LNet: Lightweight Network for Driver Attention Estimation via Scene and Gaze Consistency

Daosong Hu, Xi Li, Mingyue Cui, *Associate Member, IEEE*, and Kai Huang, *Member, IEEE*

***Abstract*—In resource-constrained vehicle systems, establishing consistency between multi-view scenes and driver gaze remains challenging. Prior methods mainly focus on cross-source data fusion, estimating gaze or attention maps through unidirectional implicit links between scene and facial features. Although bidirectional projection can correct misalignment between predictions and ground truth, the high resolution of scene images and complex semantic extraction incur heavy computational loads. To address these issues, we propose a lightweight driver-attention estimation framework that leverages geometric consistency between scene and gaze to guide feature extraction bidirectionally, thereby strengthening representation. Specifically, we first introduce a lightweight feature extraction module that captures global and local information in parallel through dual asymmetric branches to efficiently extract facial and scene features. An information cross fusion module is then designed to promote interaction between the scene and gaze streams. The multi-branch architecture extracts gaze and geometric cues at multiple scales, reducing the computational redundancy caused by mixed features when modeling geometric consistency across both views. Experiments on a large public dataset show that incorporating scene information introduces no significant computational overhead and yields a better trade-off between accuracy and efficiency. Moreover, leveraging bidirectional projection and the temporal continuity of gaze, we preliminarily explore the framework's potential for predicting attention trends.**

***Index Terms*—Driver gaze estimation, lightweight, feature fusion.**

I. Introduction

In THE past few years, distracted driving has remained a significant factor in road traffic accidents [1], [2]. Currently, autonomous driving has become a hot topic of research, and some intelligent vehicles have already deployed assisted driving functions [3], [4]. However, most vehicles are still driven by humans [5]. Before autonomous driving can be widely replaced on a large scale, the trend of human control will continue in the foreseeable future. Distracted driving refers to the driver's attention being deviated from the area to be observed for a long time, thereby failing to timely control the vehicle to avoid obstacles [6], [7]. Therefore, researching how to capture the driver's attention to monitor dangerous behaviors is valuable. Such research aims to improve road safety [8], [9] and support assisted driving decisions [10], [11].

The deviation of human attention is determined by the coupling of head pose and pupil position, and is specifically manifested as a change in gaze direction [12], [13]. Therefore, gaze tracking is widely used to capture the direction of human attention [14]. Head-mounted devices [15], [16] and additional sensors [17] are used to measure eye parameters, thereby modeling a virtual eye, with the normal vector of the pupil's surface as the gaze direction. However, such invasive devices are limited by the working distance [18]. With the development of deep learning, understanding gaze behavior based on images has become possible [19]. Cameras, as non-invasive devices, capture appearance changes at a fixed frequency, containing all the information needed for gaze estimation, such as head pose and pupil position [20].

The performance of neural networks is affected by the scale of the dataset. Since gaze is difficult to measure directly, researchers convert dense gaze vectors into sparse gaze zones and construct large-scale datasets [21]. This simplification strategy is based on an observation that the driver's attention usually focuses on fixed areas inside the car, such as the windshield and rearview mirror [22], [23]. Sparsity limits the model's ability to understand gaze behavior. To measure dense gaze directions in space, it is redefined as a unit direction vector originating from the center of the face and pointing towards the target. Depth cameras or reference objects are used to infer facial depth, thereby obtaining spatial position coordinates to achieve gaze direction measurement [24]. Compared with annotation area estimation, driver's gaze direction estimation can more effectively understand the driver's attention.

As shown in Fig. 1 (a), existing methods typically input facial images into a neural network to extract measurable gaze cues, which are then used to regress the gaze direction [25], [26]. However, these methods often overlook the gaze target. During driving, the driver's gaze is frequently disrupted by moving or salient objects, such as road markings, other vehicles, and buildings. Kasahara et al. [5] proposed a 3D geometric learning framework that models the geometric consistency between the driver's gaze and scene saliency, enabling self-supervised gaze estimation, as shown in Fig. 1 (b). Moreover, the implicit complex relationship between the scene and gaze can be used to guide the generation of attention maps [27], as shown in Fig. 1 (c). However, these approaches

Received 25 February 2025; revised 25 August 2025 and 18 October 2025; accepted 10 December 2025. Date of publication 30 December 2025; date of current version 16 January 2026. This work was supported in part by Guangxi Key Research and Development Program under Grant GuikeAB24010324 and in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2025A1515011485. The associate editor coordinating the review of this article and approving it for publication was Dr. Sebastian Bosse. *(Corresponding author: Mingyue Cui.)*

The authors are with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China (e-mail: huds@mail2.sysu.edu.cn; lixi58@mail2.sysu.edu.cn; cuimy@mail2.sysu.edu.cn; huangk36@mail.sysu.edu.cn).

Digital Object Identifier 10.1109/TIP.2025.3646893

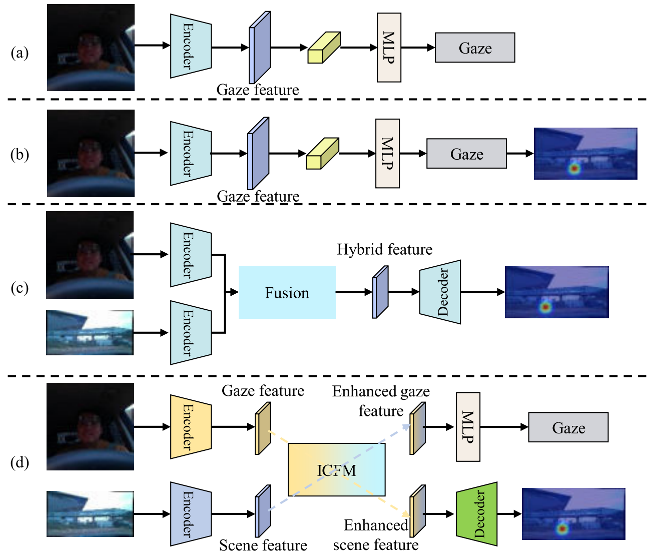


Fig. 1. Taxonomy of driver gaze estimation paradigms. (a) Single–frame gaze regression; (b) Gaze-projection to attention heatmaps; (c) Mixed guidance for attention feature extraction (gaze→scene); (d) Bidirectional guidance (gaze↔scene) adopted in this work to enforce local gaze–scene consistency.

typically focus on the unidirectional guidance of information extraction and fail to fully explore the bidirectional relationship between gaze and scene. Additionally, scene images typically have higher resolution and more complex information, which can introduce extra computational overhead. In in-vehicle systems, computational cost is a key concern, thus there is an urgent need to develop a new lightweight network architecture that balances accuracy and computational efficiency.

In this paper, we propose a novel framework, as shown in Fig. 1 (d), for estimating both the driver's gaze and attention maps. The gaze provides the precise visual attention direction of the driver, while the attention map offers a probabilistic estimate of the gaze area. Current research mainly focuses on using scene information to guide gaze estimation or using facial images to assist attention map generation, but the bidirectional relationship between the two has not been fully explored. To address this, we introduce a bidirectionally-guided lightweight network for attention estimation by leveraging the consistency between the scene and gaze. We design a lightweight feature extraction module that efficiently extracts gaze and scene features with minimal parameters. To facilitate information exchange between feature flows and model the complex cross-view relationships between the two data sources, we propose an Information Cross Fusion Module (ICFM). Based on a cross-attention mechanism, the ICFM uses a dual strategy of global modeling and local mapping to promote efficient information flow. Specifically, global feature representations are used to establish the geometric relationships between different views, while local mappings capture the relationships between gaze and local scene features. Scene saliency is used to compensate for missing gaze-related features in gaze representation extraction, and facial information is used to guide the generation of attention heatmaps. Additionally, we explore a novel direction by designing a unique ground truth for the attention trend heatmap, allowing the network to learn the ability to predict gaze trends from static facial and scene images. Specifically, our contributions are as follows:

- Based on the geometric consistency between scene and gaze, we propose a lightweight feature extraction framework that mitigates the performance drop caused by reduced parameter counts by introducing bidirectional projections. The proposed framework achieves a balance between efficiency and accuracy in attention estimation through multi-view information fusion.
- We propose an Information Cross-Fusion Module, using a dual strategy of global feature-based view modeling and local gaze mapping to facilitate efficient information exchange, thereby guiding the extraction of gaze and scene features.
- We define an attention trend heatmap to model gaze relationships across adjacent frames. The proposed bidirectional projection strategy captures gaze continuity and local scene similarity, enabling the model to infer a probabilistic distribution of attention.

II. Related Work

A. Eye Tracking

The appearance manifestation of attention change is the relative movement of the pupil position [28]. Therefore, eye tracking can be used to understand the direction of attention [29]. Eye tracking includes feature-based and appearance-based methods. For feature-based methods, gaze is usually modeled as a normal vector perpendicular to the eyeball surface with the pupil center as the origin [30]. Head-mounted devices [31], [32] are used to obtain appearance parameters. However, the working distance and limited robustness restrict the application scenarios. Appearance-based methods mainly involve inputting face or eye images into a neural network and regressing the gaze direction through a fully connected layer [33], [34]. In the early stage, people inputted eye images into a convolutional neural network to extract gaze features [19]. To further improve the accuracy of the model, additional prior information is injected into the learned gaze representation [35], [36]. To enhance the generalization ability of the model, researchers began to explore the implicit extraction of head pose from face images [37]. The rotation of the head and the displacement of the pupil can lead to the disappearance and deformation of the eyes, thereby leading to the prediction uncertainty of the model [38]. In scenarios with limited computing power, model parameters should be constrained [39]. Human subjective behavior and passive factors can cause changes in gaze. Active behavior mainly refers to the change of gaze target, while passive factors include movement and salient objects within the field of view. However, current methods have not fully explored the complex relationship between the scene and gaze.

