Look Both Ways: Self-Supervising Driver Gaze Estimation and Road Scene Saliency

Isaac Kasahara[1], Simon Stent[2], and Hyun Soo Park[1]

[1] University of Minnesota, USA
{kasah011, hspark}@umn.edu
[2] Toyota Research Institute, Cambridge, MA, USA
simon.stent@tri.global

Abstract. We present a new on-road driving dataset, called "Look Both Ways", which contains synchronized video of both driver faces and the forward road scene, along with ground truth gaze data registered from eye tracking glasses worn by the drivers. Our dataset supports the study of methods for non-intrusively estimating a driver's focus of attention while driving - an important application area in road safety. A key challenge is that this task requires accurate gaze estimation, but supervised appearance-based gaze estimation methods often do not transfer well to real driving datasets, and in-domain ground truth to supervise them is difficult to gather. We therefore propose a method for self-supervision of driver gaze, by taking advantage of the geometric consistency between the driver's gaze direction and the saliency of the scene as observed by the driver. We formulate a 3D geometric learning framework to enforce this consistency, allowing the gaze model to supervise the scene saliency model, and vice versa. We implement a prototype of our method and test it with our dataset, to show that compared to a supervised approach it can yield better gaze estimation and scene saliency estimation with no additional labels.

Keywords: Driving, 3D gaze, saliency, self-supervised learning, ADAS

1 Introduction

For the past decade, computer vision has played an increasingly important role in self-driving cars, to help them understand what is happening *outside* the vehicle (see e.g. [6, 13, 40]). But the vast majority of vehicles on the road today remain human-controlled, and will stay that way for the foreseeable future, with partially automated systems (SAE Levels 2-3 [2]) set to become the norm [1]. Given this trend, it is important to pay close attention to what is happening *inside* the vehicle: to better understand the behaviors of human drivers. As driving becomes the activity of a cooperative human-AI team, building good representations of drivers will be critical to help ensure a safe and efficient system.

This work concerns one important aspect of driver behavior: their **visual focus of attention**. Attention is a major indicator of the intent and decision-making of drivers, and humans in general [20]. An ability to precisely estimate a

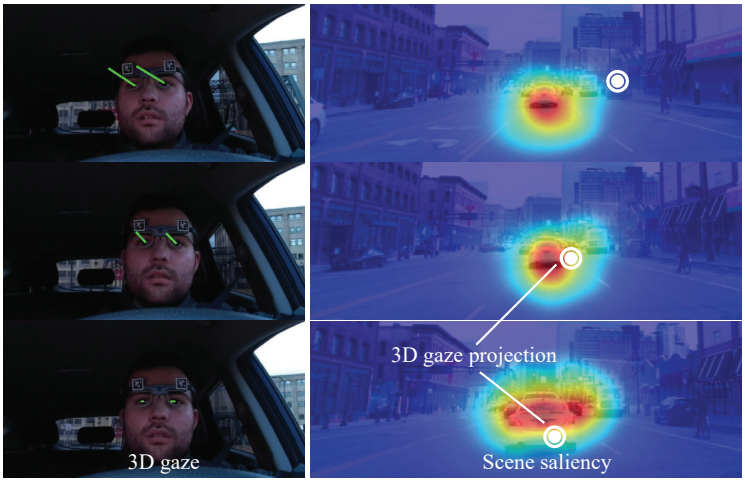


Fig. 1. Motivation. Estimating a driver's focus of attention is important to determine if they have sufficient situational awareness to drive safely. Here we show the test-time output of our method, with estimated gaze overlaid on the input face image, 3D gaze projected into the scene camera view, and saliency prediction from the scene image.

driver's attention is a stepping stone towards vehicles being able to understand the situational awareness of a driver and adapt their behavior to provide better assistance to drivers in need through warnings or interventions. For example, in Fig. 1, a vehicle which can tell whether the driver has seen the braking vehicle up ahead may be able to warn them or apply braking earlier, to help prevent a collision. Knowledge of human visual attention can also be used to help machines attend to the driving task more efficiently [3]. Due to its importance, a number of companies exist which specialize in developing driver monitoring systems centered around estimating aspects of driver state such as eye gaze. However, commercial systems, which typically (though not exclusively) use near infra-red, glint- and model-based tracking to achieve high precision, are black-box systems and require specific camera and lighting setups. For these reasons, research into open models which relax these constraints, such as appearance-based gaze models powered by deep learning [18, 49], has continued. One challenge with this work has been the lack of domain-specific data: while synthetic datasets have found some success [37, 39, 44], collecting ground truth gaze data across many subjects and conditions in the target domain of driving is expensive. As a consequence, existing appearance-based 3D gaze estimation models can be highly fragile when applied to new drivers in real driving scenes. There is a need for both a more label-efficient method to adapt gaze estimation models to drivers, and a dataset to support it.

To try to address these needs, we make the following contributions.

Firstly, we collect **a new on-road dataset for visual focus of attention estimation**. The *Look Both Ways* (LBW) dataset features seven hours of driving captured from 28 drivers with synchronized and calibrated facial RGB-D video (looking at the driver), driving scene stereo videos (looking at the road), and ground-truth gaze from eye-tracking glasses.

Secondly, we present a **self-supervised learning method to improve appearance-based driver gaze estimation** without annotations. Our main insight is that driver attention is highly correlated with 3D scene semantics (or visual saliency). For example, drivers tend to focus on vanishing points, tangent points on curving roads, pedestrians, relevant traffic signals, or approaching cars [20]. To take advantage of this, we set up our gaze learning framework to encourage geometric consistency between gaze direction and visual saliency through 3D reconstruction. Our system takes the driver face and stereo scene images as input, and outputs 3D gaze direction and an estimate of scene saliency.

Finally, we demonstrate that our self-supervised method can **improve performance** of a recent appearance-based gaze tracking method in this applied setting. As a byproduct, our method can also improve the performance of driving scene saliency. Our dataset and experimental code is available at `https://github.com/Kasai2020/look_both_ways`.

