Driver Gaze Estimation in the Real World:
Overcoming the Eyeglass Challenge

Akshay Rangesh†, Bowen Zhang‡ and Mohan M. Trivedi

Abstract—A driver’s gaze is critical for determining the driver’s attention level, state, situational awareness, and readiness to take over control from partially and fully automated vehicles. Tracking both the head and eyes (pupils) can provide reliable estimation of a driver’s gaze using face images under ideal conditions. However, the vehicular environment introduces a variety of challenges that are usually unaccounted for - harsh illumination, nighttime conditions, and reflective/dark eyeglasses. Unfortunately, relying on head pose alone under such conditions can prove to be unreliable owing to significant eye movements. In this study, we offer solutions to address these problems encountered in the real world. To solve issues with lighting, we demonstrate that using an infrared camera with suitable equalization and normalization usually suffices. To handle eyeglasses and their corresponding artifacts, we adopt the idea of image-to-image translation using generative adversarial networks (GANs) to pre-process images prior to gaze estimation. To this end, we propose the Gaze Preserving CycleGAN (GPCycleGAN). As the name suggests, this network preserves the driver’s gaze while removing potential eyeglasses from infrared face images. GPCycleGAN is based on the well-known CycleGAN approach, with the addition of a gaze classifier and a gaze consistency loss for additional supervision. Our approach exhibits improved performance and robustness on challenging real-world data spanning 13 subjects and a variety of driving conditions.

I. INTRODUCTION

Driver safety and crash risk are highly dependent on a driver’s attention levels, especially visual attention, for both traditional and self-driving vehicles [1]–[3]. For intelligent vehicles, it is essential to have the ability to monitor the driver’s gaze continuously and use this information to effectively enhance vehicle safety. Visual cues from imaging sensors and corresponding learning algorithms have proven to be effective ways of determining driver gaze zones [4]–[7]. However, the state-of-the-art for gaze estimation does not account for the complexities of the real world (for example - drivers wearing eyeglasses, harsh illumination, nighttime conditions etc.). According to the Vision Impact Institute [8], [9], one in five drivers have a vision problem and most of them wear correction eyeglasses while driving. Additionally, according to the National Safety Council, 50% of all accidents happen while driving in the dark [10]. However, it is generally hard to estimate the gaze accurately using RGB images in dark environments, primarily because RGB cameras suffer from a lack of photons captured by imaging sensors, resulting in low signal-to-noise ratios [11].

In this study, we find that using infrared cameras can mitigate some of these lighting issues; however, for face images with eyeglasses, data pre-processing or algorithmic improvements might be necessary to determine the driver’s gaze zones. Since previous works on gaze estimation [6] obtained good results (under ideal conditions) by using convolutional neural networks (CNNs), we adopt their methodology and concentrate on developing novel pre-processing approaches that can improve gaze estimation in these demanding conditions.

Generative Adversarial Networks (GANs) [12] have shown promising results on a variety of computer vision tasks like generating realistic face images [13], style transfer, image editing [14], [15], image super-resolution [16] etc. GANs are unsupervised models with two sub-models, namely a generator and a discriminator. The generator updates itself to synthesize images that are indistinguishable from real images by learning the data distribution. The discriminator is then tasked with differentiating between real and synthesized images. They are trained together in an adversarial manner until an equilibrium is attained. More recent work on GANs have focused on addressing key issues like training on large datasets, training larger models, improving stability during training, preserving finer details etc. In this study, we build on such a model called the CycleGAN [15], which achieved the state-of-the-art on image-to-image translation by using unpaired images from the source and target distributions.

Inspired by the CycleGAN architecture, we propose the Gaze Preserving CycleGAN (GPCycleGAN) which makes use of an additional gaze consistency loss. GPCycleGAN has the following advantages compared to other gaze estimation methods and glass removal techniques: First, GPCycleGAN...

