

Article

Multimodal Driver Condition Monitoring System Operating in the Far-Infrared Spectrum

Mateusz Knapik , Bogusław Cyganek *  and Tomasz Balon

Institute of Electronics, Faculty of Computer Science, Electronics and Telecommunication, AGH University of Krakow, Al. Mickiewicza 30, 30-059 Kraków, Poland; mateusz.knapik@electris.pl (M.K.); tomasz.balon@autodesk.com (T.B.)

* Correspondence: cyganek@agh.edu.pl

Abstract: Monitoring the psychophysical conditions of drivers is crucial for ensuring road safety. However, achieving real-time monitoring within a vehicle presents significant challenges due to factors such as varying lighting conditions, vehicle vibrations, limited computational resources, data privacy concerns, and the inherent variability in driver behavior. Analyzing driver states using visible spectrum imaging is particularly challenging under low-light conditions, such as at night. Additionally, relying on a single behavioral indicator often fails to provide a comprehensive assessment of the driver's condition. To address these challenges, we propose a system that operates exclusively in the far-infrared spectrum, enabling the detection of critical features such as yawning, head drooping, and head pose estimation regardless of the lighting scenario. It integrates a channel fusion module to assess the driver's state more accurately and is underpinned by our custom-developed and annotated datasets, along with a modified deep neural network designed for facial feature detection in the thermal spectrum. Furthermore, we introduce two fusion modules for synthesizing detection events into a coherent assessment of the driver's state: one based on a simple state machine and another that combines a modality encoder with a large language model. This latter approach allows for the generation of responses to queries beyond the system's explicit training. Experimental evaluations demonstrate the system's high accuracy in detecting and responding to signs of driver fatigue and distraction.

Keywords: driver's attention; real-time monitoring; YOLOv8; yawning detection; vehicle safety; driver fatigue detection; infrared imaging; automotive; ADAS; modality encoder; large language models LLM



Citation: Knapik, M.; Cyganek, B.; Balon, T. Multimodal Driver Condition Monitoring System Operating in the Far-Infrared Spectrum. *Electronics* **2024**, *13*, 3502. <https://doi.org/10.3390/electronics13173502>

Academic Editor: Xin Geng

Received: 7 July 2024

Revised: 26 August 2024

Accepted: 30 August 2024

Published: 3 September 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Ensuring the safety of drivers, passengers, and pedestrians remains a paramount concern in the rapidly advancing field of automotive technology. The development of Advanced Driver Assistance Systems (ADASs) has significantly contributed to mitigating risks on the road, yet the issue of driver fatigue and distraction persists as a critical concern. Driver fatigue, which impairs alertness and reaction times, significantly contributes to road accidents. Statistical evidence indicates that driver-related errors, including those due to inattention, accounted for over 29% of accidents in 2022 [1]. This underscores the urgent need for effective solutions to monitor and mitigate fatigue-related risks.

Although various research efforts have been undertaken, there still exists a critical need for developing more robust and practical detection systems, especially within naturalistic driving environments, where current methods often fall short, as emphasized by Koay et al. [2].

A comprehensive survey of driver facial expression recognition techniques presented by Saadi et al. [3] points out challenges such as varying illumination conditions, occlusions, and head poses that significantly hinder the accuracy of detection systems. Lambay et al. [4]

further expanded on these challenges by exploring the potential of advanced machine learning techniques in enhancing driver behavior analysis. They also underscore the considerable obstacles posed by data variability, the necessity for real-time processing, and the integration of these systems into existing infrastructures, all of which present significant barriers to widespread adoption and effectiveness.

Current approaches to addressing driver fatigue can be broadly categorized into physiological, vehicle-based methods, and facial feature-based. Physiological sensors, while effective at monitoring vital signs, require direct contact with the driver, which can be intrusive. On the other hand, vehicle-based methods rely on driving inputs like speed and lane markings, but these signals can be ambiguous and less directly correlated with the driver's state, making accurate fatigue detection difficult. Finally, facial feature-based systems analyze indicators such as yawning and blinking to assess fatigue levels. However, these systems often struggle with environmental challenges such as variable lighting conditions, vehicle vibrations, and other interferences.

Recent progress in thermal imaging technology provides a compelling alternative to the traditional methods. Unlike visible light systems, thermal cameras function effectively in total darkness and remain unaffected by extreme lighting conditions. This characteristic makes thermal imaging particularly well-suited for monitoring driver behavior in a variety of challenging environments, thereby mitigating issues related to lighting and color discrepancies. Research conducted by Knapik and Cyganek [5] has illustrated the effectiveness of thermal imaging across both visible and infrared spectra, emphasizing its benefits for background removal and feature extraction. It is also important to notice the trend of constant decline in the prices of thermal imaging cameras.

This paper introduces a novel approach toward a complex issue, which is driver's fatigue detection. With the use of far-infrared imaging cameras and our improved YOLOv8, which is the state-of-the-art object detector, to recognize the driver's condition, facial keypoints are captured and tracked in real-time. Throughout the detection process, numerous factors are taken into consideration, ranging from the most obvious ones, such as yawning, to much less straightforward ones, such as atypical head pose. However, for a comprehensive assessment of the driver's condition, individual detection signals of various phenomena must be analyzed together. The signal fusion module is used for this purpose. We use two of such modules in our system. The first one is constructed as a simple state machine with pre-defined threshold values. Its advantages include simple implementation and very fast response. Thanks to this, this module can be easily implemented in a real-time system. The second fusion module is based on AI. It combines a modality encoder with a large language model (LLM). Thanks to this, it is possible to ask questions and receive answers that the system has not been taught before. Therefore, this module can be used for a deeper analysis of the conditions of fatigue in drivers, e.g., in psychological research.

This paper makes significant contributions in the following areas:

1. Creation of a novel dataset: We present a unique dataset comprising thermal images of people situated inside a vehicle, captured with a camera positioned beneath the rear-view mirror. This dataset encompasses six individuals, including variations such as wearing glasses. Various activities associated with typical driving scenarios, including looking around, talking, and smiling, as well as signs of mental and physical fatigue, like yawning and head-drooping, are represented within the dataset. The data acquisition utilized the cost-effective FLIR E6 camera.
2. Development of a novel face and facial landmark detection model: We introduce an innovative face and facial landmark detection model optimized for thermal images. Based on enhancements to the YOLOv8-face model, which utilizes a ShuffleNet v2 [6] backbone with a simplified detection head, we introduced a convolutional block attention module [7] and a bi-directional feature pyramid network [8] into the architecture, resulting in significant performance gains with minimal computational overhead. Ablation tests were conducted to evaluate the efficacy of the modifications.

3. Introduction of a novel yawning detection model: Yawning detection, particularly in thermal imagery, presents unique challenges for deep learning methodologies due to the scarcity of training data and the temporal nature of such events. We propose a hybrid approach that combines a classic computer vision technique known as the histogram of oriented gradients (HOGs) with a long short-term memory (LSTM) recursive deep neural network. Our method demonstrates robust detection performance, as validated through four-fold cross-validation.
4. In addition to the aforementioned contributions, this paper introduces a novel data fusion technique for fatigue detection based on the large language model. Our methodology involves the integration of data from diverse submodules, including head pose estimators and event detection, such as head drop and yawning. These sequences are then fed to the LLM model for further analysis utilizing a specially crafted prompt. We propose leveraging the LLM model to predict the level of fatigue experienced by the driver. This approach offers several advantages, such as enabling zero-shot prediction utilizing general knowledge of large language models as well as facilitating very easy expansion through the inclusion of additional data or fatigue definition rules expressed in natural language. Furthermore, it opens avenues for extracting fatigue detection rules directly from scientific literature authored by researchers.

This paper is organized as follows. Section 2 contains information on important scientific works in the areas discussed in this article. The general system architecture, as well as its main modules, are presented in Section 3. These are, respectively, the image acquisition module, the face and facial landmarks detector, the head pose and event detection module, the yawning detector, and the two data fusion modules. The datasets used in our experiments are detailed in Section 4. The experiments and their results are presented in Section 5. This paper concludes with a discussion in Section 6.

2. Related Works

Addressing the critical issue of road traffic injuries requires innovative approaches to mitigate the risks associated with driver fatigue and speeding. While existing studies have explored various methods for fatigue detection and speed enforcement, gaps remain in achieving high accuracy and real-time applicability [9].

An extensive review of recent achievements in driver fatigue detection systems can be found in the paper by Sikander and Anwar [9]. As the authors indicate, continuous research is being conducted, and many promising results in constrained environments have been proposed. However, they conclude that significant progress is still necessary. In their review, they categorize driver detection methods into five main groups: subjective reporting, driver biological features, driver physical features, vehicular features, and hybrid features. For further details, interested readers are referred to the paper by Sikander and Anwar [9].

The work by Xiao et al. [10] addresses these challenges by proposing a novel driving fatigue recognition method leveraging feature parameter images and a residual Swin Transformer network. The proposed approach begins with face region detection facilitated by spatial pyramid pooling and a multi-scale feature output module, followed by the localization of 23 key facial points using a multi-scale facial landmark detector. By computing the aspect ratios of the eyes and mouth based on these keypoints, a feature parameter matrix is constructed to represent fatigue driving. Subsequently, this matrix is transformed into an image format, enabling the utilization of a residual Swin Transformer network for fatigue-driving recognition.

Another method for distracted driving detection was proposed by Mohammed et al. [11]. It uses a lightweight vision transformer trained with pseudo-label-based semi-supervised learning. They addressed the challenges of extensive data labeling and large model sizes by leveraging both labeled and unlabeled data to train the model efficiently. By incorporating a hybrid lightweight transformer model into a teacher–student network and generating

pseudo-labels from weakly augmented data, their approach achieved real-time, accurate detection of distracted drivers while maintaining a compact model size.

In the realm of fatigue detection systems, Ardabili et al. [12] presented a multi-class driver fatigue detection system based on electroencephalography (EEG) signals using deep learning networks. By integrating physiological indicators and advanced signal processing techniques, the study achieves remarkable accuracy in detecting fatigue across multiple stages, offering promising implications for real-time driving safety applications.

Similarly, Jiang et al. [13] investigated the modulating effects of olfactory stimuli on alertness within a monotonous driving context. They explored the neural responses to olfaction-modulated alertness using EEG signals and developed an objective EEG-based classification algorithm to predict alertness states induced by olfaction. The authors extended their previous work, which tracked vigilance and fatigue in driving through EEG markers. To the best of their knowledge, this represents the first effort to develop a wearable EEG-based method for characterizing olfaction-induced alertness in driving settings.

On the other hand, Abdrakhmanova et al. [14] published the SpeakingFaces dataset, which represents a significant contribution to the field of multimodal machine learning, offering researchers a publicly available resource for exploring the integration of visual, thermal, and audio data streams. With applications spanning human–computer interaction (HCI), biometric authentication, recognition systems, and speech recognition, the dataset comprises synchronized high-resolution thermal and visual image streams of fully-framed faces, accompanied by audio recordings of approximately 100 imperative phrases spoken by each subject.

In the area of data acquisition, Kuzdeuov et al. [15] introduced a thermal face dataset with annotated face bounding boxes and facial landmarks. This dataset addresses the scarcity of research in the area of facial landmark detection. The dataset comprises 2556 images of 142 individuals, each annotated with 54 landmarks across key facial features such as eyebrows, eyelids, nose, lips, chin, and face outline.

Zeng et al. [16] proposed an intelligent detection method based on a specifically tailored YOLOv8. They collected images under diverse lighting conditions and enhanced the data quality using the Laplacian image enhancement algorithm. Additionally, they incorporated the CBAM attention mechanism and the EIOU loss function to prioritize crucial features and refine box regression, respectively, resulting in improved detection accuracy.

In the paper by Cheng et al., an assessment of driver mental fatigue based on facial landmarks is presented [17]. In their work, a driving simulator-based experiment was conducted, during which 21 videos were recorded. These recordings enabled the computation of the eye and mouth aspect ratios for detecting facial landmarks. Mental fatigue detection was then conducted based on several feature candidates. However, their experiments were conducted exclusively within the visible light spectrum.

