

# FIFA: Fine-grained Inter-frame Attention for Driver’s Video Gaze Estimation

Daosong Hu<sup>1</sup> Mingyue Cui<sup>1,2,\*</sup> Kai Huang<sup>1</sup>

<sup>1</sup>School of Computer Science and Engineering, Sun Yat-sen University, China

<sup>2</sup>Key Laboratory of Machine Intelligence and Advanced Computing, Sun Yat-sen University, China

{huds, cuimy, huangk36}@mail2.sysu.edu.cn

## Abstract

Gaze direction serves as a pivotal indicator for assessing the level of driver attention. While image-based gaze estimation has been extensively researched, there has been a recent shift towards capturing gaze direction from video sequences. This approach encounters notable challenges, including the comprehension of the dynamic pupil evolution across frames and the extraction of head pose information from a relatively static background. To surmount these challenges, we introduce a dual-stream deep learning framework that explicitly models the displacement changes of the pupil through a fine-grained inter-frame attention mechanism and generates weights to adjust gaze embeddings. This technique transforms the face into a set of distinct patches and employs cross-attention to ascertain the correlation between pixel displacements in various patches and adjacent frames, thereby tracking spatial dynamics within the sequence. Our method is validated using two publicly available driver gaze datasets, and the results indicate that it achieves state-of-the-art performance or is on par with the best outcomes while reducing the parameters.

## 1. Introduction

The eyes, metaphorically described as the windows to the soul, not only reflect an individual’s emotional states and intentions but also serve as an indicator of the direction of human attention. Gaze estimation, a pivotal task within the domain of computer vision, is instrumental in a multitude of applications, including human-computer interaction [1, 23, 37], gaze object analysis [31], and safe driving [11, 15, 24]. In the context of road traffic accidents, the degree of a driver’s attention is identified as a significant factor influencing the rate of accidents, particularly within the framework of traditional intelligent vehicles [9]. Gaze direction is ascertained through the intricate interplay of head rotation and pupil displacement, phenomena that can be in-

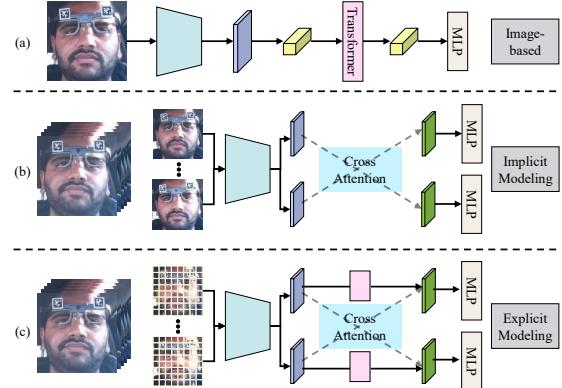


Figure 1. Illustration of various methods in gaze estimation.

ferred through physiological parameters and visual cues. To forecast a driver’s gaze direction and to monitor potential distracted behavior, supplementary devices are commonly employed to capture facial feature parameters [25, 32, 33].

Owing to the swift advancement of deep learning technology, the capacity of computers to comprehend images is markedly augmented. The non-invasive characteristic of image acquisition led to a surge in interest in image-based driver monitoring methods [15, 20]. These methodologies harness the formidable information extraction prowess of neural networks to distill quantifiable cues from facial images, thereby facilitating the prediction of gaze direction. In the assessment of predictive efficacy, models are tasked with achieving not only high precision but also with managing the number of samples within a permissible error margin to mitigate the likelihood of erroneous judgments. Furthermore, the computational expenses are a pertinent consideration, given the constraints on the computing power available within in-vehicle systems.

The acquisition of dense and precise annotations poses a significant challenge, prompting existing studies to often approximate gaze direction by estimating sparse gaze zones [27, 28]. This approximation is predicated on the assumption that a driver’s gaze predominantly focuses on specific,

\*Corresponding author

predefined areas. However, the scarcity of detailed annotations inherently restricts the model’s applicability. Consequently, with the advent of driver gaze datasets that offer precise annotations, researchers begin to explore how to predict a driver’s accurate attention direction in three-dimensional space from images [7, 20].

Current scholarly endeavors frequently conceptualize gaze as a time-independent event, as depicted in Figure 1(a), with the objective of determining a driver’s attention direction in an end-to-end manner from facial images [2, 14]. However, gaze is inherently a continuous spatial behavior, and capturing the temporal dynamics of eye movement can enhance the accuracy of gaze estimation. Existing methods demonstrated that inter-frame information can be used to model spatial dynamics [35]. By considering the temporal proximity between successive frames, gaze-related feature motion is predominantly exhibited as pupil displacement. The construction of a gaze estimation model from video streams, leveraging this pivotal feature, encounters challenges such as the interference from static background elements and the correlation of gaze alterations between adjacent frames. Certain methodologies, as illustrated in Figure 1(b), employ cross-attention to implicitly model the spatial dynamics of the face [18].

In this paper, we derive inspiration from the manifestations of gaze transformation and introduce a fine-grained inter-frame attention mechanism (FIFA) for the purpose of driver video gaze estimation, as delineated in Figure 1(c). To elaborate, adhering to the principle of fine granularity, we partition facial images into discrete patches and detect pixel alterations between frames by computing the disparities between adjacent frames. Furthermore, a cross-attention module is integrated to ascertain the contribution of each patch to gaze features, thereby promoting the exchange of information. In light of the computational overhead limitations inherent in in-vehicle systems, we have engineered a hybrid framework that amalgamates Convolutional Neural Networks (CNNs) and Transformers, with the goal of achieving model lightweighting. Given the pronounced emphasis of FIFA on pupil displacement, we propose a parallel framework. This framework extracts semantic embeddings from encoded features and assigns weights to them using attention weights derived from FIFA, thereby preserving relatively static head posture information and highlighting pupil changes. Our contributions are as follows:

- We introduce FIFA, a fine-grained inter-frame attention module that is highly sensitive to spatial changes within continuous frames, designed for capturing pupil displacement and facilitating inter-frame information exchange.
- We propose a parallel framework for concurrently modeling inter-frame changes and extracting gaze features. Given that the head tends to remain relatively stationary

between adjacent frames, this parallel design aids in the reliable extraction of head posture information.

