DashGaze: Driver Gaze Through Dashcam

Thrupthi Ann John[1], Vineeth N Balasubramanian[2] and C. V. Jawahar[3]

***Abstract*— Driver gaze monitoring is crucial for road safety, but existing methods rely on expensive, cumbersome technologies like wearable eye trackers or fixed-camera setups. To address this, we propose a low-cost approach using dashcams to capture driver gaze data. We introduce DashGaze, a large-scale dataset for training appearance-based gaze estimation models, featuring over 900,000 frames collected over 10 hours with 28 unique drivers. DashGaze includes synchronized views of the road, driver, and driver's egocentric perspective, along with the driver's gaze in both the driver and ego views. We also present DashGazeNet, a baseline model that generalizes well to unseen drivers and diverse conditions, achieving gaze angle errors within 8.5°and gaze location errors within 225 pixels. Our code and data are available at https://github.com/ThrupthiAnn/DashGaze**

I. Introduction

Driver gaze monitoring has emerged as a critical component in ensuring driving safety, with applications in reducing distraction-related risks and enhancing situational awareness. Advanced Driver Assistance Systems (ADAS) can leverage gaze data to assess driver attention and fatigue levels, prompting timely alerts for potentially hazardous events [1]. However, the practical implementation of driver gaze estimation has been hindered by the reliance on costly and cumbersome solutions, such as wearable eye trackers or calibrated camera setups. This work proposes an alternative approach, utilizing dashcams as a low-cost, easily deployable means for capturing driver gaze data, thereby addressing the need for more accessible and practical gaze monitoring systems.

To overcome the limitations of existing gaze monitoring systems, we propose a novel data collection framework that transfers the gaze from an eye tracker to the dashcam view. Using our collection framework, we propose DashGaze, a large-scale dataset specifically designed for training appearance-based models using only a single dashcam. Unlike previous dashcam datasets like DGaze [2], DashGaze captures the complexities of real-world driving scenarios with 28 unique drivers, totaling over 900,000 frames. The dataset includes synchronized views of the road scene, driver scene, and egocentric perspective, enabling correlation of driver gaze with driving context and behavior analysis. The dataset provides gaze ground truth in egocentric and road views, along with blink, fixation, and IMU data. Captured on Indian roads, DashGaze is highly realistic, encompassing diverse driver behaviors, weather, road, and traffic conditions.

Fig. 1: Is it possible to capture driver gaze through a dashcam? In this work, we present a large-scale naturalistic driver gaze dataset and propose a method for dashcam-based driver gaze estimation.

that are adaptable to diverse driving conditions.

We present DashGazeNet, our baseline dashcam-based model for driver gaze estimation. Our experiments show that this model, trained on the DashGaze dataset, performs well on unseen drivers and camera positions and generalizes effectively to the real setting, indicating the dataset's robustness. Our model is able to estimate the gaze angle within 8.5°and gaze location within 225 pixels. Additionally, our model compares favourably with popular gaze estimation algorithms. Our contributions are as follows:

- We propose a novel data collection framework to collect gaze data using dashcams. Using our framework, introduce Dash-Gaze – a novel, large-scale dataset for appearance-based gaze mapping on the road. To our knowledge, this is the largest dataset collected under real driving conditions suitable for dashcam-based driver gaze estimation.
- We introduce DashGazeNet, a baseline model for predicting driver gaze on the road from facial input. We describe calibration procedures to adapt the model to various drivers and dashcam positions. Experimental results demonstrate the generalizability of our dataset, exhibiting minimal domain gap relative to real-world conditions.

Our data and method code will be released on acceptance. We will also publicly release our entire framework's code on acceptance, for wide usage as well as extension to newer hardware.

*This work was supported by IHub-Data, IIIT Hyderabad

[1]Thrupthi Ann John and C. V. Jawahar are with the Centre for Visual Information Technology (CVIT), International Institute of Information Technology, Hyderabad, India. Email: thrupthi.ann@research.iiit.ac.in

[2]Vineeth N Balasubramanian is with the Faculty of the Department of Computer Science and Engineering, Indian Institute of Technology, Hyderabad, India

Dataset	Driving	Dashcam	Natural	Global South	Continuous	Road View	Driver View	Ego View
Gaze360 [3]	✘	✘	✔	✘	✘	✘	✘	✘
Rt-GENE [4]	✘	✘	✔	✘	✔	✘	✘	✘
MoGAZE [5]	✘	✘	✘	✘	✔	✘	✘	✘
HUMBI [6]	✘	✘	✘	✘	✔	✘	✘	✘
DG-Unicamp [7]	✔	✘	✔	✘	✘	✘	✔	✘
DMD [8]	✔	✘	✔	✘	✘	✘	✔	✘
DGW [9]	✔	✘	✔	✘	✘	✘	✔	✘
AutoPOSE [10]	✔	✘	✘	✘	✘	✘	✔	✘
LISA [11]	✔	✘	✔	✘	✘	✔	✔	✘
MDM [12]	✔	✘	✔	✘	✘	✔	✔	✘
Dr(eye)ve [13]	✔	✘	✔	✘	✔	✔	✘	✘
LBW [14]	✔	✘	✔	✘	✔	✔	✔	✘
DGAZE [2]	✔	✔	✘	✔	✔	✔	✔	✘
DashGaze	✔	✔	✔	✔	✔	✔	✔	✔

TABLE I: Comparison of the DashGaze dataset with other gaze datasets on the following parameters: 1. Is it a driving dataset; 2. Is dashcam the modality of capture; 3. Does the data represent natural behaviour or is it collected in simulation? 4. Does the dataset capture unstructured road events typical to the Global South? 6. Is the gaze ground truth continuous or categorical?; 7. Is the road view available?; 8. Is the driver view available?; 9. Is the driver's ego view available?; Evidently, our dataset scores a 'Yes' on all these parameters.