In their article, Wang et al. [18] proposed a novel class-level fatigue noise-tolerant supervised contrastive learning (cFNSCL) method to address the challenges of noise in fatigue detection caused by inherited fine-grained labels. They introduced a dynamic noise-tolerant contrastive loss (DNCL) and a class-level confidence assessment mechanism (CCAM) to select high-confidence samples, significantly enhancing model accuracy and tolerance to noise. Their approach demonstrated notable improvements in both synthetic and real-world noisy datasets.

On the other hand, Zhang et al. [19] proposed a novel framework called cross-to-merge training (C2MT) to enhance the robustness of deep neural networks trained on noisy labels. Unlike traditional sample selection methods, C2MT introduces a cross-to-merge strategy that iteratively applies cross-training and merge-training processes to two networks, effectively reducing the impact of noise and ensuring stable performance across various noise rates and types. Additionally, they introduce the median balance strategy (MBS) to further refine sample selection.

Long short-term memory (LSTM) networks, also used in our system, are a type of recurrent neural network (RNN). Due to their ability to effectively model and learn from

temporal sequences, they have been extensively studied and utilized in various applications. LSTM networks, originally introduced by Hochreiter and Schmidhuber in [20], address the vanishing gradient problem inherent in traditional RNNs, allowing them to maintain and propagate information over longer sequences.

While deep learning models like CNNs and RNNs have shown promise in facial emotion recognition, their deployment in real-world driving scenarios remains challenging due to the limitations of existing datasets and the computational demands of in-vehicle systems. Nabipour et al. [21] emphasized the need for frameworks that utilize the facial action coding system (FACS) to bridge the gap between lab-based datasets and real-world applications, ensuring more accurate and efficient emotion detection in automotive environments.

Islam et al. [22] proposed an innovative end-to-end deep learning model to recognize facial micro-expressions based on apex frames, addressing challenges such as low-intensity facial movements and the scarcity of publicly available spontaneous datasets. They utilized a two-stage transfer learning approach and fine-tuned their model on multiple benchmark datasets, achieving higher accuracy.

Ma et al. [23] introduced a feature-level fusion method, as opposed to decision-level fusion, leveraging multi-head self-attention (MHSA), which significantly outperforms conventional methods. Their system, enhanced by a novel supervised contrastive learning framework (SuMoCo), has demonstrated superior performance in detecting driver actions and improving robustness against view and modality collapses. It achieved state-of-the-art results on the DAD dataset.

Our previous publications have significantly contributed to the domains of driver fatigue recognition and thermal image processing. For instance, Knapik and Cyganek [5] proposed a system to detect driver fatigue based solely on yawning detection in thermal images. This method employs background removal using thermal thresholding and template matching to identify the facial area. Subsequently, the face is aligned by detecting the corners of the eyes. The yawning reflex is detected through a novel approach involving thermal voxel counting and dynamic threshold estimation.

The issue of the lack of large-scale thermal image datasets was addressed by Knapik and Cyganek in [24]. They proposed a novel method for eye detection in thermal images, which can also be used to bootstrap the automatic data annotation process. Their approach involves pre-processing the input thermal image with a virtual high-dynamic range algorithm, significantly enhancing the thermal image contrast. This enhancement allows for more reliable computation of sparse image descriptors. They compared the bag-of-visual-words approach with clustering and YOLOv3 for eye detection in thermal images, demonstrating the effectiveness of their methods.

In another study, Balon et al. [25] discussed object detection and classification in the thermal spectrum for automotive systems. Recognizing the limitations of visible spectrum methods under poor lighting conditions, they presented a thermal video database with thousands of annotated frames. This database was utilized to train a YOLOv5-based network optimized for thermal images. The main contributions of the paper include the thermal dataset, a pre-trained YOLOv5 model for object detection in the thermal spectrum, and an application for car speed measurement using thermal images. This system highlights the potential use of thermal imaging in advanced driver-assistance systems (ADASs) and autonomous driving, showcasing its advantages in various lighting conditions.

In a follow-up paper, Balon et al. [26] extended the Thermal Automotive Dataset introduced in their previous work by adding over 2000 new images and developing two new object detection models based on YOLOv5 and YOLOv7 architectures. Emphasizing the importance of dataset size, they compared the performance of both models to determine their reliability and effectiveness in detecting small objects in the thermal spectrum. Additionally, they analyzed the impact of preprocessing techniques on thermal imaging datasets and the models trained on them. This contribution expands the resources available for object detection research in automotive settings and provides valuable insights into optimizing thermal imaging systems for real-world applications.

It is important to note that the aforementioned works address only selected aspects necessary for a comprehensive assessment of drivers' conditions. They are often based on slightly older solutions, which may not be suitable for operation in real conditions in a moving vehicle [9]. In this paper, we aim to fill this gap by proposing a holistic AI-based system that operates exclusively with thermal images. Utilizing modern AI detectors and inference methods, our system allows for advanced situation assessments. We believe that the appropriate approach to analyzing the condition of drivers in various day and night conditions can be reliably based on the use of thermal images. This comprehensive system represents a significant advancement in driver monitoring, offering enhanced reliability and effectiveness in detecting and responding to signs of fatigue and distraction.

3. System Overview and General Assumptions

In this paper, we propose a novel system for fatigue monitoring based on the multi-modal analysis of far-infrared images. Utilizing the long-infrared spectrum, our system operates effectively in both daytime and nighttime conditions without requiring additional light sources

Figure 1 presents the block diagram of the proposed system. The process begins with image acquisition, followed by a cascade of specialized modules where each block utilizes data extracted by its predecessor. The three main components of the system are the face and facial features detector, the head pose estimator, and the yawning detector. These are followed by a fusion module for holistic driver behavior analysis. The algorithms, their modifications, and their performance are discussed in the following sections.

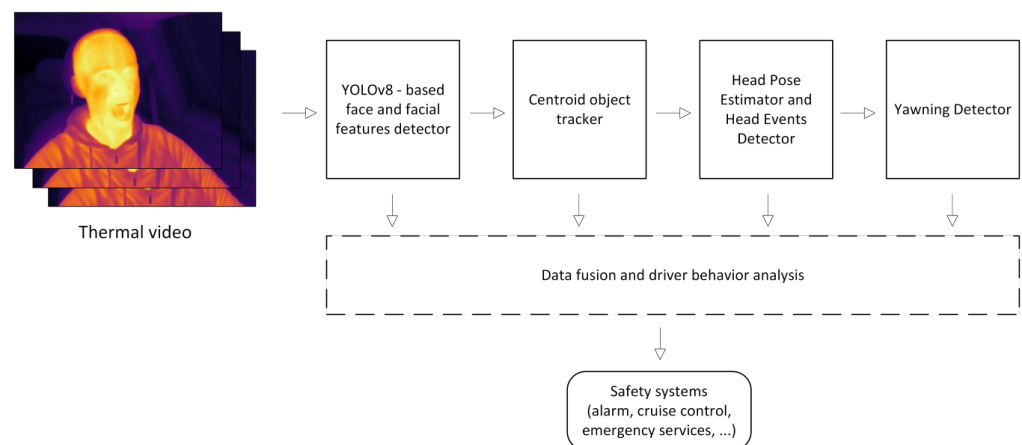


Figure 1. Block diagram of the proposed driver monitoring system.

3.1. Thermal Image Acquisition

Previously proposed methods [5,24] utilized hand-engineered features that leveraged thermal information to remove background and preprocess images. This approach presents additional challenges when integrating images from different sensors and environments, necessitating careful calibration procedures to adapt algorithms to these changes.

To overcome these limitations, our paper proposes normalizing thermal images using a procedure described below. On the other hand, we increase the augmentation of the training data to enhance the robustness of the entire processing pipeline. This approach allows for the combination of datasets from different sources, created with a wide variety of capturing devices and scenes.

Thermal camera images are captured as raw files, with temperatures represented as 14-bit linear values. To remove outliers, the images are clipped at the 1st and 99th percentiles. They are then quantized by rescaling to an 8-bit single-channel image format.

3.2. Face and Facial Landmark Detection

Accurate face and facial landmark detection is a crucial component of driver fatigue monitoring. In the long-wave infrared spectrum, previous methods have relied mostly on hand-engineered features and classical image-processing techniques due to a lack of sufficiently large datasets. This often leads to struggles with variations in camera position, occlusion, and facial expressions. Due to advancements in the deep learning field as well as the recent availability of new thermal image datasets, in this paper, we employ the YOLOv8 deep learning object detector, leveraging its robustness and efficiency in real-time object detection tasks.

YOLOv8 (You Only Look Once version 8) represents a significant advancement in the field of deep learning-based object detection. It is designed to detect objects within an image in a single forward pass, making it exceptionally fast and suitable for real-time applications. YOLOv8 combines the benefits of previous YOLO versions with improved accuracy and speed, thanks to its refined architecture and optimization techniques. However, for face localization, specialized versions employ lightweight backend networks [27,28]. It results in a very efficient algorithm but comes at a cost of reduced performance. In this work, we utilize our modified YOLOv8-face network to detect faces and facial landmarks, ensuring precise detection and localization without increasing computational burden.

The choice of YOLOv8 for face and facial landmark detection is motivated by its ability to handle complex scenarios, including varying poses, occlusions, and diverse environmental conditions. Unlike traditional methods that require extensive preprocessing and calibration, YOLOv8's end-to-end deep learning approach simplifies the detection pipeline. By training the model on a comprehensive dataset with extensive augmentation, we enhance its robustness and generalizability, enabling it to perform consistently well on real-world data.

In this section, we detail the methodology employed for modifications and training of the YOLOv8-face model, the dataset preparation, and the specific techniques used to ensure high detection accuracy. We also discuss the performance metrics used to evaluate the model and present both quantitative and qualitative results to demonstrate its effectiveness in detecting faces and facial landmarks under various conditions.

As shown in Figure 2, the proposed model takes the lightweight object detection model, YOLOv8-face, as the basic model and, consequently, enhances it with convolutional attention block modules (CBAMs) presented in [7] and a modified feature pyramid network based on bi-directional feature pyramid network (BiFPN) proposed by Tan et. al in [8].

Due to thermal imaging limitations, the image quality in comparison to visible light is inferior. Edge details are less distinguishable and often disappear due to thermal noise; texture features are lost, making the distinction between facial features low and unclear. Additionally, the temperature distribution in the facial area, especially important for the research presented in this paper, changes constantly over time due to environmental and emotional changes.

In this paper, we propose using convolutional attention blocks to enhance the feature refinement capabilities of the detection network. Specifically, we integrate CBAM, which employs both channel and spatial attention mechanisms. This dual attention approach allows the network to focus more effectively on the target of interest during the detection process.

The channel attention mechanism uses an average pooling method combined with a fully connected multilayer perceptron and sigmoid activation to determine the significance of each channel. This is then multiplied with the input feature map to highlight important channels. Conversely, the spatial attention mechanism combines pixel-wise average and max pooling, processed by a convolutional layer, to predict and emphasize important spatial features.

By incorporating CBAM into our detection network, we aim to improve the network's ability to focus on relevant features, thereby enhancing the overall detection performance.

