection pipeline, detailing the layer-wise feature propagation and the bifurcation between training (loss computation) and inference (NMS post-processing)

```
Algorithm. YOLO detection pipeline
Input: Image tensor X
Output: Detection results D
Begin
    f1 ← Conv(Conv(X))
f2 ← C2f(C2f(Conv(X)))

    p1 ← ConvModule(Conv(Concat(f1, f2)))
    p2 ← C2f(Concat(Upsample(p1), f2))
    p3 ← C2f(Concat(Upsample(C2f(Conv(p2))), f1))

    d_small ← Detect(SPPF(Conv(p3)))
    d_medium ← Detect(p2)
d_large ← Detect(p1)

    if training then
        L ← Bbox_Loss(d_small) + Bbox_Loss(d_medium) + Bbox_Loss(d_large)
        return L
    else        D ← Merge(NMS(d_small), NMS(d_medium), NMS(d_large))
        return D
    end
end
```

The training optimization of the proposed detection head is governed by a composite loss function, which integrates both classification and localization objectives to ensure precise bounding box regression and reliable confidence scoring. Specifically, the total loss L_{total} is formulated as:

$$L_{total} = \lambda_{cls} \cdot L_{cls} + \lambda_{box} \cdot L_{box} + \lambda_{dfl} \cdot L_{dfl}. \tag{4}$$

Where λ_{cls}, λ_{box} and λ_{dfl} are weighting coefficients balancing the contribution of each component. L_{cls} employs the Binary Cross-Entropy (BCE) loss for multi-label classification, defined as:

$$L_{cls} = -\sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]. \tag{5}$$

Here, $y_i \in 0,1$ denotes the ground-truth label for the i-th anchor, and $\hat{y}_i$ is the predicted probability. For bounding box regression, YOLOv8n adopts the Complete Intersection over Union (CIoU) loss, which extends the standard IoU by explicitly incorporating three geometric factors: overlap area, center-point distance, and aspect ratio consistency. The CIoU loss is formally defined as:

$$L_{box} = 1 - \mathrm{IoU} + \frac{\rho^2((b, b^{gt})}{c^2} + \alpha \cdot v. \tag{6}$$

Where b and b^{gt} denote the centroids of the predicted and ground-truth bounding boxes, respectively; $\rho(\cdot)$ is the Euclidean distance between them; c is the diagonal length of the smallest enclosing box covering both; and α is a positive trade-off parameter. The term *v* quantifies the consistency of aspect ratios:

$$v = \frac{4}{\pi^2}(\arctan\frac{\omega^{gt}}{h^{gt}} - \arctan\frac{\omega}{h})^2. \tag{7}$$

As α is computed as:

$$\alpha = \frac{v}{(1 - \mathrm{IoU}) + v}. \tag{8}$$

Additionally, L_{dfl} (Distribution Focal Loss) is incorporated to facilitate more accurate localization by focusing on the top-k probabilities of the predicted distribution. By jointly minimizing these three terms through end-to-end backpropagation, the model learns to simultaneously suppress false positives, refine spatial boundaries, and improve the confidence calibration of each detected facial instance. This mathematically grounded multi-objective optimization framework ensures that the detection module converges to a state where high recall and high precision are synergistically achieved, thereby providing reliable facial region proposals for the subsequent micro-expression recognition stage.

3.2 The Model of Micro-expression Recognition

The micro-expression recognition component of the Y8n-V Face system is constructed upon the VGGNet architectural paradigm, specifically leveraging the VGGFace variant originally developed by the Visual Geometry Group at the University of Oxford. VGGNet's foundational contribution lies in its disciplined and uniform employment of small convolutional kernels—a design principle that distinguishes it from preceding architectures such as AlexNet. In AlexNet, a heterogeneous set of kernel sizes was adopted, including larger kernels (e.g., 11×11 and 5×5) in early layers to capture coarse texture patterns from high-resolution inputs, alongside a gradually increasing proportion of 3×3 kernels in deeper layers. This mixed strategy reflected an initial emphasis on broad receptive fields for early feature abstraction but later shifted toward finer-grained modeling as spatial dimensions contracted and semantic granularity increased. In contrast, VGGNet consistently utilizes 3×3 kernels throughout the entire network, establishing a modular and scalable building-block approach.

