Appearance-Based Driver 3-D Gaze Estimation Using GRM and Mixed Loss Strategies

Taiguo Li, Yingzhi Zhang, *Member, IEEE*, and Quanqin Li

***Abstract*—Driver gaze estimation is a key technology in advanced intelligent vehicles, and it is crucial for ensuring road safety by monitoring driver visual attention. Previously, attention detection through head pose or saliency map integration only offered rudimentary estimation and was insufficient for the advanced driver assistance systems (ADASs), which require more precise gaze data. This work introduces an appearance-based method for driver 3-D gaze estimation. Initially, the Swin Transformer was used to enhance global image information processing, which enabled accurate gaze direction prediction. Furthermore, the method incorporates a gaze refinement module (GRM) as a postbackbone to optimize feature mapping, thus ensuring stable gaze direction estimation. Finally, a mixed loss function was used to improve the accuracy. This mixed loss function combines pinball loss, mean-squared error (MSE), and bias penalty. The experimental results demonstrated angular errors of 3.76° and 10.62° in the MPIIGazeFace and Gaze360 gaze estimation data sets. We inferenced the proposed method to the driver monitoring data set (DMD), and the results demonstrate the effectiveness of this work. Our code is publicly available at github.com/Rocky1salady-killer/DGE-GM.**

***Index Terms*—Driver distraction, gaze estimation, gaze refinement module (GRM), Swin Transformer.**

I. Introduction

THE FOCUS on traffic safety is a worldwide concern. According to the World Health Organization's report, annually approximately 1.35 million people perish in traffic accidents [1]. One of the primary causes is driver distraction, resulting in considerable economic and casualty losses [2]. The advancement of technology is driving deep transformations in road traffic safety. The American society of automotive engineers (SAEs) divides autonomous driving into six levels, ranging from L0 to L5. It is forecasted that by 2030, the United States, Europe, and China will have 82 million intelligent vehicles of L4 or L5 level on the roads [3]. High-level autonomous driving vehicles are anticipated to seamlessly integrate with human drivers, establishing a collaborative control mode [4]. In this mode, the vehicle can operate autonomously in most scenarios. However, rapid driver intervention is still necessary in certain emergency or abnormal situations. Continuous monitoring of the driver's condition is necessary to ensure that the vehicle can be taken over at any time [5]. Thus, the detection of driver distraction is crucial, both in preventing accidents in conventional driving and in ensuring the safe integration of emerging autonomous driving technologies [6].

The centers for disease control and prevention (CDC) [7] extensively defines distracted driving: "The driver's attention is considered diverted when it moves from the driving task to other activities." Distractions are divided into three types: cognitive, behavioral, and visual distractions [8]. Driver's decision making during driving is intimately connected with the visual information they gather. Amidst the rapid changes in complex traffic environments, drivers must swiftly process and judge various external information, with their eye movements and gaze direction often responding before their actions [9]. The driver's gaze direction is a more profound and direct measure of visual attention distraction [10]. In a basic scenario, when a driver wishes to drink water, they first direct their gaze toward the water bottle, confirm its position, and then perform the subsequent actions of grabbing and drinking. These sequential actions are closely linked to eye attention and gaze direction from the outset. Accurately and promptly estimating the driver's gaze direction can provide crucial data for driving assistance systems and is essential for assessing driver visual distraction [11].

Traditional research approaches involved researchers trying to directly link the estimation of the driver's head posture with distracted driving behavior. Zhao et al. [12] proposed a method using continuous head pose estimation to be able to identify distracted driving in real time, using computer vision to analyze head movements. However, the detection effect decreases under different lighting conditions or when the driver's face features are turned back. Concurrently, other researchers explored by segmenting the driver's visual focus across various gaze zones [13]. Palazzi et al. [13] proposed a new method for assessing driver visual attention that utilized eye tracking and scene analysis to predict driver attentional focus. However, the limitations of the model include its dependence on high-quality eye-tracking data, which may be affected by device calibration issues or individual differences in eye movements. Although these methods provide fundamental estimates of driver's gaze directions, their accuracy does not meet the requirements of complex driver distraction

Manuscript received 19 June 2024; revised 30 July 2024; accepted 15 August 2024. Date of publication 26 August 2024; date of current version 20 November 2024. This work was supported in part by the Gansu Provincial Department of Education: University Teacher Innovation Fund Project under Grant 2023A-039; in part by the Science and Technology Program of Gansu Province under Grant 24JRRA238; and in part by the Lanzhou Jiaotong University Youth Science Foundation Project under Grant 2020002. *(Corresponding author: Yingzhi Zhang.)*

Taiguo Li and Yingzhi Zhang are with the School of Automation and Electrical Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China (e-mail: leetg@mail.lzjtu.cn; 12211496@stu.lzjtu.edu.cn).

Quanqin Li is with the Children's Rehabilitation Department, Shaanxi Kangfu Hospital, Xi'an 710065, China (e-mail: 1264688978@qq.com).

Digital Object Identifier 10.1109/JIOT.2024.3449409

detection systems. The novelty of this work [14] lies in its methodological focus, exploring the impact of various methods of analysis of eye tracking data on the conclusions drawn about driver behavior and attention. However, the key limitations of this method are the invasive nature of the equipment and the dependence on the accuracy of the eye-tracking technology, which can lead to errors. Regarding precise eye-tracking data, head-mounted eye trackers can measure eye movements directly but are highly inconvenient for users due to their intrusive nature. Furthermore, eye trackers necessitate laborious calibration processes for each driver and are extremely sensitive to environmental factors. Lv et al. [15] proposed an innovative method for improving driver gaze prediction by incorporating a reinforced attention mechanism. This method utilizes reinforcement learning to enhance the accuracy of gaze prediction models, enabling more precise anticipation of where a driver will look next. Despite its advancements, the method faces limitations, including high computational demands that may affect its deployment in real-time systems. Conversely, although in-vehicle sensors offer beneficial indirect insights into the driver's gaze direction, their accuracy is potentially impacted by data noise and external interference, and increasing the uncertainty in gaze estimation. To overcome these limitations, we introduce an innovative appearance-based 3-D gaze estimation approach for drivers. Differing from conventional methods, our proposed method is nonintrusive and estimates the gaze direction by directly analyzing the driver's appearance. This approach can precisely estimate the driver's gaze direction and also employs cascading networks and arrows for visualization. It offers a more intuitive and distinct representation of the driver's visual focus.

While traditional 2-D gaze estimation techniques for drivers offer some insights in specific scenarios, they are generally unable to fully capture the driver's actual visual focal points in complex driving situations [16]. Especially when dealing with complex tasks, just using head posture or visual saliency maps might not accurately represent the driver's eye attention [17]. By comparison, 3-D gaze estimation methods offer more detailed and precise gaze data, crucial for assessing the driver's visual attention, fatigue levels, and potential driving hazards. In the field of assisted driving, there is still less work on detecting the driver's visual attention through 3-D gaze estimation, which tends to be human 3-D gaze estimation [48], [49], [50], [51]. Liu et al. [48] proposed a 3-D gaze estimation method based on autocalibration. Although the work tried to address the need for continuous calibration of head-mounted devices. However, calibration is still unavoidable, and head-mounted devices still affect the user experience. Gaze estimation is a typical regression task, and the loss function is very important as it affects the stability of the estimation results, but few researchers have done some work on it. Specifically, this research focuses on pitch and yaw angles, as they sufficiently determine the driver's gaze direction. The roll angle, which describes head rotation around the gaze axis, does not affect the gaze direction itself and is thus omitted in our estimation. This simplification allows for accurate gaze direction estimation without the computational overhead and potential confusion introduced by considering roll. Hence, this research exploratively employs a new type of 3-D gaze estimation method that outputs these key gaze data from just the driver's 2-D images. This approach also provides a more versatile and precise technological path for additional research and practical applications.

The main contributions of our work are summarized as follows.

1) In the area of driver distraction detection, this research introduces an appearance-based 3-D gaze estimation technique for drivers. The designed model provides more accurate and stable gaze estimation results compared to existing methods. The mapping function from eye appearance to gaze direction can be fitted even with rapid and large head movements by the driver, while also significantly addressing the consumption of continuous calibration and the experience-affecting issue of head-mounted devices. Our method estimates robust 3-D gaze data which is important for analyzing driver visual attention.
2) In this work, the convolutional neural network (CNN) is replaced by the Swin Transformer as a feature extractor to more accurately capture global image features. To improve the effectiveness of the representation extracted from the backbone, we introduced the gaze refinement module (GRM), which consists of some key components and a specialized prediction head to output stable gaze estimation results. Additionally, we considered the stability of the model estimation results in driving scenarios, so we designed a hybrid loss function that integrates pinball loss and mean-squared error (MSE) loss, along with an added bias penalty component.
3) The proposed model was tested on the prominent and openly accessible gaze estimation data sets MPIIGazeFace and Gaze360, with the results indicating notable performance enhancement from our proposed method, yielding angular errors of 3.76° and 10.62°, respectively. Furthermore, our model underwent additional validation through the driver monitoring data set (DMD), where the driver's gaze direction was visualized using arrows, affirming the model's generalization capability. In conclusion, despite the scarcity of 3-D gaze estimation studies for drivers in this field, a comparative analysis with other current advanced gaze estimation models illustrates the superiority and practicability of our approach.

II. Related Work

A. Driver Visual Attention Estimation

For capturing the driver's visual attention, several studies have concentrated on estimating gaze regions through head direction or uncalibrated gaze angles [18], [19], [20], [21], thereby projecting the gaze onto different areas. These approaches are based on the hypothesis linking driver behavior with the allocation of visual attention.

Jha et al. [18] designed a framework for estimating driver visual attention by analyzing head posture and eye gaze. The model, which is based on the CNN, consists of a head

posture encoder and an eye encoder. The head posture encoder evaluates six parameters to determine the head's position and orientation, transforming them into low-dimensional features.