B. Driver Attention Estimation

Driving is a complex task that involves not only controlling the movement of the vehicle but also continuously

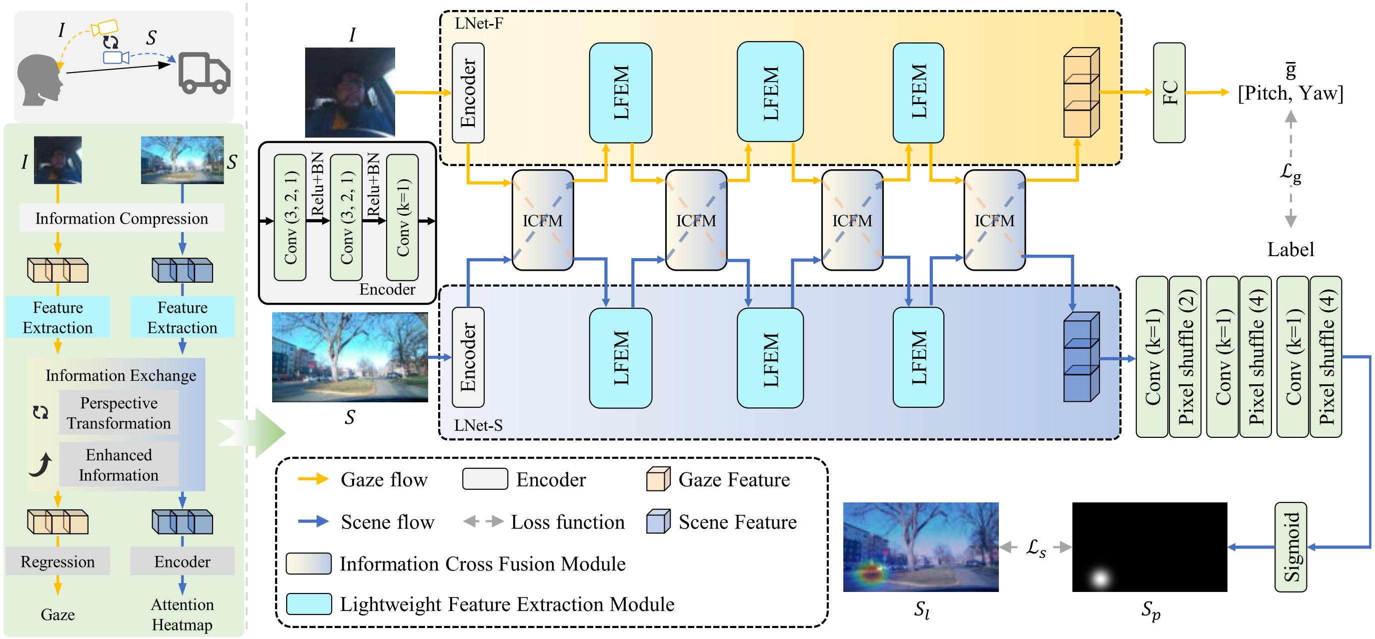


Fig. 2. The overview of our proposed method. The left subfigure illustrates the role of each module, while the right subfigure presents the detailed flow of information. LNet-F represents the gaze feature flow, and LNet-S is the scene. The model consists of a lightweight feature extraction module (LFEM), an information cross-fusion module (ICFM), and a dual-supervision head for gaze and probabilistic attention heatmap prediction. Each block is defined and color-coded for clarity, showing the data flow from face and scene inputs to final gaze estimation.

monitoring the surrounding environment [40], [41], [42]. Accurate tracking of driver attention affects road traffic safety [43], [44], [45]. In vehicles, the driver's gaze usually shows a clustering trend, mainly focusing on fixed zones such as the windshield and rearview mirror [21], [46], [47]. Therefore, previous works transform driver attention into gaze zone estimation, simplifying dense gaze vectors into sparse gaze zones to reduce the difficulty of annotation and prediction [22], [48]. However, this strategy leads to the model being unable to be directly applied to high-precision scenarios [49]. To obtain accurate gaze labels, Cheng et al. [24] proposed defining the gaze vector as a unit direction vector originating from the center of the face and pointing towards the target, and obtaining the relative coordinates in three-dimensional space through a depth camera. However, in the face of some extreme scenarios, such as lighting and occlusion leading to invisible pupils, the prediction accuracy drops significantly [50], [51]. Kasahara et al. [5] proposed that the driver's line of sight usually focuses on salient objects in the driving scene, and self-supervised learning by constructing the geometric consistency between gaze direction and scene saliency can improve the performance of the model. Zhou et al. [27] proposed that predicting the gaze direction from face images and projecting it as an attention heatmap is a two-step estimation method, and as the projection distance increases, the estimation error of gaze gradually amplifies. Therefore, an end-to-end mapping network is proposed, where face and scene images are directly input into the attention heatmap through feature extraction and information fusion. Accurate gaze estimation and probabilistic attention heatmaps have different application scenarios. The consistency between face and scene can complement each other and promote bidirectional guided feature extraction [52].

TABLE I

Feature Dimensions and Parameter Counts Per Block (Gaze & Scene branches). Yellow Represents the Gaze Feature Flow, and Blue Denotes the Scene Feature Flow

Block	Input	Output	#Param.
Encoder + ICFM	[3, 224, 224]	[24, 56, 56]	0.191M
	[3, 224, 448]	[24, 56, 112]	
2*LFEM + ICFM	[24, 56, 56]	[32, 28, 28]	0.205M
	[24, 56, 112]	[32, 28, 56]	
5*LFEM + ICFM	[32, 28, 28]	[64, 14, 14]	0.471M
	[32, 28, 56]	[64, 14, 28]	
3*LFEM + ICFM	[64, 14, 14]	[128, 7, 7]	0.940M
	[64, 14, 28]	[128, 7, 14]	
FC + Decoder	[128]	[2]	0.023M
	[128, 7, 14]	[1, 224, 448]	

In addition, under the limited computing power of the in-vehicle system, the model needs to balance accuracy and computational cost.

III. Proposed Method

In this section, we introduce LNet, a lightweight network for driver attention estimation via scene and gaze consistency, as depicted in Fig. 2. The network aims to use gaze direction and scene attention maps for bidirectional guidance. The saliency of the scene can be used to assist in facial gaze estimation, adapting to different levels of challenges such as prediction errors caused by head rotation and eye movement, thereby optimizing gaze feature representation. Facial images can be used to compensate for the lack of semantic information in scene feature extraction, such as the inconsistency between salient features and driver attention. The proposed model consists of four modules: (1) an encoder module that reduces the resolution of the input image; (2) The feature extraction module based on depthwise separable convolution; (3) The information cross fusion module that promotes bidirectional information exchange and guidance; (4) The prediction module. The input and output dimensions of the model are shown in Table I, and the detailed structure is described in the following subsections.

A. Encoder

The performance and latency of neural networks are often affected by the resolution of the input image. For gaze estimation tasks, detailed features are crucial. The coupling of pupil position and head posture determines the gaze direction, while low resolution can lead to blurred pupil position. To maintain performance, we follow the resolution limitations of existing methods for facial images (I), such as $I \in \mathbb{R}^{3\times224\times224}$. For the scene image ($S$), we set the input resolution to $S \in \mathbb{R}^{3\times224\times448}$. High resolution images have rich semantic information, but there are also some redundancies. Faced with the challenge of model parameters in the vehicle, as shown in Fig. 2, we propose to input face and scene images into an encoder, gradually reducing the spatial resolution of the images, thereby reducing the computational complexity of subsequent processing. Our proposed encoder reduces the resolution of the input image from $\mathbb{R}^{C\times H\times W}$ to $\mathbb{R}^{C\times\frac{H}{4}\times\frac{W}{4}}$ by stacking two convolutional layers. Finally, the channel is adjusted through a 1×1 convolutional layer to obtain high-level semantic features. This process can be formalized as follows:

$$F_f = \text{Encoder}(I) \tag{1}$$

$$F_s = \text{Encoder}(S) \tag{2}$$

where F_f and F_s are encoded facial and scene features.

B. Lightweight Feature Extraction Module

The Lightweight Feature Extraction Module (LFEM) is an integral part of our network architecture, as shown in Fig. 3, designed for efficient and effective feature extraction. This module is characterized by its compact design and the strategic use of residual connections to facilitate the flow of information through the network, enhancing feature propagation and training stability. Double branch depthwise separable convolutions (DWConv) are used for model lightweighting, while focusing on extracting both global and local information.

LFEM initiates with a convolutional layer with a kernel size of $k = 1$, denoted as Conv($k = 1$), and is used for linearly transform the input feature map to integrate information and adjust channels. This is followed by batch normalization (BN) and an activation function (Act):

$$F_{\text{int}} = \text{Act}(\text{BN}(\text{Conv}(F; k = 1))) \tag{3}$$

where F is the input feature, and F_{int} denotes the extracted feature.

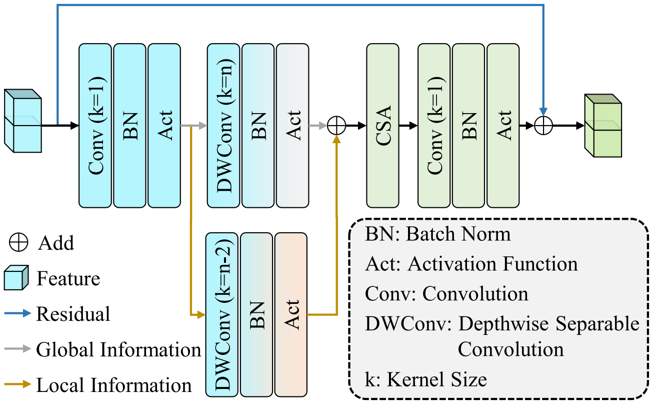


Fig. 3. LFEM details. Each block stacks depthwise separable convolutions and 1×1 projections with channel–spatial attention (CSA), performing stride-based resolution reduction while preserving local cues (e.g., pupils and lane markings).

Subsequently, a depthwise convolutional layer with a kernel size of $k = n$, DWConv($k = n$), is applied to capture spatial hierarchies across the input channels. This branch mainly extracts global information (F_g), therefore, larger kernel size ($n > 3$) is used for different network depths, as follows:

$$F_g = \text{Act}(\text{BN}(\text{DWConv}(F_{\text{int}}; k = n))) \tag{4}$$

To further refine the feature representation, another depthwise convolution with a reduced kernel size $k = n - 2$ is employed. This branch is used to extract local information (F_l), as follow:

$$F_l = \text{Act}(\text{BN}(\text{DWConv}(F_{\text{int}}; k = n - 2))) \tag{5}$$

The channel-spatial attention, CSA, consists of a series of channel and spatial attention, is integrated within this process to recalibrate the channel and spatial feature responses adaptively by emphasizing the most informative features and suppressing the less useful ones, further enhancing the representational power of the extracted features. The module concludes with a final convolutional layer with a kernel size of $k = 1$, to output the refined feature map (F_e):

$$F_e = \text{Act}(\text{BN}(\text{Conv}(\text{CSA}(F_g + F_l); k = 1))) \tag{6}$$

LFEM introduces residual connections, which directly link the output of early layers to subsequent layers. These connections aim to enable the network to learn more efficiently and reduce information loss again by retaining the information from the initial layer. The remaining connection formula in LFEM is as follows:

$$F_{out} = F_e + F \tag{7}$$

where F_{out} is the final output feature map.