†authors contributed equally
Code: https://github.com/arangesh/GPCycleGAN
The authors are with the Laboratory for Intelligent and Safe Automobiles at the University of California, San Diego, CA 92092, USA.
email: {arangesh, boz004, mtrivedi}@ucsd.edu

Fig. 1: The real world introduces variability and complexity that driver gaze estimation systems usually ignore. Some examples include - use of eyeglasses, harsh illumination, nighttime data etc.
preserves the gaze and allows for more accurate gaze estimation on images with eyeglasses. Second, unlike previous works [20], [21] on eyeglass removal, there is no need for paired images to train GP CycleGAN. Third, it works in different environments, lighting conditions, eyeglass types, and with significant variations of the head pose.

The four main contributions of this paper are: a) An in-depth analysis of traditional gaze estimation under different conditions (e.g., with and without eyeglasses, daytime, nighttime etc.), and its shortcomings, b) The GP CycleGAN model for eyeglass removal specifically optimized for the gaze classification task, c) Experimental analyses to illustrate how the GP CycleGAN model improves the accuracy over a variety of baseline models, and d) A naturalistic driving dataset with labeled gaze zones (to be released).

II. RELATED RESEARCH

We outline studies from two research areas - gaze estimation and eyeglass removal from face images. Table I lists selected contemporary research studies in two sub-tables separated by topic. For gaze estimation, we mainly focus on vision-based approaches post 2016. For earlier works on specific topics, please refer to the following studies: Vora et al. [6] for gaze estimation; Kar et al. [24] for eye-gaze estimation systems, algorithms, and performance evaluation methods in consumer platforms; a survey on driver behavior analysis for safe driving by Kaplan et al. [25]; a review on driver inattention monitoring systems by Dong et al. [26]; and a survey on head pose estimation by Murphy et al. [27]. For the second research area, we were unable to find published studies focused on eyeglass removal for drivers’ face images. We instead provide discussion on eyeglass removal for other applications.

A. Gaze Estimation

Studies on gaze estimation can be categorized in different ways. First, by the type of model used, they can be divided into convolutional neural networks [6], statistical learning models [28], or geometric approaches [5]. Second, methods can be distinguished by the cues they consider, i.e. studies that only use head pose [4] versus ones that use both head pose and eye information [6], [7], [29], [30]. Third, research can be categorized by the conditions they capture. For instance, studies with limited illumination changes [6] versus research with multiple environments and the variations that come with it [7]. Lastly, studies can be separated by the sensors they use, including RGB cameras [6], IR cameras [31], NIR cameras [17], multiple cameras [32], and wearable sensors [33]. However, only a few studies emphasize the problems associated with driving with eyeglasses or eyewear. Tawari et al. [34] and Lee et al. [35] mention the unreliability of gaze estimation for drivers with glasses, and propose methods that only rely on head pose as fallback solutions. Naqvi et al. [17] combine head pose estimation and pupil detection to determine the eye gaze, but their method only works under ideal capture conditions. Jah and Busso [4] use only head pose in their method. Wang et al. [18] combine depth images of the head and RGB images of eyes, but gaze estimation for eye images is unstable and only works with frontal images under ideal conditions. In [7], Yoon et al. collect a dataset comprising of images in daytime/nighttime, images with and without eyeglasses using two NIR cameras.