Differing from prior studies, Kasahara et al. [19] introduced a novel 3-D geometric learning framework. Their approach aims to align the driver's gaze direction with the visual prominence of the observed scene, enhancing the correlation between driver attention and 3-D scene semantics. It particularly focuses on how drivers pay attention to key elements like vanishing points, pedestrians, and traffic signals in their driving environment. Inspired by binocular asymmetry, Cheng et al. [21] proposed a face-based asymmetric regression assessment network (FARE-Net) to optimize the gaze estimation results by taking into account the differences between the left and right eyes. Their method consists of a face-based FAR-Net and an evaluation network (E-Net) for learning binocular reliability.

The prediction of driver attention in dynamic driving scenarios is a significant research subject, with Li et al. [22] introducing the ASIAF-Net for driver attention prediction at both regional and object levels. ASIAF-Net consists of three main parts: 1) attention related spatial feature encoder (AF-Encoder); 2) self-adaptive short-temporal feature extraction module (SSFE-Module); and 3) induced aware fusion network (IAF-Net). The AF-Encoder with a novel association analysis cell (AAC) is built to extract contextual spatial information.

Additionally, researchers have extensively analyzed how various driving conditions (including different landscapes, times, and weather conditions) impact driver attention estimation. This plays a vital role in understanding the dynamics of driver attention in real-world settings [22], [23]. Hu et al. [23] have developed a comprehensive and innovative framework for estimating driver attention within intelligent vehicles, leveraging a unique integration of calibration-free eye gaze and scene features. The framework meticulously extracts spatiotemporal feature maps, including low-level features, static visual saliency maps, and dynamic optical flow information from the driving scene, alongside high-level semantic descriptions. A gaze probability map, derived directly from the driver's gaze direction, is incorporated without the need for specialized eye-tracking equipment, showcasing the framework's calibration-free nature.

B. Driver Gaze Estimation

The precision of gaze estimation is important for ensuring safe driving in the domain of driver attention monitoring. Conventional gaze estimation methods typically depend on laborious calibration processes, adding to the complexity of system implementation and posing challenges in maintaining calibration continuity in real driving situations [24]. Gaze estimation accuracy is severely challenged in natural driving environments due to the variability of the driver's head and eye positions, as well as the ever-changing ambient lighting and viewpoints [25]. Thus, the development of a self-calibrating gaze estimation method that can adapt to the driver's natural behaviors [26], [27] plays a crucial role in enhancing the usability and adaptability of driver monitoring systems.

Yuan et al. [26] introduced a self-calibrating driver gaze estimation technique that innovatively applies domain knowledge of typical driver gaze patterns, eliminating the need for explicit calibration procedures. Their method employed a gaze pattern learning algorithm that learns from predefined gaze regions (side view mirrors, rearview mirrors, speedometer, and center console). The method automatically identified representative time samples and used these instances as implicit calibration points.

Moreover, certain research efforts have broken down the issue of gaze angles into two components, streamlining the calibration process and concurrently addressing variations among individuals. These studies further uncover the problem of individual biases in gaze estimation. Chen and Shi [27] introduced a gaze decomposition method that decomposes the gaze angle into a subject-independent component and a subject-dependent bias. This innovative approach acknowledges that a significant portion of estimation error arises from a consistent bias unique to each subject. Their method indicates that high-accuracy calibration can be achieved using a single gaze target and head position, with improved performance when incorporating variability in head orientation.

Following the recognition of the importance of individual biases, a unified framework proposed by some researchers has been implemented based on a thorough understanding of the interplay between facial dynamics and gaze direction [28], [29], [30]. These approaches not only markedly improve estimation accuracy but also exhibit the potential for comprehensive driver status monitoring through detailed analysis of pertinent facial characteristics [31]. Gou et al. [28] proposed a unified framework employing cascade learning for a comprehensive approach to head pose estimation, as well as eye center detection and gaze estimation. This technique efficiently utilizes the inherent correlation between facial landmarks and 3-D face model parameters The main innovation of this method lies in its ability to exploit the intricate relationship between facial dynamics and gaze direction, paving the way for more integrated driver monitoring systems.

Wei et al. [29] proposed a novel gaze estimation algorithm that incorporates a facial feature extractor (FFE) with a pyramid squeeze attention (PSA) mechanism to refine the accuracy further. By focusing on the facial features surrounding the eyes, the FFE meticulously captures the crucial data required for precise gaze estimation. The subsequent integration of the L2CSNet, which employs the PSA, effectively enhances the correlation weights related to gaze estimation in these feature areas, suppresses irrelevant weights, and extracts more fine-grained information for a more accurate gaze direction prediction.

III. Methods

This work presents a novel approach for estimating driver gaze. As shown in Fig. 1, the network architecture in this study is designed to address key challenges in driver gaze estimation. Current methods often struggle with the complex, nonlinear relationship between facial appearance and gaze direction, particularly under varying driving conditions. A deep learning approach is employed to capture subtle facial

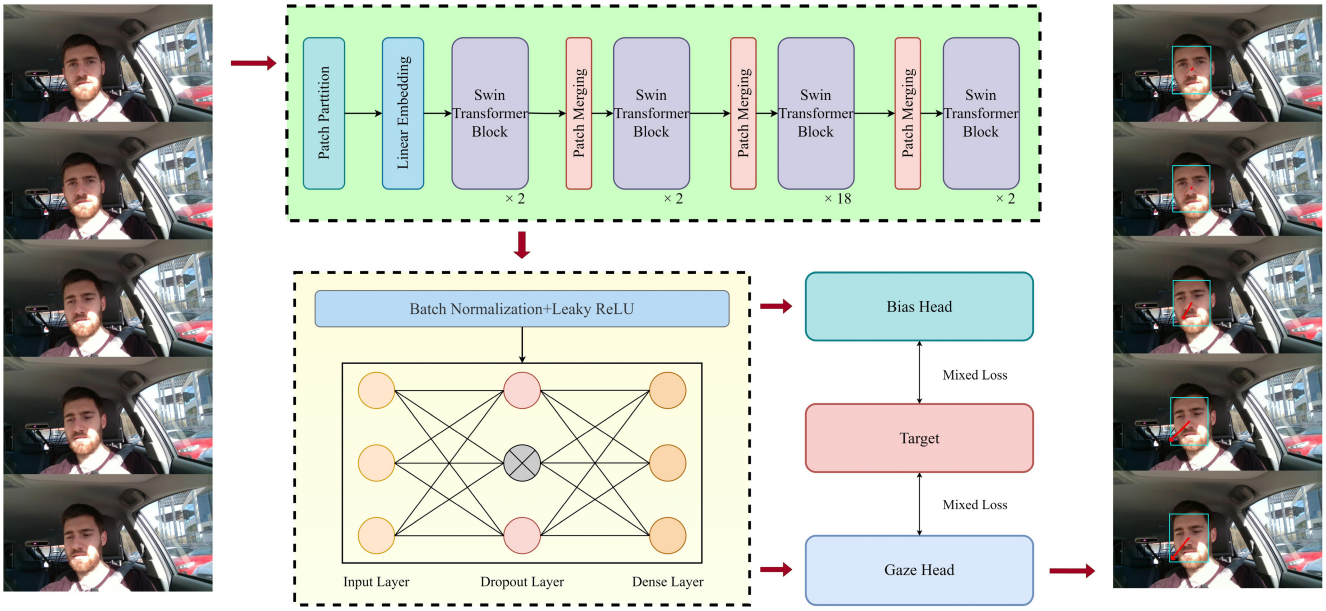

Fig. 1. Pipeline of driver gaze estimation methods proposed in this work.

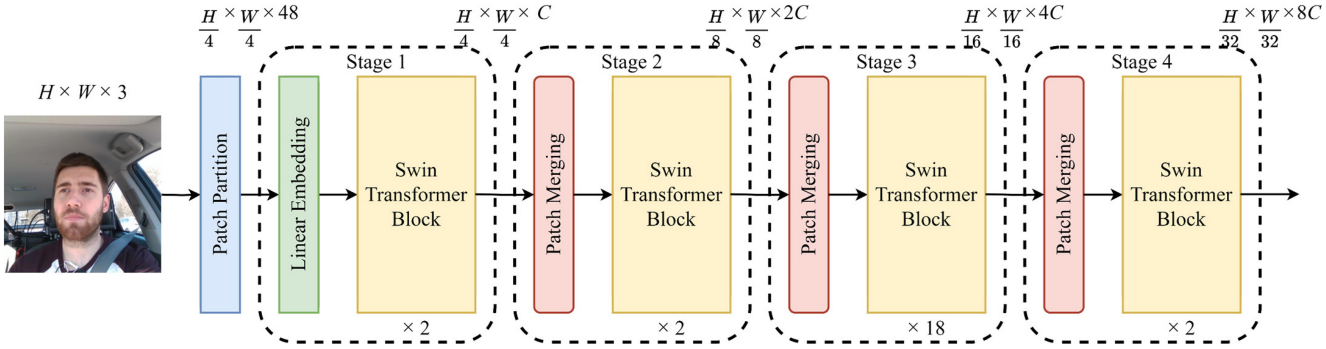

Fig. 2. Architecture of the backbone in the driver's gaze estimation framework.

feature interactions more effectively than traditional methods. The Swin Transformer is utilized as the backbone [32], leveraging its ability to efficiently model both local and global contextual information from driver images, which is a crucial aspect in gaze estimation where both fine eye details and broader facial context are equally important. Recognizing that general feature extraction may not be optimally aligned with the specific requirements of gaze estimation, the GRM is introduced. This module serves as a task-specific adaptation layer, refining features to better suit gaze estimation. To tackle the multifaceted nature of gaze estimation errors, a mixed loss function is implemented. This combines pinball loss for asymmetric error penalization, MSE for general regression accuracy, and a bias penalty term to mitigate systematic biases. These components collectively form a comprehensive solution aimed at improving the accuracy and reliability of driver gaze estimation, which is critical for advanced driver assistance systems (ADASs) and autonomous driving technologies.

The proposed method integrates several key components, each serving a specific function in the driver gaze estimation process. As illustrated in Fig. 1, the Swin Transformer backbone acts as the primary feature extractor, efficiently capturing both local and global contextual information from driver images. The GRM further adapts these features for the specific task of gaze estimation, including normalization, nonlinear activation, and a neural network that learns task-specific patterns. The final component comprises separate prediction heads for gaze direction and bias, which utilize the refined features to generate the final estimations.