2 Related Work

Appearance-based Gaze Estimation. Methods to non-intrusively estimate human gaze from facial images have been studied in computer vision for many decades [14]. Artificial neural networks were first used for appearance-based gaze estimation in the early 1990s [4], but modern deep learning techniques and the availability of larger training datasets have significantly improved their performance [18, 49, 50]. Researchers have explored techniques to further improve the data-efficiency, generalization and accuracy of these methods, for example by injecting more structure into the learned representations [9, 31], leveraging synthetic data in novel ways [37,39,44], and personalizing gaze models to individuals with minimal supervision [21, 30]. Despite this progress, supervised appearance-based models are still known to experience performance degradation at test-time, as they may struggle to transfer to new appearances (including occluders such as glasses), lighting conditions, or head poses outside of the training data. While exciting progress has been made recently in self-supervised gaze representation learning [41, 48], these methods do not yet attempt to leverage supervisory signals that may be freely available from the environment, such as the scene which the subject is looking at. Our work presents a method, dataset and evidence for how self-supervision from the environment can be used to boost appearance-based gaze estimation in an applied setting.

Leveraging the Relationship between Gaze and Saliency. Understanding where people look has long been a topic of interest in human perception and computer vision [47]. The computational modeling of visual saliency has

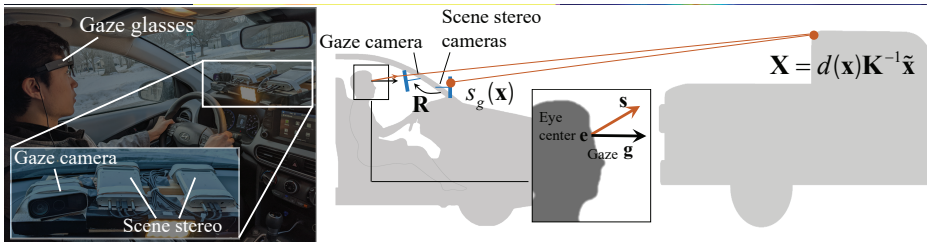
advanced significantly, moving beyond the detection of low-level salient patterns towards higher-level reasoning [5,10,26,43]. Saliency models can now serve practical purposes, for example to predict (and therefore influence) website engagement [36,51]. The idea of relating the saliency of the environment to a person's attention—which we leverage in this paper—is not new. Early work by Sugano *et al.* [38] aggregated on-screen saliency maps corresponding to similar-looking eye appearances over time to help construct an appearance-based gaze estimator in the lab without ground truth. Recasens *et al.* showed how saliency can be used to help an appearance-based gaze estimator for images containing subjects looking elsewhere within the image [33], and in follow-up work with multi-view data showed how saliency, appearance-based gaze, and geometric relationships between camera views can be solved for simultaneously using only gaze as supervision [34]. Chang *et al.* [8] propose the use of saliency information to calibrate an existing gaze model. Most related to our work, Park *et al.* [29] demonstrated a method for end-to-end video-based eye-tracking which showed that the accuracy of appearance-based gaze models for subjects watching visual stimuli on screen could be improved with knowledge of saliency. Inspired by this prior work, we explore whether saliency and gaze can be used to supervise one another outside of the screen-in-the-lab setting, and in a real, 3D driving environment.

Gaze Estimation for Driving. Driving is a complex task which requires paying attention to many different static and dynamic areas of the vehicle and road scene, and therefore involves a range of eye movement strategies [20]. Vehicles which can accurately predict driver gaze can provide assistance which is more in tune with the needs of the driver. There are numerous datasets for the study of driver behaviors, including Brain4Cars [16], BDD-A [46], DrEYEve [28], Drive&Act [24], DMD [27], DADA2000 [11], and INAGT [45]. However, no dataset exists which combines driver-facing video, scene-facing video and ground truth gaze. In Section 3, we describe our dataset contribution and how it supports the exploration of gaze and saliency self-supervision for the important task of driver gaze estimation.

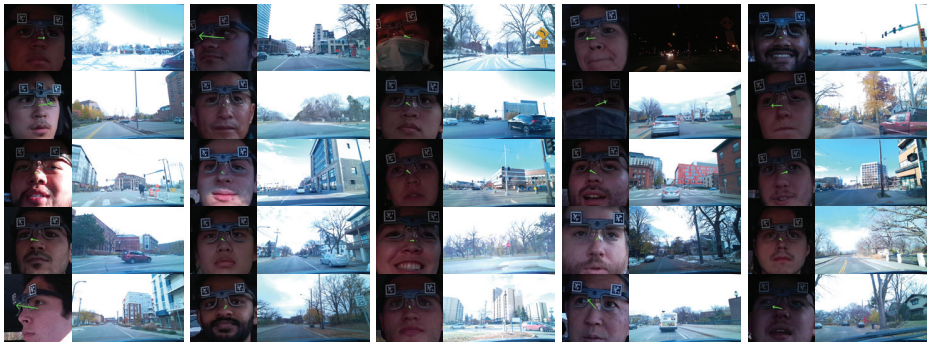
3 Look Both Ways (LBW) Dataset

To build the LBW dataset, we created a setup shown in Fig. 2(a) which captured two synchronized streams of data at 15Hz (downsampled to 5Hz for processing):

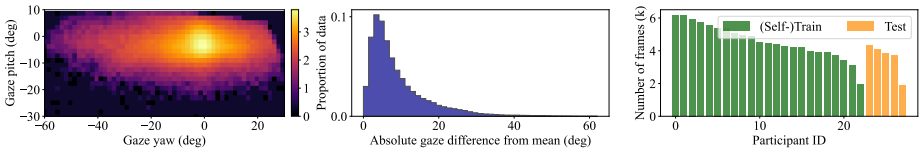
1. 3D gaze data consisting of: *(i)* face images, $\mathbf{I}_g$, *(ii)* ground truth 3D gaze directions $\mathbf{g}$, and *(iii)* 3D eye centers, $\{\mathbf{e_l}, \mathbf{e_r}\}$, with respect to the scene camera. A Kinect Azure RGB-D camera was used to capture the face image and the 3D location of eyes ("Gaze camera"). Drivers wore Tobii Pro Glasses to measure driver gaze in a glasses-centric co-ordinate system ("Gaze glasses"), which could later be transformed into ground truth gaze $\mathbf{g}$.
2. Left and right pairs of stereo scene images, $\{\mathbf{I}_s^l, \mathbf{I}_s^r\}$, captured with an additional pair of Kinect Azure cameras ("Scene stereo") synchronized with the gaze camera.