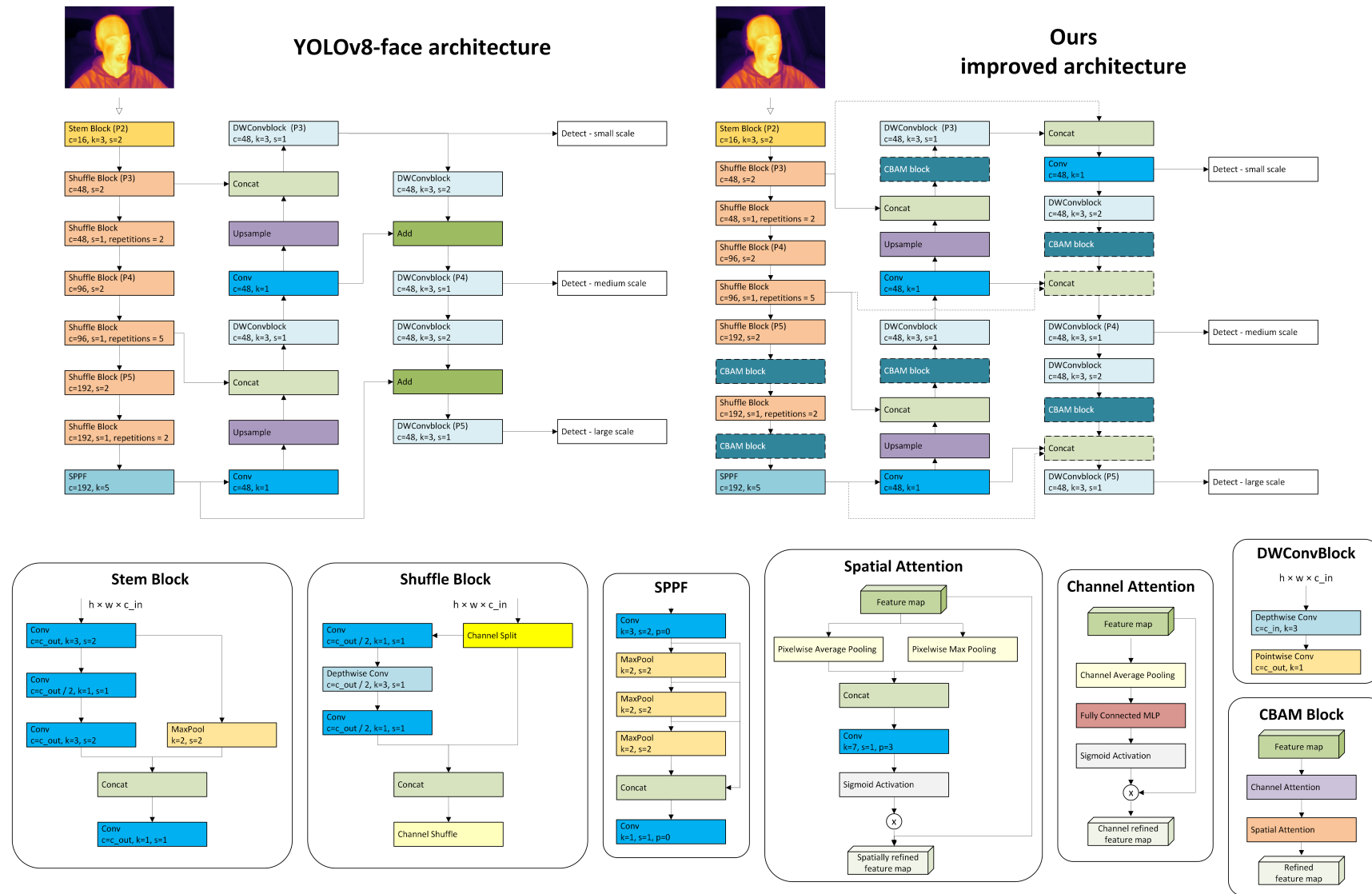


Figure 2. Block diagram and comparison of standard YOLOv8-face architecture and our improved version. Dashed lines show added blocks and connections.

Due to the inherent reduction of spatial dimensions in convolutional neural networks, many details are irreversibly lost in the deeper layers of these architectures. This loss of detail can significantly impact the performance of detection networks, especially in tasks that require high precision, such as facial feature detection. To address this issue, we incorporated concepts from the bi-directional feature pyramid network (BiFPN) as presented in the EfficientDet paper [8], which enhances feature propagation from shallow layers to deeper ones.

BiFPN is designed to improve the flow of information across different scales in the network by introducing bi-directional connections. These connections allow features to be shared more effectively between layers, facilitating better feature fusion and retention of critical information. BiFPN, in its original version, employs weighted feature fusion, where each input feature is assigned a learnable weight, enabling the network to prioritize more informative features dynamically. In our architecture, since most informative features were already filtered and selected by CBAM blocks, we opted for static, non-learnable feature concatenation.

To integrate BiFPN into our detection network, we added additional connections from the backbone network to the feature pyramid. These connections ensure that high-resolution features from the shallow layers are propagated to deeper layers, which helps maintain the integrity of spatial details throughout the network. This enhancement is particularly beneficial for thermal images, where preserving as much detail as possible is crucial due to their inherently low spatial dimensions.

The synergy between BiFPN and CBAM in our proposed architecture enables the network to retain more detailed information from the input thermal images, improving detection accuracy. By enhancing feature propagation and refining feature attention, this combined approach effectively addresses the challenges posed by low-resolution thermal images, leading to a more robust and precise detection system.

To evaluate the impact of the proposed modifications on the performance of the facial feature detector, we conducted a series of experiments using the challenging outdoor dataset, Thermal Faces in the Wild (TFW) [29]. Multiple models were trained for 300 epochs, each with an input resolution of 320×320 pixels. We also increased the pose loss gain parameter from its default value of 1.0 to 12.0 to enhance focus on facial keypoint localization. Additionally, multiplicative noise augmentation was applied to better align the TFW dataset with the characteristics of the low-cost thermal camera employed in our system.

To guide the architectural design of the final model, we performed ablation studies, progressively incorporating key components. Initially, we integrated only the convolutional block attention module (CBAM) into the model, followed by the addition of the bi-directional feature pyramid network (BiFPN) in a separate experiment. Finally, both CBAM and BiFPN were combined, resulting in a model that achieved the best performance. The outcomes of all training experiments are summarized in Section 5.

3.3. Head Pose Estimation and Event Detection

Head pose estimation plays a crucial role in driver fatigue monitoring, as it provides vital information about the driver's alertness and attentiveness. The orientation of the head can indicate signs of drowsiness, distraction, or microsleep, which are critical factors in preventing accidents and ensuring road safety. By accurately determining the head pose, systems can detect when a driver's head nods off, turns away from the road, or shows other signs of fatigue, enabling timely alerts and interventions.

In our work, we estimate head pose using 3D–2D point correspondences by leveraging Levenberg–Marquardt optimization provided by the OpenCV library. Facial keypoints detected by the YOLOv8 model are matched to a 3D face model (presented in Figure 3a) to calculate yaw, pitch, and roll angles. These angles are critical for understanding the driver's head orientation in real time.

Given that input data can be noisy, especially in real-world applications where camera vibration is present, we analyzed recorded videos of actual driving sessions to refine our

approach. Through empirical analysis, we defined six head poses commonly observed during driving: normal (en face), looking left, looking right, head up, head down, and unknown (head not detected or angles outside of normal range). These predefined poses help in mitigating the effects of noise and vibrations. The computed angles are then assigned to the closest predefined pose using the method described in Algorithm 1. Results can be seen in Figure 3.

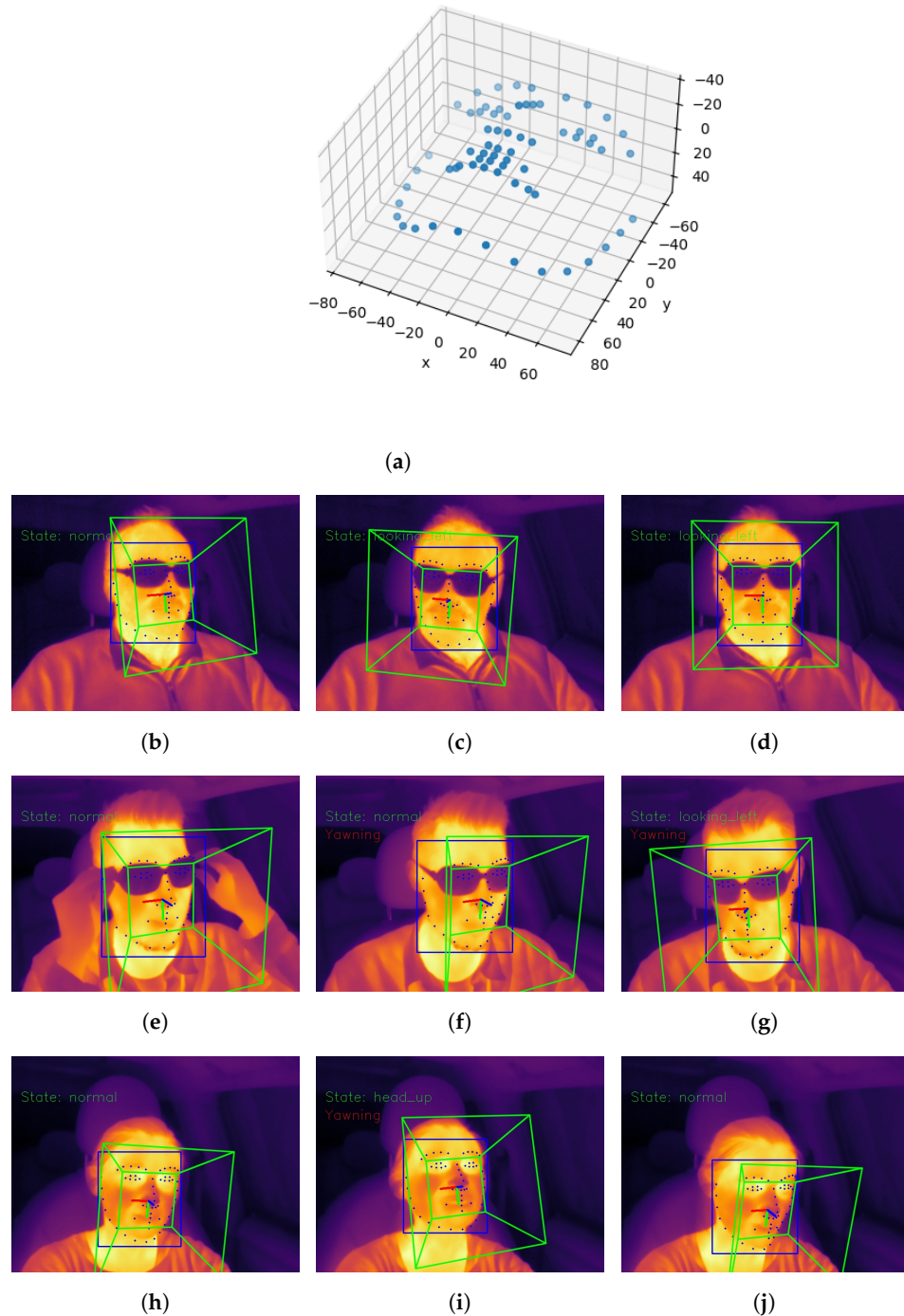


Figure 3. Visualization of a 3D face model (a) and examples of head pose estimation (b–j).

3.4. Yawning Detection

In our previous work [5], we presented a yawning detection algorithm based on thermal voxel counting to identify rapid temperature changes in the mouth region. Although this method is still viable, it has several drawbacks. Measuring small temperature changes requires a sensitive and linear imaging sensor, both spatially and temporally, which is

significantly more expensive than simpler thermal imaging sensors. Additionally, it necessitates processing raw sensor data and careful calibration, making it susceptible to ambient temperature variations and subject-specific differences.

Algorithm 1 Head pose detection.

```

1:  $X_{kpts}$ : facial landmarks from the current frame
2:  $M_{face}$ : 3D face model
3:  $c_{offset}$ : camera angle offset
4:  $Y$ : a set of pose angles  $[y_1, y_2, \dots, y_p]$ , where  $y_p = [y_{yaw}, y_{pitch}, y_{roll}]$  and each  $y_a$  is a pair of minimal and maximal angle values for the  $k$ th angle in the  $p$ th pose.
5: procedure FACEPOSEESTIMATION( $X_{kpts}, Y$ )
6:    $P \leftarrow [yaw, pitch, roll]$  - Calculate face pose from 3D–2D point correspondences using Levenberg–Marquardt optimization [30]
7:   Add  $c_{offset}$  angles to  $P$ 
8:   for each pose  $y$  in  $Y$  do
9:     if  $\forall i \in P, y_i^{min} \leq P_i \leq y_i^{max}$  then
10:       $d_y \leftarrow \min(\text{dist}(y^{min}, P), \text{dist}(y^{max}, P))$  - minimal Euclidean distance from  $P$  to  $y$ 
11:     end if
12:   end for
13:   return pose  $y$  with minimal distance
14: end procedure

```

To address these issues and create a more robust yawning detection process, we propose a novel method based on geometrical features and a recurrent neural network (RNN). Our proposed algorithm consists of a feature descriptor and a sequence classifier. For the feature descriptor, we chose the robust and well-known histogram of oriented gradients (HOG) [31].

The histogram of oriented gradients (HOG) algorithm, introduced by Dalal and Triggs in [31] for people detection, is a widely utilized feature descriptor in computer vision and image processing, particularly for object detection. It captures shape and appearance information by encoding the distribution of gradient orientations within localized regions of an image.