Fig. 2 provides a detailed view of the Swin Transformer backbone architecture. The process begins with patch partition, which divides the input image into nonoverlapping patches, followed by linear embedding which projects these patches into a higher-dimensional space. The backbone consists of four stages, each containing multiple Swin Transformer Blocks. Between stages, patch merging layers reduce spatial dimensions while increasing feature dimensions, allowing for progressive feature refinement. As shown in Fig. 3, each Swin Transformer block utilizes window-based and shifted window-based multihead self-attention (W-MSA and SW-MSA) mechanisms, along with layer normalization (LN) and multilayer perceptron (MLP) components. This structure

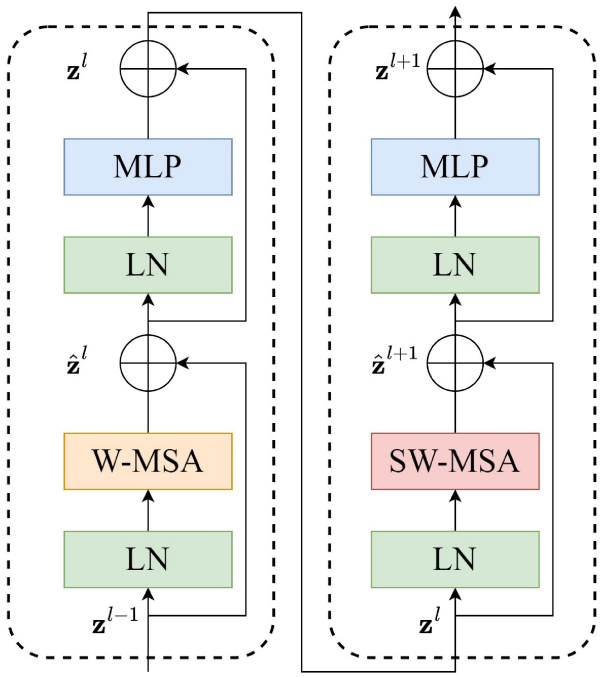

Fig. 3. Network of two Swin Transformer blocks.

enables the model to effectively capture and process complex spatial relationships within driver images at various scales.

A. Suitable Backbone for Framework

For precise estimation of driver gaze, the backbone necessitates potent representational learning abilities. Hence, in our study, we contemplate utilizing the Swin Transformer as the backbone of our detection framework.

In the driver gaze estimation framework, the backbone's architecture, as depicted in Fig. 2, begins by inputting the image into the patch partition module for breaking down into several 4 × 4 pixel patches [32]. Subsequently, these are flattened along the channel axis. For an RGB image with three channels, each 4×4 pixel patch comprises 16 pixels, each having R, G, and B values, thus flattening to $16 \times 3 = 48$. Post patch partition, the image shape transitions from [H, W, 3] to [H/4, W/4, 48]. The linear embedding layer applies a linear transformation to each pixel's channel, altering the initial count of 48 to a definable C value. Following this process, the image shape is altered to [H/4, W/4, C].

Subsequently, the image passes through four stages of processing, undergoing size changes. The linear embedding layer is exclusive to stage 1. The subsequent three stages comprise patch merging layers and differing quantities of Swin Transformer blocks. Patch merging layers facilitate downsampling. The Swin Transformer block, the primary structural component, encompasses two variants: one incorporating the W-MSA module and the other the SW-MSA module. Swin Transformer blocks are thus stacked in pairs.

As illustrated in Fig. 3, the Swin Transformer block starts with feeding the image into the first block, passing successively through a layer norm layer and a W-MSA module, with a skip connection parallel to these steps [32]. The image is then fed into another layer norm layer and an MLP module, with a corresponding skip connection. Upon completion through the first block, the image outputs to the second block.

The second block resembles the first in overall structure, however, it employs shifted windows-based multihead self-attention instead of the conventional window-based approach. The ongoing Swin Transformer block computation is as follows:

$$\hat{\mathbf{z}}^{l} = W - MSA\left(\mathrm{LN}\left(\mathbf{z}^{l-1}\right)\right) + \mathbf{z}^{l-1} \tag{1}$$

$$\mathbf{z}^{l} = \mathrm{MLP}\left(\mathrm{LN}\left(\hat{\mathbf{z}}^{l}\right)\right) + \hat{\mathbf{z}}^{l} \tag{2}$$

$$\hat{\mathbf{z}}^{l+1} = SW - MSA\left(\mathrm{LN}\left(\mathbf{z}^{l}\right)\right) + \mathbf{z}^{l} \tag{3}$$

$$\mathbf{z}^{l+1} = \mathrm{MLP}\left(\mathrm{LN}\left(\hat{\mathbf{z}}^{l+1}\right)\right) + \hat{\mathbf{z}}^{l+1} \tag{4}$$

where $\hat{\mathbf{z}}^{l}$ and $\mathbf{z}^{l}$ signify the output features from block (S)W-MSA and MLP modules. W-MSA and SW-MSA correspond to multihead self-attention configurations using regular and shifted window partitions. The Swin Transformer blocks within the encoder's four stages can be stacked in different quantities. Since each Swin Transformer block processes the image through W-MSA and SW-MSA sequentially, it comprises two layers. Therefore, the number of Swin Transformer blocks stacked must be even. In our study, the number of Swin Transformer blocks across the four stages of the backbone defaults to {2, 2, 18, 2}.

The main working modules are W-MSA and SW-MSA. In the Swin Transformer block, multihead self-attention is conducted on windows, as opposed to the original MSA. The computational load of the MSA was greatly reduced compared to the original by using window-based settings. The computational load of MSA is calculated using the following formula:

$$\Omega(\mathrm{MSA}) = 4\mathrm{hwC}^2 + 2(\mathrm{hw})^2 C \tag{5}$$

where Ω is the computation, h and w are the height and width of the image, respectively, and C is the number of channels. W-MSA module divides the feature map into a window with the width and height of M. A feature map that will get $(h/M) \times (w/M)$ windows, and then use the multiheaded self-attention module for each window. Since the window's width and height are M, bring the above formula as $4(\mathrm{MC})^2 + 2(M)^4 C$, the final W-MSA calculation is

$$\Upsilon(W - \mathrm{MSA}) = 4\mathrm{hwC}^2 + 2M^2\mathrm{hwC} \tag{6}$$

where Υ is the computation, h and w are the height and width of the image, respectively, and C is the number of channels.

B. Gaze Refinement Module

The GRM is introduced to further refine the representations extracted from the Swin Transformer to ensure that they are properly tailored to the driver gaze estimation task. This module encompasses four architectural components. Through normalization, regularization, nonlinear transformation, and specialized prediction heads, the GRM ensures that the model's outputs are both accurate and robust across varied gaze estimation scenarios.

1) *Batch Normalization [33]:* The activations are stabilized through normalization, which ensures that the mean is close to 0 and the standard deviation is close to 1. This smoothens the optimization landscape, facilitates faster and more stable training, and makes the network less sensitive to the initialization of weights. Given a mini-batch of data $B = \{x_1, x_2, \ldots, x_m\}$, batch normalization normalizes each element based on the following equations:

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i \tag{7}$$

$$\sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_B)^2 \tag{8}$$

$$\widehat{x_i} = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \tag{9}$$

where μ_B is the mini-batch mean, σ_B^2 is the mini-batch variance, and ϵ is a small constant added for numerical stability.

2) *Dropout:* Dropout is a regularization technique where [34], during training, each input unit is retained with a probability of $1 - p$ and set to zero with a probability of p. This helps in preventing overfitting by ensuring that no single neuron or feature becomes overly specialized to the training data, thus contributing to better generalization for unseen data.
Mathematically, given an input vector x, dropout is formulated as

$$y_i = \text{Bernoulli}(1-p) \times \frac{x_i}{1-p} \tag{10}$$

where Bernoulli$(1-p)$ represents a Bernoulli distribution that takes the value 1 with probability $1 - p$ and 0 with probability p (indicating the neuron is dropped) and $1/1-p$ is the scaling factor. This ensures that the model remains robust by preventing over-reliance on specific neurons.

3) *Dense Layer With Leaky ReLU:* The layer captures nonlinear relationships in the data using the Leaky ReLU activation function. The nonlinearity introduced helps the model capture complex relationships in the data. The Leaky ReLU ensures that even neurons with negative values have a gradient, which can prevent "dying" neurons and ensure consistent learning [35].
The dense layer provides additional capacity to the model, allowing it to capture intricate relationships in the data. It is represented as

$$y = \text{W}x + b \tag{11}$$

where W is the weight matrix, x is the input, and b is the bias vector. Following this transformation, the Leaky ReLU activation function is applied:

$$\text{LeakyReLU}(x) = \begin{cases} x, & \text{if } x > 0 \\ \delta x, & \text{otherwise} \end{cases} \tag{12}$$

where δ is a small constant, typically $\delta = 0.01$. This activation introduces nonlinearity while ensuring that neurons have nonzero gradients, even for negative inputs.

4) *Separate Heads for Gaze and Bias:* These heads are responsible for making the final predictions. One predicts the gaze direction, and the other predicts the gaze bias. By having distinct heads for gaze and bias, the model can learn specialized representations for each task, potentially increasing the accuracy and precision of the predictions [36].

These dedicated prediction heads generate the gaze direction and bias values, respectively. Formally represented as

$$\text{Gaze} = W_g \times \text{DenseOutput} + b_g \tag{13}$$

$$\text{Bias} = W_b \times \text{DenseOutput} + b_b \tag{14}$$

where W_g and W_b are the weight matrices for gaze and bias prediction heads, respectively, and b_g and b_b are their corresponding biases.

C. Mixed Loss Function

The primary objective of gaze estimation is not just to accurately predict gaze directions but also to ensure that the model's predictions are consistent across varying conditions and possess a level of uncertainty awareness [37]. To accomplish this, we employ a mixed loss function that synergistically combines three distinct loss components: 1) the pinball loss; 2) MSE; and 3) a penalty for bias deviations.