(a) LBW data capture setup and geometry



(b) Data samples showing a variety of driver appearances and driving conditions.



(c) Dataset statistics.

Fig. 2. The Look Both Ways (LBW) Dataset. *(a) Dataset collection setup.* We used synchronized and calibrated driver-facing monocular and scene-facing stereo cameras, and driver-mounted eye tracking glasses, in order to gather data of driver 3D gaze registered to real driving scenes. *(b) Samples from the dataset.* LBW is collected with 28 drivers who drive in various areas including urban, rural, residential, and campus, under sunny, rainy, cloudy, and snowy weather. The data includes driver face images, road-facing scene images, and 3D gaze direction from a head-mounted eye tracker. *(c) Dataset statistics.* (Left) Log-frequency histogram of ground truth gaze pitch and yaw, showing a concentration of gaze towards the road ahead. (Center) The distribution of fixations away from the mean is heavy tailed, corresponding to glances away from the forward road scene. (Right) Data was gathered from 28 subjects, of which 5 were held out from all supervised and self-supervised training as a test set.

3.1 Calibration and Pre-processing

The gaze and stereo cameras were rigidly attached to a mechanical frame as shown in Fig. 2(a), where their relative transformation $\mathbf{R}$ and intrinsic param-

Fig. 3. Ground truth gaze registration. We map the gaze focal point $\mathbf{x}_g$ from the gaze glasses camera view (left) to the scene image $\mathbf{x}_s$ (center) using a homography $\mathbf{H}$ (right) estimated by matching local image features between the two images.

eters $\mathbf{K}$ are constant. We calibrated these parameters using COLMAP, an off-the-shelf structure-from-motion algorithm [35]. Since the cameras face opposite directions where matches cannot be found, we capture a calibration sequence by rotating the mechanical frame outside the vehicle in order to acquire feature correspondences across time.

Localizing the gaze glasses in the coordinate system defined by the scene cameras is non-trivial as the glasses are in constant motion. One possible approach is to estimate the 3D rotation and translation of the glasses in the driver-facing camera by tracking AprilTags [42] which we mounted rigidly to the frame. However, the result was highly sensitive to the small noisy in the recovered rotation. Instead, similar to [28], we opted to directly register the gaze focal point projected into the glasses' own scene-facing video stream, against our own scene-facing video:

$$\mathbf{g} = \mathbf{R}\frac{\mathbf{X}_g - \mathbf{e}}{\|\mathbf{X}_g - \mathbf{e}\|}, \quad \text{where} \quad \mathbf{X}_g = d(\mathbf{x}_s)\mathbf{K}^{-1}\widetilde{\mathbf{x}}_s, \quad \widetilde{\mathbf{x}}_s \propto \mathbf{H}\widetilde{\mathbf{x}}_g, \tag{1}$$

where $\mathbf{g} \in \mathbb{S}^2$ is the 3D gaze direction, $\mathbf{x}_g \in \mathbb{R}^2$ is the gaze focal point that is measured by the gaze glasses, $\mathbf{e} \in \mathbb{R}^3$ is the center of eyes in the scene camera coordinate, $\mathbf{x}_s \in \mathbb{R}^2$ is the transferred gaze focal point in the scene image, and $\mathbf{H}$ is the homography that directly maps the gaze glass image to the scene image as shown in Fig. 3. $d(\mathbf{x}_s)$ is the depth at the gaze focal point $\mathbf{x}_s$, $\widetilde{\mathbf{x}}$ is the homogeneous representation of $\mathbf{x}$, and $\mathbf{R} \in SO(3)$ is the rotation matrix that transforms from the scene image coordinate system to the face image coordinate system. We approximate this transformation as a homography (i.e., pure rotation) assuming that the distance from the scene to the camera is sufficiently far, i.e., weak perspective. We estimate this homography by leveraging local image feature matching [23] with RANSAC [12].

We use RAFT-Stereo [22] to reconstruct scene depth from our stereo image pair, and we measure the physical baseline distance between the stereo cameras to reconstruct to metric scale, validating by capturing an object of known size.

3.2 Final Collection

We collected the data by complying with an Institutional Research Board (IRB) protocol. Each driver signed a consent form reviewed by the IRB. All drivers

Dataset	Primary Task	Driver	Scene	Scenario	Gaze	# Subj.	Size(h)
Brain4Cars [16]	Maneuver Pred.	RGB	Mono	Real	No	10	*10
Drive&Act [24]	State/Act.Rec.	RGB+D, NIR	-	Sim	No	15	12
DMD [27]	State/Act.Rec.	RGB+D, NIR	-	Both	No	37	41
INAGT [45]	HMI Timing	RGB	Mono	Real	No	46	38
BDD-A [46]	Visual FoA	-	Mono	Sim	Yes	1,232	4
DrEYEve [28]	Visual FoA	-	Mono	Real	Yes	8	6
LBW (Ours)	Visual FoA/Gaze	RGB+D	Stereo	Real	Yes	28	7

Table 1. Comparison of recent driver-facing datasets. Our dataset is unique in containing driver-facing imagery (Driver), scene-facing imagery (Scene) along with ground truth gaze fixations (Gaze). Among other things, LBW therefore supports the development of self-supervised methods to study and improve estimation of driver focus of attention (FoA) using scene saliency.

were older than 18 years old and held a US driver's license. During driving, an instructor was on board in the passenger seat to provide safety instruction and directional guidance.