The image is divided into small, non-overlapping regions known as cells. Within each cell and for every pixel, a histogram of gradient orientations is computed according to Equations (1) and (2), which provides information about the local edge directions and strengths.

$$\text{Angle} = \theta = \frac{g_x}{g_y} \quad (1)$$

$$\text{Magnitude} = \sqrt{g_x^2 + g_y^2} \quad (2)$$

$$L_2(v) \rightarrow v / \sqrt{\|v\|_2^2 + \epsilon^2} \quad (3)$$

where g_x and g_y denote the two spatial gradients, computed in the x and y directions of an image, respectively.

To enhance robustness to changes in illumination and contrast, the concatenated histogram is then normalized to reduce the impact of varying lighting conditions. This normalization is typically performed using techniques such as L2-norm (Equation (3)) or L2-Hys (L2-norm with limiting maximum values of v to 0.2) normalization.

The HOG feature descriptor is constructed by concatenating the normalized histograms from all blocks in the image. Each block contributes a feature vector based on its histogram, and these vectors are combined to form the final HOG feature vector for the entire image.

We opted for a non-learnable feature descriptor due to the limited number of yawning images available in thermal datasets.

To detect yawning, we classify changes in the geometrical features of the same facial region over multiple frames. For this purpose, we utilize the long short-term memory recurrent neural network (LSTM) introduced by Hochreiter and Schmidhuber in [20].

As already mentioned, LSTM networks are specialized forms of recurrent neural networks (RNNs) designed to effectively capture and learn long-term dependencies in sequential data. Traditional RNNs struggle with learning dependencies over long sequences due to the vanishing or exploding gradients during backpropagation. This issue hampers their ability to retain information over extended periods, limiting their effectiveness in tasks that require understanding long-range temporal dependencies.

The architecture of LSTMs is explicitly designed to address this problem. It uses a unique structure with three gating mechanisms—input, forget, and output gates—that control the flow of information, allowing the network to retain or discard information as needed. The cell state acts as a memory unit, carrying relevant information through the sequence, while the hidden state provides the output at each time step. This design enables LSTMs to perform well on tasks requiring the retention of context over extended sequences, such as natural language processing [32,33] and time-series analysis [34,35] or healthcare [36,37]. More details on the variants and operations of LSTMs can be found in [38].

A crucial aspect of our system is the ability of long short-term memory (LSTM) networks to classify temporal signals with high accuracy even when trained on relatively small datasets, as evidenced by the work of Drzazga and Cyganek in [37]. Their findings make LSTM a particularly advantageous choice in scenarios where large-scale annotated data are not readily available.

A block diagram of the proposed method is presented in Figure 4. The mouth region (defined as a rectangle bounded by the nose tip, left and right mouth corners, and chin keypoints) is resized to a rectangle of size 48×48 pixels; the HOG feature descriptor of this region is computed for each frame of the sequence. Features are then fed to the RNN classifier to obtain the final classification result, i.e., the subject is yawning or not.

We empirically defined the length of a yawning sequence to be 48 frames, which equates to 3.2 s at a frame rate of 15 frames per second (FPS). A sequence is classified as yawning if the subject is actively yawning during 24 or more of these frames. Training parameters are shown in Table 1.

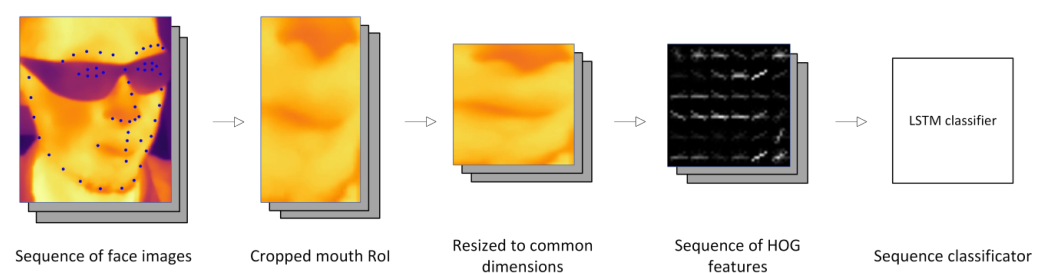


Figure 4. Block diagram of the yawning detection module.

Table 1. Training parameters for the LSTM classifier.

| Parameter | Value |
|-----------------|--------------------|
| Sequence length | 48 |
| Batch size | 32 |
| Loss function | Cross-Entropy Loss |
| Epochs | 100 |

Table 1. Cont.

| Parameter | Value |
|-------------------|-------|
| Input size | 900 |
| Hidden layer size | 600 |
| Learning rate | 0.001 |

3.5. Data Fusion Methods

As shown in Figure 1, the specialized detectors provide information on head positions and dynamic facial features, such as a head drop and detected yawning. Information from these streams is then fed to the fusion module. In our system, we prepared two versions, as follows:

1. A base algorithm that relies on empirically collected threshold values for distracting events. When these thresholds are exceeded, they indicate proper alarm conditions. This simple algorithm can be easily implemented, for example, on edge devices, and operates in real-time conditions.
2. A large language-based model that utilizes modality encodings provided by our system from the information streams. These are used with well-defined prompts for more advanced situational analysis and even zero-shot question answering. This advanced fusion module can be used, e.g., by driver psychologists to research more complex driving conditions.

These two methods are discussed in the following subsections.

3.5.1. State Machine Approach

A pseudo-code for the base solution of a simple inference module is presented in Algorithm 2. Its main task is to appropriately aggregate the occurrences of each event, related to various forms of a driver's distraction, and then trigger an alarm when an experimentally determined threshold is exceeded.

For the base fusion module, we generate three streams of the driver's activity signals:

1. PERLOOK.
2. Yawn frequency.
3. Head drop frequency.

These signals provide comprehensive insights into the driver's state, facilitating accurate detection of fatigue and distraction.

PERLOOK, a metric similar to PERCLOS [39], measures the percentage of time the driver's head is oriented away from its forward-looking normal pose. This metric is calculated over a time window, which we empirically set to 8 s. Duration and frequency of the driver's head turning away from the road indicates the distraction level.

Yawning frequency measures the rate at which yawning reflexes occur. Frequent yawning is a well-known indicator of drowsiness and fatigue, making this metric crucial for timely intervention. Similarly, head drop frequency measures the occurrence of sudden head drops, which are often associated with microsleep and severe fatigue. Both metrics are essential for a comprehensive assessment of the driver's tiredness level. In our work, we measure these metrics in events per minute. To calculate them, we divide the number of yawning events and head drop events by the length of a selected time window (minutes). To normalize the length of each event, we divide the number of reported frames for each event by the typical length of the yawning reflex and head drop event.

These three signals are combined using a weighted sum approach, as shown in Algorithm 2. The weights are determined based on the relative importance of each signal in indicating fatigue and distraction. If the combined signal exceeds a specific threshold, a distraction alert and/or fatigue alert is triggered. These alerts are reported to the control systems, which can then respond by warning the driver, prompting them to take a break or perform other safety measures.

To establish the thresholds and multipliers, several hypotheses were initially formulated regarding the impact of yawning, head drooping, and head pose on indicators of tiredness and distraction. Then the recorded data were reviewed to assess the typical durations of these sequences and their relative positions. However, it was observed that the performance of the system was not notably sensitive to these specific thresholds. This aspect could be further addressed in future research, potentially involving collaboration with physicians and psychologists, as well as the collection of additional data to ensure robust performance in real-world applications.

Additionally, we also propose a novel system based on a large language model (LLM) that eliminates hard thresholds and allows for the creation of a more sophisticated and adaptable fatigue alerting system.

Sample plots generated by the base fusion module are presented in Figure 6, while Figure 5 shows sample images with output data overlaid on the top. Threshold values and constants were established empirically using data from our dataset presented in Section 4.3 and are presented in Table 2.

Algorithm 2 Base fusion method for alarm detection

- 1: X_i : I input streams of the driver’s activity signals
 - 2: w_i^d, w_i^f : I weights for the activity signals of drivers, d —distraction, f —fatigue
 - 3: $A_d \in \{0, 1\}$: distraction alarm On/Off
 - 4: $A_f \in \{0, 1\}$: fatigue alarm On/Off
 - 5: τ_d, τ_f : distraction and fatigue alarms thresholds
 - 6: **procedure** DISTRACTIONSIGNALSFUSION(X_i, A)
 - 7: $X_d = \sum_1^I w_i^d X_i$ ▷ distraction score
 - 8: $X_f = \sum_1^I w_i^f X_i$ ▷ fatigue score
 - 9: $A_d = 0, A_f = 0$
 - 10: **if** $X_d \geq \tau_d$ **then**
 - 11: $A_d = 1$ ▷ distraction alarm detected
 - 12: **end if**
 - 13: **if** $X_f \geq \tau_f$ **then**
 - 14: $A_f = 1$ ▷ fatigue alarm detected
 - 15: **end if**
 - 16: **return** return (A_d, A_f)
 - 17: **end procedure**
-

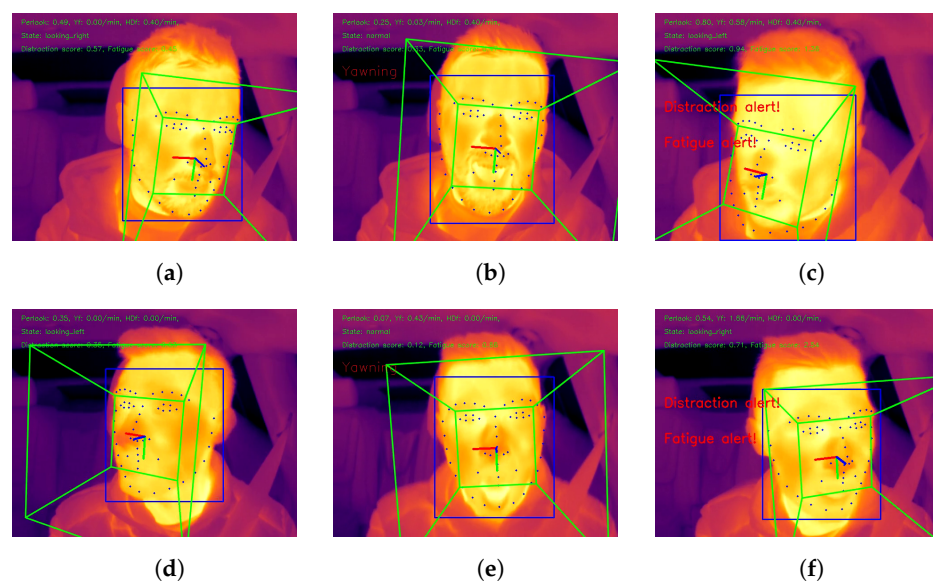
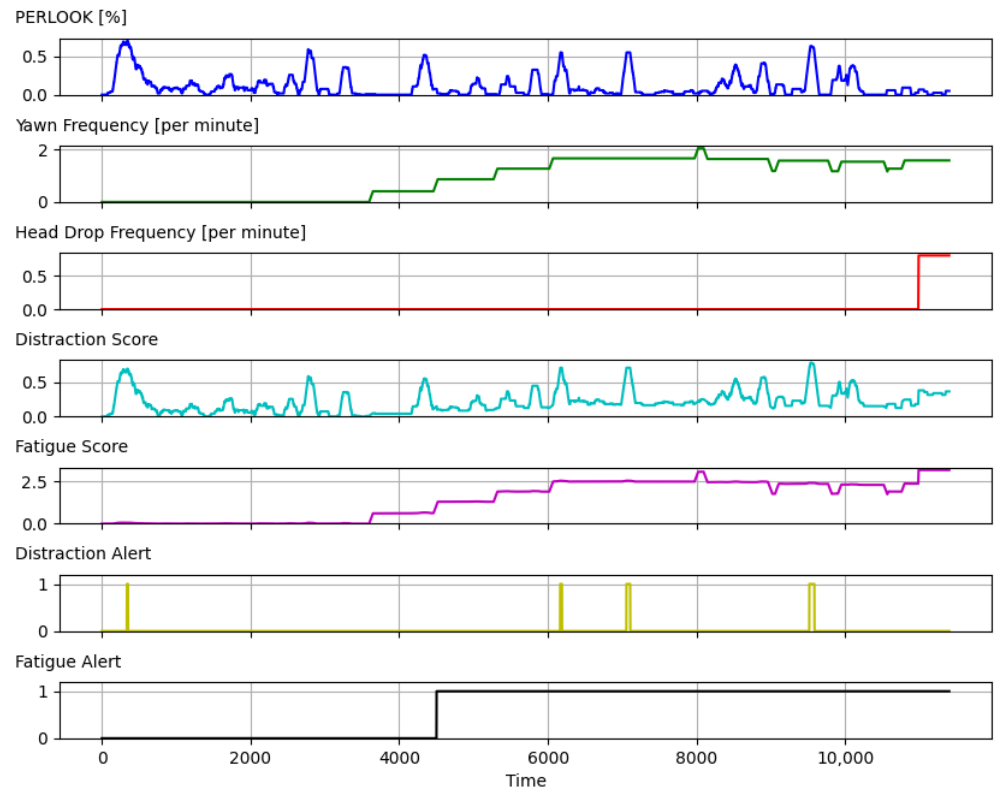


Figure 5. Sample output frames from the base fusion module. Frames with different driver actions and at various viewing angles (a–f).

