IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 5 August 2023, accepted 25 August 2023, date of publication 29 August 2023, date of current version 5 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3309810

RESEARCH ARTICLE

An Explainable Attention Zone Estimation for Level 3 Autonomous Driving

ROKSANA YAHYAABADI, (Member, IEEE), AND SOODEH NIKAN, (Member, IEEE)
Department of Electrical and Computer Engineering, Western University, London, ON N6A 3K7, Canada

Corresponding author: Roksana Yahyaabadi (ryahyaab@uwo.ca)

ABSTRACT Accurately assessing the driver's situational awareness is crucial in level 3 (L_3) autonomous driving, where the driver is in the loop. Estimating the attention zone provides essential information about the drivers' on/off-road visual attention and determines their readiness to take over the control from the autonomous agent in complicated situations. This paper proposes a double-phase pipeline to improve the explainability and accuracy of the attention zone estimation using an intermediate gaze regression layer, where the true relationships between the input images and output zone labels are interpretable. The proposed GazeMobileNet, a lightweight deep neural network, in the first phase, achieved state-of-the-art performance in estimating the gaze vector in the MPIIGaze dataset, with MAE of 2.37 degrees. The model was used to extract the corresponding gaze vectors from the LISA V2, which is a driving dataset with the in-cabin attention zone labels. As LISA V2 does not contain gaze vector labels, an unsupervised clustering approach was proposed in the second phase to categorize the driver's gaze vectors and map them to the corresponding attention zones. The proposed method demonstrated improved accuracy and robustness in the zone classification task. This model achieved the accuracies of 75.67% and 83.08% for attention zone estimation under "daytime without eyeglasses" and "nighttime without eyeglasses" capture conditions, respectively. Furthermore, the proposed model surpassed the recent research on that dataset by 73.11% and 74.02% accuracies under the "daytime with eyeglasses" and "nighttime with eyeglasses" capture conditions, respectively.

INDEX TERMS Level 3 autonomy, gaze estimation, GazeMobileNet, driver's attention zone, explainable clustering.

I. INTRODUCTION

The World Health Organization (WHO) states that the global rate of road fatalities is alarmingly high, with roughly 1.35 million people losing their lives to road accidents each year [1]. With the swift rise in vehicle numbers in recent decades, traffic accidents and congestion have become increasingly prevalent and pronounced [2]. Driver visual, auditory, bio-mechanical, and cognitive distractions account for more than half of all accidents [3]. However, in recent years, significant advancements have been made in the development of Automated Driving Systems (ADS) with the aim of enhancing road safety. It is anticipated that the integration of vehicle autonomy, specifically at level 3 or higher, as classified by the Society of Automotive Engineers (SAE), will play a pivotal role in diminishing this alarming statistic substantially [4].

In L_3 autonomy, drivers are not required to maintain constant visual attention on the road or keep their hands on the steering wheel continuously [5]. They will have the freedom to engage in secondary non-driving-related tasks (NDRTs), such as reading, writing, emailing, eating, etc. However, the drivers need to be perceptive without delay in responding to the take-over request (TOR), by the ADS, in situations beyond the capabilities of the autonomous agent. Typically, the vehicle prompts for intervention when the driving task becomes challenging and the autonomous agent encounters intricate or unpredictable situations, including abrupt obstacles, pedestrians, and temporary road work.

Safety at L_3 automation is highly correlated with the driver's attention level and readiness for taking over the vehicle's control from the autonomous agent. Eye gaze is one

The associate editor coordinating the review of this manuscript and approving it for publication was Chun-Wei Tsai.

of the most determinate tools to monitor the driver's instantaneous on/off-road visual attention. The gaze vector, derived from the driver's eye movement, offers insights into the specific areas of focus while driving. This information is critical for ensuring the safe intervention of the driver in autonomous vehicles, as it enables the vehicle to determine the driver's situational awareness before handing the control over to the driver. The direction of the driver's eye gaze can be quantified from the face/eye images captured by the sensor (camera) inside the vehicle cabin, with the aid of artificial intelligence (AI) technologies. However, in the presence of noise factors including head movements, illumination variations, occlusion, low resolution, and information loss, gaze prediction becomes a challenging task. Current methodologies in the literature, often employ a straightforward classification approach to map images directly to attention zones, lacking interpretability. Consequently, in specific safety-critical applications, this approach can lead to critical errors or inconvenient false alarms, as two classes with the highest and equal likelihood may originate from distant, completely separate zones.

The primary objective of the proposed method is to address the lack of explainability in the existing zone estimation techniques by mapping image frames to the zone labels. In the existing methodologies in the literature, the images are mapped to the attention zones directly using a simple classification without interpretability. As a result of that, for example, two classes with the highest and equal likelihood may come from two separate zones with far distances which causes critical errors in such a safety-related application or leads to inconvenient false alarms. In this study, in the first phase, we proposed GazeMobileNet and trained it on the MPIIGaze dataset. Therefore, the proposed intermediate layer estimates the gaze vector, as an interpretable feature corresponding to the attention zones in the second phase. Subsequently, since we aimed to find the attention zones, we applied the gaze vector extractor model to the LISA V2 dataset, as a driving dataset that does not contain gaze labels, like many other datasets in that context. As a result, we proposed an unsupervised method for clustering the extracted gaze vectors. After visualizing the data, we found that most gaze vectors significantly overlapped. Therefore, we chose GMM as the clustering method that can handle overlapping samples. Our main contributions are outlined below.

- In the gaze estimation phase, we introduced the GazeMobileNet model, which achieved state-of-the-art performance on the MIIGaze dataset. When head pose information was incorporated, the mean angle error (MAE) significantly decreased to 2.37°, while the head pose-free model achieved an MAE of 2.51°.
- Due to the model's sensitivity to slight changes in the head or eye direction, the driver monitoring system may mistakenly estimate a different zone when the likelihood of the attention zones is close. To address this issue and enhance the explainability of zone classification, we introduced an intermediate feature, a novel contribution to this study. This feature facilitates the mapping between the input image and the estimated attention zones by utilizing the gaze vector/angle as a significant descriptor.
- For our research, we utilized the LISA V2 attention zone dataset as the target driving dataset, which does not include the gaze vector labels. To overcome this limitation, we developed an unsupervised strategy to cluster the gaze features associated with each target video frame. By leveraging distinctive attributes of the gaze directions, we proposed a clustering model to assign each gaze vector to the corresponding attention zone.

The paper is organized as follows. In section II, we reviewed the related works. Sections III and IV describe our datasets and the proposed methodology. Section V shows the performance evaluation of the proposed models, and the results were discussed. In Section VI the paper was concluded.

II. RELATED WORK

In this work, we proposed an explainable double-phase framework to monitor the visual attention of a driver, where the input frames from the in-cabin camera are fed into the gaze estimation module to quantify the gaze information, and mapped to the region (zone) where the driver is looking. Therefore, this section is divided into reviewing three distinct areas of research. The first part pertains to the most promising investigations within the domain of gaze estimation in general. In the second part, state-of-the-art techniques for attention zone classification will be analyzed. Finally, in the third part, we will review the related works in explainable AI (XAI), followed by a discussion on its necessity in automated driving systems.

A. GAZE ESTIMATION

Vision-based gaze estimation can be classified into two main categories. 1) Model-based techniques, which estimate gaze direction by combining the geometric eye model using the eye features, such as cornea reflection and pupil center [6]. However, due to the calibration requirement, the need for dedicated infrared (IR) lighting hardware, and sensitivity to the input noise (occlusions or lighting), these models are costly prone to errors and less generalizable [7]. 2) Appearance-based methods, which regress gaze directly from the camera and map image information into the gaze angle [8]. These methods are broadly divided into two subcategories. a) Conventional machine learning models, which extract hand-crafted gaze-related features from image pixels. In a previous study [9], a multi-stream model was presented that employed features from the eyeball, iris, and the entire input image. To mitigate the effect of noise, the authors used a synthetic dataset, to train the isolation network and binary mask extraction. They achieved a mean angle error (MAE) of 4.64° on the MPIIGaze dataset. b) Deep learning approaches estimate direct mapping from images to the gaze vector quantitatively, using neural networks, which are more accurate and robust against noise factors compared to conventional machine learning. Ghosh et al. [10] introduced a multi-task gaze estimation framework where ResNet-50 was utilized to leverage pseudo-gaze, head pose, and eye orientation and achieved a MAE of 4.07° on MPIIGaze dataset.