After calibration, we filter the videos to remove missed gaze registrations, missed driver detections via OpenPose [7], and dropped frames. The final fully-annotated dataset consists of 6.8 hours of free driving on public roads. We captured 28 drivers (22 male, 6 female), totalling 123,297 synchronized driver face and stereo scene images with ground truth 3D gaze. This includes various road types (*e.g.* urban, residential and rural) and various weather conditions (sunny, cloudy, rainy, and snowy) with various lighting conditions (daytime and dusk).

Figures 2(b) and 2(c) illustrate the diversity of our dataset. Driver gaze is widely spread across yaw and pitch angles. For driving scenarios, the gaze distribution is slightly biased to the negative yaw angle because the driver seat is located on the left sides. Each participant collected more than 2,000 clean data samples. Table 1 compares LBW against recent driving datasets.

4 Self-supervised Gaze

Given a set of images that capture face appearance and the driving scene, we present a self-supervised learning framework to predict the 3D gaze direction. We represent a measure of visual saliency as a function of the 3D gaze direction, which allows us to encourage geometric consistency between the 3D gaze and visual saliency prediction.

4.1 Gaze-Driven Saliency

From an image of the driver's face, $\mathbf{I}_g \in [0,1]^{H_g \times W_g \times 3}$, where H_g and W_g are its height and width, we wish to predict the 3D gaze direction $\mathbf{g} \in \mathbb{S}^2$:

$$\mathbf{g} = f_g(\mathbf{I}_g; \boldsymbol{\theta}_g), \tag{2}$$

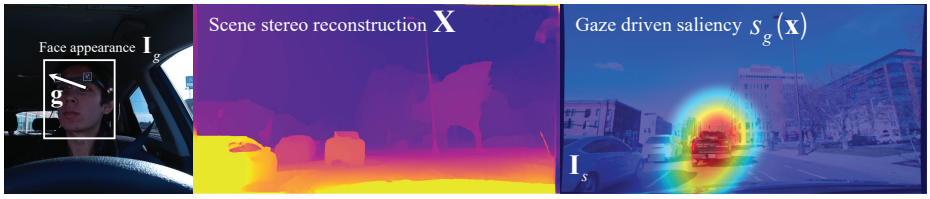


Fig. 4. Processed data sample. We represent a measure of visual saliency $s_g(\mathbf{x})$ as a function of the 3D gaze direction $\mathbf{g}$ that can be predicted by the face appearance image $\mathbf{I}_g$. Given the depth estimates from the stereo scene cameras in $\mathbf{X}$, we reconstruct 3D points and project them to the eye center $\mathbf{e}$ to form directions $\mathbf{s}$. The angular difference between $\mathbf{s}$ and the gaze direction $\mathbf{g}$ is used to model the projected scene saliency s_g.

where f_g is a learnable function, parameterized by the weights $\boldsymbol{\theta}_g$. While f_g can be learned in a supervised fashion, using a number of pairs of 3D gaze direction and appearance, the learned model may not generalize well unless the training set is very large and diverse.

Our main insight is that the 3D gaze direction is highly correlated with the visual semantics (or visual saliency) observed by the scene image $\mathbf{I}_s \in [0,1]^{H_s \times W_s \times 3}$ with its height H_s and width W_s. We represent a measure of visual saliency s_g over the scene image as follows:

$$s_g(\mathbf{x}) = \frac{\exp\left(\kappa \mathbf{g}^\mathsf{T} \mathbf{s}(\mathbf{x})\right)}{\sum_{\mathbf{x} \in \mathcal{R}(\mathbf{I}_s)} \exp(\kappa \mathbf{g}^\mathsf{T} \mathbf{s}(\mathbf{x}))}, \tag{3}$$

where $s_g(\mathbf{x}) \in [0,1]$ is the visual saliency geometrically derived from the 3D gaze direction $\mathbf{g}$, the pixel $\mathbf{x} \in [0, W_s) \times [0, H_s)$ lies in the scene image, and κ is a concentration parameter that determines the variance of salience given the 3D gaze direction[3].

The directional unit vector $\mathbf{s}(\mathbf{x}) \in \mathbb{S}^2$ corresponds to scene image point $\mathbf{x}$:

$$\mathbf{s}(\mathbf{x}) = \mathbf{R} \frac{\mathbf{X} - \mathbf{e}}{\|\mathbf{X} - \mathbf{e}\|}, \quad \mathbf{X} = d(\mathbf{x}) \mathbf{K}^{-1} \widetilde{\mathbf{x}}, \tag{4}$$

where $\mathbf{X} \in \mathbb{R}^3$ is the 3D point that is reconstructed from the scene image $\mathbf{x}$ given the intrinsic parameter $\mathbf{K}$ and the depth $d(\mathbf{x}) \in \mathbb{R}_+$ as shown in Fig. 2(a). $\mathbf{R} \in SO(3)$ is the rotation matrix that transforms from the scene image coordinate system to the face image coordinate system. $\mathbf{e} \in \mathbb{R}^3$ is the 3D location of the eye center, i.e., $\mathbf{s}(\mathbf{x})$ is the direction of the 3D point $\mathbf{X}$ corresponding to $\mathbf{x}$ seen from the eye location.

Figures 2(a) and 4 illustrate the geometry of gaze-driven visual saliency $s_g(\mathbf{x})$. The saliency at a pixel location $\mathbf{x}$ can be measured by the angle between the gaze direction $\mathbf{g}$ and the corresponding direction $\mathbf{s}$ that can be obtained by 3D

[3] We use a von Mises-Fisher density function where κ is equivalent to the standard deviation of a Gaussian density function.

reconstruction of $\mathbf{x}$, i.e., scene stereo reconstruction with the depth $d(\mathbf{x})$. To obtain eye locations $\mathbf{e}$ in 3D, we run OpenPose [7] to detect the eyes in the RGB image and read off depth at those pixels.

4.2 Losses and Network Design

In the previous section we described the computation of scene saliency from registered gaze. Visual saliency can also be predicted from the scene image directly using a saliency model:

$$s_s(\mathbf{x}) = f_s(\mathbf{x}, \mathbf{I}_s; \boldsymbol{\theta}_s), \tag{5}$$

where $s_s(\mathbf{x}) \in [0, 1]$ is the scene saliency at the pixel location $\mathbf{x}$, and f_s is a learnable function that predicts the saliency from a scene image, parametrized by the weights $\boldsymbol{\theta}_s$.