Sequence: Mateusz2



Sequence: Cinek1

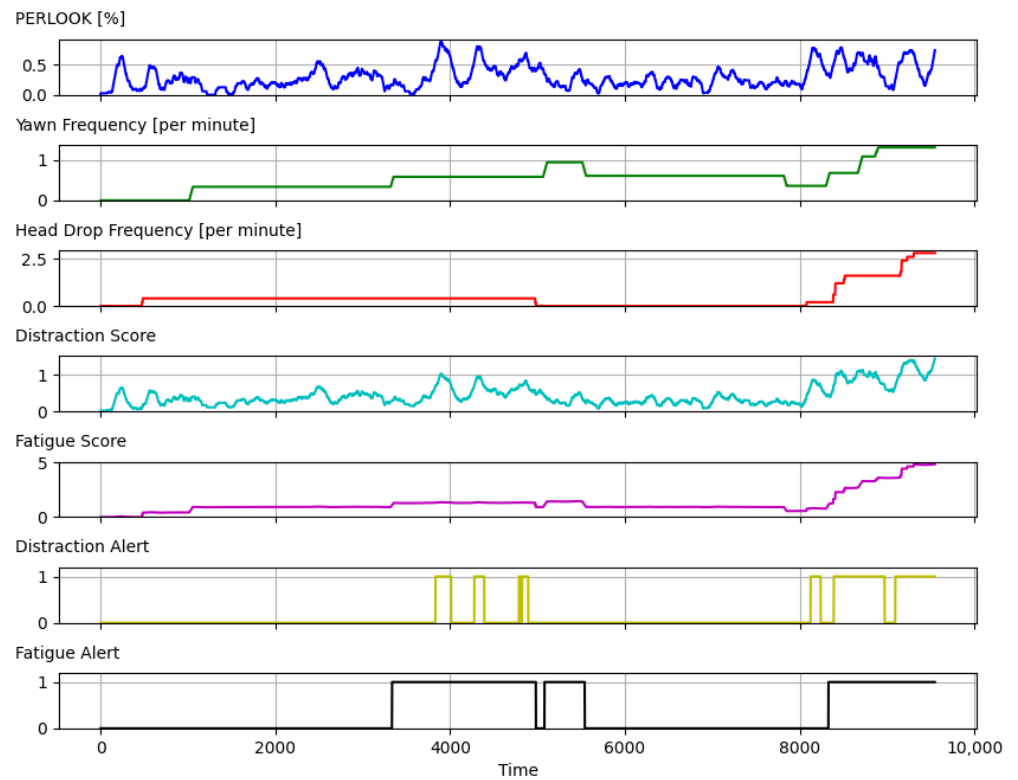


Figure 6. Output of the base fusion module.

Table 2. Basic fusion module parameters.

| Parameter | Value |
|--------------------------------|--|
| Yawning Reflex Length (frames) | 25 |
| Head Drop Length (frames) | 2 |
| Distraction Score Equation | $1.0 \cdot \text{PERLOOK} + 0.1 \cdot \text{Yawn. Freq.} + 0.2 \cdot \text{Head Drop Freq.}$ |
| Fatigue Score Equation | $0.1 \cdot \text{PERLOOK} + 1.5 \cdot \text{Yawn. Freq.} + 1.0 \cdot \text{Head Drop Freq.}$ |
| Distraction Alert Threshold | 0.7 |
| Fatigue Alert Threshold | 1.1 |

On the other hand, the LLM version—presented in the next section—takes the X_i streams and the prompts as input to detect alarm conditions. This approach allows for various interpretations of the X_i signals and inference based on user-defined prompts.

3.5.2. Modality Encoder and Large Language Model

Deep neural networks represent a significant breakthrough in the field of data classification. However, they require large datasets and tedious task-specific labeling, such as recognizing yawns. We use this type of network in the lower-level part of our system. Their outputs provide low-level information about the occurrence of a given event, such as a head drop or a yawn. In our system, lower-level information from the previously described detectors comes in the form of data frames that can be analyzed by the large language model. An exemplary data frame is shown in Table 3.

Thanks to this approach, it will be possible to formulate various questions and rules for the LLM, including those not previously considered, enabling a more extensive process of reasoning about the driver's condition. For instance, new prompts might be generated in the future based on more in-depth psychological interviews with drivers, etc. However, LLMs operate using text embeddings, while our system provides information in the form of the aforementioned data frames. Therefore, it is necessary to establish proper interfaces for the two modules to cooperate effectively. In the pattern recognition community, we can observe rapidly advancing research on vision-language models for vision tasks [40], their personalization for user-specific queries [41], and efforts to enable more general mathematical reasoning using open language models, which is by far the most difficult task [42]. Our research also follows these directions. However, we provide LLM with pre-processed data frames containing information on certain driver distraction events, such as yawning or distraction, rather than embeddings associated with bare images. Hence, this level can be seen as a modality encoder (ME). This approach provides a more domain-driven approach, which can result in more accurate and relevant responses from the upper level driven by the text LLM. As our ME provides data frames with distilled values of event detection and its frequency, we can directly use platforms already designed to interface with pre-trained LLMs, such as PandasAI [43] or LangChain [44].

In our experiments, we used the latter approach. It can be configured to use different types of LLM; in this choice, we used ChatGPT3.5-turbo, which is a reasonable compromise between the task complexity and the costs of the services. This framework uses a generative AI model for understanding and interpretation of natural language queries. These are then translated into a specific Python code. This is then used to deal with the data and return the results to the users. A block diagram of our LLM-based fusion module is shown in Figure 7. A special software agent has been developed that facilitates human control and communication through a set of so-called prompts. These include text information and questions, as well as guides that the agent encodes into a format that can be understood and processed by the LLM. An exemplary session is shown in Appendix A.

Table 3. A data frame built by the modality encoder based on information coming from event decoders.

| Timestamp | Face id | Yaw | Pitch | Roll | pose_state | pose_stable_state | head_drop_event | distracted_event | is_yawning | head_drops | PERLOOK | yawn_freq | head_drop_freq | distraction_score | distraction_alert | fatigue_score | fatigue_alert |
|-----------|---------|---------|---------|--------|--------------|-------------------|-----------------|------------------|------------|------------|---------|-----------|----------------|-------------------|-------------------|---------------|---------------|
| 150 | 0 | 195.854 | -10.348 | -4.906 | looking_left | looking_left | 0 | 1 | 0 | 0 | 0.200 | 0.000 | 0.000 | 0.200 | False | 0.020 | False |
| 151 | 0 | 195.216 | -10.432 | -5.944 | looking_left | looking_left | 0 | 1 | 0 | 0 | 0.208 | 0.000 | 0.000 | 0.208 | False | 0.021 | False |
| 152 | 0 | 194.537 | -10.205 | -6.486 | looking_left | looking_left | 0 | 1 | 0 | 0 | 0.217 | 0.000 | 0.000 | 0.217 | False | 0.022 | False |
| 153 | 0 | 196.795 | -9.995 | -7.426 | looking_left | looking_left | 0 | 1 | 0 | 0 | 0.225 | 0.000 | 0.000 | 0.225 | False | 0.023 | False |
| 154 | 0 | 197.035 | -10.334 | -4.970 | looking_left | looking_left | 0 | 1 | 0 | 0 | 0.233 | 0.000 | 0.000 | 0.233 | False | 0.023 | False |
| 155 | 0 | 197.317 | -10.270 | -4.008 | looking_left | looking_left | 0 | 1 | 0 | 0 | 0.242 | 0.000 | 0.000 | 0.242 | False | 0.024 | False |
| 156 | 0 | 196.466 | -11.638 | -5.592 | looking_left | looking_left | 0 | 1 | 0 | 0 | 0.250 | 0.000 | 0.000 | 0.250 | False | 0.025 | False |
| 157 | 0 | 194.868 | -8.738 | -6.429 | looking_left | looking_left | 0 | 1 | 0 | 0 | 0.258 | 0.000 | 0.000 | 0.258 | False | 0.026 | False |
| 158 | 0 | 193.109 | -9.717 | -5.529 | looking_left | looking_left | 0 | 1 | 0 | 0 | 0.267 | 0.000 | 0.000 | 0.267 | False | 0.027 | False |
| 159 | 0 | 191.302 | -9.910 | -4.640 | looking_left | looking_left | 0 | 1 | 0 | 0 | 0.275 | 0.000 | 0.000 | 0.275 | False | 0.028 | False |
| 160 | 0 | 194.913 | -10.140 | -5.609 | looking_left | looking_left | 0 | 1 | 0 | 0 | 0.283 | 0.000 | 0.000 | 0.283 | False | 0.028 | False |
| 161 | 0 | 195.849 | -9.790 | -7.949 | looking_left | looking_left | 0 | 1 | 0 | 0 | 0.292 | 0.000 | 0.000 | 0.292 | False | 0.029 | False |
| 162 | 0 | 196.564 | -10.252 | -8.326 | looking_left | looking_left | 0 | 1 | 0 | 0 | 0.300 | 0.000 | 0.000 | 0.300 | False | 0.030 | False |
| 163 | 0 | 196.022 | -9.023 | -9.845 | looking_left | looking_left | 0 | 1 | 0 | 0 | 0.308 | 0.000 | 0.000 | 0.308 | False | 0.031 | False |

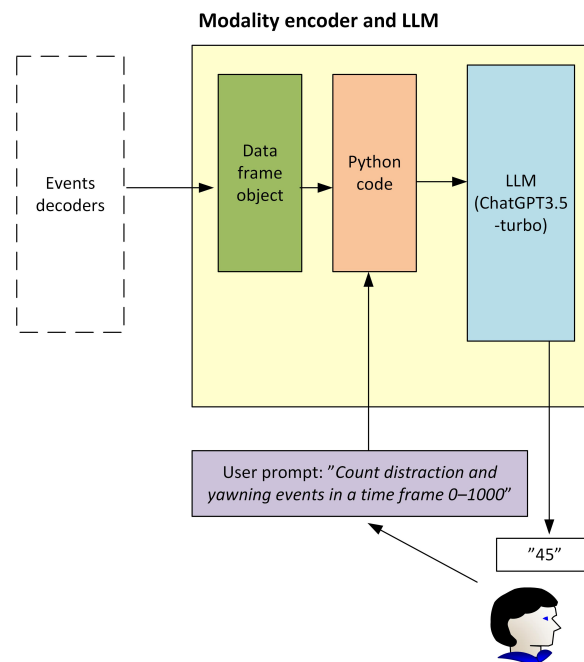


Figure 7. A fusion module built with the modality encodings coming from event decoders and LLM.

Figure A1 depicts an exemplary session in the form of a Python code that calls the LangChain agent. This is a software object that can then be provided with a number of prompts.

Apart from asking simple questions, such as how many yawning events are detected, we can also provide more elaborate prompts to fully exploit the abilities of LLM as, for example, the last prompt in Figure A1. First, it introduces a rule defining under what conditions we consider a driver to be distracted. Then, based on this rule, we ask LLM if a given driver is distracted. The answer to this question is shown in Figure A2. First, LLM correctly translated the word "distraction" into the "distracted_event" column name from the input frame shown in Table 3. Then, the LLM correctly interpreted our distraction rule, first identifying that there are 2153 distraction events, which, in accordance with the provided criteria, means that the driver is indeed distracted. This is just one example of the open questions that can be input into the LLM-based system.