TABLE I: Related research

(a) Selected research on gaze estimation in vehicular environments

<table>
<thead>
<tr>
<th>Study</th>
<th>Objective</th>
<th>Sensor</th>
<th>Features</th>
<th>Capture conditions</th>
<th>Methodology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vora et al. [6]</td>
<td>Gaze zone classification using CNNs</td>
<td>1 RGB camera</td>
<td>HP &amp; gaze</td>
<td>Daytime; w/o eyeglasses</td>
<td>CNN</td>
</tr>
<tr>
<td>Martin et al. [5]</td>
<td>Estimating gaze dynamics, glance duration and frequency</td>
<td>1 RGB camera</td>
<td>HP &amp; gaze</td>
<td>Daytime; w/o eyeglasses</td>
<td>CNN &amp; geometry</td>
</tr>
<tr>
<td>Naqvi et al. [17]</td>
<td>Gaze zone detection using NIR camera &amp; deep learning</td>
<td>1 NIR camera &amp; NIR LEDs</td>
<td>HP &amp; gaze</td>
<td>Daytime &amp; nighttime; w &amp; w/o eyeglasses</td>
<td>CNN</td>
</tr>
<tr>
<td>Yong et al. [7]</td>
<td>Gaze detection using dual NIR cameras and deep residual networks</td>
<td>2 NIR cameras &amp; NIR LEDs</td>
<td>HP &amp; gaze</td>
<td>Daytime &amp; nighttime; w &amp; w/o eyeglasses</td>
<td>CNN</td>
</tr>
<tr>
<td>Jha &amp; Busso [4]</td>
<td>Gaze region estimation using dense pixelwise predictions</td>
<td>1 RGB camera &amp; 1 headband with ApriTags</td>
<td>HP</td>
<td>Daytime; w &amp; w/o eyeglasses</td>
<td>Dense Neural Networks</td>
</tr>
<tr>
<td>Wang et al. [18]</td>
<td>Continuous gaze estimation using RGB-D camera</td>
<td>1 RGB-D camera</td>
<td>HP &amp; gaze</td>
<td>Daytime; w &amp; w/o eyeglasses</td>
<td>Feature extraction &amp; k-NN</td>
</tr>
</tbody>
</table>

(b) Selected research on eyeglass removal and related topics

<table>
<thead>
<tr>
<th>Study</th>
<th>Methodology</th>
<th>Dataset</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhu et al. [15]</td>
<td>Cycle-GAN</td>
<td>Unpaired images</td>
<td>Good performance on style transfer</td>
<td>Not realistic for eyeglass removal; does not preserve gaze direction</td>
</tr>
<tr>
<td>Hu et al. [19]</td>
<td>ER-GAN</td>
<td>Unpaired CelebA &amp; LFW</td>
<td>Good performance on style transfer and eyeglass removal</td>
<td>Eyes are often swapped; gaze is not preserved</td>
</tr>
<tr>
<td>Amadio et al. [19]</td>
<td>TraVelGAN</td>
<td>Unpaired Images from multiple domains</td>
<td>Single step, end-to-end method</td>
<td>Images need to be aligned; needs paired images; uses synthesized eyeglasses</td>
</tr>
<tr>
<td>Wang et al. [20]</td>
<td>EL-GAN &amp; LS-GAN</td>
<td>Paired CelebA</td>
<td>Better results than Cycle-GAN; improved face recognition accuracy</td>
<td>Needs an obstruction classifier</td>
</tr>
<tr>
<td>Li et al. [22]</td>
<td>DIAT-GAN</td>
<td>Aligned CelebA</td>
<td>Flexibility to modify different facial attributes</td>
<td>Generated gaze in different, blurry outputs</td>
</tr>
<tr>
<td>Shen et al. [23]</td>
<td>GAN</td>
<td>CelebA &amp; LFW</td>
<td>Capability to manipulate images with modest pixel modifications</td>
<td>Does not completely remove eyeglasses; the eye region is not preserved</td>
</tr>
</tbody>
</table>

1 Head Pose
2 Near Infra-Red
3 Generative Adversarial Networks
and NIR lights. Although they achieve good performance, they do not explicitly model the presence of eyeglasses in images. In this study, we show that such approaches tend not to generalize to different settings and usually overfit to the training set.

B. Eyeglass Removal

Eyeglass removal is never a straightforward task due to large variations in head pose, eyeglass type, and the environment, and due to the presence reflection artifacts. Most eyeglass removal studies make use of statistical learning, principal component analysis (PCA), or deep learning, including GAN based methods. Statistical learning and PCA were the primary approaches prior to the advent of deep learning. These approaches need less computational power but have limitations on eyeglasses, environmental conditions, and head poses [36], [37] they can handle reliably. On the other hand, methods using deep learning, such as ones proposed by Liang et al. [21] and Wang et al. [20] modified GANs for better results, but both these methods need paired and aligned images for training. Such datasets are expensive and tedious to collect. Different GAN architectures proposed in [15], [19], and [38] do not require paired images, but their models fail to keep gaze information and do not perform well on non-frontal images. Models developed by Li et al. [22] and Shen et al. [23] are capable of changing multiple facial attributes but do not produce satisfying results on eyeglass removal. All above methods on eyeglass removal have different and somewhat general purposes. However, none of these methods are trained to preserve the gaze of face images, and hence cannot necessarily be used out-of-the-box for gaze estimation.

