UnMoDE: Uncertainty Modeling for Driver Gaze Estimation via Feature Disentanglement

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

***Abstract*— Gaze estimation can be used for assessing the attention level of drivers. Current works predominantly focus on enhancing model accuracy, often overlooking the influence of input sample and label uncertainty. In this paper, we propose a framework for uncertainty modeling in driver gaze estimation via feature disentanglement, referred to as UnMoDE. Our approach begins by extracting facial information into distinct feature spaces using an asymmetric dual-branch encoder to obtain gaze features. Subsequently, a multi-layer perceptron (MLP) is employed to project gaze features and labels into an embedding space, representing them as Gaussian distributions. The uncertainty is described using a covariance matrix. Random sampling is applied to derive samples from the gaze embedding distribution to estimate the most probable embedding representation. This estimated representation is then used to regress the gaze direction and is projected back into the gaze feature space, along with identity information, to facilitate facial reconstruction. Extensive experimental evaluations demonstrate that UnMoDE significantly outperforms baseline and state-of-the-art methods on the latest benchmark datasets collected for drivers, particularly in reducing the number of samples with significant errors.**

***Index Terms*— Uncertainty, driver gaze, gaze estimation.**

I. Introduction

THE occurrence of road traffic accidents is significantly influenced by the attention levels of drivers, particularly within the context of traditional and semi-automated intelligent vehicles [1], [2]. Although researchers are striving to develop fully intelligent vehicles to reduce such accidents, active driver engagement remains prevalent due to legal, technical, and adoption constraints [3]. Gaze, which is closely associated with driver attention, can be directly observed through physiological parameters and visual appearance [4]. It finds applications in various scenarios, including human-vehicle interaction [5], [6] and driver monitoring [7], [8].

With the advancement of computer vision technology, particularly the powerful inference capabilities of deep learning, the understanding of images has deepened. Compared to expensive hardware devices [9], [10], [11], a more economical method involves using cameras to capture facial images and extract measurable cues to understand driver gaze behavior [12]. This method balances robustness and practicality, achieving high accuracy under unconstrained conditions. The accuracy of these methods is largely dependent on large-scale datasets [13], [14], [15]. Due to the difficulty in accurately measuring gaze, current research often simplifies driver gaze estimation to gaze zone estimation [16], [17]. These methods segment the driver's visual range into several coarse zones, such as the rearview mirror and windshield, thereby transforming the gaze estimation task into a classification task. However, this coarse segmentation can create gaps in the gaze space, leading to classification failures.

Recent studies have begun exploring new data collection methods that allow for gaze measurement in experimental environments [18], [19]. With the introduction of driver gaze datasets, it has become possible to estimate gaze under conditions of unrestricted head movement. End-to-end gaze estimation models are considered the future trend. These models incorporate additional information or new modules to enhance prediction accuracy. However, these deterministic estimation approaches, while accurate on labeled datasets, often fail to maintain accuracy in real-world scenarios. Significant prediction errors can undermine driver monitoring systems. The uncertainty in prediction results limits the applicability of gaze estimators in intelligent vehicle contexts.

Similar to other computer vision tasks, the accuracy of gaze estimators is influenced by factors such as lighting conditions, eyeglasses, and masks, which can cause key feature deformation or invisibility [20]. High-frequency behaviors such as pupillary movement and head rotation, along with motion blur, can also cause feature disappearance [21]. Consequently, the quality of images captured by cameras can vary significantly. These factors introduce uncertainty into the driver gaze estimation task. When the face is obscured but the pupil position is still visible, the gaze prediction results should be reliable. If the eyes are invisible, gaze estimators can utilize other features, such as head pose, to predict the gaze direction [22]. However, these predictions are prone to underestimation at larger gaze angles. When the gaze direction aligns with head posture, high accuracy is typically achieved; otherwise, the error margin increases. As the pupil moves back and forth between different targets, deformation of the eyes can occur, leading to estimation errors. Thus, this uncertainty primarily arises from gaze-related features and is independent of identity information.

Received 23 August 2024; revised 26 December 2024 and 14 February 2025; accepted 14 March 2025. Date of publication 14 April 2025; date of current version 1 July 2025. This work was supported in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2025A1515011485 and in part by the National Natural Science Foundation of China under Grant 61902442 and Grant 62232008. The Associate Editor for this article was J. W. Choi. *(Corresponding author: Mingyue Cui.)*

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

Digital Object Identifier 10.1109/TITS.2025.3556553

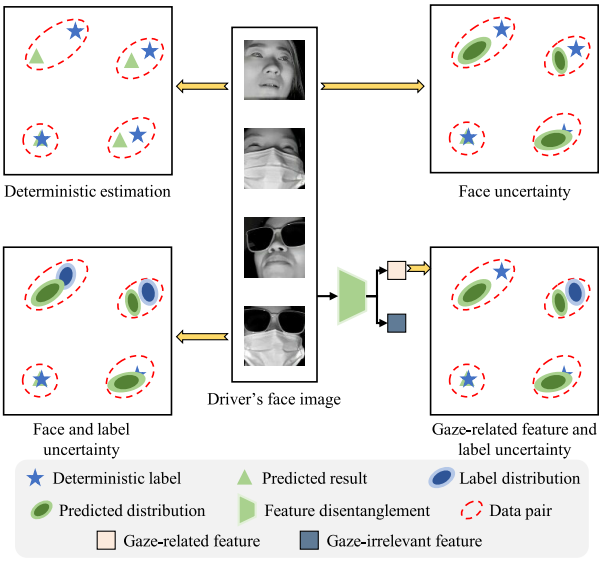

Fig. 1. Differences in methods of gaze estimation.

As shown in Fig. 1, we propose a method for modeling uncertainty in driver gaze estimation through feature disentanglement.[1] Our method differs from face-based [20] and face-and-annotation-based uncertainty modeling [21] methods, and assumes that the uncertainty in gaze prediction comes from gaze related feature and label. Initially, an asymmetric branch is used to extract gaze-related and unrelated features from the face. Cosine loss is employed to decouple gaze from identity information. For gaze-related features, probabilistic embeddings are used to model gaze uncertainty. Following existing research [21], it is noted that annotations also exhibit uncertainty; thus, a linear layer encodes the gaze vector into the label embedding space. Gaze-related features and labels are mapped into a multivariate Gaussian distribution in the embedding space, aligned using KL divergence. The mean of the multivariate Gaussian distribution represents the most likely feature, while the covariance matrix indicates uncertainty. By randomly sampling from the multivariate Gaussian distribution of gaze embeddings, the mean of these samples is considered the most probable embedding in the gaze embedding space. This gaze embedding contains all gaze-related information, which can be projected back into the feature space via a decoder. We hypothesize that the projected features, when combined with identity information, can be decoded back into the initial facial image. Additionally, we propose a gaze regression network aimed at regressing the gaze direction from the most probable gaze embedding. Specifically, our contributions are as follows:

- We propose a framework that maps gaze-related features and annotations into a multivariate Gaussian distribution to model the uncertainty of gaze and labels.
- We obtain gaze embedding samples from the multivariate Gaussian distribution and derive the most probable gaze embedding. A gaze regression network is proposed to regress the gaze direction from the estimated gaze embedding.
- To supervise the purification of identity information, we propose a gaze decoder to project the estimated gaze embedding back into the feature space. The projected gaze embedding, when connected with identity information, can be decoded through a facial reconstruction decoder to obtain the original facial image.