4. Datasets

In this section, we provide detailed information on the datasets used in our experiments. It is important to note that while numerous datasets with visible light images are publicly available, the datasets specifically containing thermal images are significantly limited. This scarcity presents a unique challenge in the development and evaluation of driver monitoring systems that rely on thermal imaging.

4.1. TFW

The TFW dataset [29] encompasses thermal images collected from both indoor and outdoor environments, facilitating research in facial detection and recognition. Exemplary images from this dataset are shown in Figure 8. The dataset contains both controlled and semi-controlled indoor environments, along with an uncontrolled outdoor setting. For the indoor dataset, thermal-visual image pairs of 142 individuals were captured from 9 different positions. This dataset contains a total of 5112 thermal images, each featuring one labeled face. On the other hand, a semi-controlled indoor dataset involves subjects walking and performing predefined commands before free movement. It includes 780 thermal-visual pairs of 9 subjects, with 1748 labeled faces in each domain. Finally, the outdoor setting comprises unconstrained outdoor locations on different days during the summer.

With a total of 4090 thermal images, this dataset provides insights into thermal imaging under real-world conditions. With 9982 images and 16,509 labeled faces, the TFW dataset offers a comprehensive resource for advancing thermal-based facial detection models.

In our research, we utilized the outdoor part of the Thermal Faces in the Wild (TFW) dataset to evaluate and compare various versions and modifications of the YOLOv8 architecture, as this dataset provided a robust foundation for assessing the effectiveness of our proposed modifications in real-world scenarios.



Figure 8. Exemplary pictures from the TFW dataset.

4.2. SF-TL54

The SF-TL54 dataset consists of 2556 thermal phase images of 142 individuals, paired with corresponding visual images, and annotated with 54 landmarks, following the Eibach landmark configuration. This dataset was utilized for landmark detection experiments, aiming to address challenges such as limited textural information and occlusion of eyes in thermal images due to glasses.

We utilized the SF-TL54 dataset to train a final version of the YOLOv8 detector for facial landmark localization, leveraging its relatively high number of images paired with extensive annotations. Examples are shown in Figure 9.

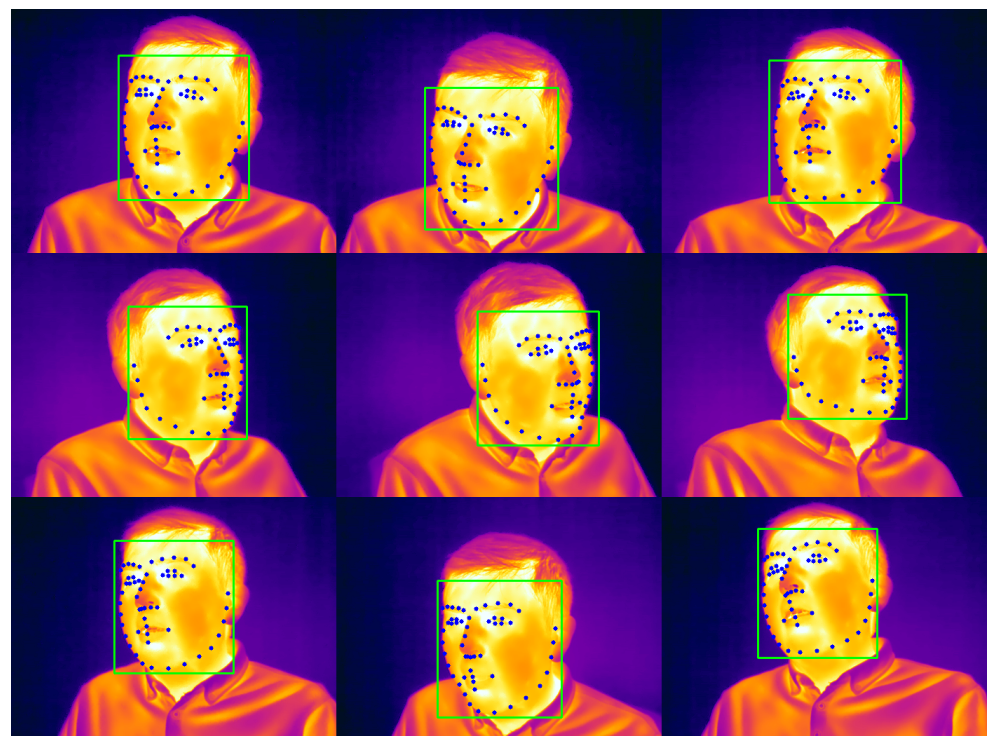


Figure 9. Exemplary pictures from the SF-TL54 dataset.

4.3. Extended ThermalYawningInCar Dataset

The original ThermalImagesInCar dataset [5] was developed in response to the lack of publicly available thermal image datasets capturing individuals in a driving scenario within a car, featuring various events such as diverse head poses, head drops, and yawning. One notable aspect of this dataset is its incorporation of video sequences, facilitating the training of models to detect temporal events such as yawning.

In this paper, we present an expanded iteration of our previous dataset. This updated version includes additional sequences and annotations featuring two new individuals previously not published, as well as fresh recording sessions with different scenarios and equipment. Sample images are presented in Figure 10.

The acquisition of new data was performed using the FLIR E6 thermal camera, which has a resolution of 240×180 pixels and captures 9 frames per second (FPS). This choice of equipment was dictated by its affordability and is more aligned with commercial automotive solutions than with research-focused thermal imaging cameras. All sequences were recorded within a prepared test scenario, comprising activities typical of a car environment, such as speaking, yawning, and typical body movements. The dataset is available online for further research and comparison with alternative methodologies.

Recordings were conducted inside a stationary vehicle during springtime, with the internal car temperature maintained at approximately 20 °C, reflective of typical modern vehicles equipped with climate control systems. A thermal imaging setup was installed under the rear-view mirror. Drivers were instructed to behave as naturally as possible, simulating a routine drive. Throughout the recording, they were asked to either remove or wear glasses.

The presence of research equipment and the awareness of participating in a scientific experiment may influence subjects' behavior, though the extent of this effect is difficult to quantify. This potential for altered behavior is a common challenge in experiments involving human subjects, as the awareness of being observed can alter natural responses. We have taken measures to minimize this impact, but it remains an inherent limitation in studies of this nature.

To address this limitation, we plan to revisit this experiment in the future with a larger number of subjects and on a broader scale. This approach will help us better understand and mitigate any potential biases introduced by the data collection process.

The annotation process for the images followed a semi-automated procedure outlined as follows:

1. Faces and facial landmarks were detected using a detection model described in Section 3.2, pre-trained on the SF-TL54 dataset.
2. Any missing or incorrect labels were manually corrected.
3. The start and end of each event, such as yawning or head drop, were manually annotated.
4. The camera angle offset was computed to compensate for head pose detection.

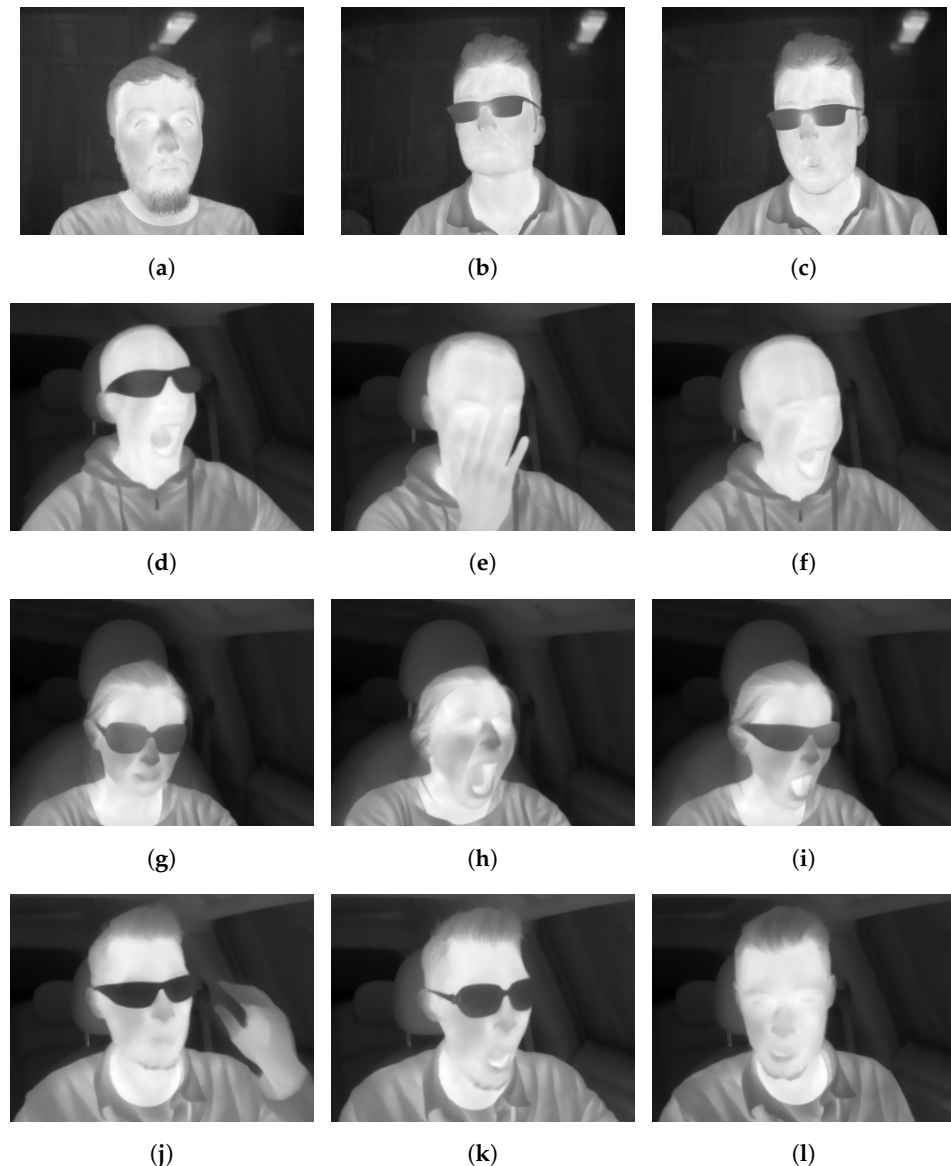


Figure 10. Sample images from an extended dataset (a–c) and a newly acquired dataset (d–l).

5. Experimental Results

In this section, the conditions of the conducted experiments, as well as their results, are presented and discussed.

5.1. Results of Face and Facial Landmark Detection

The results of the performance evaluation of our detection model are presented in Table 4, where we compare our modified YOLOv8-face architecture with its default version, as well as with standard YOLOv7 and YOLOv8 models, on the pose estimation task. The plot in Figure 11 illustrates the performance of various models as a function of the FLOPs required. As shown, our modified architecture achieves a 1.0% to 2.5% improvement in mAP@0.5 for the keypoint detection task, while requiring only 7% more GFLOPs compared to its original version. This enhancement in accuracy underscores the effectiveness of our modifications in improving the network's ability to detect facial features with greater precision. The slight increase in computational complexity is justified by the significant gains in performance, making our approach highly efficient and particularly beneficial for applications that require real-time performance on consumer-grade hardware.

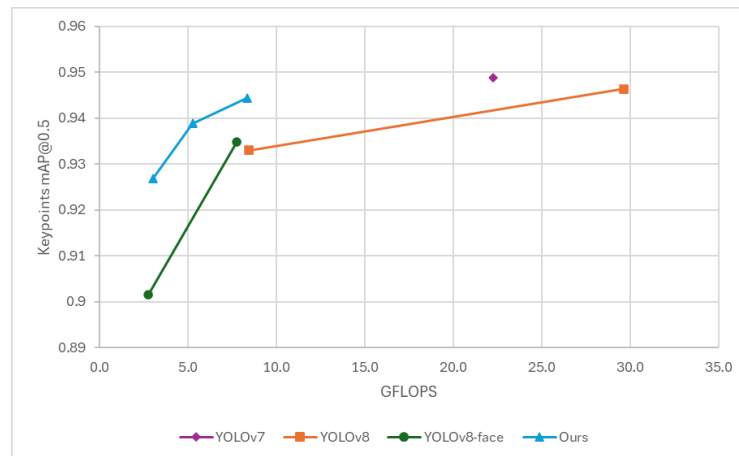


Figure 11. Model detection performance in the function of FLOPs used.