III. DATASET

Since our primary goal is to design a gaze estimation system for the real-world, we prioritized using small form-factor infrared cameras with suitable real-time performance. We decided on an Intel RealSense IR camera and enclosed it in a custom 3D printed enclosure mounted next to the rearview mirror. To ensure a good compromise between larger fields-of-view and faster processing speeds, we settled on a capture resolution of 480×480. Similar to Vora et al. [6], we divide the driver’s gaze into seven gaze zones: Eyes Closed/Lap, Forward, Left Mirror, Speedometer, Radio, Rearview, Right Mirror. Unlike previous studies, we also include gaze zones related to driver inattention or unsafe driving behaviour. Our entire dataset comprises of thirteen subjects in different lighting conditions (daytime, nighttime and harsh lighting), wearing a variety of eyeglasses. For every gaze zone, participants were instructed to keep their gaze fixed while moving their heads within reasonable limits. In total, 336177 frames of images were captured, which we split into training, validation and test sets with no overlap of subjects. Table II shows the distribution of images and subjects across different splits and capture conditions. Figure 2 depicts exemplar images from our dataset. We ensure that the dataset is suitably diverse and challenging to represent the complexities observed in the real world.

IV. METHODOLOGY

As depicted in Fig. 3, the steps involved in our proposed gaze estimation pipeline are as follows: (a) landmark detection using OpenPose [39], (b) eye image cropping, resizing, and equalization, (c) gaze estimation. We use OpenPose for landmark detection because of its high accuracy under different conditions and fast inference speed. Based on the work by Vora et al. [6], we crop the eye region using the estimated landmarks as per their conclusion that the upper half of the face as an input produced the best results for downstream gaze classification. Next, we use adaptive histogram equalization to improve the contrast and resize the images to 256×256 before feeding them to the gaze estimation models.

A. Issues with Lighting & Eyeglasses

To understand the impacts of different conditions on gaze estimation, we carry out an extensive experiment to analyze the performance of models trained with subsets of data, when they are tested on data from within and outside the training distribution. Table III shows validation accuracies of SqueezeNet-based gaze classifier models 1 - 9 (as proposed in [6]), each trained on data captured under different conditions 4 - 7. From the Table, we glean that model 5 validated on data 4, 6, and 7 produce similar accuracies, and model 9 validated on data 2, 5, and 7 also perform similarly. This demonstrates that the gaze classifier models work well on daytime and nighttime data when trained on

<table>
<thead>
<tr>
<th>Capture conditions</th>
<th>Dataset split</th>
<th>Training</th>
<th>Validation</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>daytime; w/o eyeglasses</td>
<td>(43432, 5)</td>
<td>(9062, 1)</td>
<td>(3294, 4)</td>
<td></td>
</tr>
<tr>
<td>nighttime; w/o eyeglasses</td>
<td>(59352, 9)</td>
<td>(8510, 1)</td>
<td>(2768, 4)</td>
<td></td>
</tr>
<tr>
<td>daytime; w/ eyeglasses</td>
<td>(33189, 5)</td>
<td>(8103, 1)</td>
<td>(2897, 4)</td>
<td></td>
</tr>
<tr>
<td>nighttime; w/ eyeglasses</td>
<td>(33189, 5)</td>
<td>(8103, 1)</td>
<td>(2897, 4)</td>
<td></td>
</tr>
</tbody>
</table>

Total (all conditions) | (203124, 9) | (35583, 1) | (11717, 4) |

Fig. 2: Example images from our dataset under different capture conditions.
between images with and without eyeglasses. In this study, we propose to do the latter by training an eyeglass removal network.