[1] https://github.com/Huds96/UnMoDE

II. Related Work

A. Gaze Estimation

Gaze estimation can be categorized into model-based and appearance-based methods based on feature extraction techniques. Model-based methods typically require additional equipment to capture eye features such as pupil and iris parameters [23], [24]. By modeling these parameters, the normal vector on the eye surface can be calculated to determine the gaze direction. However, the equipment used for parameter acquisition is highly sensitive to environmental conditions and usually requires a fixed relative distance to the user's eyes [25]. In contrast, appearance-based gaze estimation aims to directly predict gaze direction from facial appearance [26], [27]. Compared to model-based methods, appearance-based gaze estimation demonstrates robustness to variations in image resolution and the position of the acquisition device [28], [29], [30]. With the advent of large-scale datasets, deep learning-based gaze estimation models have gradually improved their generalization capabilities in real-world environments [31], [32]. Early methods extracted eye features directly from eye images [33]. Due to the strong correlation between head pose and gaze, subsequent works incorporated head pose vectors along with eye features into the network [34]. Recent approaches implicitly learn head pose features from facial images [35], [36], [37].

These methods perform deterministic gaze estimation from facial images [38], [39], [40], [41], [42]. However, in real-world scenarios, uncertainties in input images arise due to factors such as motion blur from head rotation and variations in lighting conditions during data acquisition. Zhang et al. [20] proposed a domain-consistent and uncertainty-aware network for generalized gaze estimation. Facial images are input into an extractor to obtain gaze features, and both gaze direction and uncertainty are regressed. Intrinsic and extrinsic uncertainty modules are defined to enhance the model's cross-domain capability. To reduce uncertainty in gaze estimation tasks, Wang et al. [43] introduced a triple-label consistency measure strategy. By modeling adjacent labels, pseudo labels, and ground truth, the quality of labels and images is assessed. Sample weighting and label correction strategies are used to mitigate the negative impact of poor-quality images and incorrect labels. Zhong et al. [21] proposed a novel uncertainty modeling strategy where facial images, cropped eye images, and landmarks are jointly used in a traditional network to extract facial features. Probabilistic embeddings based on multivariate Gaussian distribution are used to describe the uncertainty of inputs and labels. These methods directly extract semantic information from facial images to model uncertainty.

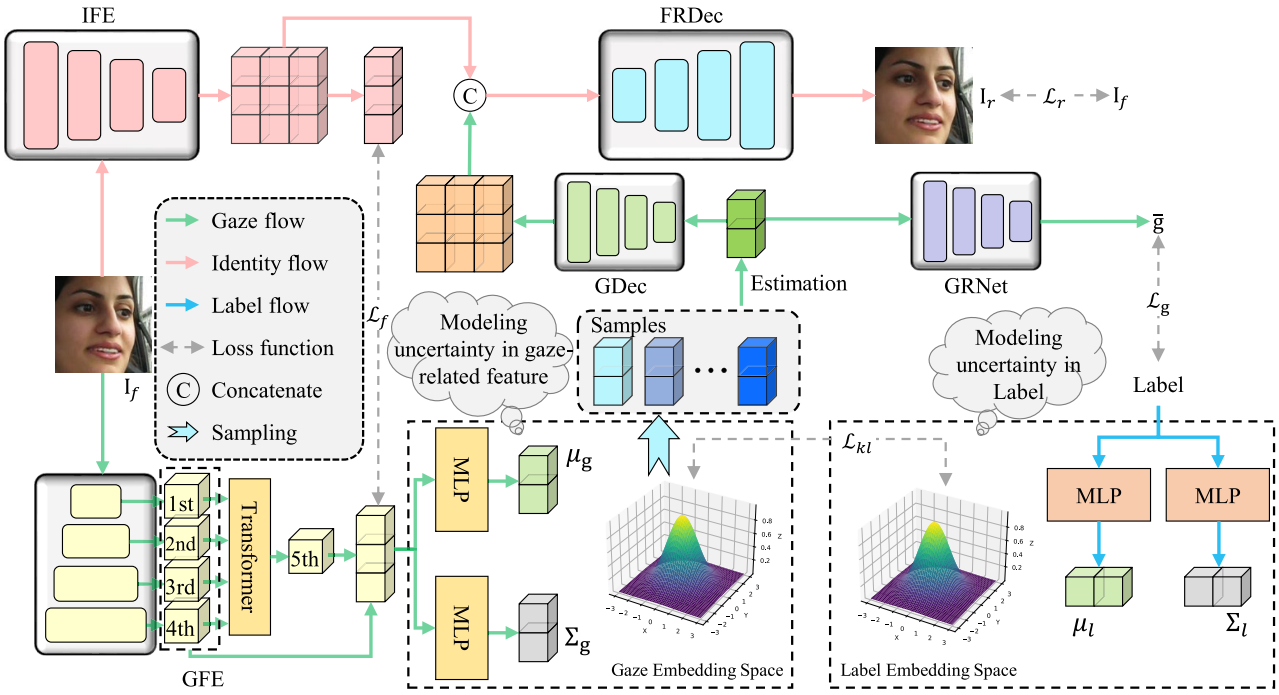

Fig. 2. The overview of our proposed method. I_f and I_r represent the input and reconstructed images, respectively. $\mathcal{L}_r$ is the reconstruction loss, $\mathcal{L}_f$ is the feature loss, $\mathcal{L}_g$ denotes the gaze loss, and $\mathcal{L}_{kl}$ represents the Kullback-Leibler divergence. 1st, 2nd, 3rd, 4th, and 5th represent different levels of gaze features $\{f_i \in \mathbb{R}^{128}\}_{i=1_{st},2_{nd},3_{rd},4_{th},5_{th}}$. μ_g denotes the mean vector and Σ_g indicates the diagonal covariance matrix for gaze features. μ_l and Σ_l is the label.

However, facial images contain both gaze-related and gaze-unrelated information. We posit that the uncertainty in gaze estimation primarily arises from gaze-related features, while the extraction of semantic information may introduce additional uncertainty.

B. Gaze Estimation for Driver

Driving is a complex task that requires constant monitoring of the surrounding environment by the driver [9], [44]. Accurate tracking of the driver's gaze can assist in assessing their attention levels and provide warnings for risky behaviors such as distracted driving [45], [46]. Since gaze direction is difficult to measure directly, previous works have simplified driver gaze estimation to gaze zone estimation [47], [48], [49]. However, gaze zone estimation often fails to adapt to the diverse scenarios presented by real-world environments and different vehicle models. To address this issue, Kasahara et al. [19] proposed a data collection strategy utilizing eye-tracking glasses worn by drivers to record ground truth. They also introduced a self-supervised training approach that leverages the geometric consistency between the driver's gaze direction and observed scene saliency. Cheng et al. [18] proposed a dual-stream pyramid transformer, validating it on an in-car gaze estimation dataset collected from 125 sub-datasets, and achieving state-of-the-art performance. The IVGaze dataset includes various attributes such as eyeglasses, sunglasses, and masks, which introduce additional input uncertainty.

Despite these advancements, measuring the uncertainty of complex facial inputs in intelligent vehicle scenarios remains a challenge. The IVGaze dataset, by including diverse attributes, highlights the need for robust methods that can handle such uncertainty effectively.