The systematic adoption of compact 3×3 kernels confers significant parametric and computational efficiencies. For instance, given a feature map with 64 input channels, a single 3×3 kernel entails only 576 parameters (3×3×64), compared to 7,424 parameters for an 11×11 kernel under the same channel configuration. This parametric economy not only reduces memory footprint and computational overhead but also mitigates overfitting risks. Furthermore, VGGNet increases representational depth not by expanding kernel dimensions but by sequentially stacking multiple 3×3 convolutional layers, each followed by a Rectified Linear Unit (ReLU) activation. This design yields deeper non-linear transformations and richer hierarchical feature representations while maintaining tractable model complexity. Notably, as formally established in the original VGG publication, a stack of two 3×3 convolutional layers can functionally substitute a single 5×5 layer while preserving the effective receptive field, and similarly, three 3×3 layers can replicate the receptive coverage of a 7×7 convolution. This substitution principle yields critical advantages: under equivalent receptive fields, the stacked small-kernel configuration introduces additional non-linear activation points, promotes more discriminative feature learning

through intermediate representations, and enhances gradient flow during backpropagation.

The superiority of stacking smaller kernels over using a single larger kernel can be formalized through parametric analysis and receptive field calculation. Let a convolutional layer with kernel size k and input channel C have $k^2 \cdot C$ parameters. Replacing a single 7×7 convolution with a stack of three 3×3 convolutions (each with the same output depth) reduces the parameter count from 49C to 3×(9C) = 27C, yielding a 45% reduction in parametric complexity. More importantly, from a representational capacity perspective, a stack of three 3×3 convolutions possesses an effective receptive field (ERF) equivalent to that of a single 7×7 convolution. The ERF after L layers of 3×3 convolutions (with stride 1) is given by:

$$ERF = 1 + L \times (3-1) = 1 + 2L\ . \tag{9}$$

For L = 3, ERF = 7, exactly matching a 7×7 kernel. However, the stacked configuration introduces three additional non-linear activation functions (ReLU), enabling the network to learn a more discriminative, non-linear function space. Given an input x, a single large kernel computes a linear weighted sum $f(x) = W_{7\times7} * x$, as the stacked small kernels approximate a more complex function:

$$g(x) = \sigma(W_3 * \sigma(W_2 * \sigma(W_1 * x))). \tag{10}$$

Where σ(·) denotes the ReLU operator. This compositional non-linearity is theoretically proven to be more expressive for modeling the subtle, localized muscle movements characteristic of micro-expressions.

From a representational perspective, the 3×3 kernel is geometrically well-suited to capturing fine-grained spatial patterns. Its compact 3×3 grid centers a focal pixel while integrating immediate neighboring context along vertical, horizontal, and diagonal orientations. When multiple such layers are cascaded, the network progressively abstracts patterns across expanding yet efficiently parameterized receptive fields. For example, stacking three 3×3 layers—effectively emulating a 7×7 receptive region—introduces two supplementary convolutional stages and corresponding ReLU activations, thereby amplifying the model's capacity to model complex, locally invariant features essential for subtle expression discrimination. This architectural choice is particularly salient in micro-expression recognition, where transient facial dynamics require sensitive yet efficient spatial modeling.

The VGGFace model, originally developed by the Visual Geometry Group at the University of Oxford, represents a seminal contribution to the field of face recognition and serves as the foundational backbone for the micro-expression recognition component in our Y8n-V Face system. Introduced in 2015 as a direct application of the VGGNet architecture to the domain of facial analysis, VGGFace was trained on a vast dataset comprising approximately 2.6 million images of over 2,600 celebrity identities, meticulously curated from publicly available web sources. The primary objective of this training regime was not expression recognition per se, but rather large-scale face identification—distinguishing between thousands of distinct individuals based on their facial appearance. This large-scale classification task, however, inadvertently compelled the network to learn highly discriminative and generalizable facial features, as the model was required to capture subtle inter-personal variations while maintaining invariance to intra-personal variations such as changes in illumination, head pose, facial expression, and aging.