Thus, we need to data with and without eyeglasses is not sufficient to ensure glasses (c) always has better accuracies when validated (x) ∈ R×W×1, H = 256, and W = 256). We denote x (x ∈ X) and y (y ∈ Y) as sample images from domains X and Y respectively.

First, we consider a baseline gaze classifier model (Fig. 4(a)), for which we use input images from both domains. We use the SqueezeNet architecture as before and denote its output gaze zone probabilities as \{p_i\}. Standard cross-entropy loss (L_{CE}) is used for training the gaze classifier. L_{CE} is defined as:

\[
L_{CE} = -\sum_i y_i \log p_i, \tag{1}
\]

where \(i\) is the class index and \(y_i\) is the ground truth class indicator.

Next, we consider a CycleGAN-based eyeglass removal network, followed by a SqueezeNet-based gaze classifier (Fig. 4(b)). The CycleGAN-based eyeglass removal model learns the mapping \(G_{w/o} : Y \rightarrow X\) (removing eyeglasses), denoted as \(G_{w/o}: Y \rightarrow X\). When trained, \(G_{w/o}\) generates images \(G_{w/o}(\cdot)\) using the input from either domain, and then passes them to the pre-trained gaze classifier to generate output probabilities \(\{p_i\}\). The gaze classifier in this model is only pre-trained using eye crop images without eyeglasses, \(x\), since we assume that \(G_{w/o}\) is able to remove eyeglasses from \(y\). In a similar manner, the mapping for adding eyeglasses on \(x\) is \(G_{w/}: X \rightarrow Y\), and it is trained simultaneously with \(G_{w/o}\). The discriminator \(D_{w}\) aims to distinguish between real images \(y\) and generated images \(G_{w/o}(x)\), whereas the discriminator \(D_{w/o}\) aims to do so between real images \(x\) and generated images \(G_{w/o}(y)\).

The losses used in this model are: cycle consistency losses [15] (Eq. 2) that encourages cycle consistency between the real images and the reconstructed images; adversarial losses [12] (Eq. 3) for making the generated images indistinguishable from real images; and identity losses [40] (Eq. 4) to ensure identity mapping when the input belongs to the target domain.

\[
\begin{align*}
L_{cyc}(G_{w/}, G_{w/o}) & = E_{x \sim P(x)} \left[ ||G_{w/o}(G_{w/}(x)) - x||_1 \right] \\
& + E_{y \sim P(y)} \left[ ||G_{w/}(G_{w/o}(y)) - y||_1 \right], \tag{2}
\end{align*}
\]

\[
\begin{align*}
L_{adv}(G_{w/}, G_{w/o}, D_{w/}, D_{w/o}) & = E_{y \sim P(y)} \left[ \log D_{w/o}(y) \right] \\
& + E_{x \sim P(x)} \left[ \log \left( 1 - D_{w/}(G_{w/}(x)) \right) \right] \\
& + E_{y \sim P(y)} \left[ \log \left( 1 - D_{w/o}(G_{w/o}(y)) \right) \right], \tag{3}
\end{align*}
\]

\[
\begin{align*}
L_{id}(G_{w/o}, D_{w/o}) & = E_{y \sim P(y)} \left[ \log D_{w/o}(y) \right] \\
& + E_{y \sim P(y)} \left[ \log \left( 1 - D_{w/o}(G_{w/o}(y)) \right) \right]. \tag{4}
\end{align*}
\]

First, we consider a baseline gaze classifier model (Fig. 4(a)), for which we use input images from both domains. We use the SqueezeNet architecture as before and denote its output gaze zone probabilities as \(\{p_i\}\). Standard cross-entropy loss (\(L_{CE}\)) is used for training the gaze classifier. \(L_{CE}\) is defined as:

\[
L_{CE} = -\sum_i y_i \log p_i, \tag{1}
\]

where \(i\) is the class index and \(y_i\) is the ground truth class indicator.