Wang et al. [11] proposed a unified framework that incorporates adversarial learning to learn gaze-responsive features and extend the point-estimation model to a Bayesian framework. They showed improved performance on benchmark datasets, and adaptation capability to new subjects/environments. However, compared to the superior improvements in other human modeling studies, with the aid of the representation power of deep neural networks (DNNs), gaze estimation has not yet achieved the same level of maturity. This is primarily due to the complex eye appearance and cognitive process in the visual system, and most importantly, the lack of sufficient annotated training datasets [7]. Ali and Kim [12] introduced a multi-stream shallow CNN that incorporates a dual spatial layer mechanism. This model individually processed each eye patch along with the head pose as the inputs, and utilized a learned regression function to predict the gaze angle. Features extracted from each eye were subsequently concatenated, and the head pose vector was appended to the final layer for the gaze estimation task. They also employed a data fusion technique, using MPIIGaze and EYEDIAP datasets to improve model generalization. This approach resulted in an accuracy of 2.60° on the MPIIGaze dataset.

B. ESTIMATING THE ATTENTION ZONE

In some of the gaze estimation-based applications, such as automotive, gaming, or virtual reality, the target is to estimate the area in 2 or 3-dimensional (2D/3D) space where the gaze is pointing to, and the time duration of the focus. For example, in driver monitoring, estimating the attention zone, such as road, mirror, steering wheel, or infotainment is an indicator of visual attention/distraction.

Vora et al. [13], proposed a systematic evaluation of various CNNs with the aim of enhancing the generalization of the attention zone classification. The study explored four different inputs including the driver's face and background, the driver's face, and the upper half of the face. They used a classification method to map from image content to the attention zone directly. However, the relationship between the input images and the output zone labels was not explicitly interpretable in their proposed method. Rangesh et al. [14] proposed a model that used an IR camera to capture images, normalized the images to manage lighting issues, and employed gaze preserving CycleGAN (GPCycleGAN) to remove eyeglasses before gaze estimation. The authors utilized squeezeNet to directly classify eye crop images into different gaze zone labels. Nonetheless, employing a direct classification approach of the attention zones from input images may lack the capability to provide a comprehensive explanation of the reasoning and decision-making process behind assigning the input image to a specific attention zone.

Yang et al. [15] proposed a model that presents a classification-based approach that assigns input images into distinct attention zones through a multistage algorithm. Initially, the detected face was fed into a facial encoding network using an attention mechanism. In the final stage, head pose information was integrated into the facial feature map, resulting in the probability distribution of different classes. However, that model also failed to elucidate the exact relationship between the input characteristics and the estimated zone.

C. EXPLAINABLE AI-BASED DRIVER MONITORING SYSTEM

As AI systems continue to evolve, the aspect of explainability has emerged as a crucial consideration. Explainable AI (XAI) addresses one of the major challenges in Machine Learning (ML) by enabling AI systems to make their operations understandable to humans during practical implementation. With the continuous advancements in AI techniques and the increasing level of automation in the field of automated driving, the significance of XAI has grown substantially. The demand for explainability has reached unprecedented levels due to the paramount importance of safety in the automated vehicle industry [16]. However, given that humans are this technology's primary social stakeholders and users, the development principles for autonomous vehicles should resonate with the target audience's requirements, incorporating their preliminary opinions and expectations. Moreover, from sociotechnical and philosophical perspectives, providing interpretations for AI decisions can yield descriptive information about the causal sequence of actions taken, especially in crucial situations.

An explainable driver monitoring system can offer considerable benefits to the societal acceptance of intelligent vehicles. Providing an intuitive human-computer interaction interface is critical to satisfying users, which includes drivers and passengers. A range of studies has leveraged human-centered XAI design that employs visual, audio, and textual modalities to communicate a vehicle's real-time decisions to its passengers [17]. Incorporating such feedback in vehicle design can significantly improve user experience, thus establishing the need for intelligent AI systems for the broader acceptance of autonomous vehicles. Furthermore, safe driving directly influences the safety of passengers and pedestrians, making it imperative to establish a reliable transportation system. However, recurrent system failures without adequate explanations can substantially erode users' trust. Therefore, the provision of detailed interpretations and building an explainable intelligent driving system are the crucial steps toward building trust in AI, which further engenders transparency and accountability within the technology [18].

Kim and Canny [19] employed visual explanations represented as real-time highlighted regions of an image, using an attention model, that exerts a causal influence on the network's output, specifically, steering control. The potential image regions impacted the network output, but those regions may contain both true influences and spurious ones. There has been limited research on the topic of explainability in AI-based driver monitoring systems. Lorente et al. [16] proposed the XRAI, an explainability technique in AI, to scrutinize the decision-making process of two DNNs in an advanced driver assistance system (ADAS). The first model, detecting the driver's mood, showed an inadequate focus on key facial features due to insufficient training. The second, identifying driver distractions, struggled to classify new data despite high training accuracy.

This paper aims to enhance the explainability of intelligent driver's attention monitoring by focusing on the specific context. Previous studies on attention zone estimation primarily utilized classification methods to allocate each sample to an attention zone. However, these methods often lacked explainability, making it difficult to establish a clear relationship between the input image and the estimated attention zone. To address this limitation, our study introduces a novel approach by incorporating gaze features as intermediate-level features. By utilizing these features, we not only improve the classification accuracy but also provide a clearer understanding of the connection between the input image and the assigned attention zone.

III. DATASET

In this study, two publicly available datasets, the MPIIGaze dataset, and the LISA V2 dataset are utilized in two phases of the proposed framework to build a model for gaze feature extraction and the attention zone estimation phase, respectively.

A. MPIIGAZE DATASET

The MPIIGaze dataset [20] includes over 213,659 facial images of 15 subjects recorded during their everyday laptop usage. This dataset is significantly challenging due to its high appearance, head movement, gaze target, and illumination variability. The number of images collected per participant ranged from 1,498 to 34,745. The "Normalized" version of the dataset contains the images of eye crops that have been normalized to cancel the scaling and rotation effects. The size of the normalized samples is a 30 × 60 grayscale eye-patch image, corresponding to each right and left eye. Figure 1. shows the RGB version of the eye-cropped images. Additionally, the dataset provides two annotations for each image: the 3D head pose and the 3D gaze direction, corresponding to the right and left eye images, respectively. Due to the availability of gaze vector information, this dataset is ideal for training and evaluating supervised gaze estimation models. It is recognized as one of the most well-known datasets, as its appearance variations, make it a valuable resource for gaze estimation research. In Table 1, the specific appearance features of the MPIIGaze dataset are shown.

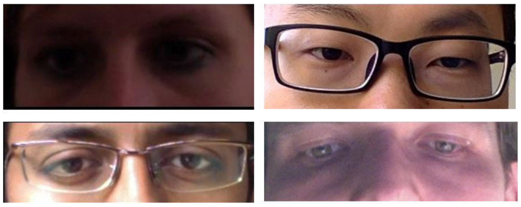


FIGURE 1. Sample images in MPIIGaze dataset.

TABLE 1. Appearance characteristics of MPIIGaze dataset per subject.

Subject	Image Count	Appearance Description
S1	29,961	Nothing special noticeable
S2	24,143	Thick rim dark colored Spectacles completely bordering eye region.
S3	28,019	Nothing special noticeable
S4	35,075	Nothing special noticeable
S5	16,831	Light colored spectacles which frequently interferes with eye region
S6	16,577	Contains blurry images.
S7	18,448	Nothing special noticeable
S8	15,509	Light colored spectacles which frequently interferes with eye region
S9	10,701	Contains blurry images
S10	7,995	Thick rim dark colored Spectacles completely bordering eye region.
S11	2,810	A significant percentage of them have spectacles.
S12	2,982	Some images are blurry with very few of them with spectacles.
S13	1,609	Nothing special noticeable
S14	1,498	Nothing special noticeable
S15	1,500	Hand and hair interfere with eye region frequently.