The architectural instantiation of VGGFace follows the canonical VGGNet-16 configuration, comprising 13 convolutional layers organized into five sequential blocks, each followed by a max-pooling operation with a 2×2 kernel and stride 2. The filter counts double progressively from 64 in Block 1 to 512 in Blocks 4 and 5, reflecting a deliberate strategy of expanding representational capacity as spatial dimensions contract. Every convolutional layer uniformly employs a 3×3 kernel with stride 1 and padding 1, preserving the spatial resolution before each pooling operation. This strict adherence to small, uniform kernels—a hallmark of the VGG design philosophy—yields two critical advantages: (1) a substantial reduction in parametric complexity compared to architectures using larger kernels, and (2) the introduction of additional non-linear activation layers, which enhance the network's capacity to model complex, non-linear facial feature interactions.

What distinguishes VGGFace from generic VGGNet is its training objective and the corresponding output layer configuration. While the convolutional stages of VGGFace mirror those of standard VGGNet-16, the classification head is adapted to produce predictions over 2,622 distinct face identities. After training, the final softmax layer is typically discarded, and the activations of the penultimate fully connected layer—a 4,096-dimensional vector—are extracted as a compact yet semantically rich facial descriptor. This descriptor has been empirically demonstrated to encode not only identity-discriminative information but also expression-

relevant attributes, as the network's internal representations naturally capture facial landmarks, muscle activation patterns, and geometric configurations that are fundamental to both identity and expression recognition. Consequently, the pre-trained VGGFace model offers a highly effective feature extractor for downstream tasks such as micro-expression recognition, particularly in scenarios where training data is limited, as the transferred features provide a strong inductive bias that accelerates fine-tuning and enhances generalization performance.

Consequently, the adoption of the VGG-based backbone in Y8n-V Face aligns with the requirements for efficient yet expressive feature extraction in embedded or real-time monitoring contexts, supporting robust micro-expression analysis while remaining amenable to deployment in resource-constrained operational environments such as in-vehicle driver monitoring systems.

4 Simulation and Results Analysis

After integrating the face detection module with the micro-expression recognition module to achieve unified face recognition functionality, this chapter systematically evaluates the comprehensive performance of the proposed Y8n-V Face model. The overarching objective of this experimental investigation is not merely to benchmark the proposed model against existing alternatives, but to rigorously scrutinize its operational viability under conditions that faithfully replicate the complexities of real-world public transport environments. To this end, the evaluation protocol is deliberately designed to encompass three distinct yet complementary tiers of analysis: (a) component-level ablation studies to isolate the individual contributions of the detection and recognition modules; (b) system-level performance benchmarking against representative baselines spanning traditional machine learning, lightweight deep learning, and full-scale deep learning approaches; and (c) operational stress testing under simulated environmental perturbations including illumination variability, occlusion density, and pose angularity. This multi-tiered strategy ensures that the assessment outcomes provide not only comparative performance indicators but also actionable diagnostic insights into the model's failure modes and robustness margins, thereby informing targeted optimization pathways for future deployment scenarios. The assessment framework is structured to examine three core dimensions: detection reliability, recognition precision, and operational efficiency under diverse scenarios. Specifically, the evaluation quantifies detection accuracy through metrics including recall, precision, and mean average precision (mAP) under varying environmental conditions such as dynamic illumination, partial occlusion, and multi-pose contexts. Recognition performance is further analyzed in terms of micro-expression classification accuracy, with emphasis on distinguishing fine-grained affective states relevant to driver monitoring—such as fatigue, distraction, and cognitive stress.

Moreover, the real-time capability of the model is rigorously assessed by measuring end-to-end inference latency and throughput under both constrained (CPU-only) and accelerated (GPU-enabled) hardware configurations, simulating real-world deployment settings in public transport vehicles. A comparative analysis is conducted against several baseline and state-of-the-art approaches, including but not limited to OpenCV-based cascade classifiers, MTCNN, and pure deep learning-based recognition pipelines, to objectively identify the relative strengths and functional limitations of the Y8n-V Face architecture. Particular attention is given to its balance between computational complexity and recognition robustness, as well as its adaptability to longitudinal monitoring scenarios requiring sustained operational stability. Through this multi-faceted evaluation, the study aims to substantiate the practical viability of the proposed model and delineate its contribution to the field of automated driver state monitoring.