Ideally, the visual saliency derived by the gaze $s_g(\mathbf{x})$ agrees with the visual scene saliency predicted from the scene image $s_s(\mathbf{x})$. We leverage this relationship to allow self-supervision of gaze estimation, by encouraging consistency between $s_g(\mathbf{x})$ and $s_s(\mathbf{x})$ through a loss term:

$$\mathcal{L}_{\text{self}}(\boldsymbol{\theta}_g, \boldsymbol{\theta}_s) = \sum_{\{\mathbf{I}_g, \mathbf{I}_s\} \in \mathcal{D}} \left(\sum_{\mathbf{x} \in \mathcal{R}(\mathbf{I}_s)} (f_s(\mathbf{x}, \mathbf{I}_s; \boldsymbol{\theta}_s) - s_g(\mathbf{x}))^2 \right), \tag{6}$$

where $\mathcal{D}$ is the set of pairs of face and scene images, and $\mathcal{R}(\mathbf{I}_s) = [0, W_s) \times [0, H_s)$. No ground truth is needed to measure $\mathcal{L}_{\text{self}}$, so a large number of unlabeled data instances may be used.

Optimizing for this self-supervised loss alone may lead to a trivial solution, such as a constant gaze prediction outside the field of view of the scene image. We therefore constrain f_g and f_s with a small set of ground truth gaze data:

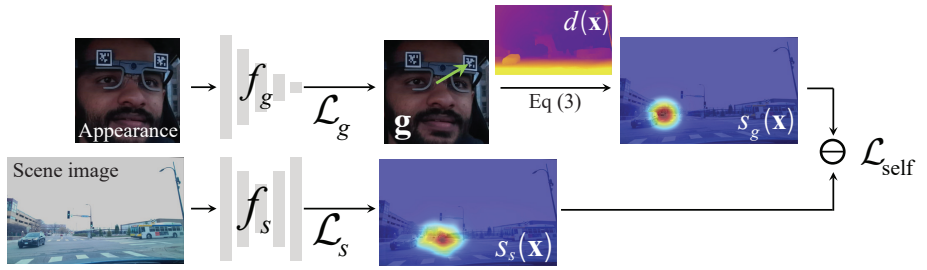
$$\mathcal{L}_g(\boldsymbol{\theta}_g) = \sum_{\{\widehat{\mathbf{g}}, \mathbf{I}_g\} \in \mathcal{D}_g} \left(1 - f_g(\mathbf{I}_g; \boldsymbol{\theta}_g)^{\mathsf{T}} \widehat{\mathbf{g}}\right)^2, \tag{7}$$

$$\mathcal{L}_s(\boldsymbol{\theta}_s) = \sum_{\{\widehat{s}, \mathbf{I}_s\} \in \mathcal{D}_s} \mathcal{L}_{\text{KL}} + \lambda_{\text{c}} \mathcal{L}_{\text{NCC}}, \tag{8}$$

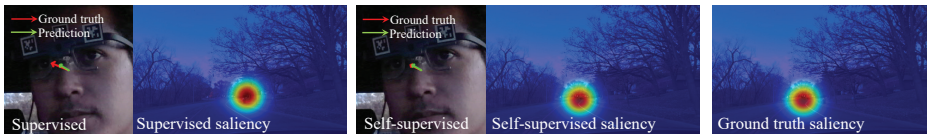
where $\mathcal{D}_g$ is the set of the ground truth pairs of the eye appearance and the 3D gaze direction where $\widehat{\mathbf{g}}$ is the ground truth 3D gaze direction, and $\mathcal{D}_s$ is the set of the ground truth pairs of the scene image and visual saliency. $\mathcal{L}_{\text{KL}}$ and $\mathcal{L}_{\text{NCC}}$ measure the Kullback-Leibler (KL) divergence and normalized cross-correlation between the visual saliency prediction f_s and the ground truth saliency, as used in [10]:

$$\mathcal{L}_{\text{KL}} = \sum_{\mathbf{x}} \widehat{s}(\mathbf{x}) \log \left(\frac{\widehat{s}(\mathbf{x})}{f_s(\mathbf{x})} \right), \quad \mathcal{L}_{\text{NCC}} = -\frac{\sum_{\mathbf{x}} \widehat{s}(\mathbf{x}) f_s(\mathbf{x})}{\sqrt{\sum_{\mathbf{x}} f_s(\mathbf{x})^2} \sqrt{\sum_{\mathbf{x}} \widehat{s}(\mathbf{x})^2}}, \tag{9}$$

where $f_s(\mathbf{x}, \mathbf{I}_s; \boldsymbol{\theta}_s)$ is denoted $f_s(\mathbf{x})$ and $\sum_{\mathbf{x} \in \mathcal{R}(\mathbf{I}_s)}$ by $\sum_{\mathbf{x}}$. λ_c is the weight to balance between KL divergence and normalized correlation, set to 0.1.



(a) Self-supervised learning network design



(b) Effect of self-supervision

Fig. 5. Model overview. (a) *Network design.* We jointly learn the gaze estimator f_g and saliency predictor f_s by enforcing the geometric consistency between them. The estimate gaze $\mathbf{g}$ is transformed to the visual saliency map $s_g(\mathbf{x})$ through Equation (3). This visual saliency map is supervised by the scene saliency prediction $s(\mathbf{x})$, i.e., minimizing $\mathcal{L}_{\text{self}}$. For the labeled data, the supervised losses $\mathcal{L}_g$ and $\mathcal{L}_s$ are used. (b) *Effect of self-supervision.* We use the geometric relationship between the gaze and scene saliency to self-supervise the gaze direction. (Left) Supervised gaze direction where the predicted gaze is deviated from the ground truth. (Middle) The self-supervision enforces the geometric consistency between the gaze and scene saliency, which improves the gaze prediction and the saliency prediction. (Right) Ground truth scene saliency.