II. Related Works

Gaze Datasets. The task of annotating eye gaze direction based solely on visual cues poses challenges for human annotators. To address this issue, most available gaze datasets are typically compiled using eye-trackers [4], [15], [16] or by instructing subjects to focus on predetermined gaze targets [3], [5], [17]. These targets may be points on a screen [17], [18], [19], [20], or objects in 3D space [3], [5], [21]. NVGaze [22] was collected by asking users to look at targets in a VR headset. The use of eye trackers in gaze data collection has the advantage of providing unconstrained behavioral data. On the other hand, object targets allow the face to remain unobstructed by the eye tracker, facilitating the training of appearance-based gaze models. Rt-GENE [4] combines the benefits of both methods by collecting data using an eye tracker and subsequently inpainting the eye tracker.

Driver Gaze Datasets. Several datasets have also been collected to particularly study driver behaviors, activities, and attention [8], [12], [23], [24], [25], [26]. Our focus is on datasets designed for gaze estimation. Few large-scale datasets concentrate on the task of "gaze zone prediction," [9], [8], [7], [11] where the interior of the car is divided into fixed "zones," such as the "rearview mirror" or "dashboard," and the goal is to predict which zone the driver's gaze falls on at any given time. As point-annotation of driver gaze is challenging, some datasets have used markers inside or outside the car [12], while others have been collected using simulators [10], [2] or parked cars [12], [9]. However, these methods do not capture realistic driving behavior. The Dr(eye)ve [13] and LBW [14] datasets address this issue by using eye trackers to collect true behavioral data while the driver is naturally driving on roads.

Appearance-based gaze estimation requires a dataset with annotated views of the driver and road. Two existing datasets, Look-Both-Ways (LBW) [14] and DGAZE [2], share similarities with ours but differ in their collection methods. LBW utilizes multiple calibrated sensors including stereo cameras and RGB-D cameras to capture facial RGB-D video , driving scene stereo videos, and ground-truth gaze from eye-tracking glasses to estimate the driver's visual focus of attention. Unlike the multi-sensor setup of LBW, our approach employs a single, uncalibrated dashcam. DGAZE uses a single dashcam setup, but the dataset is collected in simulation by instructing users to look at annotated points on a road video, and thus it does not capture realistic driving behaviour and is limited in scope. Our dataset improves on DGAZE by capturing real-world driving behavior in diverse situations on actual roads.

Driver Gaze Estimation. Methods for driver gaze estimation have been proposed in earlier literature focusing on the task of driver gaze zone estimation, where the driver's gaze is classified based on which specific 'gaze zone' it belongs to [27], [28], [29]. In some studies, depth information from RGB-D cameras has been utilized effectively for accurate gaze zone estimation [30], [31]. To address the issue of identity bias, Yu et al. [32] proposed a multimodal approach that utilized geometric features. Context-aware methods that utilize cues from the driver's environment and face for gaze zone prediction have also been proposed. For instance, Stappen et al. [33] introduced 'X-AWARE', a context-aware approach. Additionally, Yuan et al. [34] used a domain prior of typical gaze patterns to auto-calibrate gaze zone estimation. In contrast to gaze zone estimation, some works aim to predict the continuous driver gaze [35], [36], [37], [38]. Probabilistic models for gaze estimation using the driver's head pose have also been proposed by Shirpour et al. [39] and Jha et al. [40].

Recently, a few methods have been proposed for detecting driver gaze using multiple sensors. Some of these sensors include infra-red cameras [41], RGB-D cameras [30], and dual cameras [38]. Lemley et al. [42] proposed a low-cost solution for driver gaze estimation using noisy cameras. Sonom-Ochir et al. [43] proposed a dashcam-based approach that utilizes an SVM classifier on features extracted from Convolutional Neural Networks (CNNs). To address the challenge of occlusion due to eyeglasses, Rangesh et al. [44] used IR cameras and a gaze-preserving cycle Generative Adversarial Network (GAN) to remove eyeglasses and estimate the driver's gaze accurately. Unlike these methods, our gaze estimation algorithm detects continuous gaze targets using only a single-camera view of the driver.

III. The DashGaze Dataset

The DashGaze dataset is a video dataset capturing driver gaze on the road, aimed at facilitating the development of dashcam-based gaze mapping models ready for deployment. Each frame includes a triplet: (i) road view from the dash-

Fig. 2: Overview of the DashGaze dataset. Each frame consists of driver view, road view and ego view. We also provide face bounding box and fiducial points, head IMU data, driver gaze coordinates with respect to the road view, blink and fixation data.

(a) Driver view (b) Road view

Fig. 3: Data samples showing variations in the DashGaze dataset w.r.t. lighting, pose, camera position, traffic and weather conditions.

cam, (ii) driver view from the dashcam, and (iii) driver's egocentric view (see Figure 2), which captures cases where the driver's gaze falls outside the road view, like the rear view mirror or the side windows. For each frame, we provide gaze location on the ego and dash frames, azimuth and elevation of the gaze, fixation and blink information, and IMU (Inertial Measurement Unit, gyroscope, and accelerometer) data of the head. This section discusses the data acquisition procedure, dataset statistics, and analysis of driver gaze behavior.