4.1 Experimental Work

All experiments in this study were executed on a laptop equipped with an AMD Ryzen 7 6800H processor, operating under a consistent software environment built on Python. The implementation leveraged publicly available pre-trained models from the YOLO library for face detection and the DeepFace library for facial feature extraction and recognition, ensuring reproducibility and alignment with established benchmarking practices.

For dataset selection, this research adopts the Real-world Affective Faces Database (RAF-DB) [16], a widely recognized large-scale facial expression dataset designed for in-the-wild expression recognition. RAF-DB contains approximately 15,000 labeled facial images collected from the internet, of which a curated subset of 3,068 images was utilized in this study to maintain a balance between diversity and computational tractability. Each image has been pre-aligned and normalized to a fixed resolution of 100×100 pixels, though such low

resolution may impose inherent constraints on the extraction of fine-grained micro-expression features. The dataset encompasses substantial variation across multiple demographic and environmental dimensions, including age, gender, ethnicity, head pose, illumination conditions, and the presence of occlusions such as eyeglasses, facial hair, and hand gestures. This diversity makes RAF-DB particularly suitable for evaluating model robustness under realistic and heterogeneous conditions, which is essential for assessing driver monitoring systems in real-world public transport environments.

To comprehensively evaluate the recognition performance of the proposed Y8n-V Face framework, a controlled comparative analysis was conducted against several representative baseline models. The selected comparators include: (1) the standalone YOLOv8n detector, representing a modern deep learning-based object detection architecture; (2) a classical OpenCV Haar Cascade classifier, emblematic of traditional machine learning approaches with high computational efficiency on CPU; (3) MTCNN, a widely used deep learning-based face detection pipeline that employs a cascaded multi-task network; and (4) a direct facial recognition method using the VGG-Face feature extractor without prior face detection, serving as an ablation baseline to illustrate the contribution of the detection stage. Each model was applied consistently to the same test subset of RAF-DB under identical hardware and software settings, enabling a fair and reproducible assessment of detection accuracy, recognition robustness, and inference efficiency.

This experimental setup not only validates the relative merits of the proposed Y8n-V Face model but also situates its performance within the broader landscape of existing facial analysis techniques, thereby clarifying its practical applicability and limitations for real-time driver fitness monitoring.

The following section will introduce the parameters employed in the experiments as well as the computational methods.

Precision: The proportion of actual positive samples among those predicted as positive by the model, calculated as:

$$Precision = \frac{TP}{TP + FP}. \tag{11}$$

In the formula: True Positive (TP) refers to the number of samples correctly identified by the model; False Positive (FP) denotes the quantity of samples that were misclassified.

Average Processing Speed: The average number of images processed per second when recognizing all images in the database, calculated as:

$$Average\ Processing\ Speed = \frac{TP + FP}{CT}. \tag{12}$$

In the formula: Consumed Time (CT) indicates the duration required to complete the processing of all images.

Recall: A metric used to evaluate the model's accuracy in identifying all positive samples, which reflects the proportion of actual targets successfully detected by the model, calculated as:

$$Recall = \frac{TP}{TP + FN}. \tag{13}$$

In the formula: False Negative (FN) refers to the number of samples that are actually positive but are incorrectly classified as negative by the model.

4.2 Results Analysis

To better leverage these fundamental data for quantifying the performance of different models in Facial Expression Recognition, this paper employs the F1 Score as a comprehensive evaluation metric to systematically assess the performance of various models in image recognition tasks. The relevant details are provided below.

F1-Score: As the harmonic mean of precision and recall, the F1-score provides a single evaluation metric that balances both. A higher F1-score indicates that the model can effectively control the increase in false positives while maintaining a high recall rate, calculated as:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}. \quad (14)$$

The harmonic mean formulation in Eq. (14) imposes a strict penalty on extreme imbalances between Precision(P) and Recall(R). This property is mathematically expressed by the inequality $F1 \leq \sqrt{P \cdot R}$,with equality holding only when P = R. Consequently, a model achieving a high Precision but extremely low Recall (or vice versa) will yield a significantly dampened F1-Score. For instance, a hypothetical detector with P = 0.9 and R = 0.1 would produce an arithmetic mean of 0.5, but its F1-Score would be merely 2×(0.9×0.1)/(0.9+0.1) = 0.18. This conservative metric is particularly critical in safety-critical applications, such as driver fatigue monitoring, where both minimizing false alarms (high Precision) and ensuring no dangerous state goes undetected (high Recall) are equally indispensable. Therefore, maximizing the F1-Score ensures a balanced optimization of the model's sensitivity and specificity.