This design makes LFEM not only lightweight, but also sturdy and durable, and reduces information loss, enabling it to effectively capture and optimize the features of each layer of the network.

C. Information Cross Fusion Module

Facial and scene images are fed into different branches, and features are extracted through stacked LFEM. In the real world, the driver's gaze direction is consistent with the attention map of the scene, in other words, the gaze direction can be projected to the scene. In order to utilize the implicit relationship between facial images and scenes, we propose an information cross fusion module that guides feature extraction in both directions. The input features extract global and local information through two asymmetric branches, respectively. The cross attention mechanism is used to calculate the similarity between the global information of the scene and the face, in order to obtain consistency scores, enhance the most discriminative dynamic clues, and improve feature representation. The following elaborates on the role of submodules.

Firstly, the incoming features are obtained through global average pooling (Avg) to obtain a global feature representation, and the information is reinforced through two convolutional layers. In order to limit the parameters of the module, the first layer of convolution reduces the number of input feature channels to $1/r$, and then restores the channels through a convolution layer. Specifically, as follows:

$$F^g_d = \mathrm{Act}(\mathrm{BN}(\mathrm{Conv}(\mathrm{Avg}(F^{\mathrm{input}}); k = 1, c/r))) \tag{8}$$

$$F^g_u = \mathrm{Conv}(F_d; k = 1, c) \tag{9}$$

where F^{input} is the input feature map, F^g_d is the downsampled feature, F^g_u denotes the recovered feature and also the global feature, and c represents the number of channels. In this paper, r is uniformly set to 2.

Due to average pooling, some detailed features may be lost. Therefore, similar operations are used to extract local information, with the difference being that the feature map is directly manipulated without average pooling. Specifically, as follows:

$$F^l_d = \mathrm{Act}(\mathrm{BN}(\mathrm{Conv}((F_{\mathrm{input}}); k = 3, c/r))) \tag{10}$$

$$F^l_u = \mathrm{Conv}(F_d; k = 1, c) \tag{11}$$

where F^l_d is the downsampled local feature and F^l_u denotes the local information.

In order to promote the interweaving and flow of scene and facial image information in the module, and effectively calculate the consistency score between the two, we propose a cross attention submodule based on global information. Considering the trade-off between computational efficiency and performance, we further compress the global information representation through MLP. For facial information flow ($F^g_f \in \mathbb{R}^{C\times 1\times 1}$), our operation can be formulated as follows:

$$Q_f = F^g_f \cdot W_Q \tag{12}$$

$$K_s = F^g_s \cdot W_K \tag{13}$$

$$V_s = F^g_s \cdot W_V \tag{14}$$

where W_Q, W_K, and W_V are linear projection matrices. Q_f is query, K_s is key, V_s is value, and F^g_s denotes the scene global feature. By calculating the dot product of the query and keys and applying the SoftMax function, we obtain an attention map $S_{f\rightarrow s}$, which encodes the similarity between face and scene. Specifically, as follows:

$$S_{f\rightarrow s} = \mathrm{SoftMax}\left(\frac{Q_f \cdot (K_s)^{\mathrm{T}}}{\sqrt{d_k}}\right) \tag{15}$$

where $\sqrt{d_k}$ is the scale coefficient. This attention map is used to aggregate similar information from facial and scene features, and fuse it with the current feature to guide information exchange, as shown below:

$$\hat{F}^f_g = F^f_g + \mathrm{MLP}(S_{f\rightarrow s} \cdot V_s) \tag{16}$$

where $\hat{F}^f_g$ denotes the facial feature representation refined under the guidance of scene information. The MLP adaptively maps the fused attention information back to the original feature space, aligning semantic dimensions and enhancing the target feature stream. Similarly, the enhanced scene feature $\hat{F}^s_g$ is obtained through the same procedure.

The extracted local information is added to the enhanced global features to preserve detailed information. Finally, sigmoid function is applied to output features to generate probability maps, thereby guiding feature selection. This process can be formalized as follows:

$$F^{\mathrm{out}}_f = F^{\mathrm{input}}_f \otimes \mathrm{Sigmoid}(\hat{F}^f_g + F^f_l) \tag{17}$$

$$F^{\mathrm{out}}_s = F^{\mathrm{input}}_s \otimes \mathrm{Sigmoid}(\hat{F}^s_g + F^s_l) \tag{18}$$

where F^{out}_f and F^{out}_s, F^{input}_f and F^{input}_s are the output and input features of face and scene. F^f_l and F^f_l denote the local information of face and scene feature.

D. Prediction and Loss Function

After gradual extraction and guidance, we obtained enhanced facial features and scene features. For facial features, gaze related feature representations are gradually strengthened based on the saliency guidance of the scene. Adaptive pooling is used to obtain gaze embedding vectors, while fully connected layers are used to regress gaze direction ($\hat{\mathrm{g}}$) from the embedding space. This branch aligns the gaze direction of the sample with the label (g) by minimizing the gaze angle loss ($\mathcal{L}_{\mathrm{g}}$), thereby guiding the model to learn gaze-related features. The calculation of $\mathcal{L}_{\mathrm{g}}$ is as follows:

$$\mathcal{L}_{\mathrm{g}} = ||\hat{\mathrm{g}} - \mathrm{g}||_1 \tag{19}$$

For the scene branch, we use an upsampling strategy to infer the attention map (S_p) from the feature space. By stacking convolutional layers and Pixel Shuffle layers to obtain a decoder, high-dimensional scene features are mapped to two-dimensional space. Following [5], this branch minimizes Kullback-Leibler ($\mathcal{L}_{KL}$) divergence and normalized cross-correlation ($\mathcal{L}_{NCC}$) to supervise the extraction of scene information flow. The specific calculation process is as follows:

$$\mathcal{L}_s = \mathcal{L}_{KL} + \lambda\mathcal{L}_{NCC} \tag{20}$$

where $\mathcal{L}_s$ is the attention map loss, and λ the weight to balance between $\mathcal{L}_{KL}$ and $\mathcal{L}_{NCC}$, set to 0.1. For the annotation S_l

and prediction S_p of the attention map, $\mathcal{L}_{KL}$ and $\mathcal{L}_{NCC}$ are formulated as follows:

$$\mathcal{L}_{\text{KL}} = \sum_{\mathbf{x}} S_l(\mathbf{x}) \log\left(\frac{S_l(\mathbf{x})}{S_p(\mathbf{x})}\right) \tag{21}$$

$$\mathcal{L}_{\text{NCC}} = 1 - \frac{\sum_{\mathbf{x}} S_l(\mathbf{x}) S_p(\mathbf{x})}{\sqrt{\sum_{\mathbf{x}} S_p(\mathbf{x})^2}\sqrt{\sum_{\mathbf{x}} S_l(\mathbf{x})^2}} \tag{22}$$

Our overall loss ($\mathcal{L}$) is then:

$$\mathcal{L} = \alpha_1 \mathcal{L}_{\text{g}} + \alpha_2 \mathcal{L}_s \tag{23}$$

where α_1 and α_2 are adjustment factors, which are set to 2 and 1.

IV. Experiments

A. Dataset

Look Both Ways (LBW) is the only publicly available dataset that includes driving scene videos, driver facial videos, and annotations that include ground truth values of driver gaze and scene attention maps. Therefore, our work is mainly based on the LBW dataset. It includes 123297 synchronized driver facial and stereoscopic scene images. Facial landmarks are used to crop facial images, with a fixed size of 224×224. The scene images are resized to 224×448. The dataset contains 28 subjects. We divided the dataset into two subsets based on the subjects, with subjects with IDs greater than 22 as the test set and the rest as the training set.

MPIIFaceGaze The dataset is based on the MPIIGaze dataset [53]. It was collected in the real world and contains 45K images obtained from 15 people. The method of [54] is employed for preprocessing. The normalized facial image is resized to 224×224. The leave-one-person-out method is used for evaluation.

EyeDiap includes the videos of 16 participants from different races and genders [55]. the method of [56] is used to preprocess. The normalized facial image is resized to 224×224. In the benchmark test, randomly divide the subjects into four groups and implement the leave-one-out strategy for evaluation.

B. Implementation Details

The proposed method is implemented by using Pytorch, and trained for 100 epochs on V100 GPU. A batch-size of 64 is set to train, and the learning rate is set as 0.0001. The learning rate is adjusted every 25 epochs with an adjustment coefficient is 0.5.

C. Evaluation Metrics

Angle error is used to evaluate model prediction accuracy, with lower values representing better performing methods [18]. In addition, following [24], the accuracy (% within error threshold) is defined where < $k°$ means an prediction is correct if the angular error is lower than $k°$. The proportion of correct sample size to the entire dataset is used as a measure.

For attention maps, Kullback-Leibler divergence (KL), Correlation Coefficient (CC), Similarity Index (SIM), and Normalized Scanpath Saliency (NSS) are selected as evaluation metrics. Lower KL, higher CC, SIM, and NSS indicate better performance. For S_p and S_l, the specific calculation formula is as follows:

$$KL(S_p, S_l) = \sum_i S_l{}^i * \log\left(\frac{S_l{}^i}{S_{p_i}}\right) \tag{24}$$

$$SIM(S_p, S_l) = \sum_i \min\left(\frac{S_p{}^i}{\sum_{i=1}^n S_p{}^i}, \frac{S_l{}^i}{\sum_{i=1}^n S_l{}^i}\right) \tag{25}$$

$$CC(S_p, S_l) = \frac{Cov(S_p, S_l)}{\sigma(S_p) * \sigma(S_l)} \tag{26}$$

$$NSS(S_p, S_l) = \frac{1}{N} * \sum_i \frac{S_p^i - \mu(S_p)}{\sigma(S_p)} \tag{27}$$

where $Conv$ denotes the covariance function, and σ and μ are the standard deviation and mean.