Our overall loss is then:

$$\mathcal{L}(\boldsymbol{\theta}_g, \boldsymbol{\theta}_s) = \mathcal{L}_{\text{self}}(\boldsymbol{\theta}_g, \boldsymbol{\theta}_s) + \lambda_g \mathcal{L}_g(\boldsymbol{\theta}_g) + \lambda_s \mathcal{L}_s(\boldsymbol{\theta}_s), \tag{10}$$

where $\mathcal{L}_{\text{self}}$ is the self-supervised loss that ensures consistency between estimated gaze and estimated saliency without requiring ground truth data, and $\mathcal{L}_g$ and $\mathcal{L}_s$ are the supervised losses that prevent deviation from the ground truth. The hyperparameters λ_g and λ_s control the balance between supervised and self-supervised losses.

Our overall model is illustrated in Fig. 5(a). The gaze estimator f_g takes as input a face appearance image and outputs the gaze direction $\mathbf{g}$. With the reconstructed depth image $d(\mathbf{x})$, we transform the gaze direction to the scene image to form the visual saliency map $s_g(\mathbf{x})$. This saliency map is self-supervised by the saliency prediction of the scene image (and vice versa) via f_s by minimizing $\mathcal{L}_{\text{self}}$. When labeled data is available, we minimize the supervised losses $\mathcal{L}_g$ and $\mathcal{L}_s$ for the gaze and saliency, respectively. Fig. 5(b) illustrates the positive effect of self-supervision on both gaze and saliency estimation.

4.3 Implementation Details

Any end-to-end trainable gaze estimator f_g or saliency predictor f_s can be used in our framework, but we opt for simple but strong models which have pre-trained weights available. For gaze estimation we use the ETH XGaze model [49] based on a ResNet-50 [15], and for saliency estimation we use Unisal [10] (MNetV2-RNN-Decoder). Our framework is implemented in PyTorch [32], using the Adam optimizer [19] with a fixed 0.5×10^{-9} learning rate and a batch size of 6. For λ_g and λ_s a value of 2.0 was used. We will release all code, models and data.

5 Experiments

We split our data into three: supervised training, self-supervised training, and held-out testing. The supervised split includes the ground truth labels of gaze directions and saliency maps. The self-supervised split does not: self-supervised learning uses geometric consistency as described in the previous section to learn from this data without annotations. Three data split configurations are tested: {supervised, self-supervised, test} = {5%, 75%, 20%}, {20%, 60%, 20%}, {40%, 40%, 20%}, {60%, 20%, 20%}, using the same test split each time. Splits are by subject, to be able to assess generalization to new subjects.

5.1 Evaluation Metrics

We measure the mean absolute error (MAE) with its standard deviation for the gaze and saliency predictions:

$$\text{MAE}_g = \frac{1}{N} \sum_i \cos^{-1} \left(\mathbf{g}_i^\mathsf{T} \widehat{\mathbf{g}}_i \right), \tag{11}$$

$$\text{MAE}_s = \frac{1}{N} \sum_i \left(\sum_{\mathbf{x} \in \mathcal{R}(\mathbf{I}_s)} |s_i(\mathbf{x}) - \widehat{s}_i(\mathbf{x})| \right), \tag{12}$$

where $\mathbf{g}_i$ and $s_i(\mathbf{x})$ are the i^{th} predictions for gaze and saliency, and N is the number of test data samples.

5.2 Baselines

Gaze. We evaluate our self-supervised gaze estimation by comparing with strong recent appearance-based gaze estimation baselines. We note that our goal is not necessarily to target state-of-the-art accuracy, but rather to demonstrate the use of our dataset to explore saliency-based self-supervision to boost performance of a simple but strong baseline on real data.

(1) **Gaze360** [18]: we use the static model, pre-trained on a large-scale in-the-wild gaze dataset captured from 238 subjects in indoor and outdoor environments. This is an example of an off-the-shelf gaze estimator for "in-the-wild"

	{5/75/20}		{20/60/20}		{40/40/20}		{60/20/20}	
Method	Self	Test	Self	Test	Self	Test	Self	Test
Gaze360 [18]	18.7	20.3	21.4	20.3	23.0	20.3	17.2	20.3
ETH XGaze [49]	11.6	15.6	11.9	15.6	12.6	15.6	15.4	15.6
Mean	9.5	9.2	9.5	9.2	9.0	9.2	8.7	9.2
Supervised-only	9.7	**7.8**	**6.9**	6.8	8.1	7.4	7.0	**6.7**
Ours	**8.2**	**7.8**	**6.9**	**6.5**	**7.4**	**7.2**	**6.2**	**6.7**

Table 2. Gaze performance. We compare appearance based gaze estimation in MAE_g (degrees, lower is better). Our method yields small but consistent improvement over baselines. "Self" and "Test" correspond to performance on the Self-Supervised and Testing splits. We test on four sets of splits with varying levels of supervised and self-supervised training: {Supervised %, Self-supervised %, Test %}.

	{5/75/20}		{20/60/20}		{40/40/20}		{60/20/20}	
Method	Self	Test	Self	Test	Self	Test	Self	Test
Unisal [10]	1.57	1.60	1.57	1.60	1.56	1.60	1.58	1.60
Supervised-only	1.14	1.16	1.06	1.07	1.00	**1.03**	0.97	**1.03**
Ours	**1.12**	**1.14**	**1.05**	**1.06**	**0.99**	**1.03**	**0.96**	**1.03**

Table 3. Saliency performance. We compare saliency prediction in MAE_s (lower is better). Our method again yields small but consistent improvement over baselines. "Self" and "Test" correspond to performance on the Self-Supervised and Testing splits.We test on four sets of splits with varying levels of supervised and self-supervised training: {Supervised %, Self-supervised %, Test %}.

use. (2) **ETH XGaze** [49]: this is a ResNet-50 based model trained on the ETH XGaze dataset, a multi-view high-resolution gaze dataset captured in a controlled environment. We use the pre-trained model to evaluate on our dataset. (3) **Mean**: We compute the mean gaze over the entire LBW dataset and use it as a predictor. (4) **Supervised-only**: We re-train the ETH XGaze model (ResNet-50) on our LBW training dataset with the ground truth labels.