Based on the RAF-DB dataset, the proposed model was compared with other models. As shown in Table 2, the model presented in this paper outperforms the other models in terms of Precision, Recall, and F1-score, particularly demonstrating a significant advantage in Precision. The performance of the designed algorithm shows notable improvement, which confirms its effectiveness. However, in terms of processing speed, the Y8n-V Face model performs poorly in Average Processing Speed, particularly lagging significantly behind the OPENCV-based model in speed. This represents a current limitation of the proposed model and a key objective for further optimization.

Table 2. Comparative performance metrics of Y8n-V face and baseline models on the RAF-DB test subset

Model	Precision	Average Processing Speed	Recall	F1-Score
Y8n-V Face	76.0%	7.36 img/s	52.2%	0.619
OPENCV-VGG-Face	61.5%	70.97 img/s	47.4%	0.536
MTCNN-VGG-Face	57.8%	16.45 img/s	48.9%	0.528
VGG-Face	28.9%	1111.1 img/s	28.9%	0.289

The experimental results indicate that both the Y8n-V Face model and the integrated OpenCV-VGG-Face approach demonstrated recognition rates that are comparatively modest when evaluated against absolute performance benchmarks. Through systematic comparative analysis with prior studies in the domain of facial expression and micro-expression recognition, it is reasonable to hypothesize that a contributing factor to this observed performance ceiling lies in the limited spatial resolution of the RAF-DB dataset adopted in this study, wherein all facial images are normalized to a fixed resolution of merely 100×100 pixels. Such constrained resolution inherently compromises the preservation of fine-grained facial texture details and subtle transient muscle movements, which are critical for accurate and robust micro-expression discrimination—especially under suboptimal imaging conditions such as variable illumination, partial occlusion, or non-frontal head poses.

To quantitatively contextualize this speed-performance trade-off, we analyze the theoretical computational complexity of the Y8n-V Face model in terms of floating-point operations (FLOPs). For a standard convolutional layer with kernel size k×k, input channels C_{in}, output channels C_{out}, and output feature map dimensions H×W, the total number of multiply-accumulate operations is given by:

$$FLOPs_{conv} = H \times W \times C_{in} \times C_{out} \times k^2. \quad (15)$$

In the Y8n-V Face backbone, the extensive utilization of depthwise separable convolutions and the C2f module partially mitigates this burden; nevertheless, the cumulative FLOPs remain substantially higher than those of the OpenCV Haar cascade, which relies on computationally frugal integral-image-based feature extraction with O(1) complexity per feature evaluation. Conversely, while MTCNN also adopts a CNN-based architecture, its cascaded design—comprising three progressively shallower networks (P-Net, R-Net, and O-Net)—significantly reduces the average inference cost per image by quickly discarding non-face regions in early stages. Formally, the expected computational cost of MTCNN can be expressed as:

$$E[Cost_{MTCNN}] = C_P + p_1 \cdot C_R + p_2 \cdot C_O. \quad (16)$$

Where C_P, C_P, and C_O are the FLOPs of the three sub-networks, and p_1, p_2 denote the proportions of candidate windows that survive the preceding stages. Since p_1 and p_2 are typically small (e.g., $p_1 < 0.1$), the expected cost remains modest. In contrast, the Y8n-V Face model applies its full deep network uniformly to the entire input image in a single forward pass, regardless of facial presence, resulting in a constant and higher FLOPs count. This architectural distinction explains the observed speed gap. Nevertheless, the depth and representational capacity of the Y8n-V Face model, as formalized in Section 3.2, endow it with superior feature abstraction capabilities, yielding the higher Precision and F1-Score demonstrated in Table 1. This trade-off is deemed acceptable for safety-critical applications where detection reliability takes precedence over marginal throughput gains, provided that strategic optimization (e.g., Tensor RT acceleration or model pruning) is subsequently applied for real-world deployment.