- The proposed methodology is subjected to validation on publicly available driver gaze estimation datasets and demonstrated state-of-the-art performance with identical input video stream data. In addition to quantitative metrics, feature visualization further substantiates the efficacy of FIFA in video stream estimation tasks.

## 2. Related Work

### 2.1. Gaze Estimation

In the domain of gaze estimation tasks, methodologies are typically classified according to the nature of the input data into two distinct categories: image-based and video-based approaches. Image-based strategies deduce human gaze from facial images. Traditionally, these methods rely on additional devices to capture ocular features, thereby constructing eye models under the assumption that the normal vector of the eyeball surface corresponds to the gaze direction [22, 29, 30]. While these approaches exhibit robust performance in controlled settings, their susceptibility to environmental conditions constrains their applicability [13]. The emergence of deep learning has steered research towards harnessing neural networks for feature extraction and modeling, thereby obtaining gaze direction non-invasively from facial images [36, 38]. The eye images are input into convolutional neural networks to extract gaze features and subsequently regress to determine the gaze direction [6]. Furthermore, certain methodologies incorporate additional features, such as head posture, facial appearance, and the intersection of the gaze [3, 14]. Generally, the incorporation of supplementary information can significantly bolster the model’s predictive accuracy.

In contrast to image-based methodologies, video-based strategies regard gaze direction alterations as continuous within the spatial domain, necessitating the establishment of correlations between successive frames. Following the introduction of video gaze datasets, a variety of video-based gaze estimation approaches have emerged. Kellnhofer et al. [21] introduced a bidirectional LSTM that processes consecutive frames to indirectly infer changes in spatial information. Jindal et al. [18] presented a cross-attention mechanism designed to capture spatial and temporal dynamics. Guan et al. [10] proposed a multi-cue gaze estimation approach that captures spatiotemporal interactive contexts among the head, face, and eyes, thereby enhancing video gaze estimation capabilities. Nonetheless, the comprehension of the temporal dynamics of eyeball motion presents a distinct challenge for neural networks.

## 2.2. Gaze Estimation for Driver

Monitoring the behavior and attention of drivers can effectively reduce the incidence of road accidents [4]. Infrared cameras and head-mounted devices are designed to capture eye features, such as iris parameters, to infer gaze direction [8, 17, 19]. However, devices used for parameter acquisition typically require a fixed relative distance from the eyes, leading to the study of less restrictive image-based driver gaze estimation. Since gaze is difficult to measure directly, previous work has transformed gaze estimation into gaze zone estimation [16, 34]. The in-vehicle environment is divided into different regions, converting the regression task into classification. However, this strategy results in an inability to accurately capture the gaze direction. Kasahara et al. [20] proposed an in-vehicle gaze measurement method and published a dataset containing precise gaze labels. They also proposed a self-supervised framework that leverages scene saliency and the geometric consistency of gaze direction to enhance the extraction of gaze features. Cheng et al. [7] proposed a dual-stream pyramid framework based on Transformers for driver gaze estimation. However, these methods all neglect the correlation of gaze changes between two frames. Additionally, in in-vehicle systems, the number of model parameters needs to be considered simultaneously.

## 3. Proposed Method

In this section, we furnish a comprehensive delineation of the proposed gaze estimation framework tailored for video sequences, as shown in Figure 2. The framework incorporates a fine-grained inter-frame attention module designed to ascertain the regions most pertinent to detectable motion changes by examining the motion relationships between two frames. Weight coefficients are utilized to quantify the impact of various regions on the precision of estimation for adjacent frames. The framework is articulated into three distinct components: the encoding of facial images, the extraction of inner features, and the application of fine-grained inter-frame attention.

### 3.1. Encoding of Facial Image

Our initial endeavor focused on segmenting facial images into distinct regions, with the aim of enabling the model to capture the interdependencies among various areas. However, given that the pupil constitutes a relatively small portion of the facial area, lower resolutions can result in blurred pupil features. Consequently, the facial images that are typically input into the model are often of high resolution, leading to a substantial number of regions being defined on full-resolution images. An increase in the number of regions incurs a significant computational cost, which can be prohibitive. Drawing inspiration from GazeTR [5], we employed an encoder to project facial images into a feature

space, thereby mitigating the computational demands of the model.

The encoder utilizes the same architecture as ResNet18, ensuring the incorporation of pre-trained prior knowledge. As the number of encoder layers increases, the resolution of the facial images is further reduced, which introduces additional computational overhead. Balancing feature depth with resolution, we selected the residual modules from the first two layers of ResNet18, transforming the facial images from  $3 \times 224 \times 224$  to  $128 \times 28 \times 28$ . This transformation can be mathematically expressed as:

$$E_{t-1}, E_t = \mathcal{R}(x_{t-1}, x_t) \quad (1)$$

where  $E$  is the low-level feature and  $\mathcal{R}$  denotes the residual module.

### 3.2. Inner Feature Extraction

The low-level features encompass a wealth of semantic information, representing a profound integration of gaze features and appearance details. A Transformer is deployed to refine the gaze features, thereby facilitating the estimation of pupil displacement and head rotation. In particular, we segment the low-level feature map  $E$  into  $P \times P$  patches, with each patch  $E^{i,j}$  being convolved ( $Conv$ ) and flattened as a feature vector. A linear projection layer  $L$  is then employed to project these feature vectors into a higher-dimensional feature space.