Saliency. (1) **Unisal** [10]: We use the saliency model pre-trained on multiple large-scale saliency datasets including DHF1K [43], Hollywood-2 and UCF-Sports [25] and SALICON [17] to evaluate on our dataset. (2) **Supervised-only**: We re-train Unisal on our LBW training dataset with ground truth labels.

5.3 Quantitative Evaluation

Tables 2 and 3 summarize the quantitative results of our experiments on gaze and saliency estimation, comparing our self-supervised learning method against the baselines described.

Gaze. The Gaze360 and ETH XGaze baselines produce relatively higher gaze estimation error for all splits because of the train-test domain gap. One significant source of difference is that all of our quantitative evaluation data features drivers wearing gaze tracking glasses, which can sometimes partially occlude the eye region and cause spurious estimations. The appearance of LBW face data is closer to the higher-resolution ETH XGaze dataset, which may explain the improvement over Gaze360. The mean gaze predictor is a competitive baseline because the attention of the driver is highly biased to the center (forward roadway and vanishing point). The supervised-only method improves performance as it allows domain adaptation using limited labeled data. Our method, which adds in geometric self-supervision, produces consistent performance improvements of up to 10% against the supervised-only method. Interestingly, its performance on the test splits is on par with the supervised-only method, indicating that it would still benefit from further adaptation. We argue that this is possible with our self-supervised learning method as no ground truth data is needed.

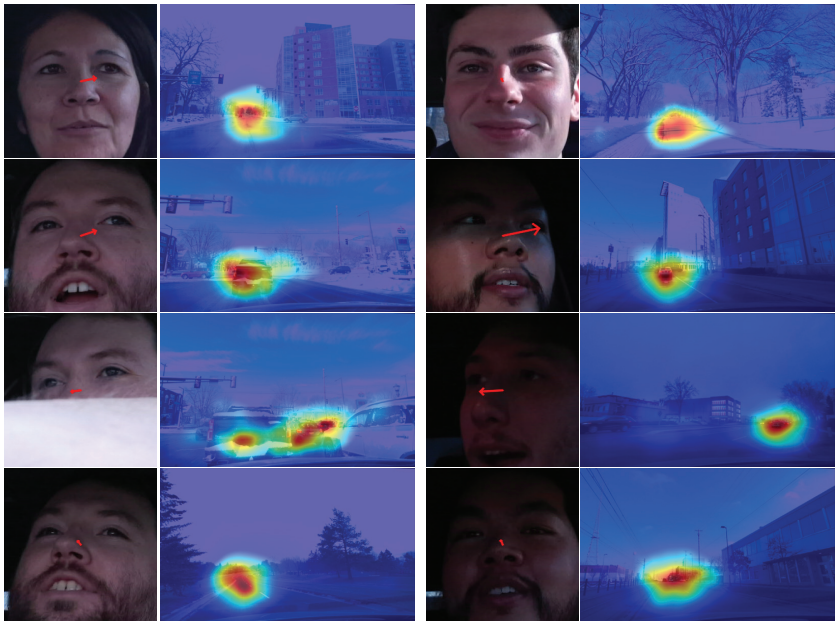
Saliency. A similar observation can be made for the saliency predictors as summarized in Table 3. Although Unisal is a competitive saliency predictor, the supervised-only method outperforms the baseline by adapting the model to the driving domain with limited labeled data. Our self-supervised learning method matches or improves on the supervised learning method consistently across splits.

5.4 Qualitative Evaluation

We show further qualitative output from our model in Fig. 6(a) over a range of drivers in the test split. It correctly predicts gaze direction in the presence of low lighting as the scene saliency provides an informative signal to refine the gaze. On the other hand, the scene saliency can sometimes mislead gaze estimation as it is biased towards the vanishing point in the scene, as shown in Fig. 6(b).

6 Discussion

We have presented a new dataset called Look Both Ways, which facilitates study into the problem of estimating a driver's focus of attention while on the road. We introduced a new approach for geometric self-supervision of 3D gaze from facial images and visual scene saliency, to take advantage of the natural relationship between the two. Using the LBW dataset, we showed that our end-to-end trained system can improve upon purely supervised methods for gaze estimation and saliency estimation, by virtue of being able to take advantage of unlabelled face and scene depth image pairs.



(a) Qualitative results for gaze and saliency using self-supervision

(b) Sample failure cases

Fig. 6. Qualitative Results. We visualize some further outputs of our model trained with self-supervised gaze and saliency. Successful results shown in (a) show the model working well in a variety of scenes and lighting conditions. Failure cases shown in (b) show how self-supervised learning can mislead the gaze estimation as the scene saliency prediction is highly biased to the vanishing point.

We believe that our dataset will be helpful for the community to further study driver attention in vehicles. Although we acknowledge that 3D gaze technology can potentially be used for surveillance applications, we hope to inspire an application that positively influences our thinking about the use of gaze estimation in vehicles, as a means to support improved assistance for drivers on the road.

Acknowledgments. This research is based on work supported by Toyota Research Institute and the NSF under IIS #1846031. The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the sponsors.