Next, we consider a CycleGAN-based eyeglass removal network, followed by a SqueezeNet-based gaze classifier (Fig. 4(b)). The CycleGAN-based eyeglass removal model learns the mapping \(G_{w/o} : Y \rightarrow X\) (removing eyeglasses), denoted as \(G_{w/o}: Y \rightarrow X\). When trained, \(G_{w/o}\) generates images \(G_{w/o}(\cdot)\) using the input from either domain, and then passes them to the pre-trained gaze classifier to generate output probabilities \(\{p_i\}\). The gaze classifier in this model is only pre-trained using eye crop images without eyeglasses, \(x\), since we assume that \(G_{w/o}\) is able to remove eyeglasses from \(y\). In a similar manner, the mapping for adding eyeglasses on \(x\) is \(G_{w/}: X \rightarrow Y\), and it is trained simultaneously with \(G_{w/o}\). The discriminator \(D_{w}\) aims to distinguish between real images \(y\) and generated images \(G_{w/o}(x)\), whereas the discriminator \(D_{w/o}\) aims to do so between real images \(x\) and generated images \(G_{w/o}(y)\).

The losses used in this model are: cycle consistency losses [15] (Eq. 2) that encourages cycle consistency between the real images and the reconstructed images; adversarial losses [12] (Eq. 3) for making the generated images indistinguishable from real images; and identity losses [40] (Eq. 4) to ensure identity mapping when the input belongs to the target domain.

\[
\begin{align*}
L_{cyc}(G_{w/}, G_{w/o}) & = E_{x \sim P(x)} \left[ ||G_{w/o}(G_{w/}(x)) - x||_1 \right] \\
& + E_{y \sim P(y)} \left[ ||G_{w/}(G_{w/o}(y)) - y||_1 \right], \tag{2}
\end{align*}
\]

\[
\begin{align*}
L_{adv}(G_{w/}, G_{w/o}, D_{w/}, D_{w/o}) & = E_{y \sim P(y)} \left[ \log D_{w/o}(y) \right] \\
& + E_{x \sim P(x)} \left[ \log \left( 1 - D_{w/}(G_{w/}(x)) \right) \right] \\
& + E_{y \sim P(y)} \left[ \log \left( 1 - D_{w/o}(G_{w/o}(y)) \right) \right], \tag{3}
\end{align*}
\]

\[
\begin{align*}
L_{id}(G_{w/o}, D_{w/o}) & = E_{y \sim P(y)} \left[ \log D_{w/o}(y) \right] \\
& + E_{y \sim P(y)} \left[ \log \left( 1 - D_{w/o}(G_{w/o}(y)) \right) \right]. \tag{4}
\end{align*}
\]
and,

\[
L_{\text{identity}}(G_{w/}, G_{w/o}) = \mathbb{E}_{y \sim P(y)} \left[ \| G_{w/}(y) - y \|_1 \right] + \mathbb{E}_{x \sim P(x)} \left[ \| G_{w/o}(x) - x \|_1 \right],
\]

where \( x \sim P(x) \) and \( y \sim P(y) \) are the image distributions.

The full objective for the CycleGAN model is the following:

\[
L_{\text{total}} = L_{\text{adv}} + \lambda_1 L_{\text{cyc}} + \lambda_2 L_{\text{identity}},
\]

where \( \lambda_1 \) is the weight of the cycle loss, and \( \lambda_2 \) is the weight of the identity loss.