During the training process, the self-attention mechanism is leveraged to produce the gaze representation. Concurrently, we incorporate the positional information of each patch by generating learnable embeddings and incorporating them into the feature matrix. The resultant feature matrix  $F$  is obtained as follows:

$$F = \sum T(L(Conv(E^{i,j})) + Pos) \quad (2)$$

### 3.3. Fine-grained Inter-frame Attention

In the context of video sequences, our focus is particularly directed towards monitoring significant shifts in pupil position. Consequently, we introduce a fine-grained inter-frame attention mechanism (FIFA) aimed at extracting and augmenting facial region appearance information from within video frames. This mechanism places special emphasis on the accurate localization and feature enhancement of the pupil, in order to tackle the challenge posed by rapid fluctuations in pupil position throughout the video sequence. Our methodology is grounded in the observation that minor appearance changes between successive frames in a video, particularly within the pupil region, harbor substantial dynamic information compared to the head. Hence, we have engineered the FIFA (Fine-grained Inter-frame Attention)

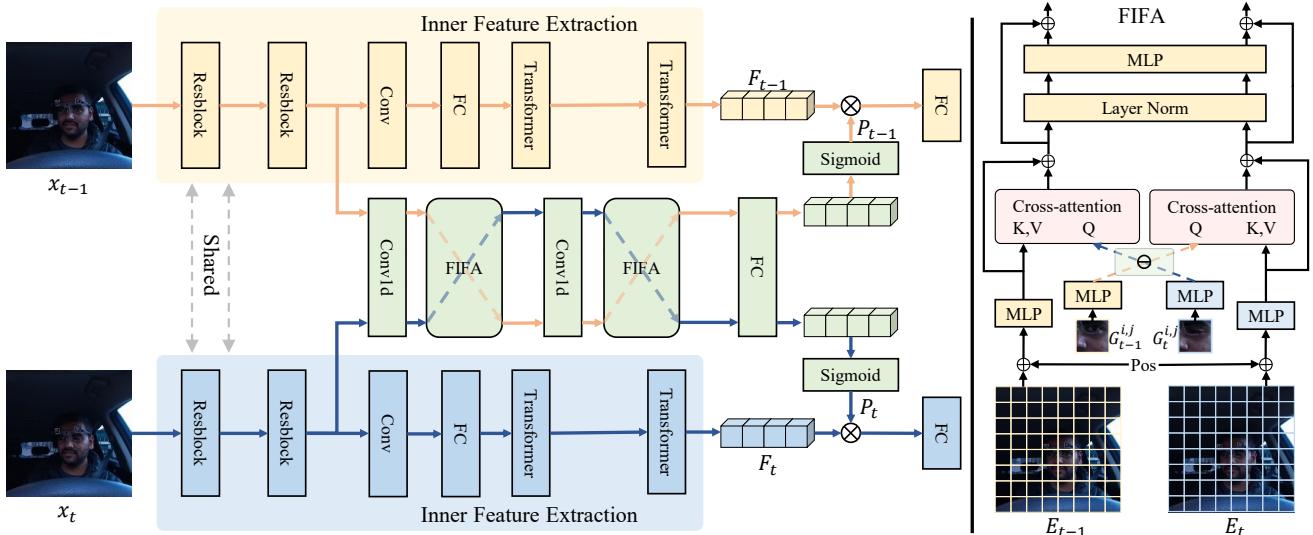


Figure 2. The overview of our proposed method.

module to amplify the appearance features of specific regions by explicitly modeling intra-frame appearance similarity.

Specifically, for a given pair of consecutive input low-level features  $E_{t-1}, E_t \in \mathbb{R}^{W \times W \times C}$ , where  $W$  is the feature size, the low-level information is represented as a collection of  $n^2$  local features through patch partitioning.  $E_{t-1}, E_t = \sum_{i=0}^n \sum_{j=0}^n G_{t-1}^{i,j}, \sum_{i=0}^n \sum_{j=0}^n G_t^{i,j}$ . A Multi-Layer Perceptron (MLP) is employed to extract representations from different regions. A cross-attention mechanism is used to aggregate similar appearance information from different regions of adjacent frames, thereby enhancing the representation of key areas in the current frame. For the differences in each region  $G_{t-1}^{i,j} - G_t^{i,j} \in \mathbb{R}^{N \times N \times C}$  in the different frames, where  $N$  denotes the patch size, it is treated as a query  $Q_{t-1}^{i,j}$ , and all regions  $G_t^{n,i,j} \in \mathbb{R}^{N \times N \times C}$  in the  $E_t$  are used as keys  $K_t^{n,i,j}$  and values  $V_t^{n,i,j}$ . The process is as follows:

$$Q_{t-1}^{i,j} = \text{Flatten}(G_{t-1}^{i,j} - G_t^{i,j}) \cdot W_Q \quad (3)$$

$$K_t^{n,i,j} = \text{Flatten}(G_t^{n,i,j} + Pos) \cdot W_K \quad (4)$$

$$V_t^{n,i,j} = \text{Flatten}(G_t^{n,i,j} + Pos) \cdot W_V \quad (5)$$

where  $W_Q$ ,  $W_K$ , and  $W_V$  are linear projection matrices,  $\text{Flatten}(\cdot)$  is the operation of flattening the features, and  $Pos$  denotes the position embedding. By calculating the dot product of the query and keys and applying the Soft-Max function, we obtain an attention map  $S_{t-1 \rightarrow t}^{i,j}$ , which encodes the similarity between different regions within the

frame. Specifically, as follows:

$$S_{t-1 \rightarrow t}^{i,j} = \text{SoftMax} \left( \frac{Q_{t-1}^{i,j} \cdot (K_t^{n,i,j})^\top}{\sqrt{d_k}} \right) \quad (6)$$

where  $\sqrt{d_k}$  is the scale coefficient. This attention map is utilized to aggregate similar appearance information from other regions within the frame and fuse it with the features of the current region, thereby enhancing the appearance features of key areas, as follow:

$$\hat{G}_{t-1}^{i,j} = G_{t-1}^{i,j} + S_{t-1 \rightarrow t}^{i,j} \cdot V_t^{n,i,j} \quad (7)$$