The pinball loss, also known as quantile regression loss [38], is instrumental in our approach for estimating the uncertainty associated with predictions. Given a model's predicted output, output_o, a true target value, target_o, and an estimated variance, var_o, we compute the pinball loss for two specific quantiles, $q_{0.1}$ and $q_{0.9}$, representing the lower and upper prediction bounds, respectively. Our customized loss function is defined as follows:

For the 10th percentile (quantile $q_{0.1}$)

$$\begin{aligned} &L_{q_{0.1}}\big(\text{output}_o, \text{target}_o, \text{var}_o\big) \\ &\quad = \max(q_{0.1} \cdot (\text{target}_o - (\text{output}_o - \text{var}_o)) \\ &\quad (q_{0.1} - 1) \cdot (\text{target}_o - (\text{output}_o - \text{var}_o))). \end{aligned} \tag{15}$$

For the 90th percentile (quantile $q_{0.9}$)

$$\begin{aligned} &L_{q_{0.9}}\big(\text{output}_o, \text{target}_o, \text{var}_o\big) \\ &\quad = \max(q_{0.9} \cdot (\text{target}_o - (\text{output}_o + \text{var}_o)) \\ &\quad (q_{0.9} - 1) \cdot (\text{target}_o - (\text{output}_o + \text{var}_o))). \end{aligned} \tag{16}$$

The final loss is the mean of the two computed losses

$$L\big(\text{output}_o, \text{target}_o, \text{var}_o\big) = \frac{L_{q_{0.1}} + L_{q_{0.9}}}{2} \tag{17}$$

where L_q is the pinball loss for the quantile q, tailored to account for the prediction variance var_o in our model. The predicted output is represented by output_o, while target_o is the true target value. Our model considers two quantiles, $q_{0.1}$ and $q_{0.9}$, to effectively capture the lower and upper bounds of the predicted gaze, integrating the variance var_o to accommodate the inherent uncertainty present in the gaze estimation task.

The MSE quantifies the average squared disparities between the predicted and actual gaze directions. For given predictions output_o and actual values target_o, the MSE is articulated as

$$\text{MSE}\left(\text{output}_o, \text{target}_o\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\text{output}_{oi} - \text{target}_{oi}\right)^2 \quad (18)$$

where MSE denotes the MSE. output_{oi} is the predicted output for the *i-th* sample, target_{oi} is the true value for the *i-th* sample, and n is the number of samples.

To deter the model from generating overly biased predictions, we introduce a penalty term that computes the absolute difference between the output and its variance. Mathematically, it is defined as

$$\text{BiasPenalty}\left(\text{output}_o, \text{var}_o\right) = \left|\text{output}_o - \text{var}_o\right| \quad (19)$$

where BiasPenalty is the penalty term for model bias, output_o is the model's predicted output, and var_o represents the variance associated with predictions.

The mixed loss function is a linear amalgamation of the aforementioned loss components, governed by two hyperparameters α and β. Formally, it is represented as

$$\text{MixedLoss} = \text{PinballLoss} + \alpha \cdot \text{MSE} + \beta \cdot \text{BiasPenalty} \quad (20)$$

where MixedLoss denotes the overall loss. α and β are the hyperparameters that balance the contributions of MSE and the bias penalty, respectively.

The mixed loss function offers a holistic assessment of the model's predictions, taking into account both accuracy and uncertainty. This multipronged strategy guarantees that the model's predictions are not only precise but also consistent and uncertainty-aware.

IV. Experiments

A. Data Sets

The data sets chiefly employed in this research are Gaze360 [38], MPIIGazeFace [39], and DMD [40]. Gaze360 is a public, large-scale data set for human gaze estimation. It is designed to enable robust 3-D gaze estimation in unconstrained imagery. The data set consists of 238 subjects in indoor and outdoor settings, marked with 3-D gazes in a variety of head poses and distances, encompassing over 1.4 million gaze samples across various environments and lighting conditions. This diversity makes Gaze360 particularly valuable for developing models that can generalize well to real-world scenarios. MPIIGazeFace is an extensively used the human gaze estimation data set featuring 3-D gaze data under various environmental conditions and head poses. It comprises 45 000 images, gathered from 15 participants.

B. Training Details

The DMD data set offers a broad, varied, and thorough approach to driver behavior monitoring. It encompasses both real and simulated driving scenarios, covering distracted driving and driver fatigue. DMD addresses the need for multifaceted DMDs, enhancing the scope of driver distraction detection.

Fig. 4. Examples of presentation from the Gaze 360 data set.

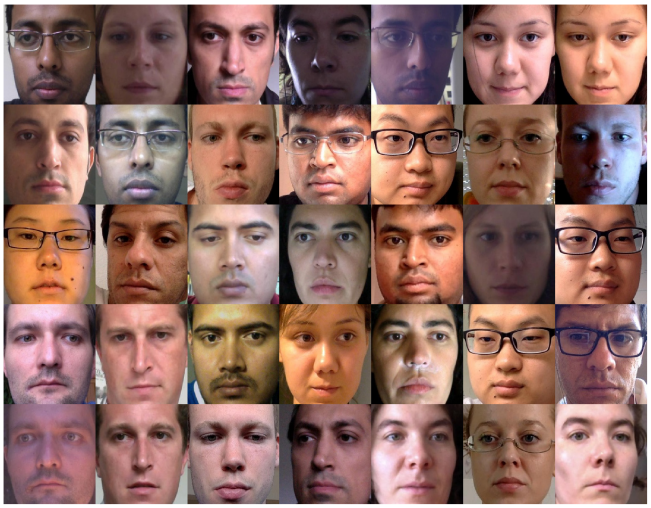

Fig. 5. Examples of presentation from the MPIIGazeFace data set.

The Gaze360 and MPIIGazeFace data sets are employed for training and testing the model to ensure it accurately learns to gaze directions across different scenarios and head poses. The DMD data set is used to verify the model's generalization in real-world settings and to visualize its inference outcomes on the DMD data set. Following the original data set partitioning of Gaze360, we used 86 000 samples as the training set and 16 031 as the test set. The MPIIGazeFace data set does not have official partitioning requirements and includes gaze data from 15 different individuals, with 3000 samples each. We utilize the leave-one-person-out (LOPO) strategy for training and testing [41], a cross-validation approach where one participant's data serves as the test set and the remaining participants' data as the training set.

This article uses the experimental environment configuration in Table I to guarantee the effectiveness of model training and testing.

For the hardware and software configuration, we employed an AMD EPYC 7453 processor with 28 cores, along with an NVIDIA RTX 4090 GPU featuring 24 G of VRAM. Regarding

TABLE I
HARDWARE AND SOFTWARE CONFIGURATION

Operating systems	Linux Ubuntu 20.04.2 LTS
CPU	AMD EPYC 7453
GPU	NVIDIA Geforce RTX 4090 (24G)
Memory	32G×2
Solid state drives	480G
Pytorch	V2.0.1
CUDA	V11.8

the Gaze360 data set, the Adam optimizer was deployed, setting epochs at 100, batch size at 80, a base learning rate of 0.0001, and maintaining weight decay at 1. With the MPIIGazeFace data set, we set epochs to 40, batch size to 80, and a learning rate of 0.00001, utilizing the AdamW optimizer. The learning rate scheduler was ReduceLROnPlateau, reducing the learning rate by a factor of 0.1 when the validation loss stopped decreasing, with a patience of 10 epochs.

C. Gaze Estimation Metrics

For gaze estimation studies, angular error is typically employed as the main metric to directly represent the discrepancy between the predicted and actual gaze directions of the model, offering a clear indication of the model's accuracy and performance. In this work, the Gaze360 and MPIIGazeFace data sets, annotated with the real gaze directions of subjects using specialized eye-tracking devices, contain labels with pitch and yaw information. Pitch and yaw refer to the angles of orientation or rotation of an object in 3-D space [42]. Pitch is the angle of rotation around the X-axis, indicating vertical tilt, while yaw is the angle of rotation around the Y-axis, indicating horizontal rotation. Pitch and yaw are instrumental in locating an individual's gaze direction in gaze estimation [17]

$$\text{pitch} = \arctan\left(\frac{y}{z}\right) \tag{21}$$

$$\text{yaw} = \arctan\left(\frac{x}{z}\right) \tag{22}$$

where x, y, and z are the coordinates of the 3-D vector from the center of the eye to the focal point. The arc tan function yields the angle associated with these coordinates.

The process for calculating angular error involves comparing the predicted gaze direction with the actual gaze direction. Initially, the 2-D gaze direction is transformed into a unit vector in 3-D space, as illustrated in the following formula:

$$x = -\cos(\text{pitch}) \cdot \sin(\text{yaw}) \tag{23}$$

$$y = -\sin(\text{pitch}) \tag{24}$$

$$z = -\cos(\text{pitch}) \cdot \cos(\text{yaw}) \tag{25}$$

where x, y, and z denote the unit vector coordinates in 3-D space. Pitch and yaw refer to the angles of gaze direction in 2-D space. The sin function calculates the sine of the given angle, while the cos function computes its cosine.

Once the actual and estimated gaze vectors are calculated, the angular error is determined using the following calculation:

TABLE II
COMPARITION OF ACTIVATION FUNCTIONS ON GAZE360 DATA SET

Method	Relu	Leaky Relu	Mean angular error
Gaze360			11.04°
Ours	√		10.86°
Ours		√	10.68°

$$\text{Angular Error} = \arccos\left(\frac{\text{PV} \cdot \text{GTV}}{||\text{PV}|| \cdot ||\text{GTV}||}\right) \times \left(\frac{180}{\pi}\right) \tag{26}$$

where PV represents the predicted gaze vector and GTV is the actual gaze vector. The magnitudes of the vectors are ||PV|| and ||GTV||. The arc cos function, the inverse of the cosine, calculates the angle from the cosine value. This is followed by converting the radians to degrees. The process of converting radians to degrees is realized by $(180/\pi)$.