A. Data Acquisition

Hardware configuration: We use the Pupil Invisible eye tracker [45] for tracking driver gaze. The eye tracker provides the 2D gaze location in the driver's ego-centric view. It also provides blink, fixation, IMU, and gaze angles. The eye tracker is connected to an Android phone (OnePlus 8), which powers the tracker and records the gaze information. This phone is mounted on the car windscreen. It simultaneously captures the road view using the rear camera and the driver view using the front camera. Figure 4 shows our hardware setup and views from the various cameras. Note that our framework can be used with newer generations of eye trackers to further improve on our dataset. To increase the variation and generality of the dataset, we moved the position of the dashcam between sessions. Figure 4 (right) shows three exemplar positions of the dashcam with respect to the driver.

Temporal synchronization: During data collection, we first turn on the dash-cam, followed by the eye tracker. We find the temporal disparity between the two streams of data using audio alignment. The ego-centric video is sampled at 30Hz,

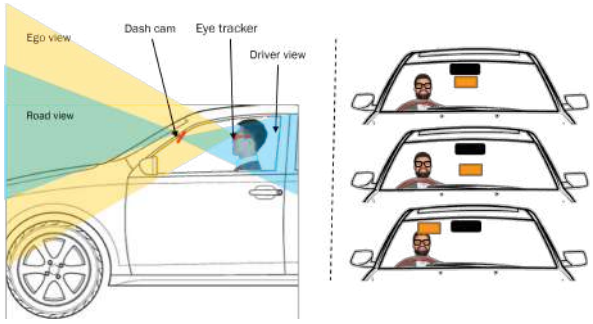

Fig. 4: Data collection setup used for the DashGaze dataset. *(LEFT)* We use three cameras in our setup. A wide-angle camera mounted on the eye-tracker captures the driver's ego view. We use OnePlus 8 Android phone as our dashcam, which is mounted on the windscreen. The front camera captures the driver while the rear camera captures the road view. *(RIGHT)* We varied the position of the dashcam in different sessions to make the dataset more general. Shown above are three positions of the dashcam (in orange): under the rear-view mirror (top), centre of the windscreen (middle) and in the top right corner of the windscreen (bottom)

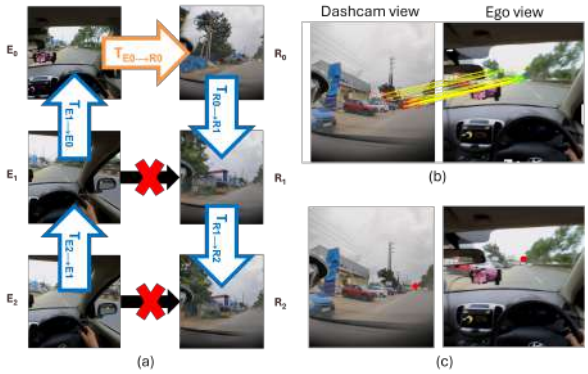

Fig. 5: (a) Illustration of our spatial alignment algorithm. Here, we could find a good homography from ego frame to road frame at timestep 0, but not at timesteps 1 and 2. We transfer the gaze from E_2 to R_2 through E_0 and R_0. (b) An example of matches between road and ego frames (c) Transfer of gaze (in red) from ego to road frame using the homography found in (b)

whereas the gaze and IMU data is sampled at 200Hz. We match each gaze and IMU sample to the nearest video frame. For each video frame, we average all the gaze samples and IMU samples corresponding to it to obtain a single value per frame, as recommended by the Pupil Invisible eye tracker team.

Spatial alignment: We describe our process to transfer egocentric gaze provided by the eye tracker to the road view below. We transfer the gaze point to the road view by computing a homography between the road view and ego-centric view using ORB keypoints [46] matched with the MLESAC algorithm [47] (a variation of RANSAC). We do not get enough matches on some road-ego frame pairs due to resolution differences, blurred frames and view differences between two frames resulting from the driver's

Dataset	Gaze	Subjects	Size
LISA [11]	Gaze zones	10	47K frames
MDM [12]	Markers	59	50.2 hrs *
Dr(eye)ve [13]	Continuous	8	0.5M frames (6 hrs)
LBW [14]	Continuous	28	0.1M frames (7 hrs)
DashGaze	Continuous	28	0.9M frames (10 hrs)

TABLE II: A comparison of recent driver gaze datasets that feature subjects driving on roads. *For MDM, only a portion of the dataset is on the road. Naturalistic gaze is not provided

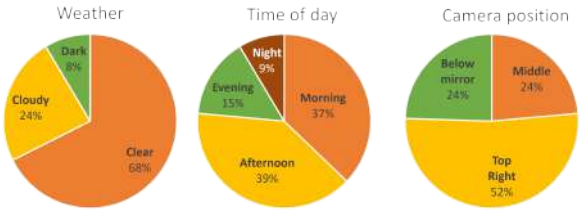

Fig. 6: Distribution of DashGaze dataset with respect to weather, time of day and camera position

head movement. It is easier to match two consecutive frames of the same stream (road or ego). We use this fact to 'propagate' the good homographic transformation obtained between road-ego frames to all frame pairs where good matches are not possible. This is illustrated in Figure 5. About 1.6% of the gaze points were obtained in this manner. If the driver gazes at an object inside the vehicle, the homography-based transformation maps the gaze point to a location in the road view where the gaze vector intersects the road scene, as the vehicle interior is not visible. Additionally, due to the narrower field of view of the road-facing camera, many gaze points fall outside its frame. In such cases, gaze coordinates are still recorded relative to the road view, even if they lie beyond its boundaries. To provide context, the egocentric view is available for all frames.