B. LISA V2 DATASET

The primary version of the LISA V2 dataset [14] was collected utilizing two cameras positioned in front of the driver, along with one outside the vehicle. In this dataset, seven classes including six distinct attention zones and one state were identified, namely "forward", "right", "left", "center console (radio)", "center rearview mirror", "speedometer", and an "eyes-closed & lap" state. LISA V2 was captured using an IR camera mounted adjacent to the rearview mirror. Furthermore, LISA V2 incorporates data about the time of driving (either daytime or nighttime) and the use of eyewear by the drivers. Figure. 2 depicts the six in-cabin attention zones, and sample images in the LISA V2 dataset. In Table 2, the number of images in the LISA V2 at different image-capturing conditions is shown. To maintain the integrity of cross-subject validation, there is no subject overlap between the training and validation sets.

As previously mentioned, the LISA V2 dataset introduces crucial real-world intricacies and variations that conventional driver gaze estimation systems frequently overlook. Noteworthy among these complexities are factors such as the presence of eyeglasses and exposure to demanding illumination conditions, both of which are commonly encountered during real driving scenarios. The dataset thoughtfully includes numerous illustrative samples, as shown in Fig. 3, that showcase these challenges, providing valuable insights into these real-world aspects.

TABLE 2. The size of LISA V2 dataset, determined by the number of images in the training, validation, and testing sets, as well as the image capturing conditions.

Capture condition	Training	Validation	Testing
Daytime; with eyeglasses	67,151	9,908	2,758
Nighttime; without eyeglasses	59,352	8,510	2,768
Daytime; with eyeglasses	43,432	9,062	3,294
Nighttime with eyeglasses	33,189	8,103	2,897

TABLE 3. Comparison of the MPIIGaze and LISA V2 datasets.

Characteristic	MPIIGaze	LISA V2
Purpose	General	Driving
No. of frames	213,659	336,177
No. of participants	15	13
No. of classes	N/A	7
Camera type	RGB	IR
Resolution	30x60	256x256
Ground truth gaze	YES	NO

IV. METHODOLOGY

As shown in Fig. 4, the proposed framework includes two main phases. 1) In the first phase, a DNN, "Gaze Estimation", learns to extract gaze vectors from the training dataset (MPIIGaze). Then the trained model extracts the gaze angles corresponding to each image in the target driving dataset (LISA V2). In the second phase, "Attention Zone Estimation", the gaze vectors from the previous step, are clustered into the predetermined attention zones using the Gaussian mixture model (GMM), which assigns each target image to an attention zone by utilizing the corresponding gaze vectors. Finally, the accuracy and robustness of the model are validated using the ground truth zone labels. The intermediate features (gaze vectors) obtained from the first phase, enhance the explainability of the mapping between the drivers' images and attention zone labels in the second phase.

A. PHASE 1: GAZE ESTIMATION

This framework aims to provide the intermediate features (gaze vectors) before mapping each image to a specific attention zone, in the next phase, to increase the interpretability of zone assignment. The implementation of this phase can be extended to other relevant gaze estimation applications. The following section provides a detailed description of each step in the first phase.

The comparison between the utilized datasets in our study (MPIIGaze and LISA V2) were summarized in Table 3.

1) GAZE ESTIMATION

Appearance-based gaze estimation using DNNs creates an end-to-end platform to estimate the gaze vector from the raw camera frames (face/eye images). Compared to traditional machine learning, deep learning-based approaches automatically extract hierarchical gaze features from high-dimensional image data and learn a direct mapping from eye appearance to the gaze vector. In this work, two setups for gaze estimation were examined: one that utilizes the eye image as the input, and the other setup which incorporates the corresponding head pose as a piece of auxiliary information.

- Head pose-free gaze estimation: Head pose-free models may be considered in certain applications where the head pose information is not readily available, or the process of collecting such information is infeasible or computationally intensive. The motivation for investigating the performance of the gaze estimation model without the head pose information emanates from the fact that in real applications, computational complexity will be increased to extract the 3D head pose information from 2D images. From a set of eye images e_i, the goal is to learn a head pose-free model f that estimates gaze angle $\alpha_\text{i} = f(e_\text{i})$ in the eye coordinate system.
- Head pose-incorporated gaze estimation: In the context of gaze estimation, both the position and orientation of a subject's head pose and eyeball may be effective for determining gaze direction [21]. However, the degree of their interaction varies per individual, due to differences in comfort and habitual posture. Therefore, while both head pose and eyeball position are important, their relative contributions to the gaze direction can fluctuate based on individual behaviors and tendencies. In this work, we examined the effect of incorporating head pose in estimating gaze direction. A model f was trained using a set of eye images e_i and their corresponding head poses h_i, with the objective of estimating gaze angles α_i in the eye coordinate system. Gaze direction depends on the position of the eyes, and a parallax effect can occur between the gaze directions of the two eyes due to pupillary distance. Therefore, the model that takes head pose information into account is expected to provide more accurate gaze estimation results.

The dataset that we utilized in this study, contained information in the form of cartesian coordinates (x, y, z) to indicate the location of a point in 3D space. The 3D gaze direction (x, y, z) can be converted to a 2D representation (θ, φ) as an angle corresponding to the i^{th} sample as follows.

$$\begin{aligned} \alpha_{p_i} &= [\theta_i, \varphi_i], \\ \theta_i &= \sin^{-1}(-y_i), \\ \varphi_i &= \text{atan2}(-x_i, -z_i). \end{aligned} \tag{1}$$

Also, 3D head rotation (x, y, z) can be converted to (θ, φ) using the following equations [22].

$$\begin{aligned} M_i &= \text{Rodrigues}((x_i, y_i, z_i)), \\ Z_{v_i} &= (\text{the third column of } M), \\ \theta_i &= \sin^{-1}(Zv_i[1]), \\ \varphi_i &= \text{atan2}(Zv_i[0], Zv_i[2]), \end{aligned} \tag{2}$$

where, Rodrigues' rotation formula transforms the rotation in 3D space, represented by Euler angles (x, y, z), into a rotation matrix M. The third column of the rotation

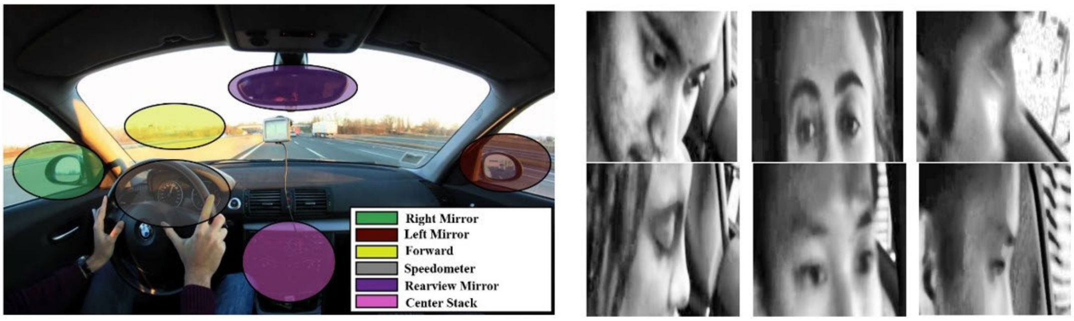


FIGURE 2. Representation of attention zones (left), and sample images (right) in LISA V2 dataset.

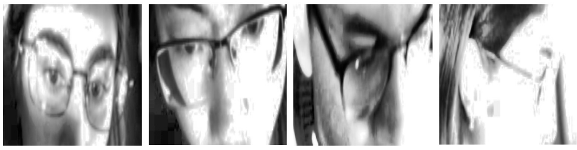


FIGURE 3. Instances exemplifying the prevalent challenges of intense illumination and eyeglass-related complexities within the daytime images of the LISA V2 dataset.

matrix M, represented by Z_v, corresponds to the unit vector in the z-axis of the rotated frame. The angle θ_i between the z-axis vector and the y-axis was computed using the arcsine function on the y-component of the z-axis vector. Lastly, φ, the angle between the z-axis vector and the x-axis (representing azimuth), was determined using the two-argument arctangent function on the x and z components of the z-axis vector to ensure the correct quadrant for φ is identified.