III. Proposed Method

In this section, compared to deterministic paradigms, we propose a probability paradigm for gaze estimation. Due to limitations in equipment and lighting conditions, captured facial images may appear occluded or blurred, resulting in inaccurate predictions. Inspired by [21], by modeling gaze uncertainty, samples can be evaluated during model training and prediction. For facial images, the uncertainty of gaze-related features can lead to a decrease in prediction accuracy, while gaze-irrelated features without an impact, such as identity information (which can reduce the model's cross domain ability). Specifically, as shown in Fig. 2, our framework is divided into four parts. Firstly, the facial image is encoded into the gaze embedding space, represented as a Gaussian distribution. Similarly, the labels are encoded into the label embedding space and aligned with the gaze features. Secondly, sampling the Gaussian distribution of gaze features can obtain gaze feature samples, thereby obtaining estimates of gaze features. The gaze features are regressed to obtain gaze estimation. In order to supervise the extraction of gaze-irrelated features, we designed a facial reconstruction process. A decoder named GDec is used to upsample gaze features and connect them with identity features, which are then fed into the face reconstruction decoder (FRDec) to reconstruct the face and supervise the extraction of identity features. The final algorithm is presented in Algorithm 1.

Algorithm 1 Operations During Each Iteration in Training

Require: Face image I_f, label g, and parameters of the network θ.

Ensure: Updated parameters θ^*

$\mathrm{F}_i = \mathrm{IFE}(I_f)$
$\mathrm{F}_\mathrm{g} = \mathrm{GFE}(I_f)$
// Uncertainty modeling
$\mu_\mathrm{g}, \Sigma_\mathrm{g} \leftarrow \mathrm{MLP}(\mathrm{F}_\mathrm{g})$ and $\mu_\mathrm{l}, \Sigma_\mathrm{l} \leftarrow \mathrm{MLP}(\mathrm{g})$
// Sampling
$z_\mathrm{g}^i = \mu_\mathrm{g} + \Sigma_\mathrm{g} \odot \epsilon_i, \epsilon_i \sim \mathcal{N}(0, I)$ *and* $i \in [1, N]$
// Estimate $\hat{\mu}_\mathrm{g}$ using the samples
$\hat{\mu}_\mathrm{g} = \frac{1}{N}\sum_{i=1}^{N} z_\mathrm{g}^i$
// Obtain $\hat{\mathrm{g}}$ via gaze regression
$\hat{\mathrm{g}} \leftarrow \mathcal{R}(\hat{\mu}_\mathrm{g})$
// Facial reconstruction
$I_r = \mathcal{F}([\![\mathrm{F_i}, \mathcal{G}(\hat{\mu}_\mathrm{g})]\!])$
According to Eq. 4, the loss of $\mathcal{L}_f$ is used to separate F_i and F_g;
According to Eq. 12, $\mathcal{L}_{kl}$ is used for aligning the embedding distribution of gaze and label;
According to Eq. 17, $\mathcal{L}_\mathrm{g}$ is used to supervise gaze embedding estimation;
According to Eq. 19, $\mathcal{L}_r$ is used for supervising facial reconstruction;
Calculate the total loss $\mathcal{L}$ by using Eq. 20;
$\theta^* \leftarrow$ Update the parameter θ by using $\mathcal{L}$;

A. Face Encoder

Following [50], facial images contain both gaze-related and gaze-irrelated features, with clear feature boundaries, and without cross. We defined two feature extractors to extract gaze-related (gaze features F_g) and gaze-irrelated (identity features F_i) information, respectively. F_i is extracted from global information and is the semantic representation of facial images. F_g is usually strongly correlated with pupil position and head posture, so extractors pay more attention to local details. ResNet18 is applied to the feature extractor. The structure of IFE is consistent with ResNet18. The framework of GFE remains consistent with [18], except that the extracted five level features as $\{f_i \in \mathbb{R}^{128}\}_{i=1_{st},2_{nd},3_{rd},4_{th},5_{th}}$ are all involved in uncertainty modeling. Hierarchical feature extraction is a strategy that progresses from coarse to fine, and finally aggregates different information through Transformer to obtain gaze features. We assume that the uncertainty modeling of gaze features follows the principle of consistency, that is, the quality of facial images not only weakens the representation of details, but also causes the blurring of global information. Therefore, we connect features of different levels and put them into the uncertainty module. This process can be formalized as:

$$f_i = \mathrm{Res}(I_f), i \in \{1_{st}, 2_{nd}, 3_{rd}, 4_{th}\} \tag{1}$$

$$f_{5_{st}} = \mathrm{T}([\![f_{1_{st}}, f_{4_{nd}}, f_{3_{rd}}, f_{4_{th}}]\!]) \tag{2}$$

$$\mathrm{F}_\mathrm{g} = [\![\{f_i\}_{i=1_{st},2_{nd},3_{rd},4_{th},5_{th}}]\!]t \tag{3}$$

where Res is the ResNet18, which inputs feature maps with scales of $64 \times 56 \times 56$, $128 \times 28 \times 28$, $256 \times 14 \times 14$, and $512 \times 7 \times 7$ into an average pooling layer and 1×1 convolutional layer to obtain four different levels of gaze features. $[\![\cdot]\!]$ denotes concatenation along channel. T is the Transformer. The cosine loss function is used to measure the distance between two sets of features. Minimizing feature distance can clearly delineate the boundaries between gaze and identity features. This process can be formulated as follows:

$$\mathrm{F_i} = \mathrm{IFE}(I_f) \tag{4}$$

$$\mathcal{L}_f = \frac{1}{2}(1 + \frac{\mathcal{A}(\mathrm{F}_i) \cdot \mathrm{F_g}}{|\mathcal{A}(\mathrm{F}_i)| \cdot |\mathrm{F_g}|}) \tag{5}$$

where I_f is the input facial image, $\mathcal{A}$ represents average pooling, and $\mathcal{L}_f$ denotes feature loss, used to measure feature distance.

B. Uncertainty Modeling

Firstly, we attempt to model the uncertainty of F_g. F_g is mapped to the probability gaze embedding space $\mathcal{Z}_\mathrm{g}$ and obtain the probability embedding z_g. In practical implementation, we represent the probability embedding as a multivariate Gaussian distribution. F_g is fed into MLP, and obtain the mean vector μ_g and covariance matrix Σ_g, which can be summarized as follows:

$$\mu_\mathrm{g} = \mathrm{MLP}(\mathrm{F_g}) \tag{6}$$

$$\Sigma_\mathrm{g} = \mathrm{MLP}(\mathrm{F_g}) \tag{7}$$

$$p(z_\mathrm{g}|I) = \mathcal{N}(z_g; \mu_\mathrm{g}, \Sigma_\mathrm{g}) \tag{8}$$

where $\mu_z \in \mathbb{R}^d$ represents the most likely feature in the gaze embedding space, and $\Sigma_\mathrm{g} \in \mathbb{R}^{d\times d}$ indicates the uncertainty of input face images.

Inspired by [51], in order to measure the accuracy of gaze embedding feature modeling, deterministic labels are also parameterized as Gaussian distributions, aligned with gaze feature embedding. Similarly, MLP is used to map labels to the label embedding space z_l. Similar to gaze embedding space, the uncertainty of labels is also modeled as follows:

$$\mu_l = \mathrm{MLP}(\mathrm{g}) \tag{9}$$

$$\Sigma_l = \mathrm{MLP}(\mathrm{g}) \tag{10}$$

$$p(g) = \mathcal{N}(z_l; \mu_l, \Sigma_l) \tag{11}$$

where g is the label. $\mu_l \in \mathbb{R}^d$ denotes the mean vector, and $\Sigma_l \in \mathbb{R}^{d\times d}$ indicates the diagonal covariance matrix. $p(\mathrm{g})$ represents the probability distribution of the label. As the covariance matrix Σ_l decreases, $p(\mathrm{g})$ degenerates into a deterministic label, while conversely, the confidence of the raw label is decreased.