Our dataset complies with the Institutional Review Board (IRB) protocols. Each driver is over 18 and has a valid driving license. Each participant signed a consent form agreeing to share their data and participate in the study. Along with the driver, an instructor occupied the front seat. The drivers were instructed to drive naturally. Some video captures have one or more passengers in the back seat to increase variation.

B. Dataset Statistics

To the best of our knowledge, the DashGaze dataset is the largest driving 'in-the-wild' dataset with driver and road-facing cameras (See Table II). Our dataset consists of 28 participants and 32 unique drives. Each drive is more than 10 minutes long and thus supports long-term driver behaviour modelling. The dataset has over 0.9M frames, corresponding to over 10 hours of video.

The DashGaze dataset is designed to be as varied and naturalistic as possible. We used 5 different cars for collecting the data. Our participants varied in age from 18 to 62. Figure 3 shows the variations in driver frames and road frames. Our dataset covers three weather conditions and dashcam positions. The dataset was collected at various

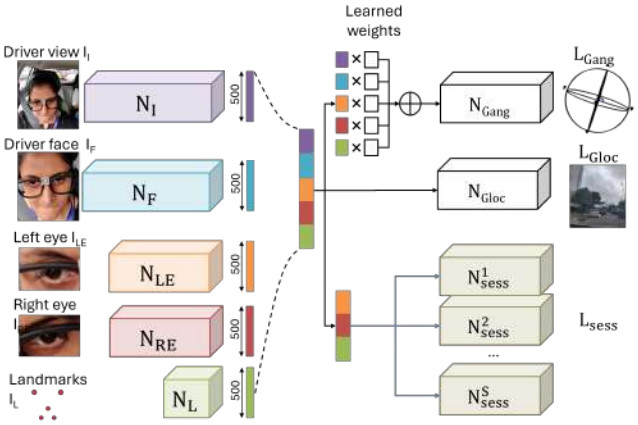

Fig. 7: The DashGazeNet architecture, depicted in the accompanying image, employs five input branches, including driver view, face, left eye, right eye, and landmarks, to estimate gaze angle and location using a multi-task learning approach. Notably, the architecture includes a separate self-supervised session branch for each dataset session, thereby enhancing the accuracy and precision of the model.

times throughout the day. The distribution of our dataset with respect to weather, time of day and camera position is given in Figure 6. In 25% of the data, the driver wore the Pupil Invisible eye tracker with prescription lenses inserted into the frame. In the remaining 75%, the eye tracker was worn without lenses.

IV. DashGazeNet: Dashcam-based Driver Gaze Estimation

We aim to estimate the gaze angle and the target location of the driver's gaze on the road using only the driver-view dashcam as input. This setting ensures maximum flexibility during deployment and prevents biases that could otherwise lead the model to predict only salient objects. Predicting both gaze angle and location is essential, as gaze angle provides a continuous representation of the driver's visual attention, independent of scene constraints, while target location helps determine which object the driver is looking at. This information is valuable for applications such as advanced driver assistance systems and driver behavior analysis. Since dashcam and head positions vary across drivers and sessions, we design our model to be calibration-free, and ensures works across most settings without per-session adjustments. Rather than optimizing for state-of-the-art accuracy, our goal is to demonstrate the feasibility of using dashcam-based data for general-purpose, appearance-based driver gaze estimation. The DashGazeNet model is a multi-branch multi-task, calibration-free model for appearance-based driver gaze prediction (see Figure 7 for the overall architecture). We describe our architecture and method below.

A. DashGazeNet Architecture

Accurate gaze estimation requires knowledge of head position, head orientation, and eye gaze direction. To capture these factors, we design our model with multiple input branches, similar to [2]. Each branch provides complementary information: the full driver-view image captures global

context, the cropped face provides head pose information, the eyes offer fine-grained gaze cues, and facial landmarks supply geometric constraints.

The first branch $N_I(I_I \in \mathbb{R}^{3\times224\times224})$ has a ResNet50 architecture and takes the entire driver-view image as input. We downsample the input image to 224×224 pixels. The face branch $N_F(I_F \in \mathbb{R}^{3\times224\times224}$ has a ResNet50 architecture and takes in an image of the driver's cropped face resized to 224×224 pixels. We detect and crop the face using MTCNN [48]. The left-eye and right-eye branches $N_{LE}(I_{LE} \in \mathbb{R}^{3\times50\times100})$ and $N_{RE}(I_{LE} \in \mathbb{R}^{3\times50\times100})$ have four convolutional layers followed by a linear layer. They are fed the left and right eye images respectively, cropped and resized to 50×100. The landmark branch $N_L(I_L \in \mathbb{R}^{10})$ is a fully connected branch that takes the x and y locations of five facial landmarks provided by MTCNN with respect to the driver view. All the input branches $N_I, N_F, N_{LE}, N_{RE}, and N_L$ output feature vectors of length 500.

The gaze location branch, N_{Gloc}, is the main output branch of DashGazeNet. It receives as input I_{Gloc}, which is the concatenation of feature outputs from the input branches: $N_F(I_F)$, $N_{LE}(I_{LE})$, $N_{RE}(I_{RE})$, $N_I(I_I)$, and $N_L(I_L)$. It then predicts the x and y coordinates of the gaze relative to the road view.

The gaze angle branch N_{Gang} learns the azimuth and elevation of the gaze vector with respect to the face (not the driver camera). Recent studies show that the appearance of the left and right eyes differ for the same angle, and thus, it is beneficial to use both eyes for gaze estimation [4], [49]. In our dataset, we notice that sometimes the left or right eye or face may not be fully visible. Hence, we combine the features of all the input branches with learned weights:

$$\begin{aligned} I_{Gang} = w_F * N_F(I_F) + w_I * N_I(I_I) + w_L * N_L(I_L) \\ + w_{LE} * N_{LE}(I_{LE}) + w_{RE} * N_{RE}(I_{RE}) \end{aligned} \quad (1)$$

where all the weights w are learned. Both N_{Gloc} and N_{Gang} are three-layer fully connected networks with the second layer having dimensions of 500×250.