2) MODEL ARCHITECTURE

In this work, three CNN-based architectures including modified LeNet (MLeNet), AlexNet, and our proposed GazeMobileNet were trained using the MPIIGaze dataset and evaluated to select the best model for estimating the gaze direction. By comparing the performance of these three different networks, we aimed to examine the influence of the depth and size of the neural network architecture on gaze estimation performance. MLeNet is an adaptation of the original LeNet architecture with some significant modifications to extract the higher-level features and improve performance. In MLeNet, the number of filters in the first and second convolutional layers was increased to 20 and 50, respectively. Additionally, the final fully connected layer of the original LeNet was removed, and a linear layer with two nodes, denoted as (θ, φ), was used instead of the softmax layer. We employed the original version of AlexNet for our application, with a few necessary modifications. Given that the minimum input size for AlexNet should be 256 × 256, the input image was upsampled to 288 × 480. Furthermore, we modified the size of the final FC layer and appended a regression layer for estimating the gaze direction parameters (θ, φ).

GazeMobileNet: A high-level overview of the GazeMobileNet architecture is shown in Fig. 5 and Table 4. As illustrated, in this architecture, the main component of GazeMobileNet is the inverted residual with a linear bottleneck. The module of the inverted residual with linear bottleneck initially accepts a compressed low-dimensional representation as input. This representation in the expansion layer was expanded to a high dimension and then subjected to a lightweight depthwise convolution for filtering. Following this in the projection layer, a linear convolution was applied to project the features back to a low-dimensional representation. The fundamental concept in the depthwise convolution involves transforming a standard convolutional operation into a factorized form, segregating it into two distinct layers [23]. The initial layer, known as depthwise convolution, executes lightweight filtering by applying a unique convolutional filter for each input channel. The subsequent layer involves a 1 × 1 convolution, termed a pointwise convolution, which constructs new features by calculating linear combinations of the input channels. The proposed GazeMobileNet was inspired by the lightweight MobileNetV2 architecture [23]. However, they differ in several ways as described below.

- Grayscale input: In MobileNetV2, the first layer expects a three-channel RGB input. However, in GazeMobileNet, the input channels in the first convolutional layer were modified from 3 to 1 to accept one-channel grayscale input images. Such adjustment was necessary for the MPIIGaze dataset, where the normalized images were grayscale. Moreover, the adaptation of the model to the target driving dataset (LISA V2) involved making adjustments due to differences in image characteristics. As a preprocessing step, grayscaling the RGB images proved somewhat beneficial in aiding the model's adaptation to the IR images present in the LISA V2 dataset.
- Reduced parameters: The last FC layer (classifier) which originally had 1000 output units (for 1000 classes in the ImageNet dataset), was replaced by a linear layer with 128 or 132 output neurons, corresponding to the head pose-free or head pose-incorporated setup, respectively, which further passes through a batch normalization (BN) layer and a ReLU activation function. By implementing this modification, there was a significant reduction in the number of parameters in the final layer, resulting in a

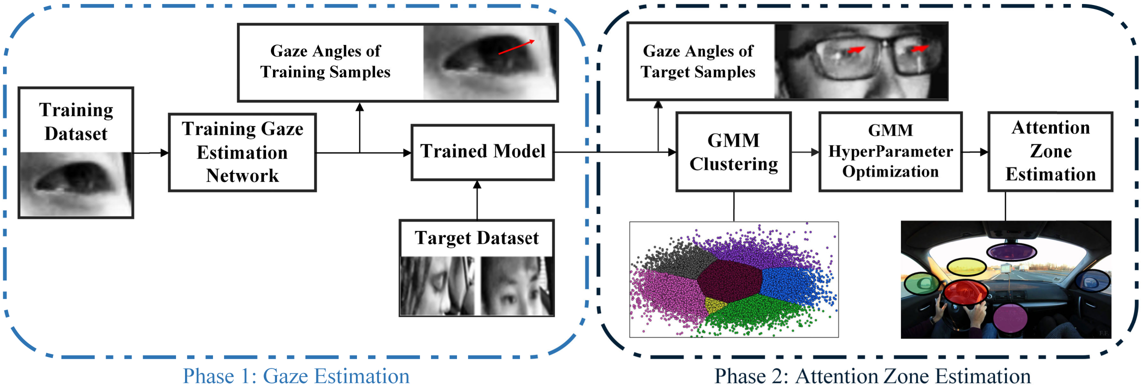

Training Dataset
Training Gaze Estimation Network
Gaze Angles of Training Samples
Trained Model
Target Dataset
Phase 1: Gaze Estimation
Gaze Angles of Target Samples
GMM Clustering
GMM HyperParameter Optimization
Attention Zone Estimation
Phase 2: Attention Zone Estimation

FIGURE 4. The proposed framework.

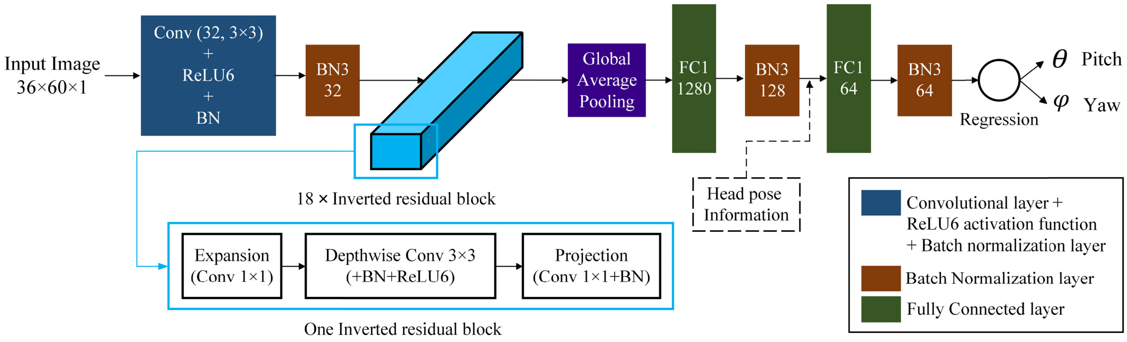


FIGURE 5. The block diagram of the proposed GazeMobileNet architecture.

more suitable network for deploying with low-latency requirements.

- Gaze regression: For the purpose of obtaining the gaze yaw and pitch angles as outputs, our proposed GazeMobileNet incorporates a gaze regression unit. This unit consists of a linear layer with 64 output neurons, followed by a BN layer, a ReLU activation function, and ultimately, a linear layer with 2 output neurons.
- Batch normalization: In contrast to the original MobileNetV2 architecture, our proposed GazeMobileNet introduces three extra BN layers. These additional BN layers are placed after the initial convolution layer and the fully connected (FC) layers. The rationale behind this modification was to harness the benefits of BN, such as accelerated learning through the reduction of internal covariate shifts, facilitation of higher learning rates, and improvement of gradient flow. Furthermore, BN decreases the network's reliance on initialization, making it less sensitive to initial weights and promoting better model generalization [24]. This was particularly important in adapting the model to the target dataset.

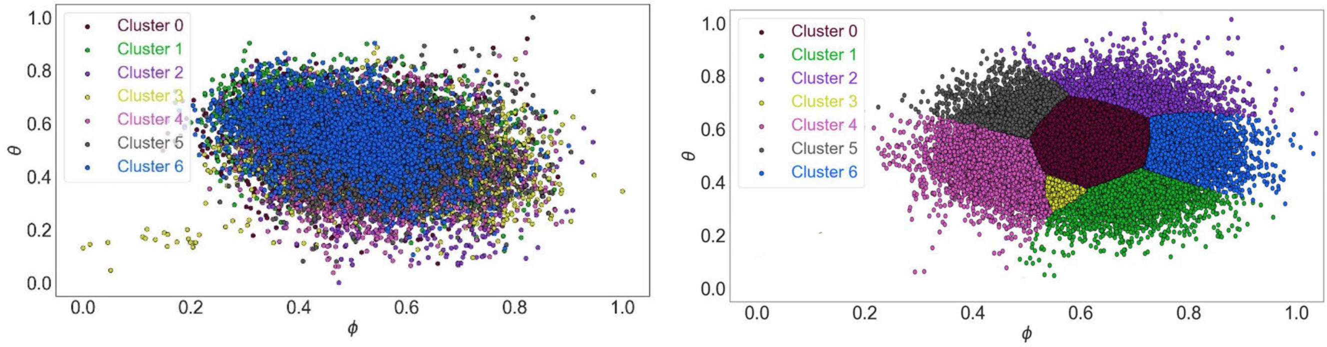
B. PHASE 2: ATTENTION ZONE ESTIMATION

In our framework, the attention zone estimation phase aimed to assign each image in the target dataset to the corresponding attention zone based on the gaze vector information obtained in phase 1. This process involved extracting gaze vectors from the target dataset and mapping them to the relevant zones. In the existing literature, the common approach to mapping images to zone labels is through a straightforward image classification method. However, this approach lacks sufficient interpretability, as it fails to provide meaningful explanations for the assigned zone labels. One reason for this is the variation of the Point of Gaze (PoG) depending on the distance between the driver and the zone plane. Additionally, in the boundaries between adjacent zones, there is a high

TABLE 4. The description of layers in GazeMobileNet.