References

1. International Data Corporation, Worldwide Autonomous Vehicle Forecast, 2020–2024 (2020)
2. SAE Levels of Driving Automation Refined for Clarity and International Audience (2021), https://www.sae.org/blog/sae-j3016-update
3. Baee, S., Pakdamanian, E., Kim, I., Feng, L., Ordonez, V., Barnes, L.: Medirl: Predicting the visual attention of drivers via maximum entropy deep inverse reinforcement learning. In: ICCV (2021)
4. Baluja, S., Pomerleau, D.: Non-intrusive gaze tracking using artificial neural networks (1993)
5. Bylinskii, Z., Recasens, A., Borji, A., Oliva, A., Torralba, A., Durand, F.: Where should saliency models look next? In: ECCV (2016)
6. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuScenes: A multimodal dataset for autonomous driving. In: CVPR (2020)
7. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: OpenPose: Realtime multi-person 2d pose estimation using part affinity fields. TPAMI (2019)
8. Chang, Z., Matias Di Martino, J., Qiu, Q., Espinosa, S., Sapiro, G.: SalGaze: Personalizing gaze estimation using visual saliency. In: ICCV Workshops (2019)
9. Deng, H., Zhu, W.: Monocular free-head 3d gaze tracking with deep learning and geometry constraints. In: ICCV (2017)
10. Droste, R., Jiao, J., Noble, J.A.: Unified image and video saliency modeling. In: ECCV (2020)
11. Fang, J., Yan, D., Qiao, J., Xue, J., Yu, H.: Dada: Driver attention prediction in driving accident scenarios. IEEE Transactions on Intelligent Transportation Systems (2021)
12. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. ACM Communications (1981)
13. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In: CVPR (2012)
14. Hansen, D.W., Ji, Q.: In the eye of the beholder: A survey of models for eyes and gaze. TPAMI (2009)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
16. Jain, A., Koppula, H.S., Raghavan, B., Soh, S., Saxena, A.: Car that knows before you do: Anticipating maneuvers via learning temporal driving models. In: ICCV (2015)
17. Jiang, M., Huang, S., Duan, J., Zhao, Q.: Salicon: Saliency in context. In: CVPR (2015)
18. Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., Torralba, A.: Gaze360: Physically unconstrained gaze estimation in the wild. In: ICCV (2019)
19. Kingma, D.P., Ba, J.L.: Adam : A method for stochastic optimization. In: arXiv (2014)
20. Land, M.F.: Eye movements and the control of actions in everyday life. Progress in retinal and eye research (2006)
21. Lindén, E., Sjostrand, J., Proutiere, A.: Learning to personalize in appearance-based gaze tracking. In: ICCV Workshops (2019)

22. Lipson, L., Teed, Z., Deng, J.: Raft-stereo: Multilevel recurrent field transforms for stereo matching. In: 23DV (2021)
23. Lowe, D.G.: Object recognition from local scale-invariant features. IJCV (1999)
24. Martin, M., Roitberg, A., Haurilet, M., Horne, M., Reiss, S., Voit, M., Stiefelhagen, R.: Drive&Act: A Multi-Modal Dataset for Fine-Grained Driver Behavior Recognition in Autonomous Vehicles. In: ICCV (2019)
25. Mathe, S., Sminchisescu, C.: Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. TPAMI (2015)
26. Min, K., Corso, J.J.: Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In: ICCV (2019)
27. Ortega, J.D., Kose, N., Cañas, P., Chao, M.a., Unnervik, A., Nieto, M., Otaegui, O., Salgado, L.: DMD : A Large-Scale Multi-Modal Driver Monitoring Dataset for Attention and Alertness Analysis. In: ECCV Workshops (2020)
28. Palazzi, A., Abati, D., Solera, F., Cucchiara, R., et al.: Predicting the Driver's Focus of Attention: the DR(eye)VE Project. TPAMI (2018)
29. Park, S., Aksan, E., Zhang, X., Hilliges, O.: Towards end-to-end video-based eye-tracking. In: ECCV (2020)
30. Park, S., Mello, S.D., Molchanov, P., Iqbal, U., Hilliges, O., Kautz, J.: Few-shot adaptive gaze estimation. In: ICCV (2019)
31. Park, S., Spurr, A., Hilliges, O.: Deep pictorial gaze estimation. In: ECCV (2018)
32. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An imperative style, high-performance deep learning library. In: NeurIPS (2019)
33. Recasens, A., Khosla, A., Vondrick, C., Torralba, A.: Where are they looking? In: NeurIPS (2015)
34. Recasens, A., Vondrick, C., Khosla, A., Torralba, A.: Following gaze in video. In: ICCV (2017)
35. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR (2016)
36. Shen, C., Zhao, Q.: Webpage saliency. In: ECCV (2014)
37. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: CVPR (2017)
38. Sugano, Y., Matsushita, Y., Sato, Y.: Appearance-based gaze estimation using visual saliency. TPAMI (2013)
39. Sugano, Y., Matsushita, Y., Sato, Y.: Learning-by-synthesis for appearance-based 3d gaze estimation. In: CVPR (2014)
40. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: CVPR (2020)
41. Sun, Y., Zeng, J., Shan, S., Chen, X.: Cross-encoder for unsupervised gaze representation learning. In: ICCV (2021)
42. Wang, J., Olson, E.: AprilTag 2: Efficient and robust fiducial detection. In: IROS (2016)
43. Wang, W., Shen, J., Xie, J., Cheng, M.M., Ling, H., Borji, A.: Revisiting video saliency prediction in the deep learning era. TPAMI (2021)
44. Wood, E., Baltrusaitis, T., Zhang, X., Sugano, Y., Robinson, P., Bulling, A.: Rendering of eyes for eye-shape registration and gaze estimation. In: ICCV (2015)

45. Wu, T., Martelaro, N., Stent, S., Ortiz, J., Ju, W.: Learning When Agents Can Talk to Drivers Using the INAGT Dataset and Multisensor Fusion. ACM Interactive, Mobile, Wearable and Ubiquitous Technologies (2021)
46. Xia, Y., Zhang, D., Kim, J., Nakayama, K., Zipser, K., Whitney, D.: Predicting driver attention in critical situations. In: ACCV (2018)
47. Yarbus, A.L.: Eye movements and vision. Springer (2013)
48. Yu, Y., Odobez, J.M.: Unsupervised representation learning for gaze estimation. In: CVPR (2020)
49. Zhang, X., Park, S., Beeler, T., Bradley, D., Tang, S., Hilliges, O.: ETH-XGaze: A Large Scale Dataset for Gaze Estimation under Extreme Head Pose and Gaze Variation. In: ECCV (2020)
50. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: CVPR (2015)
51. Zheng, Q., Jiao, J., Cao, Y., Lau, R.: Task-driven webpage saliency. In: ECCV (2018)