In order to maintain consistency between gaze features and labels in the embedding space, we penalize the difference between the two distributions by imposing a Kullback-Leibler divergence,

$$\mathcal{L}_{kl} = D_{KL}(p(z_\mathrm{g}|I)||p(\mathrm{g})) \tag{12}$$

C. Gaze Regression

Compared to alignment in probability embedding space, gaze estimation also needs to maintain consistency in real three-dimensional space. Therefore, we attempt to regress from the gaze embedding space to obtain gaze estimation. Unlike directly using μ_g, we sample the samples from a Gaussian distribution to obtain the desired results. Note that, The samples cannot be random because this would not allow computing the gradients of the distribution parameters. Therefore, we adopted the reparameterization trick. Specifically, as follows:

$$z_g^i = \mu_g + \Sigma_g \odot \epsilon_i, \epsilon_i \sim \mathcal{N}(0, I) \; and \; i \in [1, N] \tag{13}$$

where z_g^i is the samples, $\odot$ denotes the element-wise product, and ϵ_i is a normal Gaussian noise vector. N represents the number of samples. Note that, $[z_g^1, z_g^2 \ldots, z_g^n]$ are independent, that is, $z_g^i \overset{i.i.d}{\sim} \mathcal{N}(\mu_g, \Sigma_g)$. After sampling, we apply Maximum Likelihood Estimation (MLE) to calculate the most likely mean vector $\hat{\mu}_g$ as follows:

$$\begin{aligned}
\hat{\mu}_g &= \arg\max_{\mu_g} \log p(z_g|I) \\
&= \arg\max_{\mu_g} \log P(z_g; \mu_g, \Sigma_g) \\
&= \arg\min_{\mu_g} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \left(z_g^i - \mu_g\right)^T \\
&\quad \times \Sigma_g^{-1} \left(z_g^i - \mu_g\right) + \log\left|\Sigma_g\right| - \frac{1}{\sqrt{2\pi}}
\end{aligned} \tag{14}$$

In the Eq. 14, $\left|\Sigma_g\right|$ is independent of μ_g, and $\frac{1}{\sqrt{2\pi}}$ is a constant. Therefore, the formula can be simplified to $\hat{\mu}_g = \arg\min_{\mu_g} \frac{1}{N} \sum_{i=1}^{N} \left(z_g^i - \mu_g\right)^T \Sigma_g^{-1} \left(z_g^i - \mu_g\right)$. After taking the partial derivative of μ_g in this equation, we can obtain an unbiased estimate of μ_g as follows:

$$\hat{\mu}_g = \mathbb{E}_{z_g \sim p(z_g|I)}[z_g] = \frac{1}{N} \sum_{i=1}^{N} z_g^i \tag{15}$$

Then, the MLP-based gaze regression network (GRNet) is used to obtain gaze estimation from $\hat{\mu}_g$. To further supervise the extraction of gaze features, we defined gaze loss $\mathcal{L}_g$ as follows:

$$\hat{g} = \mathcal{R}(\hat{\mu}_g) \tag{16}$$

$$\mathcal{L}_g = \|\hat{g} - g\| \tag{17}$$

where $\hat{g}$ is the predicted gaze, $\mathcal{R}$ denotes the GRNet, and g is the deterministic label.

D. Face Reconstruction

$\mathcal{L}_f$ increases the distance between gaze and identity features, thus satisfying $F_i \cap F_g = \emptyset$. The gaze features are supervised by $\mathcal{L}_g$ and $\mathcal{L}_{kl}$ and can be purified during the iteration process. However, the extraction of identity information is not supervised, and its randomness can hinder the measurement of $\mathcal{L}_f$. Therefore, we defined a facial reconstruction branch to recover facial images from gaze and identity features. $\hat{\mu}_g$ contains the most likely gaze features. However, there is ambiguity between embedding space and gaze representation. As shown in Fig. 3, gaze decoder (GDec) is used to obtain a mapping from $\hat{\mu}_g$ to gaze representation, aligning with identity features. By connecting gaze and identity features, the face image is restored by feeding it into the face reconstruction decoder (FRDec), and a reconstruction loss ($\mathcal{L}_r$) optimization network is defined. This branch is used to supervise the extraction of identity features, and the entire process can be formalized as follows:

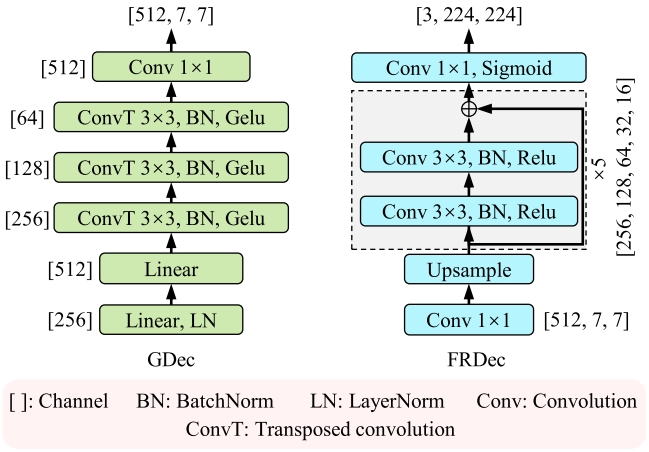

Fig. 3. The structures of GDec and FRDec.

$$I_r = \mathcal{F}([\![F_i, \mathcal{G}(\hat{\mu}_g)]\!]) \tag{18}$$

$$\mathcal{L}_r = \|I_r - I_f\| \tag{19}$$

where $\mathcal{F}$ represents FRDec, I_r is reconstructed facial image, and $\mathcal{G}$ denotes gaze decoder (GDec).

E. Loss and Uncertainty

During the training process, we optimized the model, which includes four items: feature loss $\mathcal{L}_f$, Gaussian alignment loss $\mathcal{L}_{kl}$, gaze loss $\mathcal{L}_g$, and reconstruction loss $\mathcal{L}_r$. Therefore, the overall loss function is as follows:

$$\mathcal{L} = \alpha_1 \mathcal{L}_f + \alpha_2 \mathcal{L}_{kl} + \alpha_3 \mathcal{L}_g + \alpha_4 \mathcal{L}_r \tag{20}$$

where α_1, α_2, α_3, and α_4 are the scaling factor used to balance the weights of the losses. In this paper, α_2 is set to 0.001, while the others are set to 1.

The covariance matrix of gaze embedding is used to calculate uncertainty. Since Σ_g is a diagonal covariance matrix, we use harmonic mean to measure uncertainty as follows:

$$U = \frac{D}{tr(\Sigma_g^{-1})} \tag{21}$$

where U denotes the uncertainty estimation and D is the dimension of gaze embeddings z_g. $tr(\cdot)$ denotes the trace of the matrix.

IV. Experiments

A. Datasets

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 × 224. The dataset contains 28 subjects. We divide the dataset into five subsets based on the subjects and perform five-fold cross-validation.

2) IV: IV contains 44,705 images of 125 subjects. The groundtruth includes head pose, gaze direction, gaze zone, facial attributes, and so on. To perform within-dataset evaluation, we divide the dataset into three subsets based on subjects. We follow the division of dataset from [18], and perform three-fold cross-validation.

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 0.001. 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 [37]. In addition, following [18], the average accuracy is used to describe the error distribution. The average precision 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.