D. Comparison of Results on Gaze360 and MPIIGazeFace

In order to explore the effect of the two activation functions, Leaky ReLu and ReLu on the GRM module, we performed a simple ablation experiment on the Gaze360 data set. The comparison results are shown in Table II.

Leaky ReLU introduces a small, positive slope in the nonactive phase of the neurons, which ensures a continuous flow of gradients. This attribute is particularly beneficial for maintaining gradient propagation during training, thus potentially enhancing learning efficiency and model performance. Our experimental results on the Gaze360 data set demonstrate that models utilizing Leaky ReLU achieve a lower mean angular error compared to those using ReLU, with error rates of 10.68°. In the GRM module, Leaky ReLU was chosen as the activation function over ReLU. It is notable that among the hyperparameter settings of the hybrid loss function, we set α and β to 0.8 and 0.1 as default.

This experiment was designed to systematically explore the performance impact of these parameters within specified ranges. We aim to evaluate the efficacy of different combinations of hyperparameters (α and β) in the mixed loss function to optimize the estimation of driver gaze direction. The range for α was chosen between 0.6 and 0.9 based on preliminary tests that indicated this interval provided the best balance between minimizing error and avoiding overfitting. For β, the range was set between 0.01 and 0.15 to effectively penalize minor deviations without disproportionately affecting the overall loss, ensuring that the penalties applied were stringent but fair. The experimental results are shown in Table III.

The results from our systematic hyperparameter tuning reveal the relationship: the performance of the model varies significantly with changes in α and β, demonstrating their critical role in the mixed loss function. Notably, combinations around $\alpha = 0.6$ and $\beta = 0.05$ yielded the best results, achieving the lowest angular error of 10.62°. We take the model that results in 10.62° as the best model in this work. This experiment underscores the importance of careful hyperparameter selection and its direct impact on model performance. Our findings support the use of our mixed loss

TABLE III
IMPACT OF HYPERPARAMETERS ADJUSTMENTS FOR DRIVER GAZE ESTIMATION ACCURACY ON GAZE360 DATA SET

Hyperparameter α	Hyperparameter β	Mean angular error
0	0.1	10.86°
0.8	0	10.91°
0.6	0.01	10.84°
0.6	**0.05**	**10.62°**
0.6	0.1	10.67°
0.6	0.15	10.93°
0.7	0.01	10.79°
0.7	0.05	10.77°
0.7	0.1	11.01°
0.7	0.15	10.72°
0.8	0.01	10.63°
0.8	0.05	10.75°
0.8	0.1	10.68°
0.8	0.15	10.85°
0.9	0.01	10.76°
0.9	0.05	10.69°
0.9	0.1	10.74°
0.9	0.15	10.69°

function as a robust approach to enhancing gaze estimation accuracy, providing a solid foundation for further refinements.

This section's experiments aim to verify the advancement of the methodologies presented in this study using the Gaze360 and MPIIGazeFace data sets. First, we performed ablation experiments on the Gaze360 data set. The first model is Gaze360, which is an advanced method proposed by the official team of the Gaze360 data set. The model uses ResNet for the backbone and pinball loss for the loss function. The second model is ViT [38], the loss still uses the original pinball loss. We started experimenting with using Transformer instead of cnn for feature extraction and also proved its effectiveness. For the third model, we used a transformed version of Transformer. Swin Transformer will be designed to be more suitable for computer vision tasks. SwinT+GRM and SwinT+Mixed Loss are based on the Swin Transformer using our proposed GRM and mixed loss functions. It can be seen from the table that both our proposed GRM and mixed loss give a positive impact.

In these experiments, we followed the official training settings of Gaze360, so the epoch of the Gaze360 model was 60. The remaining five models, all of which use transformers as backbones and are optimized for better feature learning, had epochs set to 100. Specific experimental results are shown in Table IV.

The initial Gaze360 model demonstrated an angular error of 11.04° in the Gaze360 data set. The performance of the model received a small improvement after using the base vision transformer as the backbone. The mean angular error of the ViT model was 10.87°, which indicates that the transformer's feature extraction is more effective in the gaze estimation task. The purpose of this improvement is to allow the model to better capture the subtle features of the eye. To further extend this advantage, we used the Swin Transformer as a backbone, and the SwinT model also obtained a good result. The GRM is designed as a small postbackbone, aiming to optimize the feature mapping of the backbone. Therefore, SwinT+GRM achieved a result of 10.72°, which also proves this point. In addition, driver gaze estimation is actually a regression task, which is different from the classification task where the loss function plays an important role during the training process and it greatly affects the performance of the model. However, as we can see from the table, the angular error is also lower in the models trained with the mixed loss function.

We would like to take a larger view of the ablation experiments and analyze the impact of these improvements, we calculated the median angular error, interquartile range (IQR) of angular error, and root mean-square error (RMSE) of the model. Their visualization results are shown in Fig. 6, in order of IQR of angular error, median angular error, and RMSE. From the visualization results, it can be clearly observed that the angular error values are gradually decreasing after the model is improved one by one. Comprehensively, we demonstrate that we proposed each point to bring effective improvement to the gaze estimation.

The test set of Gaze360 contains 16 031 samples, which we divided into groups of 400 to compute the average angular error per group. The final group comprised the remaining 31 samples. This strategy was employed to compare the baseline model (the baseline model refers to the Gaze360) and ours. The final test results, presented in Fig. 7, reveal that in 41 sample groups, the ours demonstrated a lower average angular error than the original, covering 71% of the groups. From another angle, this illustrates that the method proposed in this study is capable of correcting major deviations and reducing overall angular error, thereby enhancing the accuracy of driver gaze estimation.

In addition, we validated the method proposed in this article on another important gaze data set, MPIIGazeFace. While validating our method, we compared it to the FullFace [39] (in this section of the experiments the baseline model refers to the to the FullFace) method proposed by the MPIIGazeFace data set team. Since the MPIIGazeFace officials did not provide a data set partitioning strategy, we tested according to the LOPO strategy. The MPIIGazeFace data set contains face images and gaze data of 15 people. Each person has 3000 samples, including image and gaze data, 15 people are defined as labels 0–14, and the test results are the mean angular errors on labels 0 to 14. Fig. 8 shows the test results of the FullFace model and our model on labels 0–14 of the MPIIGazeFace data set.

The baseline model achieved an average angular error of 4.93°, while the method proposed in this article resulted in an average angular error of 3.76°. The results showed that the method proposed in this work, compared to the FullFace method, consistently maintains a lower angular error for most labels. Specifically, for labels like 05, the original method had an angular error of 5.86°, while our method achieved 3.71°. For label 09, the original method showed an error of 5.31°, where as our method resulted in 3.16°. In the test results for label 14, the original method had a substantial angular error of 7.68°, which our method significantly improved, reducing it to 5.07°.

TABLE IV
ABLATION EXPERIMENTS RESULTS ON GAZE360 DATA SET

Model	Mean angular error	Backbone	Epoch
Gaze360[38]	11.04°	ResNet	60
ViT	10.87°	Vision Transformer	100
SwinT	10.75°	Swin Transformer	100
SwinT+GRM	10.72°	Swin Transformer	100
SwinT+Mixed Loss	10.70°	Swin Transformer	100
Ours(SwinT+Mixed Loss+GRM)	**10.68°**	Swin Transformer	100

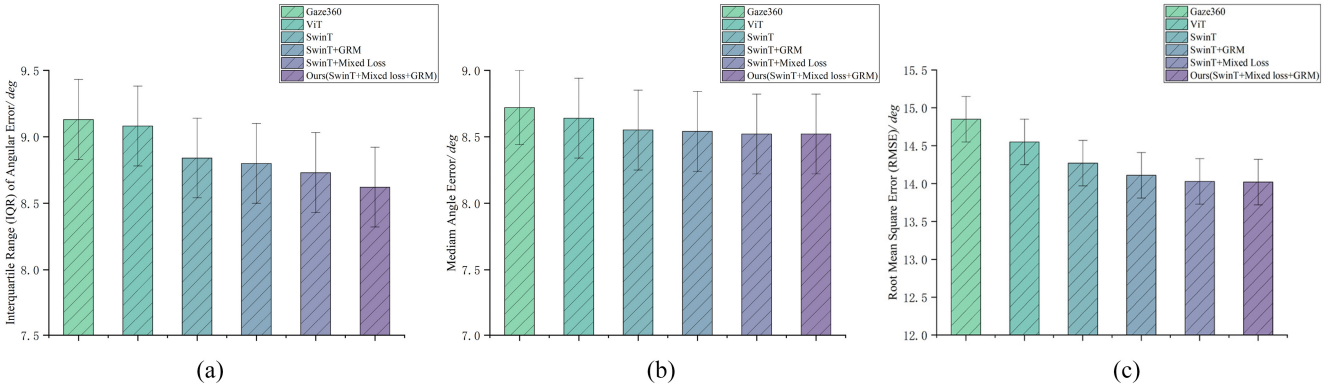

Fig. 6. Ablation experiments results on Gaze360. (a) Comparison of different models on IQR of angular error. (b) Comparison of different models on median angular error. (c) Comparison of different models on RMSE.

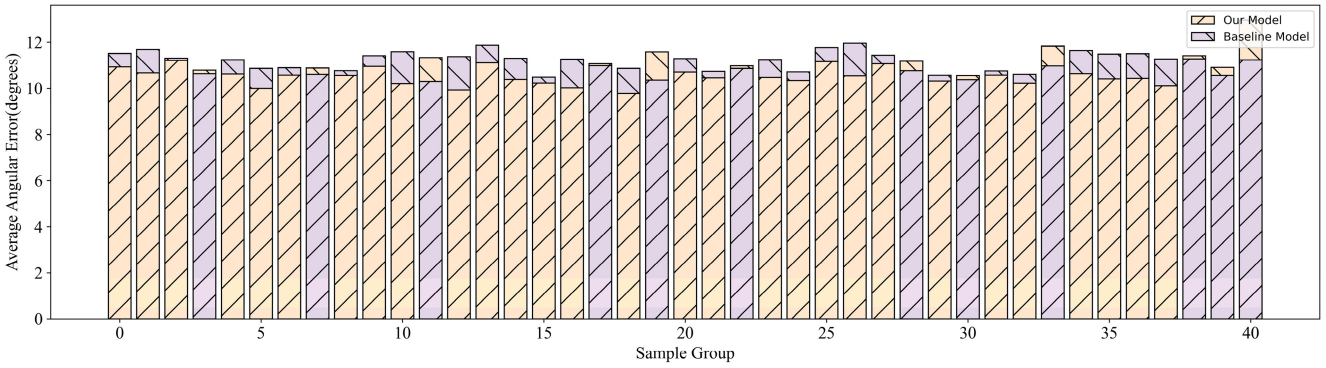

Fig. 7. Performance of the baseline model and ours on 41 sample groups of Gaze360.