The enhanced feature  $\hat{G}_{t-1}^{i,j}$  contains a mixture of similar regions between two different frames, indicating how the pupil moves between frames to explicitly model the transition of gaze. Similarly,  $\hat{G}_t^{i,j}$  can be obtained through the same pipeline. The enhanced patch features are reorganized according to positional information, resulting in a complete feature map. Specifically, as follows:

$$\hat{x}_{t-1}^{i,j} = \hat{G}_{t-1}^{i,j}, i, j \in [0, n] \quad (8)$$

$$\hat{x}_t^{i,j} = \hat{G}_t^{i,j}, i, j \in [0, n] \quad (9)$$

where  $\hat{x}_{t-1}^{i,j}$  and  $\hat{x}_t^{i,j}$  are the reinforced input, and  $i$  and  $j$  represent the index of patch.

The estimation of gaze direction is influenced by the coupling factors of pupil position and head posture. The frames reinforced by FIFA focus on the significant displacement of the pupil, while blurring the angular information of the head in space. Therefore, another branch of Inner feature extraction mainly utilizes Transformer to capture the overall semantic information of the input frame, thereby obtaining gaze embeddings  $F_{t-1}$  and  $F_t$ .  $\hat{x}_{t-1}$  and  $\hat{x}_t$  are



Figure 3. Visual comparison of predicted results at adjacent frames. The red arrow is our method, sky blue is STAGE, blue is GazePTR, and green is ground truth.

used to obtain an attention map, which is used to adjust the weights of different information embedded in gaze embeddings, thereby modeling the position changes of the pupil and preserving head posture information. The entire process is as follows:

$$P_{t-1}, P_t = \text{Sig}(\text{FC}(\text{Flatten}(\hat{x}_{t-1}, \hat{x}_t))) \quad (10)$$

$$\hat{F}_{t-1} = P_{t-1} \odot F_{t-1} \quad (11)$$

$$\hat{F}_t = P_t \odot F_t \quad (12)$$

where FC is a fully connected layer, Sig denotes the sigmoid function, and  $\odot$  represents dot product.

The gaze predictor consists of two linear layers aimed at predicting gaze direction from the enhanced gaze embedding and sharing it across all timestamps. For  $t^{\text{th}}$  frame,  $g_t$  and  $\hat{g}_t$  are the ground-truth of sequences and predicted gaze directions. We use the following objective function for training our model parameters:

$$\mathcal{L} = 0.5\|\hat{g}_t - g_t\| + 0.5\|\hat{g}_{t-1} - g_{t-1}\| \quad (13)$$

## 4. Experiments

### 4.1. Datasets

#### 4.1.1 LBW

LBW is a large-scale dataset that includes driver facial images, road facing scene images, and 3D gaze directions from head mounted eye trackers. Facial landmarks are used to crop facial images, with a fixed size of  $224 \times 224$ . 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. During training, the adjacent images are treated as adjacent frames.

#### 4.1.2 IV

The IV dataset comprises 44,705 images across 125 subjects. It encompasses a rich set of ground truth data, includ-

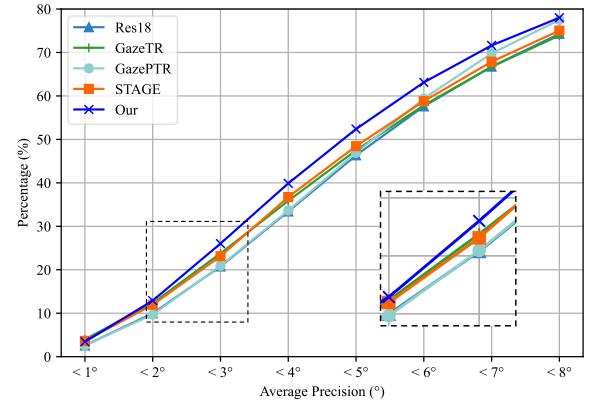


Figure 4. Error distribution at different gaze angles on LBW.

ing head pose and gaze direction. For the purpose of within-dataset evaluation, the dataset is partitioned into three subsets based on the subjects. We adhere to the dataset division strategy outlined in [7] and implement a three-fold cross-validation approach. Given the inclusion of head poses in the labels, we curate images with analogous head poses within the same subset to serve as adjacent frames. Should such images not be available, the original image is duplicated to fulfill the role of an adjacent frame.

### 4.2. Implementation Details

The proposed method is implemented by using Pytorch, and trained for 80 epochs on V100 GPU. A batch-size of 64 is set to train, and the learning rate is set as 0.001. The learning rate is adjusted every 25 epochs with an adjustment coefficient is 0.5.

### 4.3. Evaluation Metrics

The mean of angle error is used to evaluate model prediction accuracy, with lower values representing better performing methods [14]. In addition, following [7], the average accuracy is used to describe the error distribution. The average

Table 1. Comparison with State-of-the-art Methods.