To quantitatively evaluate the proposed method on the attention trend prediction task, we adopt four metrics: Temporal Consistency (TC), Next-frame Prediction Error (NFPE-KL and NFPE-MSE), and Temporal Alignment Score (TAS). TC measures the smoothness of predicted results across adjacent frames, NFPE evaluates the accuracy of next-frame attention distribution prediction, and TAS assesses temporal alignment with flexible tolerance within ±Δ frames. These indicators can be formulated as follows:

$$\text{TC} = 1 - \frac{1}{T-1} \sum_{t=1}^{T-1} \|S_p^{t+1} - S_p^t\|_2^2 \tag{28}$$

$$\text{NFPE-KL} = \frac{1}{T-1} \sum_{t=1}^{T-1} KL(S_{\text{g}}^{t+1} \parallel S_p^t) \tag{29}$$

$$\text{NFPE-MSE} = \frac{1}{T-1} \sum_{t=1}^{T-1} \|S_{\text{g}}^{t+1} - S_p^t\|_2^2 \tag{30}$$

$$\text{TAS} = \frac{1}{T-1} \sum_{t=1}^{T-1} \max_{\tau \in [-\Delta, \Delta]} CC(S_{\text{g}}^{t+1}, S_p^{t+\tau}) \tag{31}$$

D. Ground Truth of Attention Map

We conduct experiments using three types of ground truth maps: the gaze-projected heatmap, the sequential fixation heatmap, and attention trend map. To ensure the fairness of the experiment, the calculation strategy of gaze-projected heatmap follows [5], and the preprocessing method of the sequential fixation heatmap follows [27].

To verify the additional role of attention maps, we adopted a method similar to that used in the Dr(eye)ve dataset [10] and defined a attention trend map. We selected adjacent frames in the time sliding window and projected the gaze of each frame onto the scene coordinates, and used a Gaussian function with fixed parameters to probabilistically model the gaze. Specifically, our heatmap can be formulated as follows:

$$\begin{aligned} S(x, y) = \frac{1}{2\pi\sigma^2} \Bigg(& \alpha \exp\left(-\frac{(x - x_t)^2 + (y - y_t)^2}{2\sigma^2}\right) \\ & + \beta \exp\left(-\frac{(x - x_{t+1})^2 + (y - y_{t+1})^2}{2\sigma^2}\right)\Bigg) \end{aligned} \tag{32}$$

where $S(x, y)$ represents the attention trend map, (x_t, y_t) represents the coordinates of the gaze projected onto the scene

TABLE II

QUANTITATIVE COMPARISON OF GAZE ESTIMATION PERFORMANCE ACROSS DIFFERENT METHODS. MEAN ERROR IS REPORTED IN DEGREES (LOWER IS BETTER). AVERAGE PRECISION IS COMPUTED UNDER MULTIPLE THRESHOLDS (<2°, <4°, <6°, <8°; HIGHER IS BETTER), PROVIDING A COMPREHENSIVE EVALUATION OF ACCURACY

Method	Input	Mean Error	Accuracy (% within error threshold)			
			$< 2°$	$< 4°$	$< 6°$	$< 8°$
FullFace [57]	F	6.54°	11.3%	30.8%	51.0%	70.9%
Gaze360 [34]	F	6.31°	11.7%	29.5%	53.6%	71.0%
ResNet18 [58]	F	6.07°	12.4%	37.2%	60.7%	77.2%
GazeTR [25]	F	5.91°	12.9%	36.9%	61.8%	77.9%
XGaze [26]	F	6.02°	12.8%	37.9%	62.5%	78.0%
GazePTR [24]	F	**5.83°**	13.2%	**39.1%**	**63.8%**	**79.1%**
LNet-F	F	6.02°	13.1%	37.7%	63.4%	78.5%
LBW [5]	F+S	6.05°	12.9%	37.5%	62.3%	78.3%
LNet	F+S	5.89°	**13.5%**	39.0%	63.3%	78.8%

image in frame t, (x_{t+1}, y_{t+1}) is frame $t+1$, and the projection formula follows [5]. σ denotes the standard deviation of the Gaussian function and is set to 30. α and β represent weight parameters used to balance the probability distribution of gaze projection. α is set to 0.7, and β is 0.3. In our implementation, $\alpha = 0.7$ and $\beta = 0.3$ were empirically selected to balance stability and predictive capability. Larger β shifts high-attention regions toward the future frame, potentially breaking geometric consistency, whereas larger α suppresses trend prediction.

E. Gaze Error

Table II presents the quantitative test results of the proposed method for gaze estimation. Firstly, for the input of only face images, GazePTR achieves the state-of-the-art prediction performance, while LNet-F has comparable performance with XGaze based on ResNet50. In this paper, we assume that the saliency of the scene image can assist in gaze estimation. Therefore, we compared the results of inputting both scene and face images. LNet achieved an improvement in accuracy when inputting both types of information. Compared with LBW, the error of LNet is reduced by 0.16°. Under different average precision requirements, an improvement in accuracy is obtained. Compared with GazePTR, LNet achieved matching accuracy with a significant reduction in model parameters (**Subsection VI-Q**). The experimental results confirm that the consistency between the scene image and face gaze can be utilized to enhance gaze accuracy.

F. Quantitative for Attention Map

Table III shows the results of attention map estimation for scene images. Following the metric definitions in [27], we select KL, CC, SIM, and NSS as evaluation metrics, and use the sequential fixation heatmap as the ground truth for attention maps. Compared with EraW-Net with mixed input, LNet achieves breakthroughs in performance on three metrics, which validates the effectiveness of the proposed method in the task of attention map estimation.

TABLE III

COMPARISON EXPERIMENTS USING SEQUENTIAL FIXATION HEATMAPS AS GROUND TRUTH. THE PROPOSED LNET ACHIEVES COMPETITIVE ACCURACY IN ATTENTION HEATMAP ESTIMATION AT SIGNIFICANTLY REDUCED COMPUTATIONAL COST

Method	Input	KL ↓	CC ↑	SIM ↑	NSS ↑
FullFace [57]	F	3.03	0.33	0.25	2.63
GazeTR [25]	F	3.11	0.35	0.26	2.74
XGaze [26]	F	3.22	0.32	0.24	2.50
GazePTR [24]	F	3.03	0.35	0.26	2.73
LBW [5]	F+S	3.01	0.33	0.25	2.63
EraW-Net [27]	F+S	2.41	0.35	0.24	2.79
LNet	F+S	2.11	0.42	0.32	2.85

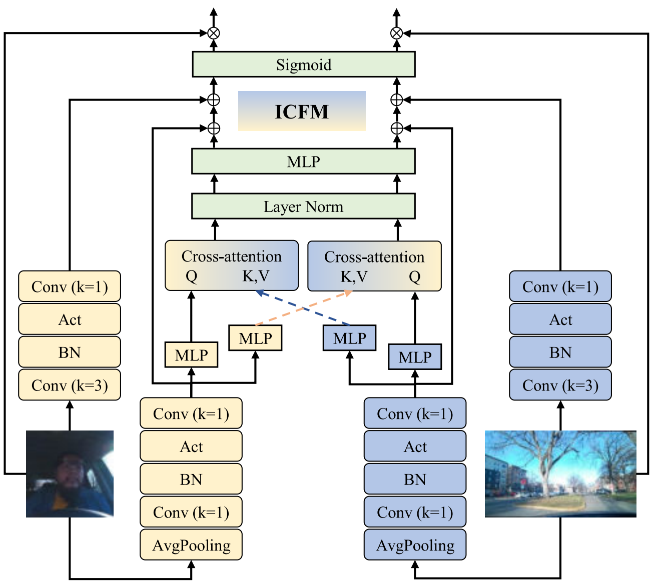


Fig. 4. ICFM details. Cross-modal attention aligns global camera geometry and local gaze cues. The adaptation MLP transforms fused attention to match and augment the original feature streams, yielding $\hat{F}_g^f$ and $\hat{F}_g^s$.

G. Qualitative for Attention Map

In addition to quantitative metrics, Fig. 5 presents the qualitative results of the proposed method in the task of attention map estimation. For the test set, we selected results with different drivers, road conditions, facial occlusions, and gaze directions for display. Firstly, in the challenges of different road conditions and different drivers, the performance of LNet is stable. For the case of facial occlusion (bottom left), the prediction of the attention map remains accurate. The second and fourth columns in the bottom right subplot represent the situations where the pupils are invisible due to head rotation and glasses occlusion, respectively, and the driver's attention is captured.

H. Feature Distribution Analysis

To further investigate why LNet can maintain stable performance even with reduced parameter quantities, we visualize the gaze features extracted under different configurations, as shown in Fig. 6. The experiment is divided into four groups: GazePTR, LNet-F, w.o. ICFM, and LNet. Here, *w.o. ICFM*

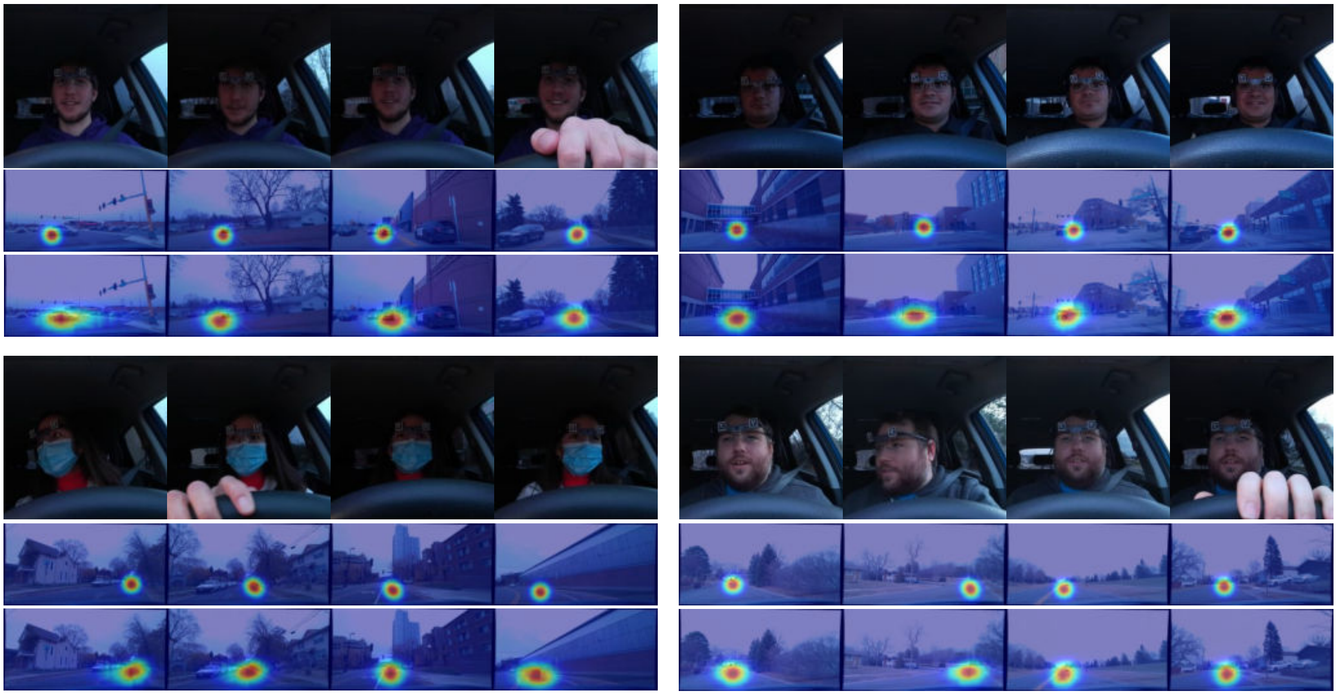
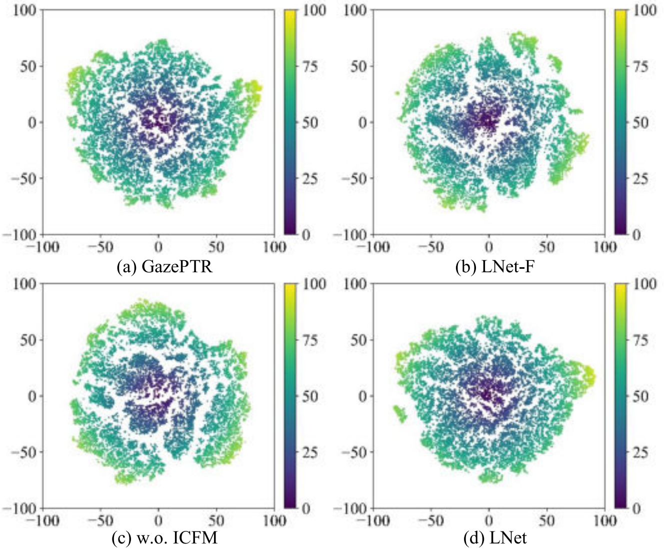