Layer Number	Description
1	Conv+BatchNorm+ReLU6
2	Additional BatchNorm
3 (Inverted Residual 1)[1]	Conv+BatchNorm+ReLU6+Conv+BatchNorm
4 (Inverted Residual 2)	Conv+BatchNorm+ReLU6+Conv+BatchNorm
…	
20 (Inverted Residual 18)	Conv+BatchNorm+ReLU6+Conv+BatchNorm
21	Global Average Pooling
22	Fully Connected
23	Additional BatchNorm
24	Fully Connected
25	Additional BatchNorm
26 (Regression)	Fully Connected

[1] Each "Inverted Residual" block is a repeated structure with some variations in the numbers of input and output channels, stride, and whether grouped convolutions are used. The sequence within each "Inverted Residual" block typically includes a series of Conv2d, BatchNorm2d, and ReLU6 layers.



(a)

(b)

FIGURE 6. Distribution of 2D gaze angles corresponding to the LISA V2 training subset of "daytime/without eyeglasses". (a) before applying GMM, and (b) after applying GMM.

TABLE 5. Comparison of GazeMobileNet's MAE (in degrees) in the presence/absence of head pose.

Head Pose Information	Average MAE (°)
Presence	2.37
Absence	2.51

level of ambiguity in zone assignment. Moreover, potential ambiguity in class assignments may arise when regions (classes) that are far apart within the vehicle cabin exhibit similar likelihoods and are consequently misclassified. This issue is particularly critical in a safety-related application such as autonomous driving, as it may lead to high-risk false negatives or inconvenient false positives. By incorporating gaze vector information prior to the zone mapping process, a higher level of precision and explainability can be achieved in assigning the zones. This level of precision and explainability is crucial for the effective implementation of autonomous driving systems.

1) UNSUPERVISED CLUSTERING

In the second phase, the gaze vectors extracted from the first phase need to be clustered in an unsupervised manner, based on the predetermined attention zones. Among various methods, such as hierarchical, K-means, density-based, and model-based clustering, the selection of an appropriate technique depends on the data distribution characteristics. Figure 6 (a), illustrates the distribution of the 2D gaze vector obtained from applying the trained gaze estimation model to the target dataset (LISA V2) with seven classes, under the "daytime/without glasses" condition. It can be observed that there is significant overlap among the majority of gaze angles, before applying the clustering method. In this particular scenario, where the gaze vectors were not well-separated and exhibited overlapping distributions, a probabilistic model-based clustering method was proposed to categorize the data samples.

2) GAUSSIAN MIXTURE MODEL

Gaussian mixture model (GMM) as a probabilistic model-based clustering approach can effectively address the extensive overlap in data distributions. GMM represents data as a combination of several Gaussian distributions. The model initializes the parameters including mean, co-variance, and mixing coefficients, and uses the Expectation-Maximization (EM) algorithm to iteratively optimize them. In the E-step, the model calculates the posterior probability that each data point belongs to each Gaussian component, and in the M-step, the model updates the parameters to maximize the likelihood of the data. The algorithm continues to alternate between the two steps until convergence. Due to its probabilistic framework in computing the likelihood data with respect to the Gaussian components, GMM proves to be a powerful tool for modeling data with complex and overlapping distributions. As demonstrated in Fig. 6 (b), applying the GMM clustering method has effectively addressed the overlapping issue, in successfully separating the gaze direction data.

In GMM, convergence tolerance is an important hyper-parameter that needs to be optimized due to its significant impact on GMM performance. Convergence tolerance plays a vital role in determining when the EM algorithm should cease. The EM algorithm is iterative, and convergence tolerance serves as a threshold to ascertain whether the iterations have achieved sufficient convergence [25].

3) CATEGORIZATION PROCESS

To enhance the interpretability of zone classification in the target driving dataset, LISA V2, this study involved the categorization of acquired gaze vectors from each image into their respective attention zones. However, due to the absence of drivers' gaze labels in the target dataset, which is a common challenge in publicly available driving datasets, we introduced an unsupervised clustering approach to categorize the obtained gaze directions, where he number of clusters is equal to the predetermined attention zones. Following the clustering process, each sample in the training subset of the target dataset was assigned a label based on its membership in the corresponding cluster. In this step, there is a $M \times n_{\text{components}}$ matrix, where M denotes the number of observations in the training subset of the target dataset, and $n_{\text{components}}$ refers to the number of clusters.

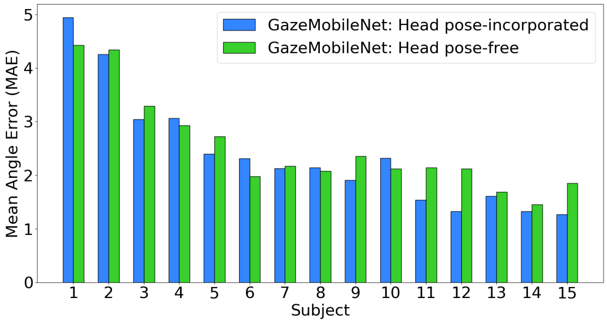


FIGURE 7. Comparison of the performance of GazeMobileNet with and without head pose information.

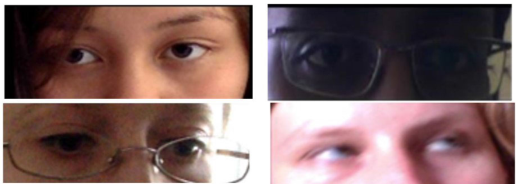


FIGURE 8. Example noise factors in MPIIGaze e.g. Spectacles, high contrasts, blurring, and Eye occlusion by hair in subjects 5, 11, 12, and 15.

Finally, the attention zone corresponding to each cluster had to be determined. To achieve an accurate assignment of the cluster labels, We adopted a comprehensive methodology to establish the correlation between the clusters and the zone labels, encompassing all potential outcomes within the sample space. We selected the optimal match based on the accuracy metric, leveraging a global approach that considers every possibility. The computational complexity of this process was $O(n_{\text{components}}!)$.

V. RESULTS AND EVALUATIONS

A. EVALUATION OF PHASE 1 - GAZE ESTIMATION

The proposed methodology was implemented on a 12th Gen Intel(R) Core(TM) i7-12700 processor, 2.10 GHz, equipped with 32.0 GB of RAM and 12.0 GB NVIDIA GeForce RTX 3060 GPU. To determine the optimal model in the first phase, we evaluated three CNN models. MLeNet, AlexNet, and GazeMobileNet. Mean Squared Error (MSE) was employed as the loss function. To compare the performance of the models, the leave-one-out cross-validation protocol was used, where the images of one subject in the MPIIGaze dataset were designated as the test set, and the images of 14 remaining subjects were adopted to train the model for 50 epochs. The performance of the trained models was evaluated using the mean angle error (MAE) metric. MAE was computed using the cosine similarity measure between the inner products of the estimated gaze angle g_{p_i} and the target gaze angle g_{t_i}.

$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{m} \arccos(\langle g_{p_i} | g_{t_i} \rangle)^2, \tag{3}$$

where m indicates the number of samples.

TABLE 6. Comparison of the performance and number of parameters of MLeNet, AlexNet, and GazeMobileNet.

Network	Average MAE (°)	# of parameters
MLeNet	4.57	$\sim$ 0.5 M
AlexNet	4.86	$\sim$ 44.5 M
GazeMobileNet	**2.51**	$\sim$ **2.5 M**

TABLE 7. Comparison of the proposed gaze estimation versus the state-of-the-art models on the MPIIGaze dataset. The best result is in bold.