We use mean squared error as a loss for gaze location and gaze angle. We learn the gaze location and the eye gaze angle (azimuth and elevation) in a multi-task manner to achieve better performance and learn more robust and universal representations [50]. Thus, the combined loss is:

$$L_{Gloc} = \sum_{i=1}^{D} (N_{Gloc}(I_{Gloc}) - \hat{g}_i)^2 \quad (2)$$

$$L_{Gang} = \sum_{i=1}^{D} (N_{Gang}(I_{Gang}) - \hat{l}_i)^2 \quad (3)$$

$$L_G = L_{Gloc} + \lambda L_{Gang} \quad (4)$$

where D is the number of samples, $\hat{g}_i$ is the ground truth gaze location, $\hat{l}_i$ is the ground truth gaze angle, and λ is a hyperparameter.

Calibration-free gaze estimation: As stated in Section III-A, the placement and angle of the dashcam vary between sessions, affecting both the road and driver views. Additionally, differences in driver height and seat adjustments introduce further variations. These factors would typically necessitate a calibration procedure at the start of each session, which is inconvenient and hinders adoption. Our goal is to develop a model that adapts to session-specific conditions without calibration or additional setup.

We achieve this through test-time training, allowing the model to adapt during inference. Session-specific variations introduce inconsistencies in the appearance-based model's inputs, making accurate gaze estimation challenging. To effectively address this, we employ different session heads, where each session head adapts to intra-session variations, such as lighting changes and head movements, while collectively, the session heads model inter-session differences, including seating position and camera angles. This design enables the model to generalize across multiple sessions and drivers without requiring explicit per-session calibration. Additionally, each session head is a lightweight, fully connected network, ensuring computational efficiency while allowing robust adaptation to session-specific conditions. To ensure robustness across sessions and drivers, the session head is designed to model session-specific features using only the left eye, right eye, and facial landmarks. We exclude the driver's face and full driver-view image to prevent the model from learning shortcuts to distinguish between sessions. Since the session head is a lightweight, fully connected network, we can maintain a separate session head for each session.

The session head is a three-layer fully connected network, with the second layer having dimensions of 500×250. Its input consists of concatenated feature outputs from N_{LE}, N_{RE}, and N_L. During training, we assign a separate session head N^s_{sess} to each session s in the training set without sharing weights. Each session head learns a self-supervised task of detecting translations in facial landmark points. We train this by alternating batches of original and augmented facial landmarks, where each sample's landmarks are randomly shifted by -20 and 20 pixels. All landmark points within a single sample are shifted by the same amount. A cross-entropy loss function is used to classify whether a sample has been translated:

$$\begin{aligned} L^s_{sess}(N^s_{sess}(I^s_{sess}), y) = -(y \log(N^s_{sess}(I^s_{sess})) \\ + (1-y)\log(1 - N^s_{sess}(I^s_{sess}))) \end{aligned} \quad (5)$$

where $y = 1$ for translated samples and $y = 0$ otherwise. The total training loss for our model is:

$$L = L_G + \sum_{s=1}^{S} L^s_{sess} \quad (6)$$

During inference, we treat all incoming samples as belonging to a new session not seen during training. The trained session heads are discarded, and a new session head N^i_{sess} is created. For each batch of test samples, we compute the gaze location and angle. We then alternate batches of translated landmarks, performing a single weight update per

Method	Location error (px)	Angular error (deg)	Azimuth (deg)	Elevation (deg)
Gaze360 [3]	N/A	36.44	29.38	32.61
ETHX-Gaze [20]	N/A	37.77	34.39	29.37
I-DGaze [2]	669.93	N/A	N/A	N/A
DashGaze	**225.06**	**8.48**	**7.92**	**6.36**

TABLE III: Comparison of gaze estimation errors across different methods. We report angular errors (eqn. 7) as well as azimuth and elevation errors in degrees for DashGazeNet, Gaze360, and ETH-XGaze. We report gaze location errors (in pixels) for DashGazeNet and I-DGaze.

batch using the loss in Equation 5. This enables the model to adapt to test-time conditions in a self-supervised manner.

V. Experiments and Results

This section comprehensively evaluates DashGazeNet's performance for gaze angle and location estimation on the DashGaze dataset. We present qualitative results for the DashGazeNet architecture, including an ablation study that assesses the various components of the model.

A. Experimental Setup

We split our data into 82% train, 10% validation and 8% test. While splitting the data, we ensured to not split a driving session into different splits. We ensured that the sessions in the test split had a variety of lightings, drivers, camera positions and cars.

Since our model outputs both gaze angles (azimuth and elevation) and gaze locations (x and y coordinates relative to the road view), we compare our results with other gaze estimation algorithms for both gaze angle and gaze location. In all cases, we quantify gaze error as the mean angular deviation (in degrees) between the predicted and ground truth gaze directions. The location error is calculated as the mean of the Euclidean distance between the predicted and ground truth coordinates.

$$E_{Ang} = \frac{1}{N}\sum_{i} \cos^{-1}(g_i^T \hat{g}_i) \tag{7}$$

$$E_{Loc} = \frac{1}{N}\sum_{i} \|l_i - \hat{l}_i\| \tag{8}$$

where g_i and l_i are the i_{th} predictions for gaze angle and location, and $\hat{g}_i$ and $\hat{l}_i$ are the ground truth. N stands for the number of samples. We also report the RMSE of azimuth and elevation in Table III.