The driver gaze estimation task is a typical regression task that requires the model to estimate specific driver gaze information. Among the test results, the closer the model estimates pitch and yaw to the ground truth, the better the estimation performance of the model. Therefore, we performed a comprehensive regression analysis to compare the baseline model with our model to get a wide range of comparisons. We selected three labels, label 0, label 6, and label 12. In Fig. 8, we can observe that the difference between the performance of the baseline model and ours on label 0 is relatively small, and the difference in the performance of label 6 is moderate, but the difference in the performance of label 12 is huge.

This analysis aims to evaluate the accuracy of the model in estimating the direction of the driver's gaze. The regression charts show the correlation between the model's estimated pitch and yaw values and the ground truth. In these diagrams, each dot signifies a sample; the horizontal axis reflects the ground truth values, and the vertical axis indicates the model's predicted values. In an ideal scenario, if the model predictions are entirely accurate, all points would align closely along the line of $y = x$.

The regression line visually represents the relationship between the predicted values and actual values. It was noted that for pitch and yaw angles, there is a significant positive correlation between predicted and actual values, demonstrating our model's effectiveness in capturing gaze direction trends. Especially notable is the our model's robust approximation performance at extreme pitch and yaw values. This aspect is crucial in gaze estimation, as it pertains to the precise identification of different angles and orientations.

From the visualization result in Fig. 9, it is easy to observe that the distribution of the samples of the baseline model on the three labels is scattered, although the overall distribution of the samples is along the line of the ground truth. In particular,

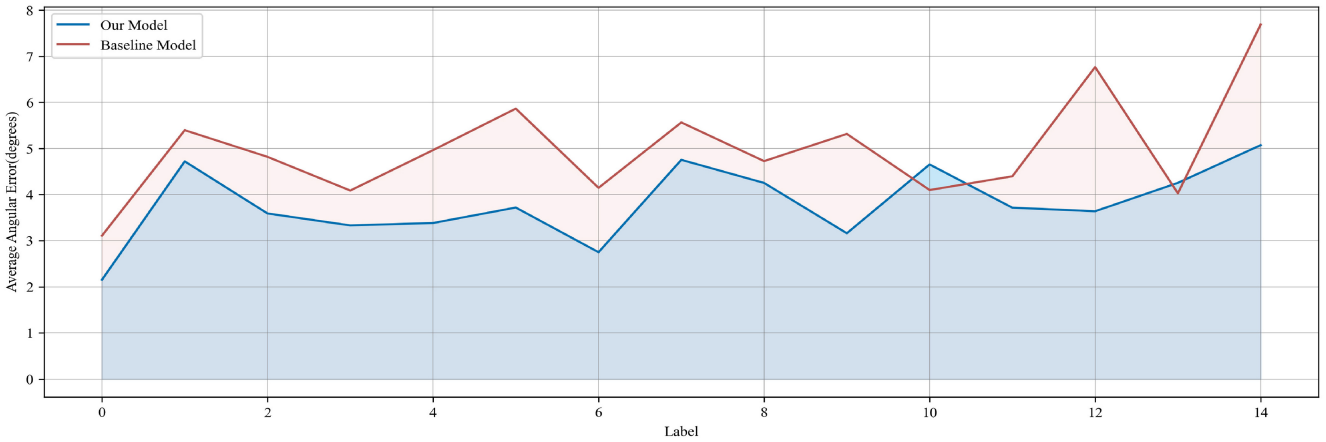

Fig. 8. Mountain peak map of the baseline model and ours on MPIIGazeFace.

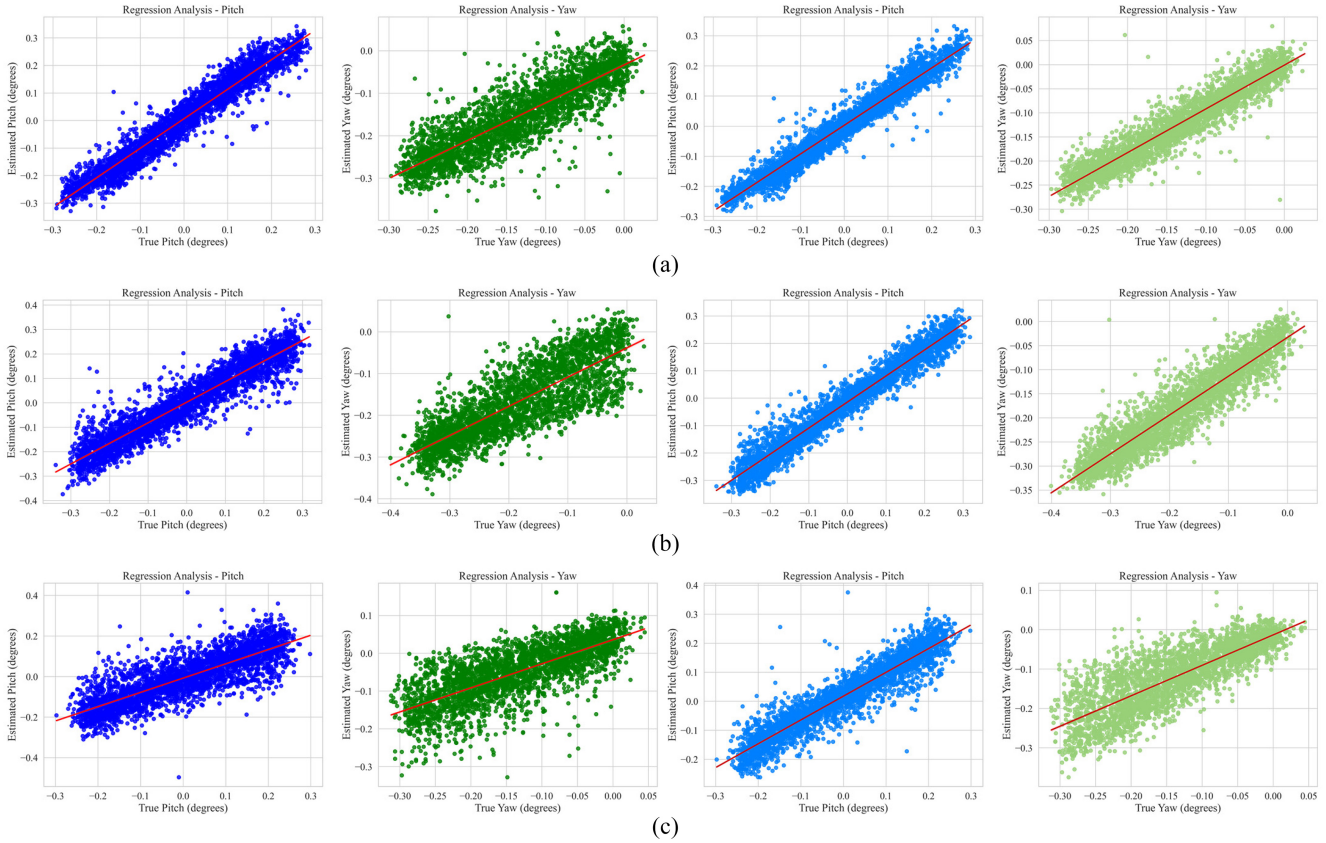

Fig. 9. Visualization of the regression analysis of our model on MPIIGazeFace. Left: denotes the regression analysis of baseline on pitch and yaw. Right: denotes the regression analysis of our method on pitch and yaw. (a) Label 0 of MPIIGazeFace. (b) Label 6 of MPIIGazeFace. (c) Label 12 of MPIIGazeFace.

in the yaw regression plot, many samples are far from the red line. On the other hand, the overall situation of our model on the three labels is better than that of the baseline, and the distribution of the samples becomes more compact and closer to the red line. Meanwhile, the baseline model has a lot of scattered samples in the bottom plot of the yaw regression plot, but our model improves this situation to a great extent.

E. Validation on DMD Data Set

To verify the generalizability of the method proposed in this article, we tested it using the DMD data set, which includes distracted driver behavior in both real-world and simulator scenarios. We selected videos from real driving scenarios, specifically those captured with RGB cameras focused on the driver's head position. Each video contains instances of normal driving behavior and distracted behavior, including talking to passengers, adjusting the radio, drinking water, etc. We chose videos of six different subjects. After processing with ours, the videos were segmented into frames for analyzing gaze estimation results from various angles and subjects.