Method	Architecture	# of Params	Average MAE (°)
Mahmud et al. [9]	U-Net+ Multi-stream wide ResNet	4 M	4.64
Wang et al. [11]	Bayesian CNN	N.A	4.30
Ghosh et al. [10]	ResNet-50	23 M	4.07
Ali and Kim [12]	Multi-stream shallow CNN	N.A	2.60
Proposed method	GazeMobileNet (Head pose-free)	2.38 M	2.51
	GazeMobileNet (Head pose-incorporated)	**2.39 M**	**2.37**

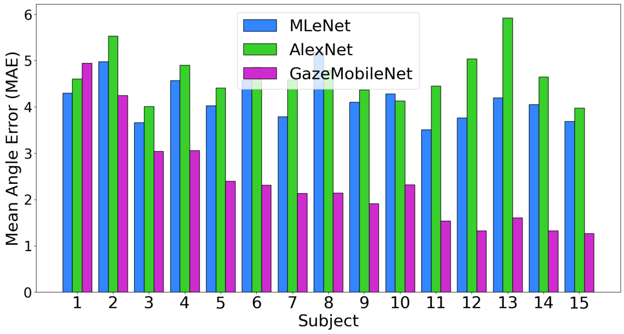


FIGURE 9. Comparison of the performance of MLeNet, AlexNet, and GazeMobileNet.

As depicted in Fig. 7 and Table 5, the head pose-incorporated GazeMobileNet showed a marginally lower MAE for most subjects compared to the head pose-free network. In addition, Table 5 provides a comparison of the average MAE between these two models across 15 subjects. The results suggested that including the head pose information could improve the model's accuracy and provide supplementary information. However, given that the normalized MPIIGaze dataset is an eye-cropped image dataset and does not contain the entire face/head image, the use of head pose information did not have a highly significant impact and, for some subjects, it was redundant information.

As depicted in Fig. 7 and Fig. 8, some differences between the results obtained from the head pose-incorporated and head pose-free models were observed in subjects 5, 11, 12, and 15. According to the specific characteristics of each subject's image in the MPIIGaze dataset, described in Table 1, these subjects were presented with instances of occlusion, such as spectacles, hand, or hair, as well as fluctuations in illumination. Consequently, while head pose information enhanced the overall efficacy of the models, its influence

TABLE 8. Per-class accuracy of attention zone estimation across four different capture conditions and seven classes in LISA V2 dataset (%).

Capture condition	Eyes-closed & Lap	Forward	Left mirror	Radio	Rearview	Right mirror	Speedometer
daytime; without eyeglasses	72.51	73.68	73.59	78.86	89.20	76.04	65.80
nighttime; without eyeglasses	69.74	89.41	77.52	93.07	89.67	87.20	75.15
daytime; with eyeglasses	72.54	76.82	75.12	75.15	77.14	68.42	64.95
nighttime; with eyeglasses	61.00	78.25	63.02	80.30	92.24	82.80	69.58

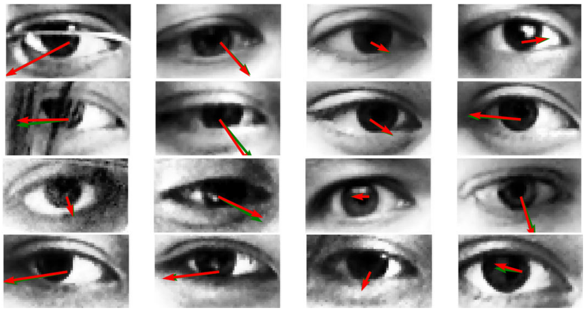


FIGURE 10. Estimated (green) and ground truth (red) gaze vectors in some MPIIGaze data samples.

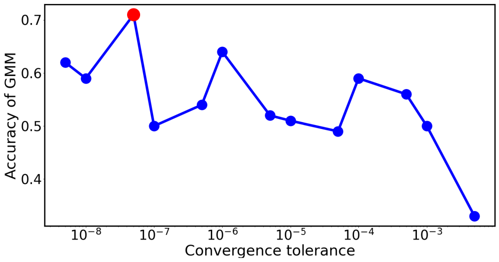


FIGURE 11. Accuracy of zone estimation versus the GMM convergence tolerance for the training subsets of the LISA V2 dataset.

was particularly pronounced in scenarios involving obscured facial or ocular regions, as well as images exhibiting varying levels of darkness or contrast. In such cases, the gaze information gleaned directly from the image may be incomplete, or insufficiently distinct, and the additional cues provided by the head pose information may be instrumental in bolstering the robustness of the model in the presence of noise factors.

In this step, three head pose-free models, MLeNet, AlexNet, and GazeMobileNet were compared to pick the best performing as the final gaze estimation model. In Fig. 9, the MAE of these three models was separately sorted by different subjects.

As shown in Fig. 9 and Table 6, GazeMobileNet had superior performance compared to the other two networks. As demonstrated in Fig. 9, GazeMobileNet significantly outperforms both AlexNet and MLeNet on other subjects, with the exception of Subject 1. There could be some reasons for this. It's possible that the initially chosen random weights for GazeMobileNet were not as suitable for Subject 1's data as those for MLeNet and AlexNet. Under these circumstances, GazeMobileNet may have struggled to converge to an optimal solution for Subject 1, leading to less satisfactory performance. Additionally, Subject 1 might exhibit unique eye characteristics or gaze patterns that the architecture of GazeMobileNet fails to effectively capture. GazeMobileNet, built for efficiency and reduced computational complexity using depthwise separable convolutions, may lack the capacity or specific layers necessary to interpret the nuances of Subject 1's gaze. On the other hand, AlexNet and MLeNet might possess architectures more adept at handling this particular subject's unique characteristics.

As presented in Table 5, GazeMobileNet achieved the state-of-the-art gaze estimation on the MPIIGaze dataset, with MAE of 2.37 ° when using head pose information, and 2.51 ° without utilizing head pose.

The superior performance of GazeMobileNet in estimating the gaze vector accurately is illustrated in Fig. 10, which depicts a cohort of randomly selected test samples. Table 7 presents a comparative analysis of the performance of our proposed GazeMobileNet model on the MPIIGaze dataset versus the state-of-the-art methods in the literature, evaluated under the leave-one-out cross-validation protocol. The proposed model not only demonstrated a significant reduction in MAE but also contained a considerably fewer number of learnable parameters, that plays a key role in determining computational efficiency. In designing an automated AI-based system, the size of the neural network is a critical factor, which impacts the feasibility of its deployment in real-world applications with low-latency requirements. As a result of depthwise convolutions in the architecture, performing separate convolutions, and applying a single filter on each input channel independently, the computational complexity of the convolutional operations was significantly reduced. The lower memory footprint and thus faster inference time in GazeMobileNet make it an appropriate choice for real-time performance and a wide range of deployment constraints in the automotive industry. Therefore, we utilized the trained GazeMobileNet on the MPIIGaze dataset for gaze estimation in the driving dataset in the next phase.

B. EVALUATION OF PHASE 2 - ATTENTION ZONE ESTIMATION

We applied GMM as the clustering method with EM optimization. An optimized convergence tolerance (tol) parameter determines the minimum change in the log-likelihood needed for the algorithm to converge. A smaller tol value leads to a more precise estimation of GMM parameters,

TABLE 9. F1-score of the attention zone estimation across four different capture conditions and seven classes in LISA V2 dataset (%).

Capture condition	Micro average F1-score	Micro average F1-score
Daytime; without eyeglasses	71.18	69.18
Nighttime; without eyeglasses	82.11	80.28
Daytime; with eyeglasses	69.15	68.12
Nighttime; with eyeglasses	70.17	70.03

TABLE 10. Comparison of attention zone estimation accuracy of the proposed method with the original method related to LISA V2 dataset (%). The best results have been bold.