Finally, our proposed GPCycleGAN model with the gaze classifier (Fig. 4(c)) builds on the previous model, with the addition of a gaze identity loss \( L_{\text{gaze}} \). To preserve the gaze features during eyeglass removal, we use the trained gaze classifier from step 1 and compute the gaze consistency loss using its Class Activation Maps (CAMs). The proposed gaze identity loss is defined as follows:

\[
L_{\text{gaze}}(G_{w/}, G_{w/o}) = \frac{1}{N |X|} \sum_{i=1}^{N} \sum_{(x,y) \in X} (A_i^{w/o,real}(x,y) - A_i^{w/o,rec}(x,y))^2,
\]

where \( N \) is total the number of classes, \( i \) is the class index, and \( A_i \)'s are the CAMs from the trained gaze classifier. Subscript \( w/o, real \) denotes CAMs obtained from a real eye image crop without eyeglasses, and subscript \( w/o, rec \) denotes CAMs corresponding to reconstructed images \( G_{w/o}(\cdot) \) using the same eye image crop. The nature of the proposed gaze loss necessitates the use of a cyclic structure, as we do not have perfectly paired samples of images with and without glasses to enforce gaze consistency. The full objective for the proposed GPCycleGAN model is:

\[
L_{\text{total}} = L_{\text{adv}} + \lambda_1 L_{\text{cyc}} + \lambda_2 L_{\text{identity}} + \lambda_3 L_{\text{gaze}},
\]

where \( \lambda_3 \) is the weight for the gaze identity loss.

Fig. 4: Training setup and architectures for different gaze zone estimation models.
C. Implementation Details

The training and inference architectures are the same for the baseline gaze classification models 1-9, but different for models A and B. Models 1-9 are adopted from Vora et al. [6], and differ in the data they were trained on (see Table III). Models A and B are identical in terms of their architectures, and only differ in their training setup and losses (i.e., the addition of a gaze consistency loss in model B). Both models consist of generators with 9 residual blocks and 70 × 70 PatchGANs [15] as discriminators. As illustrated in Fig. 4, training models A and B proceeds in 3 steps: Step 1 involves training a SqueezeNet gaze classifier using eye image crops from domain X; Step 2 is for training the generator G_{w/o} in CycleGAN/GPCycleGAN; Step 3 is to fine-tune the gaze classifier from step 1 using generated images G_{w/o}(x, y). The inference for models 1-9 (Fig. 5(a)) is a simple forward propagation through the gaze classifier to obtain the gaze zone probabilities. Models A and B have the same inference setup (Fig. 5(b)), where eye crop images are first passed to the generator G_{w/o} for eyeglass removal, after which they are fed to the gaze classifier which outputs the gaze zone probabilities.

We choose \( \lambda_1 = 10, \lambda_2 = 5 \) in Eq. 5, and \( \lambda_3 = 10 \) in Eq. 7. We use a learning rate of 0.0005 with an SGD optimizer for training the SqueezeNet classifiers, and a learning rate of 0.0002 with an Adam optimizer for CycleGAN/GPCycleGAN. The gaze classifiers are trained for a total of 30 epochs, while the GANs are trained for 15 epochs.

V. EXPERIMENTAL ANALYSIS

Table IV presents the cross-subject test accuracies for 6 different models. We use two metrics, namely the macro-average and micro-average accuracy, defined as follows:

\[
\text{Micro-average accuracy} = \frac{\sum_{i=1}^{N} (\text{True positives})_i}{\sum_{i=1}^{N} (\text{Total population})_i}, \quad (9)
\]

where \( N \) is the number of classes. Micro-average accuracy represents the overall percentage of correct predictions, while macro-average accuracy represents the average of all per-class accuracies.

First, we consider model 5 trained only on images without eyeglasses to illustrate the domain gap between images with and without eyeglasses. As can be seen, this model performs poorly on a test set containing images outside its training distribution. Next, model 9 represents the scenario where the eyeglasses are not explicitly modelled. Although it is trained on the entire training set, the resulting accuracies indicate a performance penalty, especially when tested on images with eyeglasses.

Next, we see that adding a pre-processing network to remove eyeglasses such as in models A and B improves the accuracies over the baseline model 9. Nonetheless, the improvement is meager, with model B producing accuracies slightly higher than the baseline and model A. However, after fine-tuning using the entire training set, the model B accuracies increase considerably, while that of model A remain relatively fixed. In conclusion, our proposed model B demonstrates significant improvement over both the baseline model 9 and the vanilla CycleGAN-based model A after fine-tuning. The above evidence implies the benefits of our proposed gaze consistency loss, and demonstrates that the generator resulting from the GPCycleGAN model acts effectively as a pre-processing step for the downstream task of gaze estimation.