Notably, our method can directly estimate the driver's 3-D gaze direction from their appearance, which is crucial for gaze data. However, in 2-D images or videos, we rely on

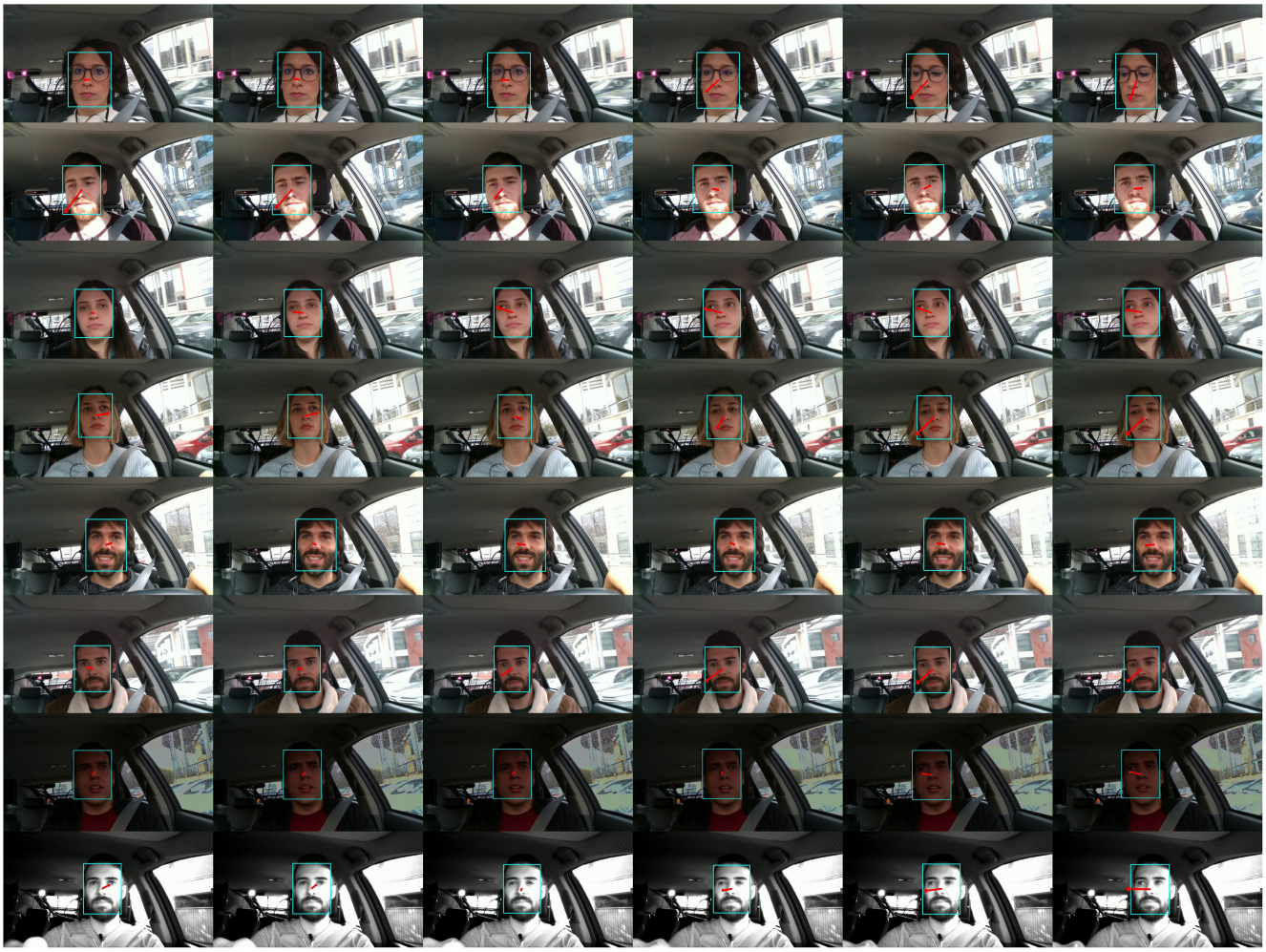

Fig. 10. Visualization of driver gaze estimation.

arrows for visualization. Thus, the direction of the arrows in the visualization represents the direction of the driver's gaze. For example, if the driver is looking forward, the arrow may appear as a point or it may be shortened, indicating that the driver is looking forward.

The specific gaze estimation visualization results are shown in Fig. 10. It is clear from the figure that our model's gaze estimation is quite accurate, whether the driver is driving normally or distracted. This demonstrates that although our method was trained on the general gaze estimation data sets, it has strong generalizability in real driving scenarios. For example, the fifth driver had just finished drinking water, focusing their attention on the water bottle, and our model accurately estimated the driver's line of sight. Additionally, considering various driver conditions, we specifically selected a video of a driver wearing glasses for validation. Wearing glasses poses additional challenges for appearance-based gaze estimation methods, especially due to reflections from the glasses, but our model still performs well in terms of generalizability.

Additionally, validation of the proposed method was conducted using driver traveling video footage from infrared cameras under dark night conditions. Furthermore, weak light situations caused by varying times of day and weather conditions, such as evening or overcast days, were also considered to assess the method's performance across a range of challenging visibility scenarios.

The test results from the infrared camera video demonstrated relatively accurate driver gaze estimation overall. The infrared camera effectively captured the driver's gaze characteristics at night, maintaining good eye movement detail. Consequently, the proposed method still achieved good estimation results under these conditions. For weak light video tests, footage captured under extremely low light conditions was utilized, particularly focusing on environments within the driver's cab where illumination was severely limited. This increased the validation difficulty, with periods where the human eye was also unable to discern the driver's line of sight. Despite these challenges, the method performed robustly. While there were instances of lost or biased estimation results for brief durations, the overall estimation accuracy exceeded initial expectations.

F. Comparison With Related Works

In order to verify the detection capabilities of the method presented in this study, a comparison was made between our optimally performing ours and earlier advanced SOTA

TABLE V
COMPARISON RESULTS WITH DIFFERENT MODELS ON THE MPIIGAZEFACE AND GAZE360 DATA SETS

Method	Backbone	MPIIGazeFace	Gaze360
Gaze360 [38]	ResNet	4.06°	11.04°
FullFace [39]	AlexNet	4.93°	14.99°
Mnist [41]	Multimodal CNN	6.39°	N/A
GazeNet [43]	VGG-16	5.76°	N/A
Dilated-Net[45]	VGG with Dilated Convolutions	4.42°	13.73°
RT-Gene [46]	MTCNN with Custom Networks	4.66°	12.26°
CA-Net[47]	CNN with Attention Mechanism	4.27°	11.20°
GazeTR-Pure[44]	Vision Transformer	4.74°	13.58°
GazeTR-Hybrid [44]	Vision Transformer+ResNet18	4.00°	10.62°
Ours	Swin Transformer	**3.76°**	**10.62°**

"N/A" No data available

models, using the Gaze360 and MPIIGazeFace data sets. Given the scarcity of gaze estimation research specifically directed at driver attention estimation, our method was contrasted with previously established general gaze estimation techniques. Detailed comparative results are depicted in Table V.

In the table, the ours denotes the approach introduced in this study. The Gaze360 and Fullface methods were developed by the teams of the Gaze360 and MPIIGazeFace data sets, respectively, and were used as baseline models in Section III-D of the experiment. Mnist and GazeNet [43] are seminal works in the early phase of gaze estimation, employing CNN networks as backbones, with Mnist simply using stacked CNNs. As a result, Mnist and GazeNet exhibited the highest angular errors, yet they significantly advanced the field of appearance-based gaze estimation. Innovations and improvements on CNN networks were made by later researchers.

GazeTR-Pure [44], relying solely on transformers as the main network, lacks supplemental network designs. This method, with suboptimal feature extraction and feature mapping handling, only managed to attain results of 4.74° and 13.58°. GazeTR-Hybrid, integrating ResNet and using CNNs for initial feature extraction before processing feature mappings with ViT, achieved results of 4.00° and 10.62°. In contrast, the method introduced in this article, based entirely on Swin Transformer for feature extraction, fully exploits the benefits of transformers. Additionally, it incorporates extra network structures to effectively manage feature mappings. The mixed loss function design further improves the model's generalizability, surpassing GazeTR-Hybrid in cross-data set performance. Our method secured a result of 10.62° on the Gaze360 data set and excelled on the MPIIGazeFace data set with a mean angular error of 3.76°.

V. CONCLUSION

In the realm of driver visual attention detection research, this study introduces an appearance-based method for estimating a driver's 3-D gaze. This approach substantially enhances the precision and efficiency of driver attention state monitoring in intricate driving environments. A new framework has been devised to improve the driver's 3-D gaze estimation and to augment the adaptability and robustness of gaze estimation within driving contexts.

1) Utilizing Swin Transformer as the backbone facilitates more accurate capturing and processing of both local and global information within image data, thereby enabling the model to swiftly and precisely predict the driver's gaze direction.
2) GRM is introduced, leveraging a fusion strategy for spatial and temporal features, which bolsters the continuity of gaze estimation across successive frames and markedly enhances the estimation's accuracy and stability.
3) A mixed loss function with an integrated weight design has been formulated. This compound loss function amalgamates pinball loss with MSE loss, supplemented by a bias penalty term, offering comprehensive error feedback to the model.

In future work, we will explore the following two aspects.

1) We will prioritize enhancing our model's adaptability to diverse environmental conditions, particularly varying lighting scenarios encountered in real-world driving. We plan to develop a dynamic adaptation mechanism that can adjust the model's parameters in real-time based on input data characteristics. This will involve implementing a reinforcement learning algorithm capable of swiftly adapting to changing environmental conditions, including different lighting, nighttime driving, and infrared camera inputs. We plan to expand the training data set to encompass a broader range of lighting conditions and explore light-invariant feature extraction techniques to improve robustness across different scenarios.
2) Additionally, considering the reduction of model complexity and the edge computational resource limitations. Our main approach will be to explore knowledge distillation techniques to create a more efficient model. This process aims to significantly reduce the model's parameters and computational requirements while maintaining its accuracy. We will investigate methods to identify and eliminate redundant components, streamline the architecture, and compress the model without compromising its performance.

REFERENCES

[1] *Global Status Report on Road Safety 2023*, World Health Org., Geneva, Switzerland, 2023.

[2] (World Health Org., Geneva, Switzerland). *Global Status Report on Road Safety*. (Jan. 2018), [Online]. Available: https://www.who.int/publications/i/item/global-status-report-on-road-safety-2018

[3] L. Chen et al., "Milestones in autonomous driving and intelligent vehicles: Survey of surveys," *IEEE Trans. Intell. Veh.*, vol. 8, no. 2, pp. 1046–1056, Feb. 2023.

[4] C. Huang, H. Huang, J. Zhang, P. Hang, Z. Hu, and C. Lv, "Human-machine cooperative trajectory planning and tracking for safe automated driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 12050–12063, Aug. 2022.