Capture condition	Rangesh et al. [14]	Proposed method
Daytime; without eyeglasses	**81.00**	75.67
Nighttime; without eyeglasses	**87.00**	83.08
Daytime; with eyeglasses	70.63	**73.11**
Nighttime; with eyeglasses	64.81	**74.02**

at the cost of more iterations and computational resources. By adjusting tol, we can control the accuracy of the clustering model. Figure 11 shows the results for 13 different tol values, versus the GMM accuracy. We chose the GMM model with tol $= 5 \times 10^{-8}$ as the attention zone estimation model on the training subset in the LISA V2 dataset and validated with the validation subset.

We evaluated our proposed attention zone estimation approach on four main image-capturing conditions of the LISA V2 dataset, and the per-class accuracies corresponding to six in-cabin zones and the "eyes-closed & lap" state are shown in Table 8. As can be seen in Table 8, the best accuracies in both "daytime without eyeglasses" and "nighttime without eyeglasses" conditions belong to the "radio" and "rearview" classes which achieved accuracies of 78.86% and 89.20%, respectively. Most of the samples from these classes are front-facing views with the entire face and both eyes visible. Hence, the trained network had ample information for a more accurate gaze vector estimation, making the target LISA V2 dataset highly compatible with the training dataset (MPIIGaze) in this context. The lower recognition rates for the "right" and "forward" zones are due to the camera direction; in some captured images, where both eyes were not visible. In the "left mirror" attention zone, most samples contain a side view of the face, where at least one eye and other facial features, such as the nose, and eyebrows, were lost. Consequently, the recognition rate for this class was comparatively lower than that of other classes. The lowest accuracy in the "daytime/without eyeglasses" condition, belongs to the "speedometer" zone and "eyes-closed & lap" state. This phenomenon was expected because, in the "speedometer" class, the eyes of the subjects were not fully visible in most samples. The upper eyelid obscured the sclera, iris, and pupil, making them difficult to recognize. Moreover, subjects in the "eyes-closed & lap" state had their eyes closed. Consequently, the model trained on the MPIIGaze dataset, which was based on open eyes, encountered difficulties in detecting gaze-related features from those samples.

The situation differed slightly for "daytime with eyeglasses" and "nighttime with eyeglasses", likely due to the effect of eyeglasses as an occlusion factor and their impact on the image texture. The highest accuracy in both "daytime with eyeglasses" and "nighttime with eyeglasses" conditions belongs to the "rearview" class, with accuracies of 77.14% and 92.24%, respectively. The "right mirror" class in the "daytime with eyeglasses" condition had the second worst accuracy, while it had the second best accuracy in the "nighttime with eyeglasses". In the "daytime with eyeglasses" condition, there were no significant or meaningful differences among per-class accuracies. Upon examining the samples, it is evident that the number of side-view faces in the "right mirror" class of "daytime with eyeglasses" exceeded the same class in the "nighttime with eyeglasses". Furthermore, daytime introduces the noise factor of illumination variation and sunlight reflection in the eyeglasses, which is absent during nighttime.

Furthermore, we assessed our model's performance using the Micro and Macro F1-score metrics as shown in Table 9 and were calculated as follows.

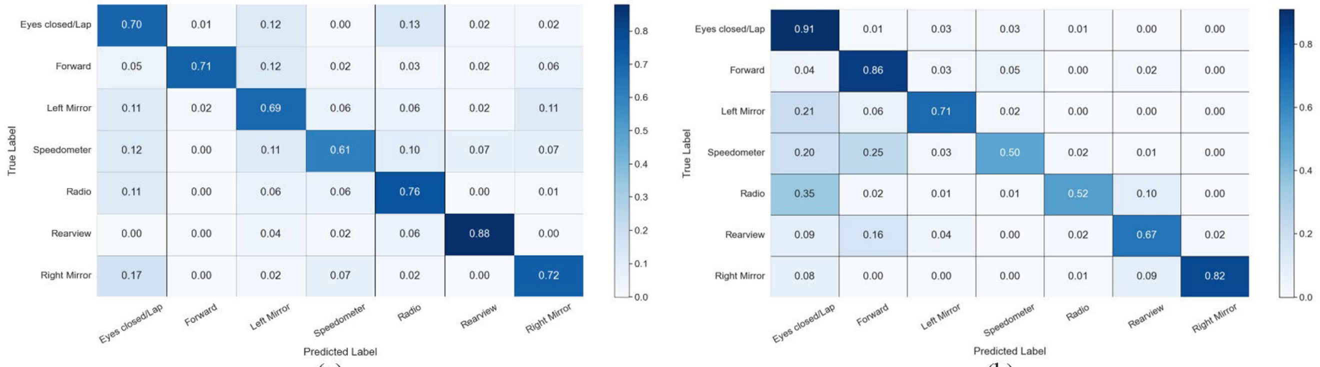
$$\text{Micro average F1-score} = \frac{\sum_{i=1}^{N} 2 \times \mathrm{TP}_i}{\sum_{i=1}^{N} 2 \times \mathrm{TP}_i + \mathrm{FP}_i + \mathrm{FN}_i}, \quad (4)$$

$$\text{Macro average F1-score} = \frac{1}{N} \sum_{i=1}^{N} \frac{2 \times \mathrm{TP}_i}{2 \times \mathrm{TP}_i + \mathrm{FP}_i + \mathrm{FN}_i}, \quad (5)$$

where, TP, FP, and FN represent True Positive, False Positive, and False Negative respectively and N is the number of classes. Micro F1-score aggregates the contributions of all classes to compute the average metric. In other words, it calculates the F1-score by counting the total TP, FN, and FP across all classes while the Macro F1-score, estimates the per-class F1-score independently for and takes the average (hence treating all classes equally), irrespective of the class imbalance.

The discrepancy between Micro- and Macro-averaged F1-scores is caused by the existing class imbalance in the dataset. Under all capture conditions, the Micro F1-score marginally supersedes the Macro F1-score. This finding implies that the model shows better performance in classes that contain a greater number of instances.

As real-world variability and complexity in driver gaze estimation systems are often ignored, we made use of the LISA V2 dataset in our research. This dataset includes examples that encompass a wide range of factors such as the use of eyeglasses, harsh illumination, nighttime data, and more. The diverse characteristics of the LISA V2 dataset make



(a) (b)

FIGURE 12. Normalized confusion matrix for testing subset in "daytime without eyeglasses" capture condition. (a) Proposed model, and (b) proposed model in [14].

it both challenging and general, enabling us to assess the generalization capabilities of our proposed model effectively. Furthermore, the LISA V2 dataset stands out as one of the few publicly available resources specifically designed for in-cabin attention zone evaluation during driving tasks, which adds significant value to our research. Regarding previous research that utilized the LISA V2 dataset for comparison purposes, the [14] is the only study to date that has employed the LISA V2 dataset for their investigations.

Table 10 compares the results of the proposed method with that presented in [14] for LISA V2, following the same hypotheses and settings in this work. As shown in Table 10, our proposed method has been evaluated in the night condition and day condition that contains different lighting and where the performance on the "nighttime" images outperformed the "daytime" cases, in both "with" and "without" eyeglasses. This trend can be attributed to a significant factor: daytime images within this dataset often exhibit intense illumination due to sunlight, leading to increased noise and intricacy in the samples. Harsh illumination in these conditions can result in diverse effects on IR images, including overexposure and diminished contrast. In contrast, nighttime images are inherently shielded from the challenges of harsh illumination.

Also, the proposed model excelled over the method in [14] with samples featuring eyeglasses. Removing eyeglasses can negatively impact image quality, causing the gaze estimation model to be unable to estimate accurate gaze-related appearance features from the eye region. Therefore, we can conclude that removing eyeglasses does not necessarily improve the accuracy of the attention zone classification. Our explainable model, which uses intermediate gaze features, showed robustness in dealing with the eyeglasses as an occlusion noise factor, effectively. One reason is that the gaze estimation model was trained on the images of subjects with eyeglasses in the MPIIGaze dataset.

C. EVALUATION OF EXPLAINABILITY

As observed in Fig. 12, when employing direct classification as the reference method [14], the accuracy of the "eyes-closed & lap" state exceeds that of the proposed method which utilizes the intermediary gaze angle features. This occurs because when the driver's eyes are closed or in the lap position, the gaze direction could align with any of the other zones. As a result, the accuracy of the "eyes-closed & lap" state is distributed among other zones, notably the "left mirror" and "radio" classes.