Method	Input	Error		Specification	
		IV	LBW	FLOPs	#Param.
Gaze360 [21]	Seq.	8.53°	7.04°	12.78G	14.6M
ResNet18 [12]	Seq.	8.05°	6.41°	3.65G	11.25M
GazeTR [5]	Seq.	7.31°	6.19°	3.68G	11.39M
XGaze [39]	Seq.	7.37°	6.25°	8.26G	23.64M
GazePTR [7]	Seq.	7.33°	6.16°	3.69G	11.52M
STAGE [18]	Seq.	7.38°	6.11°	3.67G	11.32M
Ours	Seq.	<b>7.29°</b>	<b>5.84°</b>	<b>2.70G</b>	<b>5.87M</b>

precision is defined where  $< k^\circ$  means an prediction is correct if the angular error is lower than  $k^\circ$ . The proportion of correct sample size to the entire dataset is used as a measure.

#### 4.4. Comparison with State-of-the-art Methods

In Table 1, we present a comparative analysis of our method’s performance against state-of-the-art techniques. Initially, in the IV dataset, our method manifests an improvement with an error of  $7.29^\circ$ , representing a reduction of  $0.76^\circ$  relative to ResNet18’s error of  $8.05^\circ$ . The performance of our proposed method is on par with GazeTR, which amalgamates CNN and Transformer architectures. For the LBW dataset, our method similarly exhibits superior performance, attaining an error of  $5.84^\circ$ , which signifies a decrease of  $0.57^\circ$  from ResNet18’s error of  $6.41^\circ$ . In comparison to GazeTR, our proposed method enhances performance by 4.8%. These outcomes demonstrate that our method outperforms current technologies in terms of accuracy.

For FLOPs, our method attains  $2.70G$ , which is markedly lower than other methodologies, including ResNet18 and GazeTR. This indicates that our model offers greater computational efficiency while sustaining minimal error rates. Moreover, our method encompasses only  $5.87M$  parameters, a significantly smaller number compared to Gaze360’s  $14.6M$  and XGaze’s  $23.64M$ , rendering the model more suitable for deployment in resource-limited settings.

#### 4.5. Visualization of Prediction Results

Fig. 3 illustrates the performance of different estimation strategies on adjacent frames. For minor facial occlusions and exposure, all methods demonstrate a certain degree of robustness. However, when the occlusion area is large, FIFA leverages information from the previous frame for prediction and achieves accurate results. This outcome indicates that inter-frame information can be utilized to assist in prediction.

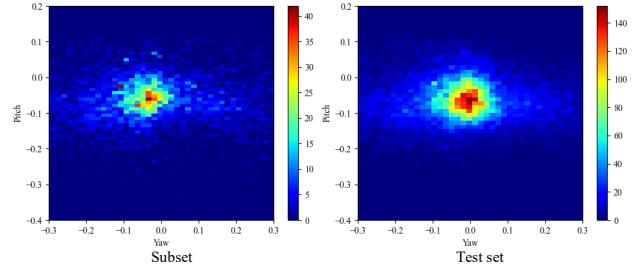


Figure 5. Ground truth distribution of subset and test set on LBW.

#### 4.6. Average Precision

The findings presented in Figure 4 highlight the utilization of average precision within the LBW dataset to assess the model’s performance across varying levels of permissible error. In the domain of driver attention monitoring, it is acknowledged that the model’s predictions may deviate from actual values within a defined margin of error. Essentially, the preference is for the model’s prediction error to fall within an acceptable range, even if absolute precision is not achieved. For instances classified as having low predictive difficulty, all models delivered exemplary outcomes; specifically, the count of samples accurately predicted was nearly identical for errors less than  $2^\circ$ . Nevertheless, as the permissible error margin expands, our method displays a more pronounced increase, suggesting a higher concentration of samples within these bounds. At an allowable error threshold of  $< 8^\circ$ , the performance of GazePTR aligns with that of our proposed method. It is our conviction that the error threshold should not exceed  $8^\circ$ , as surpassing this limit could substantially compromise the model’s monitoring precision. To encapsulate, our method ensures that a greater number of samples maintain prediction accuracy within the designated error limits, thereby validating its effectiveness in the context of driver monitoring.

#### 4.7. Embedding Visualization

To substantiate the efficacy of the proposed FIFA mechanism, we present a visualization of the mean gaze embeddings  $F$  from the test dataset. For comparative purposes, we also depict the enhanced gaze embeddings  $\hat{F}$ . As illustrated in Figure 5, it is noteworthy that within the LBW subset, the predominant gaze direction of drivers is directed forward, with a few samples on the sides. The t-SNE algorithm [26], initialized with principal component analysis, is employed to render the embedding space in a two-dimensional format. As depicted in Figure 6, we conduct dimensionality reduction on various frames of data within both the subset and the comprehensive test set. Utilizing FIFA, we observe that the embeddings for both  $F$  and  $\hat{F}$  exhibit more refined distributions. The alignment of the enhanced gaze embeddings with the subset labels suggests that our proposed model