Fig. 5. Qualitative for Attention Map. The first row represents the input face image, the second row is the ground truth, and the last row denotes the prediction result.



(a) GazePTR

(b) LNet-F

(c) w.o. ICFM

(d) LNet

Fig. 6. t-SNE visualization of gaze feature embedding: Naive fusion (without ICFM) and LNet-F suffer from uneven feature distribution due to insufficient representation capabilities. LNet guides feature extraction step by step through gaze-scene geometric consistency, significantly reducing the number of parameters while maintaining performance.

denotes the variant where gaze and scene features are directly concatenated and regressed into gaze direction without the proposed Information Cross Fusion Module.

For regression tasks, where labels are continuous, the ideal feature distribution should also exhibit smooth continuity. The features extracted by GazePTR display such global continuity, consistent with its strong performance. In contrast, LNet-F, due to the significant reduction in parameter quantities, suffers from weaker representation ability, resulting in fragmented and less coherent feature distributions. In the case of *w.o. ICFM*, the use of only scene information leads to localized clustering of features, which negatively impacts generalization.

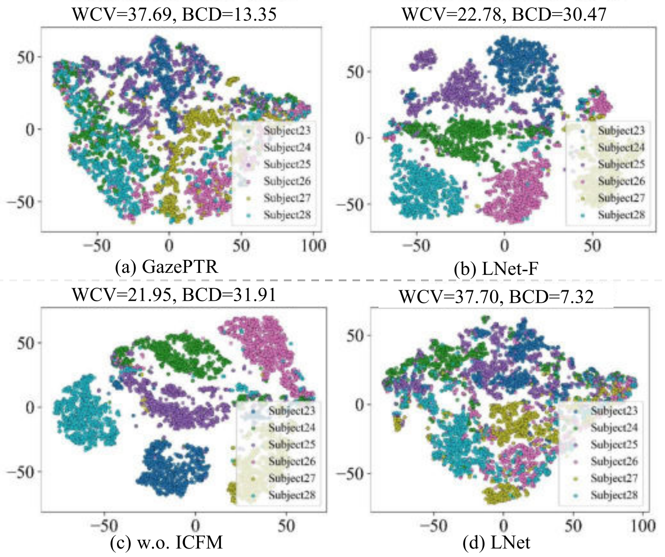
Compared with LNet-F and *w.o. ICFM*, LNet achieves the most reasonable feature distribution, closely approximating that of GazePTR. This change in feature continuity demonstrates that ICFM effectively leverages scene information to guide the learning of gaze features, thereby compensating for the lightweight design of LNet-F. Overall, this analysis highlights the critical role of ICFM in ensuring both compactness and discriminative power in the learned representations.

I. Analysis of Feature Distribution

To investigate the effect of scene information on gaze feature extraction, we visualized gaze features by randomly sampling 1,000 instances from the subjects and applying t-SNE for dimensionality reduction, as shown in Fig. 7. Prior studies [59], [60] indicate that identity information is unrelated to gaze and can impair model generalization. Given the consistent data acquisition protocol, gaze distributions across subjects are similar, and as reported in [61], gaze features should align with label distributions. In theory, gaze-related features should be identity-independent and uniformly distributed in the feature space. Accordingly, we evaluate feature dispersion using two metrics: between-class distance (BCD) and within-class variance (WCV). The calculation is as follows:

$$\mathrm{WCV} = \frac{1}{n}\sum_{k=1}^{K}\sum_{i:y_i=k}\|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2 \quad \text{with} \quad \boldsymbol{\mu}_k = \frac{1}{n_k}\sum_{i:y_i=k}\mathbf{x}_i \tag{33}$$

$$\mathrm{BCD} = \min_{1\le k<l\le K}\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_l\| \tag{34}$$



(a) GazePTR (b) LNet-F (c) w.o. ICFM (d) LNet

Fig. 7. t-SNE visualization of extracted gaze features. Different colors represent different drivers, showing how identity-related information clusters in feature space. A well-generalized gaze estimation model should disentangle gaze-related features from identity, resulting in dispersed distributions rather than tightly clustered identity features. WCV represents the within-class variance, and BCD represents the between-class distance.

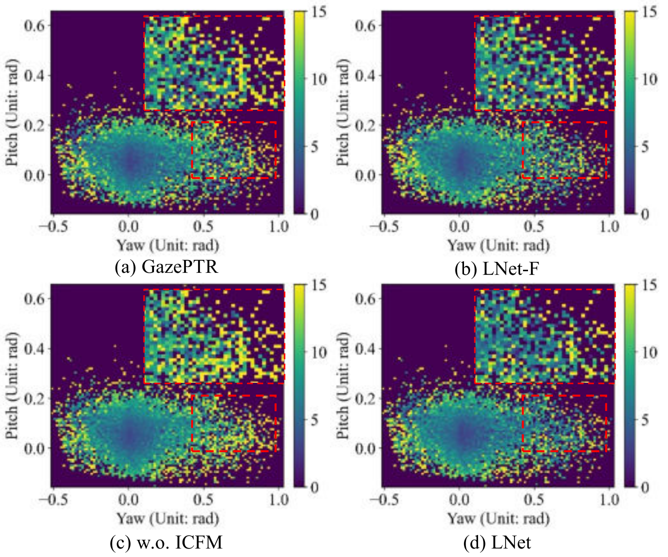
where $\mathbf{x}_i$ represents the two-dimensional feature points, y_i is the label, K denotes the number of subjects, $\boldsymbol{\mu}_k$ represents the class center, and n_k is the number of samples for each subject.

Results show that GazePTR produces highly mixed gaze features, while LNet-F exhibits pronounced subject-specific clustering (lower WCV, higher BCD) due to limited representational capacity; directly introducing scene features does not mitigate this trend. In contrast, LNet achieves WCV comparable to GazePTR but with lower BCD, indicating reduced identity leakage. These results suggest that the bidirectional guidance strategy in ICFM enables more effective extraction of gaze-related discriminative features, alleviating the representational degradation caused by parameter reduction.

J. Error Analysis at Different Gaze Angles

Fig. 8 reports the average estimation errors of different models across varying gaze angles. When the gaze angle becomes large (highlighted by the red dashed box), drivers usually rotate their heads to shift their field of view. This head rotation often causes the pupils to either disappear from the camera's perspective or become geometrically deformed, thereby reducing the accuracy of gaze estimation. The experimental results confirm this observation, as most models exhibit higher errors under large head poses.

Nevertheless, we observe that LNet consistently outperforms GazePTR in such challenging scenarios. A plausible explanation is that LNet leverages salient scene information to compensate for the degraded quality of ocular cues, thus maintaining robust performance. This finding highlights the importance of integrating scene information as an auxiliary signal to enhance gaze estimation accuracy under extreme conditions.



(a) GazePTR (b) LNet-F (c) w.o. ICFM (d) LNet

Fig. 8. Average gaze estimation errors at different gaze angles. The red dashed box indicates the challenging case of large head rotations, where ocular features are degraded. LNet shows superior robustness compared to GazePTR, suggesting that scene saliency effectively compensates for pupil deformation.

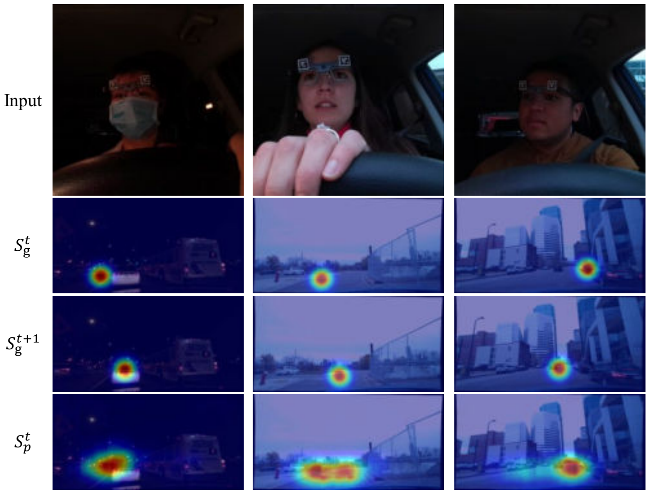


Fig. 9. Attention trend heatmap prediction results. The first row is the input t-frame facial image, where S_{g}^{t} represents the t-frame gaze projection heatmap, S_{g}^{t+1} is $t+1$ frame, and S_{p}^{t} represents the predicted attention trend map.

K. Attention Trend Prediction

In addition to serving as supervision for enforcing the consistency between scene and gaze, the attention map can also be exploited to predict the direction of future gaze. This relies on the assumption that human gaze direction evolves continuously in space and time due to physiological constraints of eye movements. Furthermore, gaze shifts are often driven by salient elements in the scene (e.g., pedestrians, traffic lights, vehicles), which provide contextual cues for predicting future attention targets.