Regarding explainability, it is evident that, for instance, when considering the "speedometer" attention zone, the method in [14] allocates 25% of probability to the "forward/normal driving" zone, even though the true zone is "speedometer". This means that the automated system has a 25% chance of incorrectly identifying a distracted zone as a normal one, increasing the rate of false negatives. Conversely, in the proposed method, the "forward" zone is integrated, ensuring that the ADAS does not mistakenly perceive an abnormal situation as normal driving. In other words, a distractive zone is not erroneously classified as a normal zone.

VI. CONCLUSION

In this paper, we present a two-phase framework for estimating the driver's gaze direction and attention zone to detect their visual focus. This is explicitly crucial in improving the safety aspects in the intermediate level of autonomy (level 3). We proposed a GazeMobileNet network in the training phase of our proposed platform to estimate the gaze vectors corresponding to each driver's image in the target LISA V2 dataset, as the intermediate features to improve the explainability of the zone classification. In addition to obtaining the state-of-the-art MAE in gaze angle prediction on the MPIIGaze dataset, the network showed sufficient efficiency to be applied in the real-time analysis of the driver's video frames and low-latency deployment requirements. The GMM clustering approach, which was proposed in the second phase, with an optimized tolerance level, increased the accuracy of zone mapping from the highly overlapping distribution of the gaze angle data. Our proposed method demonstrated superior performance in estimating attention zones within the LISA V2 dataset, especially when subjects wearing eyeglasses. This platform offers explainable, robust, and generalizable predictions adequate for safety-related applications.

Additionally, the influence of the head pose information on gaze estimation was examined. We concluded that for normalized data in the MPIIGaze dataset, where images were cropped around the eyes region, and facial features are not readily extractible, head pose information did not significantly enhance the model's performance. This holds true except in instances where noise factors, such as occlusion, lighting, and eyeglasses are present. Moreover, we found that our proposed method excels in estimating attention zones under certain conditions, where both eyes are visible and the face direction is primarily front view, such as "forward", "rearview", and "radio" zones.

In future research, we will utilize a dataset with a broader range of gaze angles, which fully represents the driving gaze distribution based on Original Equipment Manufacturer (OEM) requirements. Additionally, we will estimate the Aleatoric/Epistemic uncertainty and out-of-distribution robustness of our framework which are crucial in safety-related applications.

REFERENCES

[1] K. Hazaymeh, A. Almagbile, and A. H. Alomari, "Spatiotemporal analysis of traffic accidents hotspots based on geospatial techniques," *ISPRS Int. J. Geo-Inf.*, vol. 11, no. 4, p. 260, Apr. 2022.

[2] B. Janakiraman, S. Shanmugam, R. P. De Prado, and M. Wozniak, "3D road lane classification with improved texture patterns and optimized deep classifier," *Sensors*, vol. 23, no. 11, p. 5358, Jun. 2023.

[3] S. Ghosh, A. Dhall, M. Hayat, J. Knibbe, and Q. Ji, "Automatic gaze analysis: A survey of deep learning based approaches," 2021, *arXiv:2108.05479*.

[4] S. Riedmaier, T. Ponn, D. Ludwig, B. Schick, and F. Diermeyer, "Survey on scenario-based safety assessment of automated vehicles," *IEEE Access*, vol. 8, pp. 87456–87477, 2020.

[5] J. Lee and J. H. Yang, "Analysis of driver's EEG given take-over alarm in SAE level 3 automated driving in a simulated environment," *Int. J. Automot. Technol.*, vol. 21, no. 3, pp. 719–728, Jun. 2020.

[6] A. A. Akinyelu and P. Blignaut, "Convolutional neural network-based methods for eye gaze estimation: A survey," *IEEE Access*, vol. 8, pp. 142581–142605, 2020.

[7] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, "Gaze360: Physically unconstrained gaze estimation in the wild," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6911–6920.

[8] Y. Cheng, H. Wang, Y. Bao, and F. Lu, "Appearance-based gaze estimation with deep learning: A review and benchmark," 2021, *arXiv:2104.12668*.

[9] Z. Mahmud, P. Hungler, and A. Etemad, "Multistream gaze estimation with anatomical eye region isolation by synthetic to real transfer learning," 2022, *arXiv:2206.09256*.

[10] S. Ghosh, M. Hayat, A. Dhall, and J. Knibbe, "MTGLS: Multi-task gaze estimation with limited supervision," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1161–1172.

[11] K. Wang, R. Zhao, H. Su, and Q. Ji, "Generalizing eye tracking with Bayesian adversarial learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11899–11908.

[12] A. Ali and Y.-G. Kim, "Deep fusion for 3D gaze estimation from natural face images using multi-stream CNNs," *IEEE Access*, vol. 8, pp. 69212–69221, 2020.

[13] S. Vora, A. Rangesh, and M. M. Trivedi, "Driver gaze zone estimation using convolutional neural networks: A general framework and ablative analysis," *IEEE Trans. Intell. Vehicles*, vol. 3, no. 3, pp. 254–265, Sep. 2018.

[14] A. Rangesh, B. Zhang, and M. M. Trivedi, "Driver gaze estimation in the real world: Overcoming the eyeglass challenge," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Oct. 2020, pp. 1054–1059.

[15] Y. Yang, C. Liu, F. Chang, Y. Lu, and H. Liu, "Driver gaze zone estimation via head pose fusion assisted supervision and eye region weighted encoding," *IEEE Trans. Consum. Electron.*, vol. 67, no. 4, pp. 275–284, Nov. 2021.

[16] M. P. S. Lorente, E. M. Lopez, L. A. Florez, A. L. Espino, J. A. I. Martínez, and A. S. de Miguel, "Explaining deep learning-based driver models," *Appl. Sci.*, vol. 11, no. 8, p. 3321, Apr. 2021.

[17] N. Gang, S. Sibi, R. Michon, B. Mok, C. Chafe, and W. Ju, "Don't be alarmed: Sonifying autonomous vehicle perception to increase situation awareness," in *Proc. 10th Int. Conf. Automot. User Interfaces Interact. Veh. Appl.*, Sep. 2018, pp. 237–246.

[18] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, "Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions," 2021, *arXiv:2112.11561*.

[19] J. Kim and J. Canny, "Interpretable learning for self-driving cars by visualizing causal attention," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[20] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4511–4520.

[21] I. Schmitz and W. Einhäuser, "Gaze estimation in videoconferencing settings," *Comput. Hum. Behav.*, vol. 139, Feb. 2023, Art. no. 107517.

[22] Y. Sugano, Y. Matsushita, and Y. Sato, "Learning-by-synthesis for appearance-based 3D gaze estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1821–1828.

[23] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[24] C. Garbin, X. Zhu, and O. Marques, "Dropout vs. batch normalization: An empirical study of their impact to deep learning," *Multimedia Tools Appl.*, vol. 79, nos. 19–20, pp. 12777–12815, May 2020.

[25] J. A. Bilmes, "A gentle tutorial of the em algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," *Int. Comput. Sci. Inst.*, vol. 4, no. 510, p. 126, Apr. 1998.

ROKSANA YAHYAABADI (Member, IEEE) received the B.S. and M.S. degrees in electronic engineering from the University of Isfahan, Isfahan, Iran, in 2016 and 2019, respectively, and the Ph.D. degree from Western University, London, ON, Canada, in 2022. Her M.S. thesis was focused on face detection algorithms and the implementation of machine learning algorithms on FPGA. Her doctoral research primarily concentrates on improving level 3 autonomous driving through extensive research on various "gaze estimation" and "driver's action recognition" models. Since 2022, she has been an active member of the IEEE Young Professionals, the IEEE Membership Program, and the IEEE Women in Engineering Initiative.

SOODEH NIKAN (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Windsor, in 2014. She is currently an Assistant Professor in software engineering with the Department of ECE, Western University, Canada. Her research interests include artificial intelligence, machine learning, computer vision, data analytics, and signal processing. She has made significant contributions to optimized deep/machine learning-based technologies for highly demanding and safety-critical areas. She has an extensive academic and industry portfolio in AI and automotive research through her research in autonomous driving with Ford Motor Company and Western University. She is the Counselor for the IEEE London Ontario Section Branch. She has been serving on the Technical Program Committee of the International Conference on 6G Net. She has been a reviewer for several electrical and computer engineering journals.