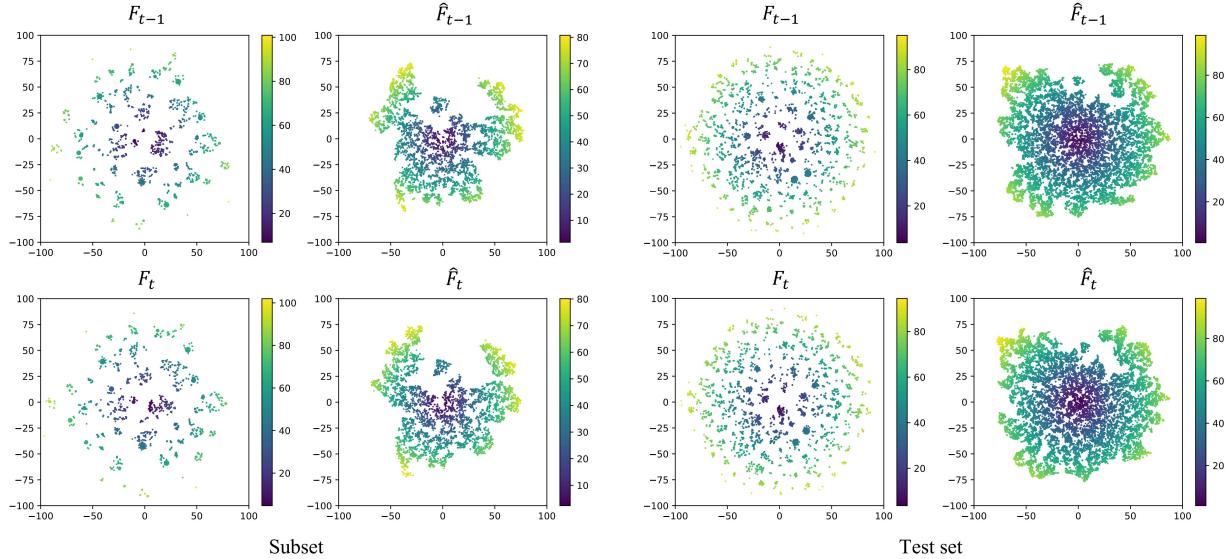


Figure 6. Visualization of the embeddings  $F$  and  $\hat{F}$  on the LBW dataset.

adeptly captures features pertinent to the task. Moreover, the enhanced embeddings exhibit a more uniform distribution across the entire embedding space of the test set. The congruence between the uniform distribution of annotations in the LBW label space and the distribution of the enhanced embeddings validates the rationale behind the smoothing features introduced by FIFA. We posit that the clustering patterns observed in the gaze embedding  $F$  are influenced by factors such as personalization, which may diminish the network’s capacity to accurately extract gaze embeddings.

#### 4.8. Prediction Robustness

During the driving process, the driver’s gaze primarily involves changes in the yaw angle, consistent with Fig. 5. Therefore, we statistically analyze the changes in the pitch angle along the temporal axis. The average absolute error between the predicted pitch angle and the ground truth is used as a metric to describe the model’s stability. As shown in Fig. 7, compared to STAGE’s 0.088 and GazePTR’s 0.085, FIFA reduces the error to 0.054. This experimental result demonstrates that inter-frame information can reduce prediction errors between adjacent frames, especially when there are significant changes in the yaw angle.

#### 4.9. Error Distribution

When driving a vehicle, the gaze direction is usually fixed in a few areas. According to the results shown in Figure 5, most of the samples in LBW are looking forward. We visualized the error distribution of different gaze directions in the LBW dataset, as shown in Figure 8. ResNet18 has a large average error both in large gaze angles and in ar-

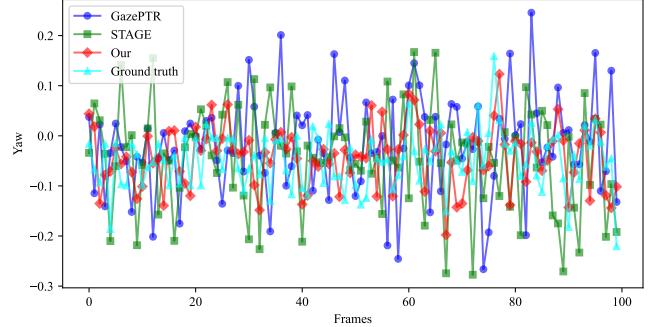


Figure 7. Comparison of yaw angles at different time frames. The average absolute difference from the ground truth is 0.054 for Our, 0.088 for STAGE, and 0.085 for GazePTR

eas where samples are concentrated. GazePTR, STAGE and we rely on a hybrid structure and obtain matching accuracy in areas with large gaze angles, but are better than ResNet18 (fewer red points). In areas where samples are concentrated, our method has a more ideal prediction accuracy (the blue area is larger). We guess that in areas where samples are concentrated, some of the eyes of the samples are occluded, and the proposed method can obtain more accurate gaze estimation by modeling the correlation between the two frames.

#### 4.10. Scalable Capability of FIFA

To ascertain the scalability of FIFA, we have substituted the Inner feature extractor with several prevalent backbone architectures, specifically VGG16, ResNet18, and ResNet50. Notably, the role of the backbone network is to generate

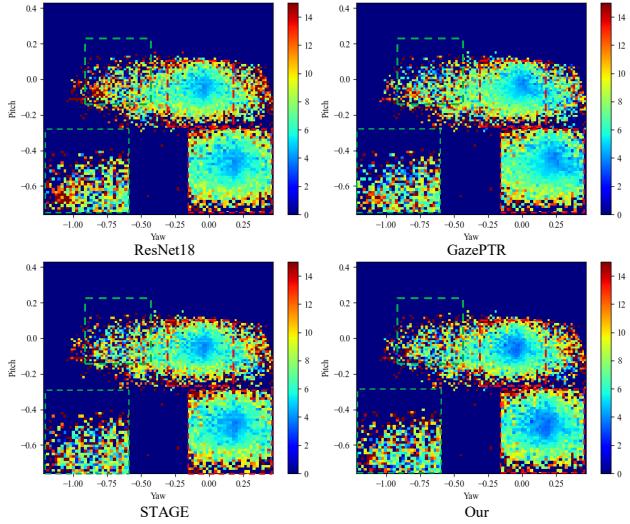


Figure 8. Ground truth distribution of subset and test set on LBW. The red dashed box indicates the sample concentration area, and the green indicates the area with larger gaze angles.

Table 2. The scalable capability of FIFA.