D. Comparison With Methods

The experimental results demonstrate the effectiveness of our proposed method in improving gaze estimation accuracy. On the IV dataset, our method achieves an error of 6.90°, which is an improvement of approximately 1.85% over the lowest baseline error of 7.03° (GazeDPTR). Additionally, our method shows significant improvements in average precision at higher angular thresholds: for $< 4°$, 36.5% compared to 34.6%; for $< 6°$, 52.3% compared to 49.8%; and for $< 8°$, 70.8% compared to 67.2%. On the LBW dataset, our method achieves an error of 5.95°, an improvement of approximately 2.14% over the lowest baseline error of 6.08° (GazePTR). The average precision improvements for $< 6°$ and $< 8°$ thresholds are notable: 63.4% compared to 61.5% and 78.3% compared to 76.4%, respectively. These results suggest that our probabilistic framework effectively captures and models input and annotation uncertainties, leading to enhanced overall performance, particularly at larger angular thresholds. However, the accuracy of $< 2°$ and $< 4°$ slightly decreased, which may be due to samples with partially occluded eyes having gaze consistent with head posture, resulting in higher accuracy. However, in our probability embedding, its uncertainty is relatively high. Our method compresses the number of samples with significant errors, which is more meaningful for driver monitoring that allows sacrificing some accuracy.

E. The Impact of Face Accessories

The experimental results presented in Table II indicate the accuracy of gaze estimation under various conditions, specifically evaluating the impact of wearing glasses, masks, and sunglasses. Our method demonstrates improved performance across most conditions compared to other baseline methods.

For the glasses condition: - With glasses: Our method achieves an error of 7.01°, showing a slight improvement over the best baseline method, GazeDPTR*, which has an error of 7.10°. - Without glasses: Our method achieves an error of 6.90°, marginally better than GazePTR's 6.90°.

For the mask condition: - With a mask: Our method achieves an error of 7.41°, which is better than the best baseline method, GazeDPTR*, with an error of 7.61°. - Without a mask: Our method achieves an error of 6.81°, outperforming GazeDPTR* which has an error of 6.88°.

For the sunglasses condition: - With sunglasses: Our method achieves an error of 16.34°, which is an improvement over the best baseline method, XGaze, which has an error of 15.15°.

Overall, our method shows a consistent improvement in accuracy, particularly in the presence of glasses and masks, and maintains competitive performance with sunglasses. This

TABLE I
THE COMPARISON OF GAZE ESTIMATION. IN ORDER TO MAINTAIN INPUT CONSISTENCY, * INDICATES THAT GAZEDPTR WITHOUT CAMERA POSE

	Method	Error	Average Precision			
			$< 2°$	$< 4°$	$< 6°$	$< 8°$
IV	FullFace	13.67°	2.3%	8.8%	17.8%	28%
	DGW	8.82°	6.6%	21.7%	38.1%	53.2%
	Gaze360	8.15°	9.2%	27.3%	44.6%	58.9%
	ResNet18	7.48°	14.2%	31.1%	46.7%	63.1%
	GazeTR	7.33°	17%	32.8%	47.5%	64.7%
	XGaze	7.06°	11.7%	32.7%	51.5%	66.7%
	GazePTR	7.04°	17.6%	34%	49.3%	66.7%
	GazeDPTR*	7.03°	**18.3%**	34.6%	49.8%	67.2%
	Ours	**6.90°**	14.3%	**36.5%**	**52.3%**	**70.8%**
LBW	FullFace	7.16°	8.4%	25.8%	47.1%	60.9%
	Gaze360	7.01°	10.3%	29.7%	53.3%	65.2%
	ResNet18	6.16°	11.4%	36.2%	59.8%	76.4%
	GazeTR	6.20°	11.5%	35.9%	59.4%	75.9%
	XGaze	6.11°	**12.5%**	36.3%	61.5%	75.7%
	GazePTR	6.08°	12.3%	**38.1%**	61.3%	76.1%
	Ours	**5.95°**	12.2%	37.8%	**63.4%**	**78.3%**

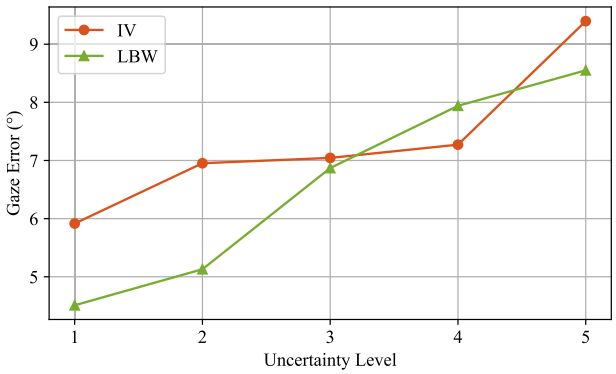

Fig. 4. The mean gaze estimation accuracy in different uncertainty level.

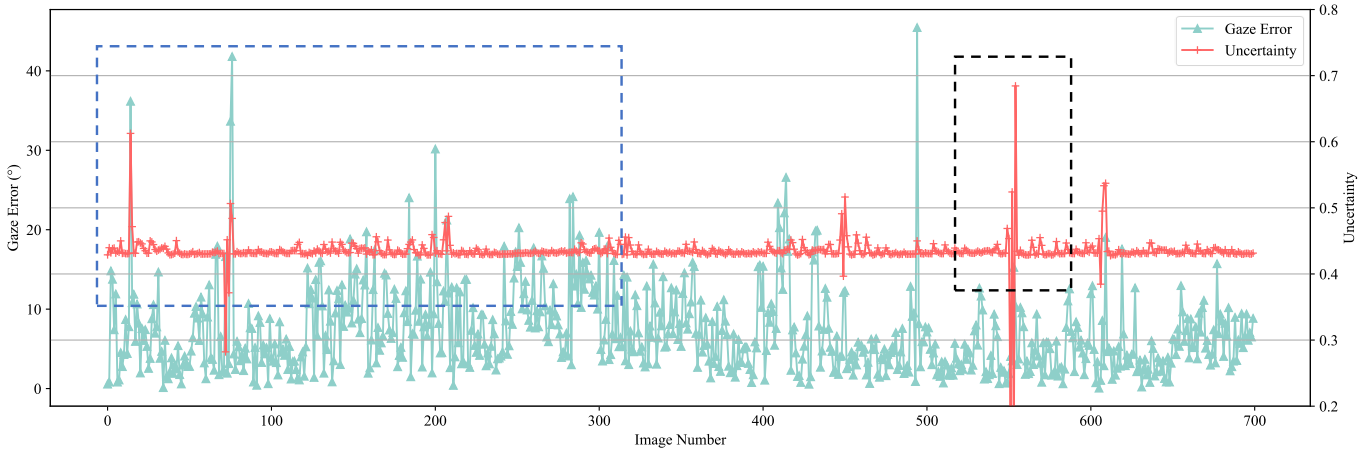

Fig. 5. The relationship between uncertainty and gaze error.

TABLE II

THE COMPARISON OF GAZE ESTIMATION FOR DIFFERENT FACE ACCESSORIES. IN ORDER TO MAINTAIN INPUT CONSISTENCY, * INDICATES THAT GAZEDPTR WITHOUT CAMERA POSE

Method	Glasses		Mask		Sunglasses
	with	w.o.	with	w.o.	with
FullFace	14.43°	12.40°	15.20°	13.35°	21.39°
DGW	9.20°	8.19°	9.43°	8.69°	17.43°
Gaze360	8.30°	7.91°	8.95°	7.99°	17.99°
ResNet18	7.59°	7.30°	8.37°	7.30°	16.50°
GazeTR	7.40°	7.22°	8.12°	7.17°	17.49°
XGaze	7.07°	7.03°	7.80°	6.90°	15.15°
GazePTR	7.13°	6.90°	7.78°	6.89°	16.54°
GazeDPTR*	7.10°	**6.81°**	7.61°	6.88°	16.51°
Ours	**7.01°**	6.90°	**7.41°**	**6.81°**	**16.34°**

indicates that our probabilistic framework effectively models input uncertainty under various occlusion conditions, leading to enhanced robustness in gaze estimation. The improvements, though sometimes marginal, suggest that our approach can handle real-world scenarios with different facial accessories better than the baseline methods.