Despite these inherent dataset limitations, the Y8n-V Face model exhibited statistically consistent recognition accuracy and stable inference latency across multiple independent experimental trials. The model's performance remained robust without significant deviation or degradation when evaluated repeatedly on identical sample sets, underscoring its operational reliability and reproducibility under controlled test conditions. This consistency in both accuracy and processing speed suggests that the proposed architecture maintains stable feature extraction and classification capabilities, even when confronted with the inherent constraints of low-resolution input data. Thus, while the absolute recognition rates may be tempered by dataset characteristics, the model demonstrates promising operational stability—a desirable attribute for real-world deployment where consistent performance is often prioritized over peak laboratory accuracy.

Notably, the Y8n-V Face model demonstrated a distinct and consistent advantage in precision over comparative models, as quantitatively validated in Table 1. This suggests that its integrated face detection and region-cropping mechanism provides high-fidelity facial region extraction, which critically supports subsequent expression feature encoding. In practical deployment, implementing a frame-filtering mechanism to discard detection failures or low-confidence outputs could further enhance the reliability of the recognition pipeline, thereby improving the utility and robustness of downstream fitness-for-duty analytic modules.

In terms of computational performance, the Y8n-V Face model achieved superior detection quality but exhibited lower processing speed compared to traditional methods such as the OpenCV-based cascade classifier. This performance trade-off stems from fundamental algorithmic differences: OpenCV's Haar cascade or LBP classifiers are highly optimized for CPU execution through handcrafted feature extraction and efficient sliding-window search, whereas the Y8n-V Face model, as a deep learning-based architecture, relies on parallelized GPU computation to realize its full throughput potential. In the absence of GPU acceleration in this experimental setup, the model's CPU-bound inference efficiency was consequently constrained. Similarly, MTCNN—though also a CNN-based detector—utilizes a shallower cascaded structure and smaller model footprint, enabling relatively faster CPU execution despite also facing computational bottlenecks compared to classical methods.

To quantify this limitation, we formalize the real-time processing constraint as follows. Let f_{input} denote the input video frame rate (typically 30 fps for standard surveillance cameras), and let f_{proc} denote the model's processing throughput (7.36 img/s in our CPU-only configuration). The sampling ratio $r_s = f_{proc}/ f_{input} \approx 0.245$ indicates that the system can process approximately one in every four frames, yielding a temporal gap of $\Delta t = 1/ f_{proc} \approx 136$ms between successive processed frames. For micro-expression detection, where expressions typically last 40–200 ms, this effective sampling frequency of approximately 7.36 Hz may still capture at least one frame within most micro-expression events—specifically, the probability of capturing at least one frame within an event of duration T_{ME} is $P_{capture} = 1 - (1-r_s)^{\lfloor T_{ME} \cdot f_{input} \rfloor}$. For $T_{ME} = 100$ms, $P_{capture} = 1 -(1-0.245)^3 \approx 0.57$, indicating that approximately 57% of micro-expressions would be captured. However, for $T_{ME} < 60$ms, the capture probability drops below 40%, indicating a substantial risk of missing the briefest expressions. This formal analysis establishes a quantitative basis for determining the minimum acceptable processing speed required for reliable micro-expression detection in specific operational contexts, thereby guiding hardware selection and optimization priorities.

From a practical deployment perspective, widespread GPU deployment entails substantial hardware and energy costs, particularly in distributed vehicle-based monitoring systems. Hence, this evaluation adopted a CPU-only configuration to reflect resource-constrained operational scenarios. Under such conditions, the real-time frame processing capability of the Y8n-V Face model may limit the frequency of valid expression samples in continuous monitoring applications. To mitigate this, strategic deployment of selective GPU-accelerated nodes for high-density processing, coupled with model lightweighting techniques—such as network pruning,

quantization, or knowledge distillation—could enhance the effective frames-per-second rate while preserving detection accuracy.