The ablation studies conducted allowed us to systematically evaluate the impact of integrating the CBAM and BiFPN modules individually.

The results demonstrated that both the CBAM-only and BiFPN-only variants resulted in performance improvements over the baseline model with minimal additional computational overhead. Specifically, the CBAM-only model achieved a 0.7 percentage point increase in mAP@0.5 for bounding box regression and a 0.4 percentage point increase in mAP@0.5 for keypoint regression, while the BiFPN-only variant improved performance by nearly 1 percentage point over the baseline in both metrics.

Moreover, when these modifications were applied in combination, the model exhibited an even greater enhancement in performance, scoring almost 2.5 percentage points in both mAP@0.5 for bounding box and keypoint regression. The results of the intermediary models are also presented in Table 4.

Furthermore, our model achieves results that are comparable to the original YOLOv7 and YOLOv8 models. Despite the similarities in detection accuracy, our modified architecture requires nearly three times less computational power. This efficiency not only facilitates real-time operation but also extends the potential use cases of our system, enabling its application in a wider range of devices and scenarios. The presented architecture stands out as a robust solution for facial feature detection in low-resolution thermal images.

However, due to differences in the training dataset (SF-TL54) and target image acquisition system, like resolution, imaging device, and head pose range, detecting facial features is not always perfect. In Figure 12, we present examples where the proposed detection model fails to correctly detect and/or align facial features.

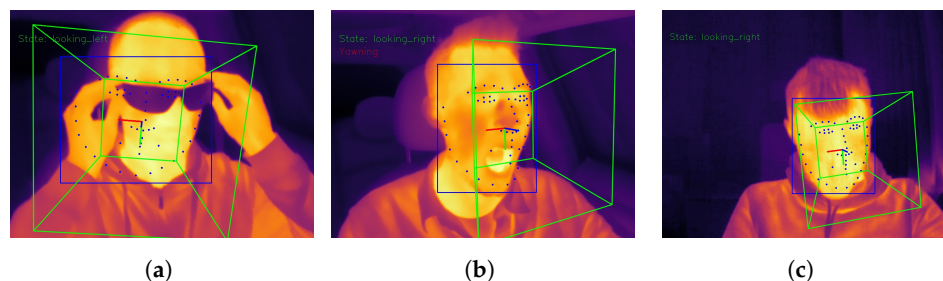


Figure 12. Examples of detection model failures. (a) Face detection contains too much background. (b) Facial feature keypoints not aligned with the actual face. (c) False positive detection when the head pose is outside of the training dataset distribution.

Table 4. Comparison of real-time state-of-the-art facial landmark detection architectures with our modified YOLOv8-face model, including ablation study results that guided the final architecture’s design.

| Model Family | Size | Backbone | Params [M] | FLOPs [B] | Box mAP@0.5 (%) | Box mAP@0.5–0.95 (%) | Keypoint mAP@0.5 (%) | Keypoints mAP@0.5–0.95 (%) |
|---------------------|-------|------------------------------|------------|-----------|-----------------|----------------------|----------------------|----------------------------|
| YOLOv7 | tiny | CBL. MCB. MP modules | 8.4 | 22.3 | 0.9532 | 0.6383 | 0.9488 | 0.8771 |
| YOLOv8 | nano | CSPDarkNet | 3.1 | 8.4 | 0.9409 | 0.6251 | 0.9330 | 0.8527 |
| | small | | 11.4 | 29.6 | 0.9447 | 0.6435 | 0.9463 | 0.8803 |
| YOLOv8-face | tiny | ShuffleNet v2 | 0.6 | 2.8 | 0.9024 | 0.5737 | 0.9016 | 0.7861 |
| | small | | 1.9 | 7.8 | 0.9342 | 0.6309 | 0.9348 | 0.8632 |
| Ours Attention Only | tiny | ShuffleNet v2 + CBAM | 0.7 | 2.9 | 0.9095 | 0.5820 | 0.9057 | 0.7960 |
| Ours BiFPN Only | tiny | ShuffleNet v2 + BiFPN | 0.6 | 2.9 | 0.9123 | 0.5762 | 0.9125 | 0.8040 |
| Ours | tiny | ShuffleNet v2 + CBAM + BiFPN | 0.8 | 3.0 | 0.9271 | 0.6011 | 0.9268 | 0.8222 |
| | small | | 1.5 | 5.3 | 0.9331 | 0.6244 | 0.9389 | 0.8634 |
| | large | | 2.5 | 8.4 | 0.9439 | 0.6406 | 0.9444 | 0.8757 |

5.2. Results of Yawning Detection

For the yawning detection evaluation, we employed four-fold cross-validation. In each fold, 25% of the test subjects were selected for the validation set, ensuring that the training data did not contain any images of individuals from the validation set (yawning or not). This approach eliminates data validation data leakage into the training phase.

On average, our model achieves an F1 score of 85%, with precision and recall values of 98% and 87%, respectively. These metrics indicate that the model is highly precise at identifying yawning events and has a strong ability to recall actual yawning instances. It is important to note that there is some ambiguity in the labeled data regarding the exact start and end points of a yawning reflex, which could affect the precision of our measurements.

Additionally, validation was conducted using a fixed time window length and a fixed stride (the number of frames skipped between each window in the sequence). This methodology ensures consistency in evaluation but may not fully capture the variability of yawning durations in real-world scenarios. However, in practical applications, the exact duration of a detected yawning reflex is less critical than the accurate detection of the event itself.

Moreover, there is an inherent delay in the response of the recurrent neural network, as it requires sufficient data to accurately distinguish between true yawning events and false positives. This delay represents a necessary trade-off for achieving higher accuracy in event detection. Despite this, our proposed novel detection method demonstrates excellent performance, making it suitable for real-time driver fatigue monitoring systems.

Results for each fold as well as average classification performance are presented in Table 5.

Table 5. Results of the yawning classification model.

| Fold Number | F1 Score | Precision | Recall |
|--------------------|----------|-----------|--------|
| 1 | 0.8943 | 0.9880 | 0.9001 |
| 2 | 0.9569 | 0.9684 | 0.9667 |
| 3 | 0.8381 | 0.9566 | 0.8102 |
| 4 | 0.7240 | 0.9881 | 0.7935 |
| Average Value | 0.8533 | 0.9753 | 0.8676 |
| Standard Deviation | 0.0989 | 0.0155 | 0.0810 |

5.3. Results of Distraction and Fatigue Detection

In our work, we propose two distinct approaches for data fusion to enhance the detection and analysis of driver fatigue and distraction. The first approach is a basic fusion model that leverages empirically selected multipliers and thresholds to combine various driver activity signals. This model integrates metrics such as PERLOOK, yawning frequency, and head drop frequency using predefined weights to generate alerts when specific conditions are met. The empirical nature of this model ensures that it is tuned to respond to the most critical indicators of driver fatigue and distraction.

The second approach is a novel data analysis method based on large language models (LLMs). This advanced method allows for the formulation of meta-rules and more sophisticated reasoning about the driver's condition. By utilizing LLMs, we can incorporate a broader range of contextual information and more nuanced interpretations of the driver's behavior, leading to a more comprehensive assessment of their state. This method not only detects fatigue and distraction events but also provides insights into the underlying patterns and trends that may indicate emerging risks.

To evaluate the effectiveness of these approaches, we selected two of the longest sequences recorded during real-world driving sessions. These sequences provided a diverse set of scenarios and behaviors for testing. Both methods successfully recognized events of distracted driving, such as prolonged head turns and lack of focus on the road. Additionally, they raised appropriate alarms when fatigue symptoms, such as yawning and sudden head drops, began to escalate.

The basic fusion model demonstrated its reliability in identifying critical events through its straightforward yet effective use of empirically derived thresholds. Meanwhile, the LLM-based data analysis approach showcased its ability to offer deeper insights and more adaptive responses by interpreting complex patterns in the driver's behavior. Together, these approaches represent a significant advancement in driver monitoring systems, providing robust and adaptable solutions for enhancing road safety.

Overall, our proposed methods highlight the potential of combining empirical models with advanced machine learning techniques to create more effective and intelligent driver fatigue and distraction detection systems. By addressing both immediate and long-term indicators of driver state, these approaches can help mitigate risks and improve overall driving safety.

6. Conclusions

In this study, we present a comprehensive system for driver fatigue and distraction monitoring that integrates advanced methods for facial feature detection, head pose estimation, and yawning detection using thermal imaging. Our approach leverages the YOLOv8 model enhanced with convolutional attention blocks (CABs) and the bi-directional feature pyramid network (BiFPN) to improve accuracy and efficiency, addressing the challenges posed by the low resolution of thermal images.

For yawning detection, we propose a method that computes histograms of oriented gradients (HOG) in thermal images, which are then classified using a long short-term memory (LSTM) recurrent neural network (RNN). Despite the limited number of yawning images in the available datasets, this approach demonstrated high precision and recall. Our assessment, conducted using four-fold cross-validation on a challenging dataset, indicates that our models achieved high detection performance.

The integration of head pose estimation and yawning detection into our system enhances its robustness and reliability, making it a practical solution for real-time driver monitoring. The use of thermal imaging ensures operation in various lighting conditions, further increasing the system's applicability in real conditions.

Furthermore, we introduce two approaches for data fusion: a basic model using empirically selected multipliers and thresholds, as well as an advanced method based on large language models (LLMs). Both methods effectively recognized distracted driving events and raised appropriate alarms for fatigue symptoms, validating their effectiveness in real-world scenarios. However, the LLM method enables the asking of open questions and facilitates a holistic analysis of system operations. Undoubtedly, the ongoing development of LLMs will drive further progress in systems that integrate low-level detection modules with high-level rule systems and user dialogue systems.

The proposed system offers several advantages over systems employing other modalities:

- **Thermal imaging:** Allows the system to operate in pitch-black darkness without any source of light, as well as in challenging lighting scenarios like sunset, sunrise, and reflections. It is also immune to variations in skin color.
- **Computational efficiency:** The use of a highly computationally efficient detection model makes the system well-suited for edge devices, enabling real-time performance.
- **Flexibility with LLM:** The integration of a large language model enhances the system's flexibility, allowing for open-ended analysis. This flexibility means the system can be improved and evolved by the end user through prompt engineering.

However, thermal imaging also introduces certain limitations that affect the system:

- Eye state detection: The system cannot reliably detect whether the human eyes are open or closed, as eyes are not well visible in thermal imaging. If the subject is wearing glasses, even corrective ones, the eyes are completely obstructed because glass is opaque in the LWIR spectrum.
- Head pose estimation: The system relies on facial feature detection to estimate head pose. Due to the limited amount of publicly available training data, the system currently only supports typical driver behavior.

Our work underscores the importance of advanced machine learning techniques in developing intelligent and adaptive driver monitoring systems. These systems can significantly contribute to road safety by providing timely alerts and interventions to prevent accidents caused by driver fatigue and distraction.

Future work will focus on further refining the models and exploring additional features to enhance detection accuracy. An interesting direction involves computing sparse features from a CNN or ViT, operating with thermal images, which are highly efficient in classification while maintaining very small sizes [45]. We also aim to conduct extensive field trials to validate the system's performance under diverse driving conditions and environments. It will be particularly interesting to further explore the possibilities for cooperation with the next generations of LLMs. The ultimate goal is to integrate this system into commercial driver assistance systems, thereby contributing to safer and more reliable transportation.

Author Contributions: Conceptualization, M.K. and B.C.; methodology, M.K. and B.C.; software, M.K. and B.C.; validation, M.K., B.C. and T.B.; data curation, M.K. and T.B.; writing, M.K., B.C. and T.B.; visualization, M.K.; supervision, B.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: Both the dataset and code used in the presented experiments are available on our GitHub repository: <https://github.com/mat02/ThermalImagesDataset> (accessed 29 August 2024).

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

Appendix A. Example of the LLM Application Code

Figure A1 shows an exemplary session in the form of Python code that calls the LangChain agent. This is a software object that can communicate via a number of prompts.