F. Gaze Error Distribution

Uncertainty is used to describe the quality of input samples. Similar to [21], uncertainty is non-uniformly divided into 5 levels. The average error of different levels is depicted in Fig. 4. We observed that as the level of uncertainty increases, the gaze error gradually increases. This result indicates a certain correlation between uncertainty and gaze error, and is consistent with our hypothesis that uncertainty can be used to describe the average error caused by eye invisible.

G. Discussion on Uncertainty

This experiment is mainly used to observe whether gaze error and uncertainty are completely correlated. However, according to our experimental results, this is not absolute. We perform random sampling in IVGaze to obtain its gaze error and uncertainty values. As shown in Fig. 5, the blue dashed box represents consistent large errors and uncertainties, indicating that as key facial features are obscured, the performance of the gaze estimator will decrease. However, uncertainty and gaze error are not linearly correlated. In the black dashed box, when the uncertainty is high, the gaze error is low. This individual ambiguity is easily overlooked in terms of average error, but uncertainty still represents the average quality of similar sample sets.

Fig. 6. Classification results of uncertainty levels. Each image is labeled with 'Level-Gaze Error'.

As shown in Fig. 6. For example, '1-0.23' indicates an uncertainty level of 1 and a gaze error of 0.23°. We find a certain correlation between uncertainty and eye occlusion. With the obstruction of the eyes, the uncertainty level will gradually increase. For level 4, it also shows that the pupils are invisible, so the difficulty of evaluation is similar. However, different gaze errors once again confirm that uncertainty is not completely correlated with gaze errors. We speculate that when the eyes are invisible, due to the strong correlation between head posture and gaze, the gaze estimator will mainly extract head posture features to output gaze. Therefore, when the gaze direction is consistent with the head posture and the eyes are obstructed, higher accuracy can also be obtained. However, the improvement brought by uncertainty to the model and the description of the input image are equally valuable.

H. Performance Distribution

We demonstrated the performance distribution of the proposed method in different head posture ranges. Due to the

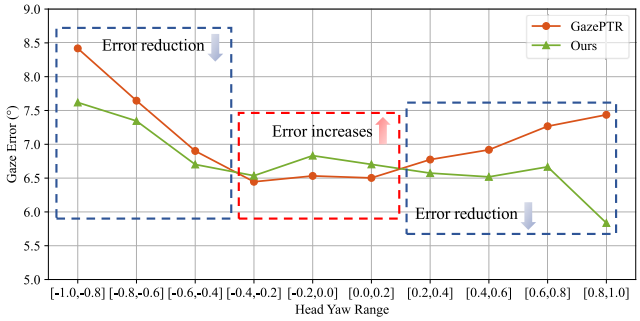

Fig. 7. The mean gaze estimation accuracy in different head posture ranges. The head yaw angle is expressed in radians and ranges from $[-1, 1]$.

imbalance in the distribution of head pose yaw angles in the IV dataset, we present the average prediction error within the range of [-1, 1] to intuitively compare the performance of UnMoDE under symmetric head poses. As shown in Fig. 7, when the head deflection is small, the introduction of uncertainty will lead to a slight decrease in performance. This is because in the IV dataset, samples with eye occlusion and blurring dominate, posing a challenge to the performance of gaze estimators. During smaller head rotations, people are usually accustomed to turning their eyes to modify their gaze, resulting in a misalignment of head posture and gaze direction. Therefore, in the face of this situation, the introduction of uncertainty disturbs the estimation of gaze embedding. When the rotation angle of the head is large, the pupil is usually located at the center of the eye, and its gaze is close to the direction of the head. Therefore, modeling uncertainty can weaken the impact of facial occlusion. In summary, UnMoDE achieves consistency improvement for large head pose challenges while sacrificing a small amount of accuracy.

Compared to the range of [-1.0, -0.8], why does the error of UnMoDE decrease more significantly within the range of [0.8, 1.0]? Since UnMoDE enhances the model's gaze estimation in extreme head poses, and as shown in Fig. 8, there are more samples in the IV dataset with head poses within the range of [0.8, 1.0], leading to a faster reduction in the average error. The difference between these two ranges once again confirms the robustness of UnMoDE's framework design for larger head rotations.

I. Average Error Distribution

A large gaze angle can cause deformation of the pupils or rotation of the head, thereby increasing the difficulty of the model's prediction. Fig. 8 shows the average error distribution of GazePTR and UnMoDE at different gaze angles. In IV, the red dashed box represents a larger yaw angle, and the yellow dashed box denotes a larger pitch angle. In both cases, the blue area of UnMoDE is larger, indicating higher prediction accuracy. In LBW, the consistency results demonstrate the power of UnMoDE.

J. Visualization of Gaze Features

UnMoDE proposes a disentanglement framework for obtaining gaze related features. To verify the necessity of

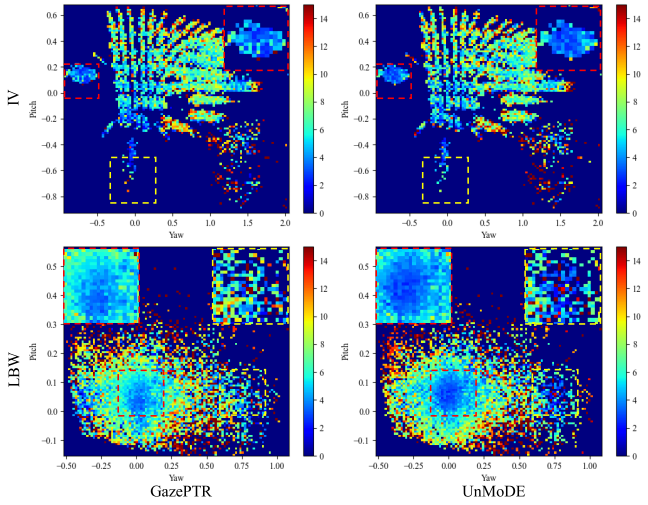

Fig. 8. Average error distribution of different gaze angles. The colorbar represents the mean error, and the unit is degree. The x and y axes respectively represent the pitch and yaw angles of the gaze ground truth.

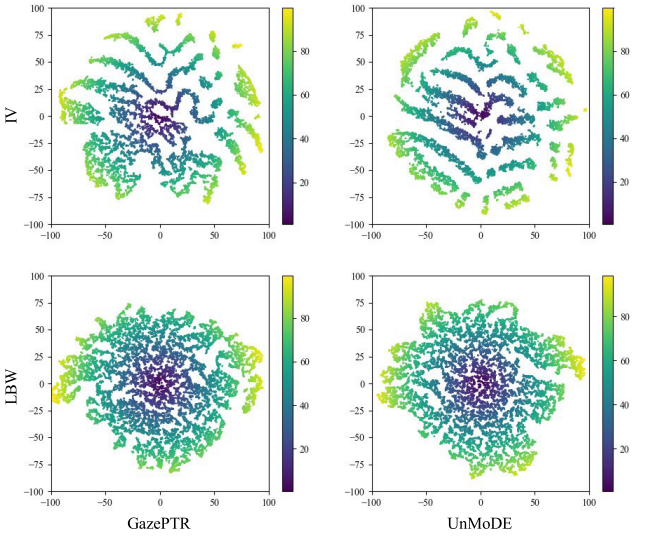

Fig. 9. Visualization of gaze features in different test subset. The color of the dots indicates the respective distance of the samples to the camera sensor.

feature disentanglement, we visualize the gaze features of GazePTR and UnMoDE using t-SNE [52] and embedded them for visualization by principal component analysis (PCA). The results are shown in Fig. 9. For LBW, GazePTR and UnMoDE all obtained smooth embedding distributions. In IV, as shown in Fig. 8, the label distribution appears as a bar shape, and the gaze embedding distribution extracted by GazePTR and UnMoDE is consistent with the label. However, due to the mixing of identity information, GazePTR has experienced confusion in some areas (Bottom left corner). Due to the separate design of the disentanglement module, UnMoDE effectively learns consistent and task related features.