B. Baseline Gaze Estimation Results

We compare our gaze angle estimates with Gaze360 [3] and ETHX-Gaze [20], two popular off-the-shelf gaze estimators designed for 'in-the-wild' scenarios. We use the static models provided by the authors. Since our model outputs gaze angles relative to the face rather than the driver's view, we employ a head pose estimator [52], [53] to estimate the driver's head pose and adjust our gaze angles accordingly. Table III presents the results of this experiment. The angular error across the entire test set for DashGazeNet is 8.48°, compared to 37.77°for ETHX-Gaze and 36.44°for Gaze360.

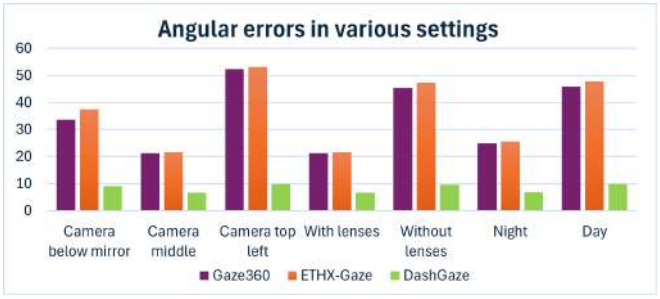

Fig. 8: Graph comparing the angular error (in degrees) of gaze estimation across different settings. Errors are shown for three different camera placements, as well as for subsets where the driver is wearing lenses, not wearing lenses, during daytime sessions, and nighttime sessions. (lower values indicate better performance)

Fig. 9: Qualitative results of DashGazeNet gaze angle prediction on randomly selected samples. Arrow colors represent the following: red: Gaze360 [3], yellow: ETHX-Gaze [20], green: DashGazeNet, and white: ground truth.

We compare our gaze location estimation with the I-DGaze model [2]. Like DashGazeNet, I-DGaze uses only the driver-side video to infer the two-dimensional gaze target location on the road video. We trained the I-DGaze model on the Dash-Gaze train set with the same hyperparameters as in the original work. We see from Table III that our model significantly outperforms I-DGaze. The lower performance of I-DGaze may be attributed to its limited ability to adapt to the wide variability present in the DashGaze dataset. This shows that the Dash-Gaze dataset is collected in a more unconstrained setting than DGaze and requires more contextual information to estimate gaze on the Dash-Gaze dataset.

Figure 8 compares the angular error of Gaze360, ETHX-Gaze and DashGazeNet over different settings. For camera placement, we observe that placing the dashcam on the top left of the windscreen gives the worst performance across the three methods. The best results for Gaze360 and DashGazeNet are obtained when the dashcam is placed in the middle of the windscreen, while ETHX-Gaze gets the best results with the dashcam placed directly under the rear-view mirror.

Figure 9 shows some qualitative comparison of gaze angle prediction on random samples. We see that the gaze angles predicted by DashGazeNet is close to the ground truth gaze in most cases, despite different drivers and dashcam positions. We also note that the gaze prediction is robust to various lighting and occlusions.

	Variation	Azimuth	Elevation	Location
1	No TTT	8.12°	6.15°	319.32pix
2	TTT all features	8.13°	6.80°	296.07pix
3	eyes + landmarks	9.67°	6.37°	287.69pix
4	weighted features	8.07°	**6.14°**	1047.22pix
5	**DashGazeNet**	**7.92°**	6.36°	**225.06pix**

TABLE IV: Results of ablation study on the DashGazeNet model. The numbers are errors (lower is better)

C. Ablation Studies

We examines the impact of various parts of the DashGazeNet architecture on the model's performance. In Table IV, we see the results of four variants of our model. The first row shows the results of a variant that does not use test time adaptation. This model cannot predict the gaze location well on unseen sessions. The second row uses a variant of test time training where the session branch was asked to predict which session a sample came from during training. During inference, a new session branch predicted if a sample was translated. The session branch took the concatenated feature vectors of all input branches as input. The third-row variant is the same as the second row except that only the eye and landmark features were used for training the session model. We see a marked improvement in the gaze location prediction in this variation. The fourth model variant learned weights for each input branch and combined the input branch features using weighted addition. While this improves the performance of gaze angle estimation, gaze location prediction does not work. Finally, we have our model, the DashGazeNet. It is similar to the third variant, except it uses weighted addition to combine the features for the gaze angle branch. In addition, one session branch for each session during training predicts whether a sample has been translated. This architecture gives the best results for gaze angle and gaze location predictions.

VI. Discussion: Potential Societal Impact

Driver gaze tracking offers significant benefits for creating advanced driver assistance systems (ADAS), driver education, and in designing better roads and infrastructure with minimal distractions. However, gaze-tracking technology raises significant privacy concerns, particularly regarding appearance-based methods that can be deployed without a person's consent. Unauthorized use of gaze tracking can lead to severe privacy violations, especially if the data is exploited for market research, capturing personal interests, or targeting advertisements without the individual's knowledge. In the context of driving, malicious actors could misuse gaze tracking data to design distracting signboards, intentionally diverting drivers' attention and increasing the risk of accidents. These potential abuses highlight the need for stringent regulations and ethical guidelines to protect individuals' privacy and ensure that gaze-tracking technology is used responsibly and transparently.