Motivated by these observations, we modified the annotation of sequential fixation heatmaps by redefining the pair as the current frame and the subsequent frame. The experimental results are shown in Fig. 9, where the face image of frame

TABLE IV

ATTENTION TREND PREDICTION: TC (↓), NFPE-KL (↓), NFPE-MSE (↓), AND TAS WITH ±Δ-FRAME ELASTIC ALIGNMENT (↑). COMPARED TO GAZEPTR (GAZE-ONLY PROJECTION), LNET (BIDIRECTIONAL PROJECTION) SIGNIFICANTLY IMPROVES BOTH THE ACCURACY AND STABILITY OF TEMPORAL GAZE PREDICTION, DEMONSTRATING THAT SCENE ELEMENTS ENCODE CUES THAT IMPLICITLY SIGNAL UPCOMING GAZE SHIFTS

	TC ↓	NFPE-KL ↓	NFPE-MSE ↓	TAS ↑
GazePTR	10.36	15.29	2.07	0.49
Our	2.15	5.10	1.04	1.01

TABLE V

SCALABILITY OF LNET-F ON MPIIFACEGAZE AND EYEDIAP. PERFORMANCE VS. GAZETR WITH FLOPS (G) AND PARAMS (M)

	MPIIFacaGaze	EyeDiap	FLOPs	#Param.
GazeTR	4.02	5.26	1.83G	11.39M
LNet-F+TR	4.10 (↑ 1.99%)	5.63 (↑ 7.03%)	0.072G (↓ 96.07%)	0.36M (↓ 96.84%)

t is used as input, and the network outputs the attention map. It can be observed that the predicted attention map (S_p^t) not only captures the current gaze position but also encodes the tendency toward future gaze shifts.

These results demonstrate that, compared with independently modeling gaze direction or attention maps, a multi-task learning paradigm that leverages both facial dynamics and scene saliency provides richer supervisory signals. This enables the model to anticipate gaze trends by combining the continuity of eye movement with contextual drivers present in the scene.

L. Attention Trend Prediction Error

To quantitatively analyze the performance of the model, we use four indicators to evaluate the smoothness, accuracy, and robustness of the model in predicting attention trends. The results are shown in Table IV. Experimental results show that our method consistently outperforms GazePTR in all key temporal metrics: TC is reduced from 10.36 to 2.15, indicating smoother and more continuous temporal changes; NFPE-KL and NFPE-MSE decrease by 66.6% and 49.8%, respectively, reflecting more accurate next-frame predictions and improved temporal consistency; TAS increases from 0.49 to 1.01, demonstrating better alignment between predicted and ground-truth trends over time. These gains are likely due to the enhanced modeling of temporal dependencies between gaze and scene, particularly through fine-grained features such as pupil positions and scene elements. The global–local information allocation mechanism in ICFM strengthens the mapping between fine-grained features and global semantics, thereby substantially improving both the accuracy and stability of temporal predictions.

M. Scalable Capability

To further analyze the performance of the feature extraction module, we evaluate the scalable capability of LNet-F. As shown in Table V, LNet-F is concatenated with a Transformer and compared with GazeTR under the same framework. On the MPIIFaceGaze and EyeDiap datasets, the model parameters are reduced by 96.84%, while accuracy shows only a 1.99% and 7.03% trade-off, respectively. These results demonstrate that the proposed framework achieves a substantial improvement in efficiency while maintaining competitive accuracy.

Beyond the numerical improvements, two key implications are highlighted. First, the significant reduction in FLOPs and parameters indicates that LNet-F can be seamlessly deployed in computationally constrained environments, such as embedded in-vehicle systems, where hardware resources are typically limited. In such scenarios, achieving comparable accuracy at a fraction of the computational cost is critical for real-time applications. Second, the scalability of LNet-F shows that its feature extractor is architecture-agnostic, meaning it can be flexibly integrated into different backbone models or combined with emerging architectures while preserving efficiency advantages.

It is worth noting that the higher error observed on EyeDiap may be related to lower image acquisition quality and larger head pose variations. This suggests a potential trade-off between efficiency and robustness under varying resolution and viewpoint conditions. Consequently, future work could explore **hybrid strategies**, such as dynamically loading models of different scales according to available onboard computational power, to better balance scalability and accuracy in diverse driving scenarios.

N. Gaze-Guided Attention Map

To further investigate the role of facial images in guiding attention and predicting future gaze behavior, we extend our analysis to temporal sequences. Fig. 10 illustrates the evolution of attention maps from frame t to the subsequent three frames $(t+1, t+2, t+3)$.

In the first two columns, when approaching an intersection, LNet leverages facial cues and scene elements to predict the next-frame attention. Due to the vehicle's orientation bias, the high-probability regions of the attention map initially shift, indicating that gaze trend prediction may be disturbed by visually similar scene elements. As the vehicle moves forward and updated facial inputs are incorporated, the attention heatmap gradually converges toward the correct driving direction. This highlights the temporal continuity of gaze embedded in facial dynamics. In the third column, low-probability regions progressively diminish as irrelevant scene elements recede, further demonstrating the importance of scene context in refining gaze trend predictions.

When multiple salient objects are present, LNet exhibits consistent behavior: the generated attention maps remain focused on task-relevant areas, while the predicted future maps progressively refine toward the intended gaze target. This validates two key aspects: (1) facial features serve as an effective prior for attention map generation, and (2) scene elements provide complementary cues that enhance gaze trend prediction.

Overall, this experiment confirms the effectiveness of combining facial and scene information to generate temporally

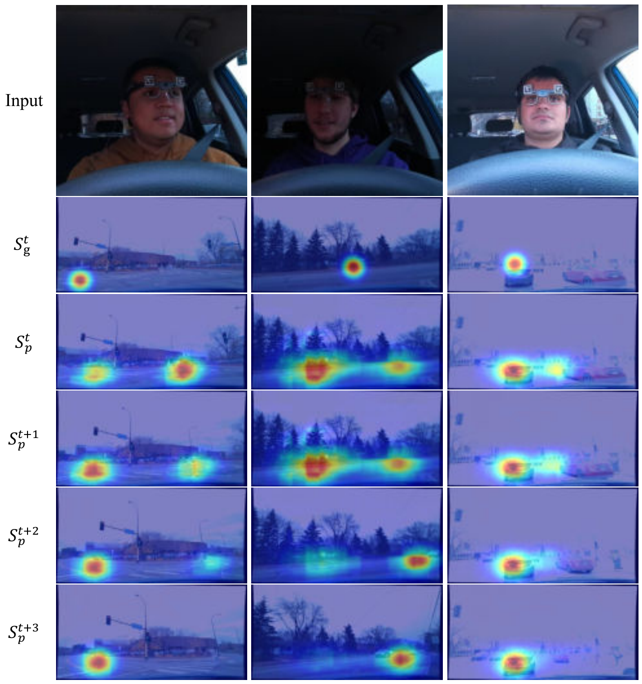


Fig. 10. Generation of gaze guided probability attention maps. The first row is the input t-frame facial image, and S_g^t represents the t-frame gaze projection heatmap. S_p^t is the predicted result of t frame, and S_p^{t+1}, S_p^{t+2} and S_p^{t+3} denote the frame $t+1$, $t+2$, and $t+3$. The predicted future attention maps by LNet progressively refine toward the intended gaze target.

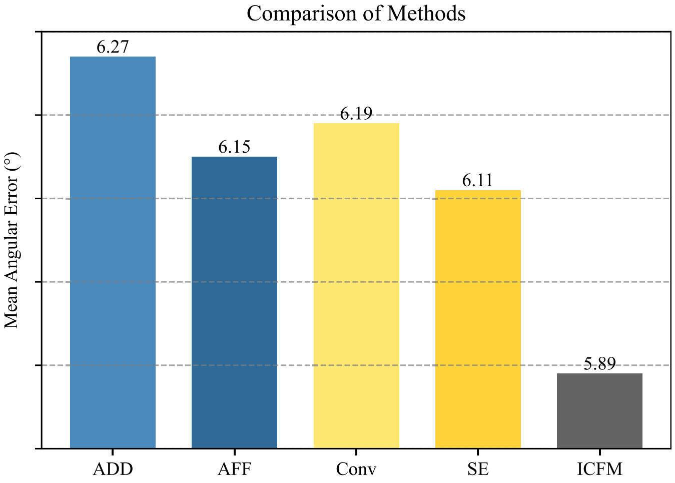


Fig. 11. Comparison for feature fusion. By fully exploiting the information latent in perspective transformation, the ICFM enhances overall model performance and facilitates bidirectional information flow.

TABLE VI

Fusion Strategy Comparison. Naive Concatenation or Plain Cross-Attention Underutilizes Heterogeneous Cues, Whereas ICFM's global–local Bidirectional Design Better Aligns Gaze With Scene Targets and Stabilizes Temporal Trends

	GazeError ↓	KL ↓	TC ↓	NFPE-KL ↓
CrossAttention	6.01	2.42	4.16	12.53
ICFM	5.89	2.11	2.15	5.10

coherent attention maps, thereby improving both predictive accuracy and robustness in dynamic driving environments.

O. Comparison for Feature Fusion

This subsection evaluates the effectiveness of the proposed Information Cross-Fusion Module (ICFM) in promoting bidirectional information fusion between gaze and scene features. To this end, four control groups are considered: ADD, AFF [62], Conv, and SE [63]. Specifically, **ADD** denotes the direct summation of gaze and scene features, while **Conv** indicates feature concatenation followed by a 1×1 convolution. **AFF** is a multi-scale channel attention-based fusion strategy, and **SE** employs channel attention for feature reweighting.

The experimental results are presented in Fig. 11. Both ADD and Conv perform poorly, as such simple fusion operations cannot capture the complex geometric transformations between gaze and scene modalities. In contrast, ICFM achieves the best results, surpassing SE (the second-best method) by 3.6%. This demonstrates that ICFM not only facilitates more effective bidirectional guidance but also significantly improves the overall accuracy of gaze estimation. These results confirm the advantage of explicitly modeling consistency between gaze and scene features through the proposed ICFM.

P. Information Exchange

To clarify the contribution of ICFM, we conducted an ablation study in which the ICFM module in LNet was replaced with a CrossAttention module, as shown in Table VI. Specifically, gaze and scene features were average-pooled and fed into CrossAttention to model multi-view geometric relationships. Results show that ICFM consistently improves performance across all four metrics, which we attribute to its multi-branch design. Based on an in-depth analysis of gaze behavior, we argue that geometric variations from camera viewpoints should be captured by global features, while the consistency between gaze and scene targets depends on local features such as pupil position and the attended object. Notably, ICFM reduces the TC metric by 48.32%, indicating a substantial enhancement in gaze continuity between adjacent frames. Furthermore, as shown in Table IV, gaze projection introduces continuity fluctuations, and CrossAttention performs between ICFM and GazePTR, further demonstrating that ICFM facilitates more accurate information flow.