In addition to the test set accuracies, we also present the corresponding confusion matrices of all 6 models in Fig. 7. The error modes of our best performing model (Fig. 7f) can mostly be attributed to confusion between gaze zones close in physical space (for example, Forward versus Speedometer), occlusion of the eyes by the eyeglass frame and glare, and the inability to distinguish between looking downwards versus closed eyes.

To carry out qualitative comparison between different GAN variants, we also show 10 examples of eyeglass removal on real images using CycleGAN and GPCycleGAN in Fig. 6. In columns i, iii, v, vi, and viii, GPCycleGAN not only removes the eyeglasses, but also removes the glare resulting from it; whereas CycleGAN perceives the glare as part of the sclera. Glare removal is essential for gaze estimation because glare from glasses is a relatively common occurrence in the real world, and often occludes the eyes, making it harder for

![Fig. 5: Inference setup and architectures for different gaze zone estimation models.](image)

![Fig. 6: Inference setup and architectures for different gaze zone estimation models.](image)

<table>
<thead>
<tr>
<th>Model</th>
<th>Micro-average accuracy(%)</th>
<th>Macro-average accuracy(%)</th>
<th>Confusion matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>59.83</td>
<td>56.71</td>
<td>Fig. 7a</td>
</tr>
<tr>
<td>B</td>
<td>73.45</td>
<td>72.57</td>
<td>Fig. 7b</td>
</tr>
<tr>
<td>A with fine-tuning</td>
<td>72.82</td>
<td>72.14</td>
<td>Fig. 7c</td>
</tr>
<tr>
<td>B with fine-tuning</td>
<td>74.92</td>
<td>73.00</td>
<td>Fig. 7d</td>
</tr>
<tr>
<td>B with fine-tuning</td>
<td>80.49</td>
<td>79.00</td>
<td>Fig. 7e</td>
</tr>
</tbody>
</table>
models to learn discriminative gaze features. In columns ii, iv, vi, and vii, the images generated from GPCycleGAN are realistic and preserve the gaze more accurately. Columns ii, iii, iv, vi, and vii show that our model does not only work with frontal face images but also performs well for a variety of head poses. Column ix is an example where both models perform well because the gaze is clear and not occluded by the frame or glare. The last column (x) depicts a failure case for both models. The models fail because the frame of the eyeglass is too thick, and both the frame and glare occlude the eye regions severely. These problems could potentially be solved by collecting more data with thicker eyeglass frames, increasing the image resolution, and/or by designing better GANs.

VI. CONCLUDING REMARKS

Reliable and robust gaze estimation on real-world data is essential yet hard to accomplish. A driver’s gaze is especially important in the age of partially automated vehicles as a cue for gauging driver state/readiness. In this study, we improved the robustness and generalization of gaze estimation on real-world data captured under extreme conditions, such as data with the presence of eyeglasses, harsh illumination, nighttime driving, significant variations of head poses, etc. For dealing with issues arising from bad lighting, we demonstrate that using an IR camera with suitable equalization/normalization suffices. For images that includes eyeglasses, we present eyeglass removal as a pre-processing step using our proposed Gaze Preserving CycleGAN (GPCycleGAN). The GPCycleGAN enables us to train a generator that is capable of
removing eyeglasses while retaining the gaze of the original image. This ensures accurate gaze zone classification by a downstream SqueezeNet model. We show that this combined model exceeds the baseline approach by 10.5% on micro-average accuracy and 8.9% on macro-average accuracy, and it outperforms the vanilla CycleGAN + SqueezeNet model by 1.7% on micro-average accuracy and 9.5% on macro-average accuracy. Future work entails improving on the architectures of different components like the generator, discriminator and the gaze classifier.

VII. ACKNOWLEDGMENTS

We thank our industry sponsors Toyota Collaborative Safety Research Center (CSRC) for their continued support.

REFERENCES