K. Visualization of Gaze and Identity Flow

According to the definition in [50], gaze and identity features are non-intersect. We visualize the gaze (F_g) and identity

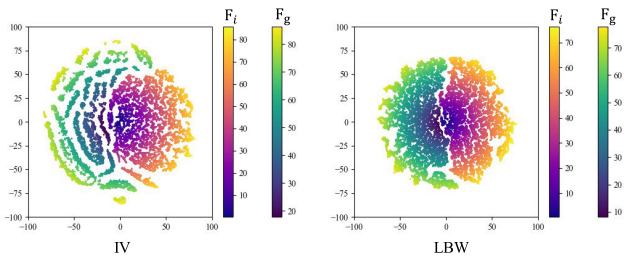

Fig. 10. Visualization of gaze and identity features in different test subset. The color of the dots indicates the respective distance of the samples to the camera sensor.

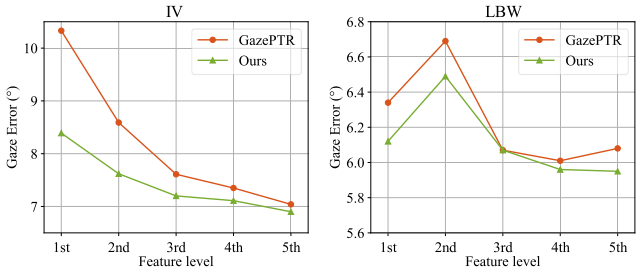

Fig. 11. The impact of uncertainty on the performance at each level of the feature.

(F_i) flow, and t-SNE with the same settings is used to map F_g and F_i to a two-dimensional space, as shown in Fig. 10. In two datasets, identity and gaze features are completely isolated, indicating that UnMoDE divides the embedding space into two parts, corresponding to gaze and identity information, respectively. This result is consistent with the previous definition [50]. In addition, in IV, the consistency between the stripe distribution of gaze and the label distribution validates the effective extraction of gaze embeddings. In summary, UnMoDE achieves the disentanglement of gaze flow and identity flow.

L. Ablation Study

1) The Impact of Uncertainty: To verify the impact of introducing uncertainty on the gaze feature extractor, we conducted ablation studies. We extracted multiple layers of features from the gaze feature extractor and regressed them to obtain the gaze direction. Gaze errors at different levels across different datasets are shown in Fig. 11. The results indicate that compared to Gaze PTR with the same architecture, the multi-level features have shown improved consistency. This suggests that uncertainty aids the network in feature extraction. In the LBW dataset, the highest-level features in GazePTR cause performance degradation. The proposed method achieved excellent results with the highest-level features by more accurately extracting gaze features. Additionally, in the IV dataset, the model's predictive capability increased with the feature level, as anticipated. However, we observe a decline in the performance of the 2nd-level features within the LBW dataset, which contradicts intuition. Given that a consistent anomaly is observed in both GazePTR and UnMoDE, we surmise that this phenomenon does not stem from the uncertainty framework or the feature disentanglement module. Therefore, we speculate that the anomaly originates from the LBW dataset itself.

2) Evaluation of Contributions: To validate our contribution, we combine the designed loss functions and verify their performance individually. The proposed gaze feature extractor (GFE) is evaluated using three different backbones: GazePTR, ResNet18, and GazeTR. The baseline uses GFE to obtain gaze information and regress to predict the gaze direction, with training supervised by $\mathcal{L}_g$. From the experimental results shown in Table III, we draw four important conclusions. (1) The combination of $\mathcal{L}_r$ and $\mathcal{L}_f$ better purifies gaze features. (2) Representing gaze features as probabilistic embeddings and using KL divergence $\mathcal{L}_{kl}$ for supervision improves performance compared to using only $\mathcal{L}_g$ for supervision. (3) The combination of all loss functions achieved consistent performance improvements, indicating that uncertainty modeling for gaze features, rather than semantic information of faces, better enhances model performance. (4) Uncertainty modeling ($\mathcal{L}_{kl}$) improves performance across different backbones, suggesting that the introduction of uncertainty enhances the model's generalization.

TABLE III
COMPARISON OF DIFFERENT LOSS FUNCTIONS IN DIFFERENT BACKBONES

	GazePTR		ResNet18		GazeTR	
	IV	LBW	IV	LBW	IV	LBW
Baseline	7.05	6.07	7.41	6.16	7.33	6.20
$+\mathcal{L}_r$	6.95	6.02	7.31	6.08	7.27	6.14
$+\mathcal{L}_f$	6.97	6.03	7.39	6.11	7.25	6.17
$+\mathcal{L}_r+\mathcal{L}_f$	6.93	6.00	7.35	6.07	7.20	6.09
$+\mathcal{L}_{kl}$	6.98	6.01	7.33	6.07	7.19	6.09
$+\mathcal{L}_{kl}+\mathcal{L}_r$	7.05	6.05	7.45	6.13	7.27	6.13
$+\mathcal{L}_r+\mathcal{L}_f+\mathcal{L}_{kl}$	**6.90**	**5.95**	**7.29**	**6.01**	**7.11**	**6.01**

TABLE IV
COMPARISON OF DIFFERENT NUMBER OF SAMPLES

	Number of samples (N)					
	16	32	64	128	256	512
IV	7.01	6.97	**6.90**	6.93	6.91	6.97
LBW	6.07	**5.94**	5.95	5.96	5.95	5.97

3) Parameter Settings: We conducted an ablation study on the selection of parameters. First, we test the number of samples for embedding gaze distribution. The experimental results are shown in Table IV. When the number of samples is small, the estimation of $\hat{\mu}_g$ is random, leading to fluctuations in model performance. As the number of samples increases, performance improves. We observed that when $N = 512$, model performance degraded on the IV dataset. We speculate that for datasets with significant eye occlusion, rough estimation may disturb the embedding, leading to better performance. Thus, N is set to 64.

For the embedding dimension d, experimental results are shown in Table V. Smaller dimensions limit the representational power of gaze and label embeddings, while larger dimensions can lead to resource waste. According to the experimental results, the most suitable embedding size is 128.

4) Uncertainty Modeling: Uncertainty arises from the quality of the input image. As a result, features extracted from

TABLE V
COMPARISON OF DIFFERENT EMBEDDING DIMENSIONS

	Embedding Dimensions (d)					
	16	32	64	128	256	512
IV	7.07	7.01	6.93	**6.90**	6.91	6.90
LBW	6.05	6.03	5.96	**5.95**	5.96	5.95

TABLE VI
ANALYSIS OF UNCERTAINTY MODELING FOR DIFFERENT FEATURE LEVELS

	IV			LBW		
	Mean	$< 4°$	$< 8°$	Mean	$< 4°$	$< 8°$
ResNet18	7.31	32.5%	65.3%	6.11	37.1%	76.9%
GazePTR	7.01	34.7%	67.5%	6.07	**38.1%**	77.5%
Ours	**6.90**	**36.5%**	**70.8%**	**5.95**	37.8%	**78.3%**

facial images at different levels exhibit consistent uncertainty. We conducted experiments to test this hypothesis, modeling the uncertainty of features at different levels (with GFE replaced), and observed the changes in the errors. The experimental results are shown in Table VI. Compared to ResNet18, GazePTR consistently achieves performance gains by aggregating four levels of features using a Transformer. Our method performs uncertainty modeling at five levels. Compared to GazePTR, the average error is reduced by 0.11° and 0.12° on the IV and LBW datasets, respectively. For $< 8°$, our method successfully reduces the uncertainty in predictions, allowing more samples to fall within a smaller error range. In summary, image quality significantly affects the entire process of feature extraction.