In summary, the experimental evidence indicates that the Y8n-V Face model achieves a favorable balance between high-precision facial expression state extraction and practically feasible processing latency. Its demonstrated reliability in capturing discriminative facial dynamics establishes a technically sound foundation for subsequent continuous assessment of driver fitness for duty, while also delineating clear pathways for further optimization toward real-world, scalable deployment in public transport safety systems.

5 Future Work

It should be acknowledged that the present study employs the RAF-DB dataset, which is primarily designed for macro-expression recognition, whereas our ultimate target is micro-expression-based driver state monitoring. The inherent discrepancies in expression duration, intensity, and annotation granularity may limit the generalizability of our current performance metrics. To address this gap, future work will prioritize two complementary efforts: (1) validation and fine-tuning of the Y8n-V Face model on established micro-expression benchmarks, including CASME II, SAMM, and SMIC; and (2) collection of an in-vehicle micro-expression dataset using high-frame-rate cameras installed in actual bus cabins, with frame-level FACS-based annotation. This multi-dataset strategy will provide a more rigorous foundation for evaluating the model's fitness-for-duty monitoring capability in real-world transit operations.

To achieve objective and reliable assessment of public transport drivers' fitness for duty, this study proposes several directions for future research:

First, the accuracy and robustness of facial expression recognition models in real operational environments need to be further enhanced. This includes retraining the Y8n-V Face model with driving scenario-specific datasets to improve its adaptability to complex in-vehicle conditions, such as varying illumination, partial facial occlusion, and driver posture changes. Additionally, exploring GPU acceleration and model lightweighting techniques could significantly improve real-time processing capabilities, making the system more suitable for large-scale deployment.

Second, a key research direction lies in multimodal data fusion. Facial expression data should be integrated with other physiological indicators—such as electroencephalogram (EEG), electrocardiogram (ECG), eye-tracking metrics, and vehicle operation parameters—to establish a comprehensive driver state assessment framework. By constructing mapping models between multi-source physiological signals and fitness-for-duty states, the system can achieve more accurate and holistic driver monitoring. Future work should also investigate the dynamic correlations between different data modalities and develop fusion strategies that enhance both interpretability and predictive performance.

Third, this study underscores the need for longitudinal tracking and dynamic modeling of driver states. Future systems should incorporate continuous data collection and time-series analysis to capture the evolution of driver states over extended periods. This would enable the early detection of fatigue accumulation, stress trends, and other risk-related patterns, providing a proactive basis for scheduling optimization and timely intervention.

Furthermore, to improve the reliability of the assessment system, error-correction mechanisms and data quality control protocols should be introduced. For instance, invalid or low-confidence detection frames can be filtered or reprocessed, while contextual information—such as shift schedules, road conditions, and weather—could be incorporated to refine state judgments.

Finally, future research should focus on the interpretability and practical applicability of the model. By combining domain knowledge from transportation psychology and human factors engineering, the system can not only output state predictions but also provide actionable insights and management recommendations. This will help public transport operators implement targeted driver support measures, enhance occupational health management, and ultimately improve overall traffic safety and service quality.

To translate the aforementioned research directions into tangible progress, we envision a phased implementation roadmap spanning three temporal horizons. In the short term, priority will be devoted to model optimization and dataset development: specifically, completing the collection and annotation of the in-vehicle facial expression dataset, conducting systematic model pruning and quantization experiments to achieve sub-50 ms inference latency on edge devices, and validating the optimized system through controlled field trials with at least 20 volunteer drivers across diverse routes and shift patterns. In the medium term, the focus will shift toward multimodal integration and longitudinal modeling: deploying synchronized EEG, ECG, and eye-tracking sensors

in test vehicles, constructing the unified driver state knowledge graph, and developing the sequence-to-sequence predictive architecture with early warning capabilities. We aim to achieve a prediction accuracy of at least 85% for fatigue onset within a 30-minute forecast horizon. In the long term, the system will be integrated with existing transit operation management platforms through standardized APIs, enabling seamless data exchange with scheduling systems, vehicle telematics, and driver health records. Pilot deployments in at least three mid-sized cities are planned to evaluate scalability and operational sustainability. This phased approach ensures that each milestone delivers demonstrable safety benefits while systematically laying the groundwork for subsequent advancements, balancing technological ambition with practical feasibility.