Figure A2 displays the answers output by the LLM agent in response to several rules and questions provided by the user. The rule defines conditions corresponding to a distracted driver. Endowed with this definition, the LLM is asked to assess the state of a given driver. In the displayed answer, the LLM correctly identified the measurements corresponding to the word "distraction", then correctly counted the number of such events, and ultimately confirmed that the driver was indeed distracted.

```

import pandas as pd

df = pd.read_csv( 'csv_Cinek1.csv', sep=';' )

from langchain_openai import OpenAI
from langchain_experimental.agents.agent_toolkits import create_pandas_dataframe_agent
from langchain.agents.agent_types import AgentType
from langchain_openai import ChatOpenAI

my_openai_key = "..." # enter your key

bc_agent = create_pandas_dataframe_agent(
    ChatOpenAI(temperature=0, openai_api_key=my_openai_key, model="gpt-3.5-turbo-0125"),
    df,
    verbose=True,
    agent_type=AgentType.OPENAI_FUNCTIONS,
    allow_dangerous_code=True
)

bc_agent.invoke( "How many rows are there?" )

bc_agent.invoke( "How many rows have head position dropped?" )

bc_agent.invoke( "Please, print all timestamps when yawning is detected." )

bc_agent.invoke( "A driver is tired if there are more than 25 head drops or more than 10 yawnings. Is this driver tired?" )

bc_agent.invoke( "A driver is distracted if there are more than 40 distraction events. Is this driver distracted and why?" )

```

Figure A1. Exemplary code with LLM prompts.

```

Index(['timestamp ', 'face id ', 'yaw          ', 'pitch          ', 'roll          ',
      'pose_state      ', 'pose_stable_state ', 'head_drop_event ',
      'distracted_event ', 'is_yawning      ', 'head_drops      ', 'perlook      ',
      'yawn_freq      ', 'head_drop_freq ', 'distraction_score ',
      'distraction_alert ', 'fatigue_score ', 'fatigue_alert'],
      dtype='object')
Invoking: `python REPL` with `{'query': "df['distracted_event '].sum()}`
late the total number of distraction events in the dataframe.

The total number of distraction events for this driver is 2153, which is more than 40. Therefore, based on the
criteria that a driver is distracted if there are more than 40 distraction events, this driver is distracted.

```

Figure A2. LLM answer to the last prompt that contains a rule to evaluate a driver's fatigue condition.

References

1. The Department of Transportation's National Highway Traffic Safety Administration (NHTSA). *Distracted Driving in 2022*; NHTSA's National Center for Statistics and Analysis: DOT HS 813 559, 2024. Available online: <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813559> (accessed on 29 August 2024).
2. Koay, H.V.; Chuah, J.H.; Chow, C.O.; Chang, Y.L. Detecting and recognizing driver distraction through various data modality using machine learning: A review, recent advances, simplified framework and open challenges (2014–2021). *Eng. Appl. Artif. Intell.* **2022**, *115*, 105309. [\[CrossRef\]](#)
3. Saadi, I.; cunningham, D.W.; Taleb-Ahmed, A.; Hadid, A.; Hillali, Y.E. Driver's facial expression recognition: A comprehensive survey. *Expert Syst. Appl.* **2024**, *242*, 122784. [\[CrossRef\]](#)
4. Lambay, A.; Liu, Y.; Morgan, P.L.; Ji, Z. Machine learning assisted human fatigue detection, monitoring, and recovery: A Review. *Digit. Eng.* **2024**, *1*, 100004. [\[CrossRef\]](#)
5. Knapik, M.; Cyganek, B. Driver's fatigue recognition based on yawn detection in thermal images. *Neurocomputing* **2019**, *338*, 274–292. [\[CrossRef\]](#)
6. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. *arXiv* **2018**, arXiv:1807.11164.
7. Woo, S.; Park, J.; Lee, J.; Kweon, I.S. CBAM: Convolutional Block Attention Module. *arXiv* **2018**, arXiv:1807.06521.
8. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. *arXiv* **2019**, arXiv:1911.09070.
9. Sikander, G.; Anwar, S. Driver Fatigue Detection Systems: A Review. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 2339–2352. [\[CrossRef\]](#)
10. Xiao, W.; Liu, H.; Ma, Z.; Chen, W.; Hou, J. FPIRST: Fatigue Driving Recognition Method Based on Feature Parameter Images and a Residual Swin Transformer. *Sensors* **2024**, *24*, 636. [\[CrossRef\]](#)
11. Mohammed, A.A.; Geng, X.; Wang, J.; Ali, Z. Driver distraction detection using semi-supervised lightweight vision transformer. *Eng. Appl. Artif. Intell.* **2024**, *129*, 107618. [\[CrossRef\]](#)
12. Ardabili, S.Z.; Bahmani, S.; Lahijan, L.Z.; Khaleghi, N.; Sheykhivand, S.; Danishvar, S. A Novel Approach for Automatic Detection of Driver Fatigue Using EEG Signals Based on Graph Convolutional Networks. *Sensors* **2024**, *24*, 364. [\[CrossRef\]](#)

13. Jiang, M.; Chaichanasittikarn, O.; Seet, M.; Ng, D.; Vyas, R.; Saini, G.; Dragomir, A. Modulating Driver Alertness via Ambient Olfactory Stimulation: A Wearable Electroencephalography Study. *Sensors* **2024**, *24*, 1203. [[CrossRef](#)] [[PubMed](#)]
14. Abdрахmanova, M.; Kuzdeuov, A.; Jarju, S.; Khassanov, Y.; Lewis, M.; Varol, H.A. SpeakingFaces: A Large-Scale Multimodal Dataset of Voice Commands with Visual and Thermal Video Streams. *Sensors* **2021**, *21*, 3465. [[CrossRef](#)] [[PubMed](#)]
15. Kuzdeuov, A.; Koishigarina, D.; Aubakirova, D.; Abushakimova, S.; Varol, H.A. SF-TL54: A Thermal Facial Landmark Dataset with Visual Pairs. In Proceedings of the 2022 IEEE/SICE International Symposium on System Integration (SII), Narvik, Norway, 9–12 January 2022; pp. 748–753. [[CrossRef](#)]
16. Zeng, Q.; Zhou, G.; Wan, L.; Wang, L.; Xuan, G.; Shao, Y. Detection of Coal and Gangue Based on Improved YOLOv8. *Sensors* **2024**, *24*, 1246. [[CrossRef](#)]
17. Cheng, Q.; Wang, W.; Jiang, X.; Hou, S.; Qin, Y. Assessment of Driver Mental Fatigue Using Facial Landmarks. *IEEE Access* **2019**, *7*, 150423–150434. [[CrossRef](#)]
18. Wang, M.; Hu, R.; Zhu, X.; Zhu, D.; Wang, X. Learning with noisy labels for robust fatigue detection. *Knowl.-Based Syst.* **2024**, *300*, 112199. [[CrossRef](#)]
19. Zhang, Q.; Zhu, Y.; Yang, M.; Jin, G.; Zhu, Y.; Chen, Q. Cross-to-merge training with class balance strategy for learning with noisy labels. *Expert Syst. Appl.* **2024**, *249*, 123846. [[CrossRef](#)]
20. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)]
21. Nabipour, M.; Nikan, S. Action Unit Analysis for Monitoring Drivers' Emotional States. *IEEE Sens. J.* **2024**, *24*, 24758–24769. [[CrossRef](#)]
22. Sajjatul Islam, M.; Jiang, W.; Lv, J.; Mohammed, A.A.; Sang, Y. Effective DemeapexNet: Revealing Spontaneous Facial Micro-Expressions. In Proceedings of the Proceedings of the 2022 6th International Conference on Compute and Data Analysis (ICCD '22), Shanghai, China 25–27 February 2022; pp. 81–90. [[CrossRef](#)]
23. Ma, Y.; Sanchez, V.; Nikan, S.; Upadhyay, D.; Atote, B.; Guha, T. Robust Multiview Multimodal Driver Monitoring System Using Masked Multi-Head Self-Attention. *arXiv* **2023**, arXiv:2304.06370.
24. Knapik, M.; Cyganek, B. Fast eyes detection in thermal images. *Multimed. Tools Appl.* **2021**, *80*, 3601–3621. [[CrossRef](#)]
25. Balon, T.; Knapik, M.; Cyganek, B. New Thermal Automotive Dataset for Object Detection. In Proceedings of the 17th Conference on Computer Science and Intelligence Systems, ACSIS, Sofia, Bulgaria, 4–7 September 2022; Volume 31, pp. 43–48. [[CrossRef](#)]
26. Balon, T.; Knapik, M.; Cyganek, B. Real-Time Detection of Small Objects in Automotive Thermal Images with Modern Deep Neural Architectures. *Ann. Comput. Sci. Inf. Syst.* **2023**, *37*, 29–35.
27. Qi, D.; Tan, W.; Yao, Q.; Liu, J. *YOLO5Face: Why Reinventing a Face Detector*; Springer Nature: Cham, Switzerland, 2021.
28. Qi, D.; Tan, W.; Yao, Q.; Liu, J. YOLOv8-Face. 2024. Available online: <https://github.com/derronqi/yolov8-face> (accessed on 20 April 2024).
29. Kuzdeuov, A.; Aubakirova, D.; Koishigarina, D.; Varol, A. TFW: Annotated Thermal Faces in the Wild Dataset. *IEEE Trans. Inf. Forensics Secur.* **2021**, *17*, 2084–2094. [[CrossRef](#)]
30. Moré, J.J. The Levenberg-Marquardt algorithm: Implementation and theory. In Proceedings of the Numerical Analysis, Dundee, Scotland, 28 June–1 July 1977; Watson, G.A., Ed.; Springer: Berlin/Heidelberg, Germany, 1978; pp. 105–116.
31. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893. [[CrossRef](#)]
32. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv* **2016**, arXiv:1609.08144.
33. Yang, M.; Tu, W.; Wang, J.; Xu, F.; Chen, X. Attention-based LSTM for target-dependent sentiment classification. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17), San Francisco, CA, USA, 4–9 February 2017; AAAI Press: Menlo Park, CA, USA, 2017; pp. 5013–5014.
34. Wu, Y.X.; Wu, Q.B.; Zhu, J.Q. Improved EEMD-based crude oil price forecasting using LSTM networks. *Phys. A Stat. Mech. Its Appl.* **2019**, *516*, 114–124. [[CrossRef](#)]
35. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *arXiv* **2015**, arXiv:1506.04214.
36. Zhang, W.; Han, J.; Deng, S.W. Abnormal heart sound detection using temporal quasi-periodic features and long short-term memory without segmentation. *Biomed. Signal Process. Control* **2019**, *53*, 101560. [[CrossRef](#)]
37. Drzazga, J.; Cyganek, B. An LSTM Network for Apnea and Hypopnea Episodes Detection in Respiratory Signals. *Sensors* **2021**, *21*, 5858. [[CrossRef](#)]
38. Wikipedia. Long Short-Term Memory. 2024. Available online: https://en.wikipedia.org/wiki/Long_short-term_memory (accessed on 30 August 2024)
39. Dinges, D.; Grace, R. *PERCLOS: A Valid Psychophysiological Measure of Alertness as Assessed by Psychomotor Vigilance*; Federal Highway Administration: Washington, DC, USA, 1998.
40. Zhang, J.; Huang, J.; Jin, S.; Lu, S. Vision-Language Models for Vision Tasks: A Survey. *arXiv* **2024**, arXiv:2304.00685. [[CrossRef](#)]
41. Alaluf, Y.; Richardson, E.; Tulyakov, S.; Aberman, K.; Cohen-Or, D. MyVLM: Personalizing VLMs for User-Specific Queries. *arXiv* **2024**, arXiv:2403.14599.

42. Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.K.; Wu, Y.; et al. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv* **2024**, arXiv:2402.03300.
43. PandasAI. PandasAI Library. 2024. Available online: <https://docs.pandas-ai.com> (accessed on 30 August 2024).
44. LangChain. LangChain library. 2024. Available online: <https://python.langchain.com> (accessed on 30 August 2024).
45. Łazewski, S.; Cyganek, B. Highly compressed image representation for classification and content retrieval. *Integr. Comput.-Aided Eng.* **2024**, *31*, 267–284. [[CrossRef](#)]

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