Q. Ablation Study

To evaluate the contribution of each proposed component, we conduct an ablation study, with results shown in Table VII. Compared with GazePTR, LNet-F (face branch only) significantly reduces parameters and FLOPs, albeit with a small accuracy drop, highlighting the efficiency advantage. However, naïvely adding scene features without dedicated fusion (LNet-F+Scene) further degrades performance from 6.02° to 6.27°. This indicates that direct concatenation of heterogeneous information sources is not only ineffective but can be harmful, possibly due to the offset between gaze and scene saliency.

TABLE VII

Ablation Study of Different LNet Components. Gaze Error (°; Lower Is better), FLOPs (Lower Is better), and Parameter Counts (Lower Is Better) Are Reported. Results Show That Naive Fusion of Scene Features Degrades Performance, While the Proposed ICFM Enables Effective Cross-Modal Integration and Achieves the Best Trade-off Between Accuracy and Efficiency

	Input	Gaze Error	FLOPs	#Param.
GazePTR	F	5.83	1.88G	12.09M
LNet-F	F	6.02	0.063G	0.46M
LNet-F+Scene	F+S	6.27	0.19G	0.92M
LNet-F+Scene+ICFM	F+S	5.95	0.207G	1.81M
LNet	F+S	5.89	0.21G	1.83M

TABLE VIII

Effect of Resolution Compression Strategies. ✥ Represents Direct Input at Low Resolution, and ✓ Denotes Using Encoder Compression. Encoder-Based Compression Preserves Fine-Grained Cues (Pupil, Lane Markings) Better Than Direct Low-Resolution Inputs, Improving Gaze Error and Temporal Metrics

Face	Scene	GazeError ↓	KL ↓	TC ↓	NFPE-KL ↓
×	×	6.59	3.16	7.35	21.16
✓	×	6.15	2.25	2.76	9.77
×	✓	6.25	2.73	3.10	7.14
✓	✓	5.89	2.11	2.15	5.10

The proposed Information Cross-Fusion Module (ICFM) addresses this issue by enabling bi-directional guidance between gaze and scene branches, ensuring semantic and geometric consistency. Incorporating ICFM (LNet-F+Scene+ICFM) recovers the lost performance and improves accuracy to 5.95°, demonstrating its ability to facilitate meaningful cross-modal interaction. The complete LNet, supervised jointly by gaze and attention heatmaps, achieves 5.89°, matching the accuracy of GazePTR while using only $\frac{1}{7}$ of the parameters and $\frac{1}{9}$ of the FLOPs. These results validate that effective fusion of heterogeneous features is critical, and that ICFM plays a central role in achieving the balance between efficiency and accuracy.

R. Face Encoder

Table VIII compares the effects of encoder-based compression and direct low-resolution input for both face and scene streams. Results show that encoder compression outperforms direct low-resolution input across all metrics, with the best performance achieved when both inputs use encoder compression (GazeError reduced from 6.59° to 5.89°, KL from 3.16 to 2.11, along with notable reductions in TC and NFPE-KL). For face images, gaze estimation relies on the tightly coupled relationship between pupil position and head pose, where pupil position is a fine-grained feature that is easily lost at lower resolutions, leading to higher gaze errors. For scene images, driver attention is often influenced by salient objects such as buildings, lane markings, and traffic lights, where certain geometric cues (e.g., lane markings) may disappear at low resolution, impairing geometric consistency modeling. Encoder compression helps preserve such fine details during downsampling, thereby improving performance. Moreover, maintaining high resolution in either the face or scene stream notably reduces TC and NFPE-KL, indicating the complementary potential of gaze and scene features for temporal information modeling.

S. Latency Analysis

To simulate computation-constrained in-vehicle systems, Table IX reports the training and inference times, along with the corresponding inference FPS, for different configurations on two platforms (with or without GPU). Although LNet achieves substantial reductions in FLOPs and parameters, its runtime improvements are not uniform. In the single-input (F) mode, LNet-F delivers significant speedups over GazePTR, achieving up to 58.38 FPS on the GTX 960M. However, in the dual-input (F+S) mode, training time increases markedly (e.g., from 135.57 ms to 264.51 ms on the GTX 960M), while inference latency still shows moderate improvements (e.g., 14.01 FPS vs. 16.57 FPS for GazePTR and LNet, respectively). As noted in [64], such discrepancies may arise from limited backend optimization for depthwise-separable convolutions, which can increase memory access costs and kernel launch frequency. For inference latency and FPS, our method consistently outperforms the baseline on both platforms, especially in the single-input mode. These results further highlight that theoretical reductions in computational complexity do not necessarily yield proportional wall-clock acceleration or FPS gains, though they constrain the upper bound of inference latency. In practical in-vehicle systems, hardware-level optimizations are still required to fully approach the theoretical computational efficiency.

V. Discussion

A. Limitations

The validation of LNet primarily relies on the LBW dataset, which is currently the only publicly available high-quality dataset providing synchronized scene and facial videos for gaze-related research. This single-source dependency limits the comprehensiveness of the model's generalization evaluation. Without synchronized scene data, it is difficult to directly assess the model's robustness in more complex or diverse driving environments. Moreover, the fixed capture settings and vehicle configurations of LBW may cause performance degradation when transferring the model across devices or environments. To address these limitations, our future work will proceed in three directions: (1) exploring semi-supervised and self-supervised learning approaches to leverage large-scale, weakly annotated or single-modality data, enhancing generalization through cross-modal consistency constraints or pseudo-label generation; (2) developing synthetic data strategies using generative models to learn geometric mappings from gaze to scene, enabling realistic scene image synthesis from existing gaze datasets; and (3) collecting more real-world driving data under varied lighting, weather, and

TABLE IX

Runtime and FPS on Compute-Constrained Platforms: GTX 960M and Intel i5-6300HQ. Training Vs. Inference Latency. Although Parameter Count Is Markedly Reduced, Inference Latency Does Not Scale Commensurately; Additional Hardware-Level Optimizations Are Required to Fully Realize the Theoretical Computational Gains

	Input	GTX 960M (2G)			Intel i5-6300HQ		
		Training (ms)	Inference (ms)	FPS	Training (ms)	Inference (ms)	FPS
GazePTR	F	91.63	30.80	32.47	374.32	101.03	9.89
LNet-F	F	69.62	17.13	58.38	121.59	32.21	31.04
GazePTR	F+S	135.57	71.37	14.01	543.33	248.92	4.02
LNet	F+S	264.51	60.34	16.57	485.61	193.09	5.18

road conditions to strengthen robustness during training. In addition, although LNet achieves significant reductions in parameters and computational cost, the current implementation is not fully optimized for specific hardware, leading to a gap between theoretical complexity reduction and actual speedup. Future deployment will require hardware-level optimizations to approach theoretical performance. Despite these limitations, such challenges also open up rich opportunities for future research and technological innovation, paving the way for LNet to demonstrate its potential in a broader range of applications.

B. Applications

The superior performance of LNet in driver attention modeling and prediction suggests strong potential for integration into various in-vehicle intelligent systems. In driver monitoring systems (DMS), LNet can capture real-time gaze trajectories and attention distributions, linking them with scene context and, in combination with object detection, enabling accurate detection of distraction, fatigue, or risky behaviors. In assisted and autonomous driving scenarios, the model's next-frame gaze prediction capability can be leveraged for intention inference and interaction optimization in shared-control driving modes, enhancing both safety and comfort. Furthermore, the lightweight design and low computational complexity of LNet allow deployment on resource-constrained embedded hardware (e.g., automotive MCUs or low-power GPUs), with the possibility of achieving theoretical real-time performance through hardware optimization. Importantly, the bidirectional guidance mechanism in LNet is not limited to driving contexts and can be extended to other gaze-related tasks, such as cross-camera attention prediction in surveillance videos, where geometric consistency between pedestrian gaze and target location is critical. In addition, most existing gaze estimation methods lack depth modeling capabilities, which is particularly problematic in driving scenarios where pedestrians and buildings overlap along the line of sight. By incorporating depth information, LNet's bidirectional guidance strategy can be extended to accurately distinguish between multiple objects along the gaze vector, based on the physical assumption that binocular gaze vectors converge at a single target point in 3D space.

VI. Conclusion

In this paper, we propose a lightweight framework (LNet) designed to capture drivers' attention through bidirectional guidance from gaze and attention maps while controlling computational overhead. We introduce a lightweight feature extraction module aimed at efficiently extracting both gaze and scene features. Additionally, based on the consistency principle between gaze and scene, we propose an Information Cross Fusion Module (ICFM), which facilitates the exchange of information between the scene and gaze streams, implicitly modeling the complex cross-view relationships between them. This enables bidirectional guidance during the feature extraction process. Experimental results show that the incorporation of scene information not only smooths and stabilizes the extracted gaze features but also reduces the interference of identity information on the model's generalization ability. To explore the role of attention maps and verify that they are not merely projections of gaze, we introduce the concept of gaze trends. By designing the ground truth calculation for attention heatmaps, LNet is capable of predicting future gaze trends. Importantly, our results reveal that gaze trends are jointly influenced by both facial cues and scene elements, enabling the model to anticipate upcoming gaze shifts in advance. This predictive ability further enhances monitoring efficiency in real-world driving scenarios. Experimental results demonstrate that LNet significantly improves computational efficiency while maintaining negligible statistical differences in accuracy, making it a viable candidate for deployment in vehicular systems. In the future, we will continue to explore the application potential of the proposed method in multi-source information fusion scenarios, such as modeling the consistency between gaze focus and depth information (RGBD and LiDAR point clouds).

References

[1] T. A. Dingus et al., "Driver crash risk factors and prevalence evaluation using naturalistic driving data," *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 10, pp. 2636–2641, Mar. 2016.

[2] Y. Ma, R. Du, A. Abdelraouf, K. Han, R. Gupta, and Z. Wang, "Driver digital twin for online recognition of distracted driving behaviors," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 2, pp. 3168–3180, Feb. 2024.

[3] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *IEEE Access*, vol. 8, pp. 58443–58469, 2020.

[4] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger, and H. Li, "End-to-end autonomous driving: Challenges and frontiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 10164–10183, Dec. 2024.