In summary, advancing facial expression recognition technology, promoting multimodal data fusion, strengthening longitudinal analysis capabilities, and enhancing system interpretability constitute the main pathways for future research. Through interdisciplinary collaboration and continuous technological innovation, a more intelligent, reliable, and practical driver fitness-for-duty monitoring system can be developed, contributing to the safety and sustainability of public transportation systems.

6 Acknowledgement

This work was supported by Beijing Natural Science Foundation under Grant 9252013.

References

[1] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 39(6)(2017) 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

[2] J. Dai, Y. Li, K. He, J. Sun, R-FCN: Object Detection via Region-based Fully Convolutional Networks, in: Proc. 2016 Advances in Neural Information Processing Systems, 2016. https://proceedings.neurips.cc/paper/2016/hash/577ef1154f3240ad5b9b413aa7346a1e-Abstract.html

[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, SSD: Single Shot MultiBox Detector, in: Proc. 2016 European Conference on Computer Vision, 2016. https://doi.org/10.1007/978-3-319-46448-0_2

[4] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You Only Look Once: Unified, Real-Time Object Detection, in: Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016. https://doi.org/10.1109/CVPR.2016.91

[5] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, in: Proc. 2015 International Conference on Learning Representations, 2015. https://mlanthology.org/iclr/2015/simonyan2015iclr-very/

[6] J. Hu, L. Shen, G. Sun, Squeeze-and-Excitation Networks, in: Proc. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018. https://doi.org/10.1109/CVPR.2018.00745

[7] C. Nakagawa, N. Akiu, T. Watanabe, T. Kojima, S. Yoshie, A. Suzuki, Basic Study on Assessment of the Driver's Condition Based on Physiological Indices, Quarterly Report of RTRI 60(4)(2019) 281-285. https://doi.org/10.2219/rtriqr.60.4_281

[8] M. Mavrin Jeličić, T.J. Mlinarić, P. Zajac, Assessment and Relationship of Work Ability and the Level of Physical Activity Among Professional Bus Drivers, Promet – Traffic & Transportation 35(2)(2023) 175-183. https://doi.org/10.7307/ptt.v35i2.158

[9] E.K. Chung, S.Y. Park, Y.W. Sohn, Effects of occupational drivers' traffic accident history on risk perception, work attitudes, and unstable driving, Transportation Research Record 2675(10)(2021) 734-744. https://doi.org/10.1177/03611981211012695

[10] L. Ye, S.Z. Liu, M. Guo, 5D Psychological Quality Model and Its Application to Key HSR Posts, China Railway (12) (2018) 80-85. https://doi.org/10.19549/j.issn.1001-683x.2018.12.080

[11] Y. Wang, Evaluation of Bus Driver's Driving Behavior based on Entropy Weight TOPSIS Model, Highlights in Science, Engineering and Technology 35(2023) 261-269. https://doi.org/10.54097/hset.v35i.7205

[12] Y.Y. Liu, Development and Preliminary Application of Occupational Psychological Health Assessment Scale for High-speed Railway Drivers, [dissertation] Chengdu: Southwest Jiaotong University, 2020. https://doi.org/10.27414/d.cnki.gxnju.2020.001195

[13] Z.X. Li, G.J. Chen, P.F. Gao, H.D. Liu, Review and analysis of dominant factors influencing the fitness-for-duty state of bus drivers, in: Proc. 2022 World Transport Convention (Transportation Planning and Interdisciplinary Studies), 2022. https://doi.org/10.26914/c.cnkihy.2022.019312
[14] J. Fan, S. Deng, X. Song, J. Liu, Y. Sun, A gradient-based lightweight network automated design method for facial expression recognition, Expert Systems with Applications 296(Part C)(2026) 129130. https://doi.org/10.1016/j.eswa.2025.129130
[15] F. Wang, J. Li, C. Qi, L. Wang, P. Wang, A micro-expression recognition algorithm fusing visual information with textual semantics, Expert Systems with Applications 296(Part C)(2026) 129000. https://doi.org/10.1016/j.eswa.2025.129000
[16] S. Li, W. Deng, J. Du, Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild, in: Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017. https://doi.org/10.1109/CVPR.2017.277