[5] Y. Zhang, T. Li, C. Li, and X. Zhou, "A novel driver distraction behavior detection method based on self-supervised learning with masked image modeling," *IEEE Internet Things J.*, vol. 11, no. 4, pp. 6056–6071, Feb. 2024.

[6] K. Muhammad, A. Ullah, J. Lloret, J. D. Ser, and V. H. C. de Albuquerque, "Deep learning for safe autonomous driving: Current challenges and future directions," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4316–4336, Jul. 2021.

[7] "Distracted driving," 2024. [Online]. Available: https://www.cdc.gov/motorvehiclesafety/distracted_driving/

[8] Y. Zhang and D. Kaber, "Evaluation of strategies for integrated classification of visual-manual and cognitive distractions in driving," *Hum. Factors*, vol. 58, no. 6, pp. 944–958, Sep. 2016.

[9] J. A. Abbasi, D. Mullins, N. Ringelstein, P. Reilhac, E. Jones, and M. Glavin, "An analysis of driver gaze behaviour at roundabouts," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 8715–8724, Jul. 2022.

[10] L. Alam, M. M. Hoque, M. A. A. Dewan, N. Siddique, I. Rano, and I. H. Sarker, "Active vision-based attention monitoring system for non-distracted driving," *IEEE Access*, vol. 9, pp. 28540–28557, 2021.

[11] I. Kotseruba and J. K. Tsotsos, "Attention for vision-based assistive and automated driving: A review of algorithms and datasets," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 19907–19928, Nov. 2022.

[12] Z. Zhao et al., "Driver distraction detection method based on continuous head pose estimation," *Comput. Intell. Neurosci.*, vol. 2020, pp. 1–10, Nov. 2020.

[13] A. Palazzi et al., "Predicting the driver's focus of attention: The DR (eye) VE project," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1720–1733, Jul. 2019.

[14] J. Li et al., "Appearance-based gaze estimation for ASD diagnosis," *IEEE Trans. Cybern.*, vol. 52, no. 7, pp. 6504–6517, Jul. 2022.

[15] K. Lv, H. Sheng, Z. Xiong, W. Li, and L. Zheng, "Improving driver gaze prediction with reinforced attention," *IEEE Trans. Multimedia*, vol. 23, pp. 4198–4207, 2021.

[16] D. He, B. Donmez, C. C. Liu, and K. N. Plataniotis, "High cognitive load assessment in drivers through wireless electroencephalography and the validation of a modified N-back task," *IEEE Trans. Human-Mach. Syst.*, vol. 49, no. 4, pp. 362–371, Aug. 2019.

[17] P. Pathirana, S. Senarath, D. Meedeniya, and S. Jayarathna, "Eye gaze estimation: A survey on deep learning-based approaches," *Expert Syst. Appl.*, vol. 199, Aug. 2022, Art. no. 116894.

[18] S. Jha, N. Al-Dhahir, and C. Busso, "Driver visual attention estimation using head pose and eye appearance information," *IEEE Open J. Intell. Transp. Syst.*, vol. 4, pp. 216–231, 2023.

[19] I. Kasahara, S. Stent, and H. S. Park, "Look both ways: Self-supervising driver gaze estimation and road scene saliency," in *Proc. 17th Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 126–142.

[20] L. Dai, J. Liu, Z. Ju, and Y. Gao, "Attention-mechanism-based real-time gaze tracking in natural scenes with residual blocks," *IEEE Trans. Cogn. Develop. Syst.*, vol. 14, no. 2, pp. 696–707, Jun. 2022.

[21] Y. Cheng, X. Zhang, F. Lu, and Y. Sato, "Gaze estimation by exploring two-eye asymmetry," *IEEE Trans. Image Process.*, vol. 29, pp. 5259–5272, 2020.

[22] Q. Li, C. Liu, F. Chang, S. Li, H. Liu, and Z. Liu, "Adaptive short-temporal induced aware fusion network for predicting attention regions like a driver," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 18695–18706, Oct. 2022.

[23] Z. Hu, C. Lv, P. Hang, C. Huang, and Y. Xing, "Data-driven estimation of driver attention using calibration-free eye gaze and scene features," *IEEE Trans. Ind. Electron.*, vol. 69, no. 2, pp. 1800–1808, Feb. 2022.

[24] J. Guo et al., "A novel robotic guidance system with eye-gaze tracking control for needle-based interventions," *IEEE Trans. Cogn. Develop. Syst.*, vol. 13, no. 1, pp. 179–188, Mar. 2021.

[25] Z. S. Zhe, D. Zhang, C. L. Chi, M. Li, and D. J. Lee, "A complementary dual-branch network for appearance-based gaze estimation from lowres-olution facial image," *IEEE Trans. Cogn. Develop. Syst.*, vol. 15, no. 3, pp. 1323–1334, Sep. 2023.

[26] G. Yuan, Y. Wang, H. Yan, and X. Fu, "Self-calibrated driver gaze estimation via gaze pattern learning," *Knowl.-Based Syst.*, vol. 235, Jan. 2022, Art. no. 107630.

[27] Z. Chen and B. E. Shi, "Towards high performance low complexity calibration in appearance based gaze estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 1174–1188, Jan. 2023.

[28] C. Gou, Y. Zhou, Y. Xiao, X. Wang, and H. Yu, "Cascade learning for driver facial monitoring," *IEEE Trans. Intell. Veh.*, vol. 8, no. 1, pp. 404–412, Jan. 2023.

[29] J. Wei, H. Wu, Q. Wu, Y. Iwahori, X. Yu, and A. Wang, "Gaze estimation method combining facial feature extractor with pyramid squeeze attention mechanism," *Electronics*, vol. 12, no. 14, p. 3104, 2023.

[30] J. Fang, D. Yan, J. Qiao, J. Xue, and H. Yu, "DADA: Driver attention prediction in driving accident scenarios," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 4959–4971, Jun. 2022.

[31] F.-Y. Wang, "A new phase of IEEE transactions on intelligent vehicles: Being smart becoming active and believing intelligent vehicles," *IEEE Trans. Intell. Veh.*, vol. 8, no. 1, pp. 3–15, Jan. 2023.

[32] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[33] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[34] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[35] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," 2015, *arXiv:1505.00853*.

[36] Y. Cheng, H. Wang, Y. Bao, and F. Lu, "Appearance-based gaze estimation with deep learning: A review and benchmark," *IEEE Trans. Pattern Anal. Mach. Intell*, early access, Apr. 25, 2024, doi: 10.1109/TPAMI.2024.3393571.

[37] Y. Rong, C. Han, C. Hellert, A. Loyal, and E. Kasneci, "Artificial intelligence methods in in-cabin use cases: A survey," *IEEE Intell. Transp. Syst. Mag.*, vol. 14, no. 3, pp. 132–145, May/Jun. 2021.

[38] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, "Gaze360: Physically unconstrained gaze estimation in the wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6912–6921.

[39] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "It's written all over your face: Full-face appearance-based gaze estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops)*, 2017, pp. 51–60.

[40] J. D. Ortega et al., "DMD: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2020, pp. 387–405.

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

[42] S. J. Lee, J. Jo, H. G. Jung, K. R. Park, and J. Kim, "Real-time gaze estimator based on driver's head orientation for forward collision warning system," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, pp. 254–267, Mar. 2011.

[43] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "MPIIGaze: Real-world dataset and deep appearance-based gaze estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 162–175, Jan. 2019.

[44] Y. Cheng and F. Lu, "Gaze estimation using transformer," in *Proc. Int. Conf. Pattern Recognit.*, 2022, pp. 3341–3347.

[45] Z. Chen and B. E. Shi, "Appearance-based gaze estimation using dilated-convolutions," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 309–324.

[46] T. Fischer, H. Chang, and Y. Demiris, "RT-GENE: Real-time eye gaze estimation in natural environments," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 334–352.

[47] Y. Cheng, S. Huang, F. Wang, C. Qian, and F. Lu, "A coarse-to-fine adaptive network for appearance-based gaze estimation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 10623–10630.

[48] M. Liu, Y. Li, and H. Liu, "3D gaze estimation for head-mounted eye tracking system with auto-calibration method," *IEEE Access*, vol. 8, pp. 104207–104215, 2020.

[49] Y. Xia, J. Liang, Q. Li, P. Xin, and N. Zhang, "High-accuracy 3D gaze estimation with efficient recalibration for head-mounted gaze tracking systems," *Sensors*, vol. 22, no. 12, p. 4357, 2022.

[50] M. Mokatren, T. Kuflik, and I. Shimshoni, "3D gaze estimation using RGB-IR cameras," *Sensors*, vol. 23, no. 1, p. 381, 2022.
[51] X. Zhou, J. Lin, Z. Zhang, Z. Shao, S. Chen, and H. Liu, "Improved Itracker combined with bidirectional long short-term memory for 3D gaze estimation using appearance cues," *Neurocomputing*, vol. 390, pp. 217–225, May 2020.

Taiguo Li received the B.Eng degree in computer science and technology and the M.Eng. degree in computer software and theory from Wuhan University, Wuhan, China, in 2008 and 2010, respectively, and the Ph.D. degree in electronic science and technology from Lanzhou Institute of Physics, China Academy of Space Technology, Beijing, China, in 2017.

He is currently an Assistant Professor with Lanzhou Jiaotong University, Lanzhou, China. His research interests include electronic science and technology, computer vision, and driver behavior analysis.

Yingzhi Zhang (Member, IEEE) was born in Laiwu, Shandong, China, in 1998. He received the B.Eng. degree in automation from Jiangsu Normal University, Xuzhou, China, in 2021. He is currently pursuing the master's degree in electrical engineering with Lanzhou Jiaotong University, Lanzhou, China.

He works as a Student Assistant with Hong Kong Polytechnic University, Hong Kong. His research interests include driver behavior, machine learning, and human factors.

Quanqin Li received the B.N. degree in rehabilitation therapeutics from Shaanxi University of Chinese Medicine, Xi'an, China, in 2015.

She is currently a Medical Care Staff with Shanxi Provincial Rehabilitation Hospital, Xi'an. Her research interests include rehabilitation treatment and mental state evaluation.