VII. Conclusion

This work presented the DashGaze dataset, which is a largescale driver gaze dataset that offers both road view and driver view to enable training of appearance-based driver gaze estimation models using only a dashcam. Our dataset is the most comprehensive of its kind and offers diverse variations in pose, lighting, roads, weather, and traffic conditions. We also introduced DashGazeNet, an appearance-based driver gaze estimation model trained on the DashGaze dataset. Our experiments showed the model's generalizability and potential to be implemented in real-world scenarios.

The proposed method has far-reaching implications for various downstream applications, including enhanced ADAS functionality, improved driver behavior analysis, optimized road sign and obstruction design, and more effective driver training programs. As a result, we believe that our work will contribute to safer roads by enabling the development of more accurate and reliable gaze-based systems, ultimately saving lives and reducing the risk of accidents. In future studies, we plan to streamline the method to work on edge devices in real time.

References

[1] M. C. Catalbas, T. Cegovnik, J. Sodnik, and A. Gulten, "Driver fatigue detection based on saccadic eye movements," in *2017 10th International Conference on Electrical and Electronics Engineering (ELECO)*. IEEE, 2017, pp. 913–917.

[2] I. Dua, T. A. John, R. Gupta, and C. Jawahar, "Dgaze: Driver gaze mapping on road," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2020*, 2020.

[3] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, "Gaze360: Physically unconstrained gaze estimation in the wild," in *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[4] T. Fischer, H. J. Chang, and Y. Demiris, "RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments," in *European Conference on Computer Vision*, 2018.

[5] P. Kratzer, S. Bihlmaier, N. Balachandra Midlagajni, R. Prakash, M. Toussaint, and J. Mainprice, "Mogaze: A dataset of full-body motions that includes workspace geometry and eye-gaze," *IEEE Robotics and Automation Letters (RAL)*, 2020.

[6] Z. Yu, J. S. Yoon, I. K. Lee, P. Venkatesh, J. Park, J. Yu, and H. S. Park, "Humbi: A large multiview dataset of human body expressions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[7] R. F. Ribeiro and P. D. Costa, "Driver gaze zone dataset with depth data," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–5.

[8] J. D. Ortega, N. Kose, P. Cañas, M.-a. Chao, A. Unnervik, M. Nieto, O. Otaegui, and L. Salgado, "DMD : A Large-Scale Multi-Modal Driver Monitoring Dataset for Attention and Alertness Analysis," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops (Accepted)*, 2020.

[9] S. Ghosh, A. Dhall, G. Sharma, S. Gupta, and N. Sebe, "Speak2label: Using domain knowledge for creating a large scale driver gaze zone estimation dataset," *ICCVW*, 2021.

[10] M. Selim, A. Firintepe, A. Pagani, and D. Stricker, "Autopose: Large-scale automotive driver head pose and gaze dataset with deep head orientation baseline," in *International Conference on Computer Vision Theory and Applications (VISAPP)*, 2020. [Online]. Available: http://autopose.dfki.de

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

[12] S. Jha, M. F. Marzban, T. Hu, M. H. Mahmoud, N. Al-Dhahir, and C. Busso, "The multimodal driver monitoring database: A naturalistic corpus to study driver attention," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[13] A. Palazzi, D. Abati, S. Calderara, F. Solera, and R. Cucchiara, "Predicting the driver's focus of attention: the dr(eye)ve project," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[14] I. Kasahara, S. Stent, and H. S. Park, "Look both ways: Self-supervising driver gaze estimation and road scene saliency," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 126–142.

[15] A. Rudenko, T. P. Kucner, C. S. Swaminathan, R. T. Chadalavada, K. O. Arras, and A. J. Lilienthal, "Thör: Human-robot navigation data collection and accurate motion trajectories dataset," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 676–682, 2020.

[16] R. Kothari, Z. Yang, C. Kanan, R. Bailey, J. B. Pelz, and G. J. Diaz, "Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities," *Scientific reports*, vol. 10, no. 1, p. 2539, 2020.

[17] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2176–2184.

[18] S. He, H. R. Tavakoli, A. Borji, and N. Pugeault, "Human attention in image captioning: Dataset and analysis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8529–8538.

[19] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Mpiigaze: Real-world dataset and deep appearance-based gaze estimation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 1, pp. 162–175, 2017.

[20] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 2020, pp. 365–381.

[21] H. Tomas, M. Reyes, R. Dionido, M. Ty, J. Casimiro, R. Atienza, and R. Guinto, "Goo: A dataset for gaze object prediction in retail environments," in *CVPR Workshops (CVPRW)*, 2021.

[22] J. Kim, M. Stengel, A. Majercik, S. De Mello, D. Dunn, S. Laine, M. McGuire, and D. Luebke, "Nvgaze: An anatomically-informed dataset for low-latency, near-eye gaze estimation," in *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019, pp. 1–12.

[23] A. Jain, H. S. Koppula, S. Soh, B. Raghavan, A. Singh, and A. Saxena, "Brain4cars: Car that knows before you do via sensory-fusion deep learning architecture," *arXiv preprint arXiv:1601.00740*, 2016.

[24] M. Martin, A. Roitberg, M. Haurilet, M. Horne, S. Reiß, M. Voit, and R. Stiefelhagen, "Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2801–2810.

[25] J. Fang, D. Yan, J. Qiao, J. Xue, H. Wang, and S. Li, "Dada-2000: Can driving accident be predicted by driver attention analyzed by a benchmark," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 4303–4309.

[26] T. Wu, N. Martelaro, S. Stent, J. Ortiz, and W. Ju, "Learning when agents can talk to drivers using the inagt dataset and multisensor fusion," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 3, pp. 1–28, 2021.