Method	Pretraining Dataset	w.o. FIFA		w. FIFA	
		IV	LBW	IV	LBW
ResNet18	N/A	8.63°	6.52°	7.93°	6.30°
ResNet50	N/A	7.61°	6.40°	7.42°	6.21°
VGG-16	N/A	8.77°	6.98°	7.81°	6.69°
ResNet18	ImageNet-1k	8.05°	6.41°	7.65°	6.22°
ResNet50	ImageNet-1k	7.37°	6.25°	7.31°	6.10°
VGG-16	ImageNet-1k	8.49°	6.65°	7.96°	6.45°

gaze embeddings. To uphold the principle of experimental fairness, features of a  $28 \times 28$  dimension are duplicated and fed into FIFA to determine the attention weights. As presented in Table 2, the experimental outcomes are bifurcated into two categories, contingent upon whether pre-trained parameters are incorporated. The integration of FIFA has consistently yielded performance enhancements, suggesting that FIFA is capable of effectively discerning gaze transitions across various frameworks, thereby explicitly capturing pupil displacement and mitigating the impact of gaze-independent features.

#### 4.11. Ablation Study

The ablation study has been meticulously crafted to assess the individual contributions of each module, with the findings presented in Table 3. The mean error and the allowable error threshold ( $< 5^\circ$ ) serve as the benchmarks for

Table 3. Ablation study.

Method	IV		LBW	
	Mean	$< 5^\circ$	Mean	$< 5^\circ$
Res18	8.05°	38.70%	6.41°	46.34%
$\mathcal{R}$	10.61°	25.15%	7.81°	31.55%
$\mathcal{R} + T$	8.11°	37.67%	6.42°	47.17%
$\mathcal{R} + \text{FIFA}$	7.58°	41.18%	6.31°	48.35%
Our	<b>7.29°</b>	<b>47.04%</b>	<b>5.84°</b>	<b>52.38%</b>

gauging the model’s performance. Initially, we conducted a comparison between two configurations: utilizing solely the encoder ( $\mathcal{R}$ ) and employing ResNet18. Owing to the constrained number of layers within  $\mathcal{R}$ , there is a marked deterioration in performance across both datasets. FIFA and Transformer share analogous architectural frameworks, with the primary distinction being that FIFA incorporates cross-attention mechanisms internally for facilitating information exchange. By serially integrating these two structures subsequent to  $\mathcal{R}$ , we observed an average accuracy enhancement of 26.06% in the IV dataset and 18.51% in the LBW dataset. Moreover, it is noteworthy that FIFA outperformed Transformer in both metrics, underscoring the efficacy of the fine-grained information exchange between consecutive frames. Ultimately, our comprehensive framework, which encompasses both Transformer and FIFA components, delivered the most remarkable performance. The proposed method ensures that a greater number of samples maintain their error within the permissible limits, which is particularly beneficial for driver monitoring tasks.

## 5. Conclusion

This paper introduces a fine-grained inter-frame attention module (FIFA) to model the correlation between successive frames, thereby enhancing the efficacy of video-based gaze estimation. The module explicitly captures the significant displacement of the pupil between adjacent frames. Grounded in the observation that within the brief temporal proximity of consecutive frames, the primary manifestation of gaze alterations is the pupil’s position shift, FIFA allows for explicit modeling of the pupil’s dynamic evolution across two frames and its consequential impact on the gaze estimator. Given the relative immobility of the head between frames, a parallel branch is dedicated to extracting gaze embeddings. Experimental validation has confirmed the performance enhancement of the proposed framework, attributable primarily to the model’s consideration of gaze continuity along the time axis. Moreover, the model’s constrained parameter count endows it with broader application potential in intelligent vehicle settings.

## 6. Acknowledgments:

This work was supported in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2025A1515011485, in part by the National Natural Science Foundation of China under Grant 62232008, and in part by the National Natural Science Foundation of China under Grant 61902442.