V. DISCUSSION

Compared to model-based gaze estimation, appearance-based methods are more intensively studied due to their non-intrusive data acquisition. However, image quality limits their accuracy. In vehicles, cameras are affected by factors such as illumination, masks, and motion blur, leading to low-quality data. The proposed method can partially mitigate the impact of uncertainty on model predictions, but it cannot completely resolve the issue. Specialized commercial sensor devices can overcome these challenges. However, their limited working distance hinders widespread application. With the rapid development of multimodal fusion, data from a single sensor is no longer sufficient for precise driver monitoring. In the future, we plan to integrate multi-sensor data, such as infrared, laser, and scene data, to address the shortcomings of both methods and develop a more robust driver monitoring model.

VI. CONCLUSION

This paper introduces the concept of uncertainty into driver gaze estimation. This approach uses uncertainty to describe sample quality and estimate the most probable gaze embedding, enhancing the model's generalization. Specifically, our framework, UnMoDE, consists of four modules: face encoding, uncertainty modeling, gaze regression, and face reconstruction. Facial images are first processed by asymmetric dual-branch encoders to extract gaze-related and identity-related features. Feature separation is supervised using cosine distance to ensure distinct representation of gaze-related and identity-related information. Gaze features and labels are projected onto an embedding space using a multi-layer perceptron (MLP) and aligned with Kullback-Leibler (KL) divergence. A random sampling strategy is used to obtain representations from Gaussian distributions, estimating the expected gaze embedding. We hypothesize that the expectation of the Gaussian distribution contains all relevant gaze information, which can be obtained through regression. The estimated gaze embedding is projected back into the feature space through a gaze decoder (GDec) and combined with identity information for face reconstruction. Experimental results show that introducing uncertainty effectively reduces samples with significant errors, making our method more suitable for intelligent vehicle applications.

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] A. Rangesh et al., "Exploring the situational awareness of humans inside autonomous vehicles," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 190–197.

[3] 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.

[4] M. F. Land, "Eye movements and the control of actions in everyday life," *Prog. Retinal Eye Res.*, vol. 25, no. 3, pp. 296–324, May 2006.

[5] A. R. Aftab, "Multimodal driver interaction with gesture, gaze and speech," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2019, pp. 487–492.

[6] P. K. Murali, M. Kaboli, and R. Dahiya, "Intelligent in-vehicle interaction technologies," *Adv. Intell. Syst.*, vol. 4, no. 2, Feb. 2022, Art. no. 2100122.

[7] A. C. Hayley, B. Shiferaw, B. Aitken, F. Vinckenbosch, T. L. Brown, and L. A. Downey, "Driver monitoring systems (DMS): The future of impaired driving management?" *Traffic Injury Prevention*, vol. 22, no. 4, pp. 313–317, May 2021.

[8] M. Khanv and S. Lee, "A comprehensive survey of driving monitoring and assistance systems," *Sensors*, vol. 19, no. 11, p. 2574, 2019.

[9] Z. Zhu and Q. Ji, "Eye gaze tracking under natural head movements," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 918–923.

[10] 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.

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

[12] K. Krafka et al., "Eye tracking for everyone," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2176–2184.

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

[14] 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.

[15] 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. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Cham, Switzerland: Springer, May 2020, pp. 365–381.

[16] 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.

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

[18] 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.

[19] 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.* Cham, Switzerland: Springer, Jan. 2022, pp. 126–142.

[20] S. Zhang, Y. Tian, Y. Zhang, M. Tian, and Y. Huang, "Domain-consistent and uncertainty-aware network for generalizable gaze estimation," *IEEE Trans. Multimedia*, vol. 26, pp. 6996–7011, 2024.

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

[22] L. Fridman, J. Lee, B. Reimer, and T. Victor, "'Owl' and 'Lizard': Patterns of head pose and eye pose in driver gaze classification," *IET Comput. Vis.*, vol. 10, no. 4, pp. 308–314, Jun. 2016.

[23] 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.

[24] J. Li and S. Li, "Gaze estimation from color image based on the eye model with known head pose," *IEEE Trans. Hum.-Mach. Syst.*, vol. 46, no. 3, pp. 414–423, Jun. 2016.

[25] 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.

[26] G. Liu, Y. Yu, K. A. F. Mora, and J.-M. Odobez, "A differential approach for gaze estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 1092–1099, Mar. 2021.

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

[28] 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.

[29] 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.

[30] X. Wang et al., "Dual regression-enhanced gaze target detection in the wild," *IEEE Trans. Cybern.*, vol. 54, no. 1, pp. 219–229, Jan. 2023.

[31] X. Zhang, "Appearance-based gaze estimation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Aug. 2015, pp. 4511–4520.

[32] Y. Cheng and F. Lu, "DVGaze: Dual-view gaze estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Aug. 2023, pp. 20632–20641.

[33] 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.

[34] Y. Cheng, F. Lu, and X. Zhang, "Appearance-based gaze estimation via evaluation-guided asymmetric regression," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 100–115.

[35] Y. Bao, Y. Liu, H. Wang, and F. Lu, "Generalizing gaze estimation with rotation consistency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4197–4206.

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

[37] 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.

[38] 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.

[39] P. Biswas, "Appearance-based gaze estimation using attention and difference mechanism," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3143–3152.

[40] M. Xu, H. Wang, and F. Lu, "Learning a generalized gaze estimator from gaze-consistent feature," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, pp. 3027–3035.

[41] Y. Yu and J.-M. Odobez, "Unsupervised representation learning for gaze estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7314–7324.

[42] S. Ghosh, M. Hayat, A. Dhall, and J. Knibbe, "MTGLS: Multi-task gaze estimation with limited supervision," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Oct. 2022, pp. 3223–3234.

[43] S. Wang and Y. Huang, "Suppressing uncertainty in gaze estimation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, 2024, pp. 5581–5589.

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

[45] 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.

[46] L. Yang, K. Dong, A. J. Dmitruk, J. Brighton, and Y. Zhao, "A dual-cameras-based driver gaze mapping system with an application on non-driving activities monitoring," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 10, pp. 4318–4327, Oct. 2020.

[47] 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)*, Qingdao, China, Oct. 2014, pp. 988–994.

[48] 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*, Glasgow, U.K. Cham, Switzerland: Springer, 2020, pp. 387–405.

[49] M. Lundgren, L. Hammarstrand, and T. McKelvey, "Driver-gaze zone estimation using Bayesian filtering and Gaussian processes," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 10, pp. 2739–2750, Oct. 2016.

[50] 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.

[51] A. Rahman, J. M. J. Valanarasu, I. Hacihaliloglu, and V. M. Patel, "Ambiguous medical image segmentation using diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 11536–11546.

[52] L. Van der Maaten and G. E. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, Jan. 2008.

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 the School of Computer Science and Engineering, Sun Yat-sen University, China. His research interests include computer vision and medical imaging.

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 (November 2021–November 2022). He is currently a joint Post-Doctoral Fellow with the School of Computer Science and Engineering, Sun Yat-sen University. 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 the University of Leiden in 2005, and the Ph.D. degree from ETH Zürich in 2010. He joined Sun Yat-sen University as a Professor in 2015. 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 a numbers of conferences.