[27] Y. Yang, C. Liu, F. Chang, Y. Lu, and H. Liu, "Driver gaze zone estimation via head pose fusion assisted supervision and eye region weighted encoding," *IEEE Transactions on Consumer Electronics*, vol. 67, pp. 275–284, 2021.

[28] S. Vora, A. Rangesh, and M. M. Trivedi, "On generalizing driver gaze zone estimation using convolutional neural networks," *2017 IEEE Intelligent Vehicles Symposium (IV)*, pp. 849–854, 2017.

[29] Y. Zhang, X. Yang, and Z. Ma, "Driver's gaze zone estimation method: A four-channel convolutional neural network model," *Proceedings of the 2020 2nd International Conference on Big-data Service and Intelligent Computation*, 2020.

[30] Y. Wang, G. Yuan, and X. Fu, "Driver's head pose and gaze zone estimation based on multi-zone templates registration and multi-frame point cloud fusion," *Sensors (Basel, Switzerland)*, vol. 22, 2022.

[31] Y. Wang, G. Yuan, Z. Mi, J. Peng, X. Ding, Z. Liang, and X. Fu, "Continuous driver's gaze zone estimation using rgb-d camera," *Sensors (Basel, Switzerland)*, vol. 19, 2019.

[32] Z. Yu, X. Huang, X. Zhang, H. Shen, Q. Li, W. Deng, J.-B. Tang, Y. Yang, and J. Ye, "A multi-modal approach for driver gaze prediction to remove identity bias," *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020.

[33] L. Stappen, G. Rizos, and B. Schuller, "X-aware: Context-aware human-environment attention fusion for driver gaze prediction in the wild," *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020.

[34] G. Yuan, Y. Wang, H. Yan, and X. Fu, "Self-calibrated driver gaze estimation via gaze pattern learning," *Knowledge-Based Systems*, vol. 235, p. 107630, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705121008923

[35] S. M. Shah, Z.-L. Sun, K. Zaman, A. Hussain, M. Shoaib, and L. Pei, "A driver gaze estimation method based on deep learning," *Sensors (Basel, Switzerland)*, vol. 22, 2022.

[36] B. Vasli, S. Martin, and M. M. Trivedi, "On driver gaze estimation: Explorations and fusion of geometric and data driven approaches," *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 655–660, 2016.

[37] 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 Transactions on Industrial Electronics*, vol. 69, pp. 1800–1808, 2021.

[38] 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 Transactions on Intelligent Transportation Systems*, vol. 21, pp. 4318–4327, 2020.

[39] M. Shirpour, S. S. Beauchemin, and M. A. Bauer, "A probabilistic model for visual driver gaze approximation from head pose estimation," *2020 IEEE 3rd Connected and Automated Vehicles Symposium (CAVS)*, pp. 1–6, 2020.

[40] S. K. Jha and C. Busso, "Probabilistic estimation of the driver's gaze from head orientation and position," *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1–6, 2017.

[41] S. Mohan and M. R. Phirke, "Eye gaze estimation invisible and ir spectrum for driver monitoring system," *Signal & Image Processing : An International Journal*, vol. 11, pp. 1–20, 2020.

[42] J. Lemley, A. Kar, A. Drimbarean, and P. M. Corcoran, "Convolutional neural network implementation for eye-gaze estimation on low-quality consumer imaging systems," *IEEE Transactions on Consumer Electronics*, vol. 65, pp. 179a–187, 2019.

[43] U. Sonom-Ochir, S. Karungaru, K. Terada, and A. Ayush, "Appearance-based driver's gaze mapping using a dash camera," *2022 Joint 12th International Conference on Soft Computing and Intelligent Systems and 23rd International Symposium on Advanced Intelligent Systems (SCIS&ISIS)*, pp. 1–5, 2022.

[44] A. Rangesh, B. Zhang, and M. M. Trivedi, "Gaze preserving cyclegans for eyeglass removal and persistent gaze estimation," *IEEE Transactions on Intelligent Vehicles*, vol. 7, pp. 377–386, 2020.

[45] M. Tonsen, C. K. Baumann, and K. Dierkes, "A high-level description and performance evaluation of pupil invisible," 2020.

[46] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International Conference on Computer Vision*, 2011, pp. 2564–2571.

[47] P. H. S. Torr and A. Zisserman, "Mlesac: A new robust estimator with application to estimating image geometry," *Comput. Vis. Image Underst.*, vol. 78, pp. 138–156, 2000.

[48] B. Jiang, Q. Ren, F. Dai, J. Xiong, J. Yang, and G. Gui, "Multi-task cascaded convolutional neural networks for real-time dynamic face recognition method," in *Communications, Signal Processing, and Systems: Proceedings of the 2018 CSPS Volume III: Systems 7th*. Springer, 2020, pp. 59–66.

[49] Y. Cheng, F. Lu, and X. Zhang, "Appearance-based gaze estimation via evaluation-guided asymmetric regression," in *European Conference on Computer Vision*, 2018.

[50] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 12, pp. 5586–5609, 2021.

[51] Y. Sun, X. Wang, Z. Liu, J. Miller, A. Efros, and M. Hardt, "Test-time training with self-supervision for generalization under distribution shifts," in *International conference on machine learning*. PMLR, 2020, pp. 9229–9248.

[52] T. Hempel, A. A. Abdelrahman, and A. Al-Hamadi, "Toward robust and unconstrained full range of rotation head pose estimation," *IEEE Transactions on Image Processing*, vol. 33, pp. 2377–2387, 2024.

[53] ——, "6d rotation representation for unconstrained head pose estimation," in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 2496–2500.