## References

- [1] Abdul Rafey Aftab. Multimodal driver interaction with gesture, gaze and speech. In *2019 International Conference on Multimodal Interaction*, pages 487–492, 2019. [1](#)
- [2] Yiwei Bao, Yunfei Liu, Haofei Wang, and Feng Lu. Generalizing gaze estimation with rotation consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4207–4216, 2022. [2](#)
- [3] Pradipta Biswas et al. Appearance-based gaze estimation using attention and difference mechanism. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3143–3152, 2021. [2](#)
- [4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liang, Qiang Xu, Anush Krishnan, Yu Pan, Giacomo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. [3](#)
- [5] Yihua Cheng and Feng Lu. Gaze estimation using transformer. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 3341–3347. IEEE, 2022. [3, 6](#)
- [6] Yihua Cheng, Feng Lu, and Xucong Zhang. Appearance-based gaze estimation via evaluation-guided asymmetric regression. In *Proceedings of the European conference on computer vision (ECCV)*, pages 100–115, 2018. [2](#)
- [7] Yihua Cheng, Yaning Zhu, Zongji Wang, Hongquan Hao, Yongwei Liu, Shiqing Cheng, Xi Wang, and Hyung Jin Chang. What do you see in vehicle? comprehensive vision solution for in-vehicle gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1556–1565, 2024. [2, 3, 5, 6](#)
- [8] Mungyeong Choe, Yeongcheol Choi, Jaehyun Park, and Jungyoon Kim. Head mounted imu-based driver’s gaze zone estimation using machine learning algorithm. *International Journal of Human–Computer Interaction*, pages 1–12, 2023. [3](#)
- [9] Thomas A Dingus, Feng Guo, Suzie Lee, Jonathan F Antin, Miguel Perez, Mindy Buchanan-King, and Jonathan Hankey. Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences*, 113(10):2636–2641, 2016. [1](#)
- [10] Yiran Guan, Zhuoguang Chen, Wenzheng Zeng, Zhiguo Cao, and Yang Xiao. End-to-end video gaze estimation via capturing head-face-eye spatial-temporal interaction context. *IEEE Signal Processing Letters*, 30:1687–1691, 2023. [2](#)
- [11] Amie C Hayley, Brook Shiferaw, Blair Aitken, Frederick Vinckenbosch, Timothy L Brown, and Luke A Downey. Driver monitoring systems (dms): The future of impaired driving management? *Traffic injury prevention*, 22(4):313–317, 2021. [1](#)
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [6](#)
- [13] Daosong Hu and Kai Huang. Gfnet: Gaze focus network using attention for gaze estimation. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2399–2404. IEEE, 2023. [2](#)
- [14] Daosong Hu and Kai Huang. Semi-supervised multitask learning using gaze focus for gaze estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. [2, 5](#)
- [15] Zhongxu Hu, Chen Lv, Peng Hang, Chao Huang, and Yang Xing. Data-driven estimation of driver attention using calibration-free eye gaze and scene features. *IEEE Transactions on Industrial Electronics*, 69(2):1800–1808, 2021. [1](#)
- [16] Zhongxu Hu, Yuxin Cai, Qinghua Li, Kui Su, and Chen Lv. Context-aware driver attention estimation using multi-hierarchy saliency fusion with gaze tracking. *IEEE Transactions on Intelligent Transportation Systems*, 2024. [3](#)
- [17] Sumit Jha and Carlos Busso. Challenges in head pose estimation of drivers in naturalistic recordings using existing tools. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–6. IEEE, 2017. [3](#)
- [18] Swati Jindal, Mohit Yadav, and Roberto Manduchi. Spatio-temporal attention and gaussian processes for personalized video gaze estimation. In *Proceedings of the CVPRW*, pages 604–614, 2024. [2, 6](#)
- [19] Murray W Johns, Andrew Tucker, Robert Chapman, Kate Crowley, and Natalie Michael. Monitoring eye and eyelid movements by infrared reflectance oculography to measure drowsiness in drivers. *Somnologie*, 11(4):234–242, 2007. [3](#)
- [20] Isaac Kasahara, Simon Stent, and Hyun Soo Park. Look both ways: Self-supervising driver gaze estimation and road scene saliency. In *European Conference on Computer Vision*, pages 126–142. Springer, 2022. [1, 2, 3](#)
- [21] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6912–6921, 2019. [2, 6](#)
- [22] Jianfeng Li and Shigang Li. Gaze estimation from color image based on the eye model with known head pose. *IEEE Transactions on Human-Machine Systems*, 46(3):414–423, 2015. [2](#)
- [23] Prajval Kumar Murali, Mohsen Kaboli, and Ravinder Dahiya. Intelligent in-vehicle interaction technologies. *Advanced Intelligent Systems*, 4(2):2100122, 2022. [1](#)
- [24] Akshay Ranges, Bowen Zhang, and Mohan M Trivedi. Driver gaze estimation in the real world: Overcoming the eyeglass challenge. In *2020 IEEE Intelligent vehicles symposium (IV)*, pages 1054–1059. IEEE, 2020. [1](#)
- [25] David A Robinson. A method of measuring eye movement using a scleral search coil in a magnetic field. *IEEE Transactions on bio-medical electronics*, 10(4):137–145, 1963. [1](#)

- [26] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 6
- [27] Sourabh Vora, Akshay Rangesh, and Mohan M Trivedi. On generalizing driver gaze zone estimation using convolutional neural networks. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 849–854. IEEE, 2017. 1
- [28] Sourabh Vora, Akshay Rangesh, and Mohan Manubhai Trivedi. Driver gaze zone estimation using convolutional neural networks: A general framework and ablative analysis. *IEEE Transactions on Intelligent Vehicles*, 3(3):254–265, 2018. 1
- [29] Zhonghua Wan, Caihua Xiong, Wenbin Chen, Hanyuan Zhang, and Shiqian Wu. Pupil-contour-based gaze estimation with real pupil axes for head-mounted eye tracking. *IEEE Transactions on Industrial Informatics*, 18(6):3640–3650, 2021. 2
- [30] Zhong-Hua Wan, Cai-Hua Xiong, Wen-Bin Chen, and Han-Yuan Zhang. Robust and accurate pupil detection for head-mounted eye tracking. *Computers & Electrical Engineering*, 93:107193, 2021. 2
- [31] Binglu Wang, Tao Hu, Baoshan Li, Xiaojuan Chen, and Zhi-jie Zhang. Getectr: A unified framework for gaze object prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19588–19597, 2022. 1
- [32] Zhimin Wang, Yuxin Zhao, Yunfei Liu, and Feng Lu. Edge-guided near-eye image analysis for head mounted displays. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 11–20. IEEE, 2021. 1
- [33] Dongshi Xia and Zongcai Ruan. Ir image based eye gaze estimation. In *Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD 2007)*, pages 220–224. IEEE, 2007. 1
- [34] Yirong Yang, Chunsheng Liu, Faliang Chang, Yansha Lu, and Hui Liu. Driver gaze zone estimation via head pose fusion assisted supervision and eye region weighted encoding. *IEEE Transactions on Consumer Electronics*, 67(4):275–284, 2021. 3
- [35] Guozhen Zhang, Yuhan Zhu, Haonan Wang, Youxin Chen, Gangshan Wu, and Limin Wang. Extracting motion and appearance via inter-frame attention for efficient video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5682–5692, 2023. 2
- [36] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4511–4520, 2015. 2
- [37] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Everyday eye contact detection using unsupervised gaze target discovery. In *Proceedings of the 30th annual ACM symposium on user interface software and technology*, pages 193–203, 2017. 1
- [38] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):162–175, 2017. 2
- [39] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 365–381. Springer, 2020. 6