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

[6] P. Papantoniou, E. Papadimitriou, and G. Yannis, "Review of driving performance parameters critical for distracted driving research," *Transp. Res. Proc.*, vol. 25, pp. 1796–1805, May 2017.

[7] Y. Du, X. Liu, Y. Yi, and K. Wei, "Incorporating bidirectional feature pyramid network and lightweight network: A YOLOv5-GBC distracted driving behavior detection model," *Neural Comput. Appl.*, vol. 36, no. 17, pp. 9903–9917, Jun. 2024.

[8] J. Xu, S. H. Park, X. Zhang, and J. Hu, "The improvement of road driving safety guided by visual inattentional blindness," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 4972–4981, Jun. 2022.

[9] 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.

[10] S. Alletto, A. Palazzi, F. Solera, S. Calderara, and R. Cucchiara, "DR (eye) VE: A dataset for attention-based tasks with applications to autonomous and assisted driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 54–60.

[11] D. Viet Thanh Nguyen, A. Tran, H. Nam Vu, C. Pham, and M. Hoai, "Driver attention tracking and analysis," 2024, *arXiv:2404.07122*.

[12] P. K. Sharma and P. Chakraborty, "A review of driver gaze estimation and application in gaze behavior understanding," *Eng. Appl. Artif. Intell.*, vol. 133, Jul. 2024, Art. no. 108117.

[13] R. Valenti, N. Sebe, and T. Gevers, "Combining head pose and eye location information for gaze estimation," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 802–815, Feb. 2012.

[14] D. Hu and K. Huang, "GFNet: Gaze focus network using attention for gaze estimation," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2023, pp. 2399–2404.

[15] Z.-H. Wan, C.-H. Xiong, W.-B. Chen, and H.-Y. Zhang, "Robust and accurate pupil detection for head-mounted eye tracking," *Comput. Electr. Eng.*, vol. 93, Jul. 2021, Art. no. 107193.

[16] A. N. Angelopoulos, J. N. P. Martel, A. P. S. Kohli, J. Conradt, and G. Wetzstein, "Event based, near eye gaze tracking beyond 10,000Hz," 2020, *arXiv:2004.03577*.

[17] D. Xia and Z. Ruan, "IR image based eye gaze estimation," in *Proc. 8th ACIS Int. Conf. Softw. Eng., Artif. Intell., Netw., Parallel/Distrib. Comput. (SNPD)*, Jul. 2007, pp. 220–224.

[18] D. Hu and K. Huang, "Semi-supervised multitask learning using gaze focus for gaze estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 9, pp. 7935–7946, Sep. 2024.

[19] 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.

[20] I. Martinikorena, A. Larumbe-Bergera, M. Ariz, S. Porta, R. Cabeza, and A. Villanueva, "Low cost gaze estimation: Knowledge-based solutions," *IEEE Trans. Image Process.*, vol. 29, pp. 2328–2343, 2020.

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

[22] 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.

[23] M. Choe, Y. Choi, J. Park, and J. Kim, "Head mounted IMU-based driver's gaze zone estimation using machine learning algorithm," *Int. J. Human–Comput. Interact.*, vol. 40, no. 23, pp. 7970–7981, Dec. 2024.

[24] Y. Cheng et al., "What do you see in vehicle? Comprehensive vision solution for in-vehicle gaze estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 1556–1565.

[25] Y. Cheng and F. Lu, "Gaze estimation using transformer," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 3341–3347.

[26] X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang, and O. Hilliges, "ETH-XGaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 365–381.

[27] J. Zhou et al., "EraW-Net: Enhance-refine-align W-Net for scene-associated driver attention estimation," 2024, *arXiv:2408.08570*.

[28] G. Huang et al., "Gaze estimation by attention-induced hierarchical variational auto-encoder," *IEEE Trans. Cybern.*, vol. 54, no. 4, pp. 2592–2605, Apr. 2024.

[29] Y. Zhou, L. Liu, and C. Gou, "Learning from observer gaze: Zero-shot attention prediction oriented by human-object interaction recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 28390–28400.

[30] R. S. Kothari, A. K. Chaudhary, R. J. Bailey, J. B. Pelz, and G. J. Diaz, "EllSeg: An ellipse segmentation framework for robust gaze tracking," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 5, pp. 2757–2767, May 2021.

[31] Z. Wang, Y. Zhao, Y. Liu, and F. Lu, "Edge-guided near-eye image analysis for head mounted displays," in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*, Oct. 2021, pp. 11–20.

[32] Y. Li et al., "Real-time gaze tracking via head-eye cues on head mounted devices," *IEEE Trans. Mobile Comput.*, vol. 23, no. 12, pp. 13292–13309, Dec. 2024.

[33] M. Zhang, Y. Liu, and F. Lu, "GazeOnce: Real-time multi-person gaze estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 4197–4206.

[34] 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. 6912–6921.

[35] 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.

[36] W. Zhu and H. Deng, "Monocular free-head 3D gaze tracking with deep learning and geometry constraints," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3143–3152.

[37] 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.*, vol. 46, no. 12, pp. 7509–7528, Dec. 2024.

[38] W. Zhong, C. Xia, D. Zhang, and J. Han, "Uncertainty modeling for gaze estimation," *IEEE Trans. Image Process.*, vol. 33, pp. 2851–2866, 2024.

[39] X. Chen, M. Chen, Y. Chen, Y. Lin, B. Ke, and B. Ni, "Large generative model impulsed lightweight gaze estimator via deformable approximate large kernel pursuit," *IEEE Trans. Image Process.*, vol. 34, pp. 1149–1162, 2025.

[40] 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.

[41] Z. Zhu and Q. Ji, "Eye gaze tracking under natural head movements," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 918–923.

[42] T. Huang et al., "Driver lane change intention prediction based on topological graph constructed by driver behaviors and traffic context for human-machine co-driving system," *Transp. Res. C, Emerg. Technol.*, vol. 160, Mar. 2024, Art. no. 104497.

[43] Z. Hu, Y. Cai, Q. Li, K. Su, and C. Lv, "Context-aware driver attention estimation using multi-hierarchy saliency fusion with gaze tracking," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 8, pp. 8602–8614, Aug. 2024.

[44] A. Tawari, K. H. Chen, and M. M. Trivedi, "Where is the driver looking: Analysis of head, eye and iris for robust gaze zone estimation," in *Proc. Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2014, pp. 988–994.

[45] H. Yang, J. Wu, Z. Hu, and C. Lv, "Real-time driver cognitive workload recognition: Attention-enabled learning with multimodal information fusion," *IEEE Trans. Ind. Electron.*, vol. 71, no. 5, pp. 4999–5009, May 2024.

[46] I.-H. Choi, S. Kyung Hong, and Y.-G. Kim, "Real-time categorization of driver's gaze zone using the deep learning techniques," in *Proc. Int. Conf. Big Data Smart Comput. (BigComp)*, Jan. 2016, pp. 143–148.

[47] S. Ghosh, A. Dhall, G. Sharma, S. Gupta, and N. Sebe, "Speak2Label: Using domain knowledge for creating a large scale driver gaze zone estimation dataset," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Montreal, BC, Canada, Oct. 2021, pp. 2896–2905.

[48] Y. Wang et al., "Continuous driver's gaze zone estimation using RGB-D camera," *Sensors*, vol. 19, no. 6, p. 1287, Mar. 2019.

[49] 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.

[50] R. Liu, Y. Liu, H. Wang, and F. Lu, "PnP-GA+: Plug-and-play domain adaptation for gaze estimation using model variants," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 5, pp. 3707–3721, May 2024.

[51] Y. Bao and F. Lu, "Unsupervised gaze representation learning from multi-view face images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 1419–1428.

[52] Y. Yi, C. Lu, B. Wang, L. Cheng, Z. Li, and J. Gong, "Fusion of gaze and scene information for driving behaviour recognition: A graph-neural-network- based framework," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 8, pp. 8109–8120, Aug. 2023.

[53] 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.

[54] X. Zhang, Y. Sugano, and A. Bulling, "Revisiting data normalization for appearance-based gaze estimation," in *Proc. ACM Symp. Eye Track. Res. Appl.*, 2018, pp. 1–9.

[55] K. A. Funes Mora, F. Monay, and J.-M. Odobez, "EYEDIAP: A database for the development and evaluation of gaze estimation algorithms from RGB and RGB-D cameras," in *Proc. Symp. Eye Tracking Res. Appl.*, 2014, pp. 255–258.

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

[57] 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. Workshops (CVPRW)*, Jul. 2017, pp. 51–60.

[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[59] Y. Cheng, Y. Bao, and F. Lu, "Puregaze: Purifying gaze feature for generalizable gaze estimation," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 436–443.

[60] Z. Liang, Y. Bao, and F. Lu, "De-confounded gaze estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 219–235.

[61] Y. Bao and F. Lu, "From feature to gaze: A generalizable replacement of linear layer for gaze estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 1409–1418.

[62] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3560–3569.

[63] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2018, pp. 7132–7141.

[64] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 116–131.

Daosong Hu received the B.S. degree in mechanical design, manufacturing and automation and the M.S. degree in mechanical engineering from China University of Geosciences (CUG), Wuhan, China, in 2019 and 2022, respectively. He is currently pursuing the Ph.D. degree with School of Computer Science and Engineering, Sun Yat-sen University, China. His research interests include gaze estimation and medical imaging.

Xi Li is currently pursuing the bachelor's degree in computer science with Sun Yat-sen University, China. Her research interests primarily focus on gaze estimation.

Mingyue Cui (Associate Member, IEEE) received the B.Sc. degree in software engineering from Chongqing Normal University in 2014, and the M.Sc. degree in software engineering and the Ph.D. degree in computer science from Sun Yat-sen University, Guangzhou, Guangdong, China, in 2017 and 2022, respectively. He was a Visiting Student with the Technical University of Munich, Germany, from November 2021 to November 2022. He is currently a joint Postdoctoral Fellow with the School of Computer Science and Engineering, Sun Yat-Sen University, China. His research interests are in the area of 3D vision and entropy coding.

Kai Huang (Member, IEEE) received the B.Sc. degree from Fudan University in 1999, the M.Sc. degree from University Leiden in 2005, and the Ph.D. degree from ETH Zürich in 2010. In 2015, he joined Sun Yat-sen University as a Professor. He was appointed as the Director of the Institute of Artificial Intelligence and Unmanned Systems, School of Computer Science, in 2020. His research interests include techniques for the analysis, design, and optimization of embedded/CPS systems, particularly in the automotive, medical, and robotic domains. He was a recipient of best paper awards/candidates for several conferences.