Gaze Preserving CycleGANs for Eyeglass Removal & Persistent Gaze Estimation

Akshay Rangesh†, Bowen Zhang† and Mohan M. Trivedi
Laboratory for Intelligent & Safe Automobiles, UC San Diego
{arangesh, boz004, mtrivedi}@ucsd.edu

Abstract—A driver’s gaze is critical for determining their attention, state, situational awareness, and readiness to take over control from partially automated vehicles. Estimating the gaze direction is the most obvious way to gauge a driver’s state under ideal conditions when limited to using non-intrusive imaging sensors. Unfortunately, the vehicular environment introduces a variety of challenges that are usually unaccounted for - harsh illumination, nighttime conditions, and reflective eyeglasses. Relying on head pose alone under such conditions can prove to be unreliable and erroneous. 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 suffices. To handle eyeglasses and their corresponding artifacts, we adopt image-to-image translation using generative adversarial networks to pre-process images prior to gaze estimation. Our proposed Gaze Preserving CycleGAN (GPCycleGAN) is trained to preserve the driver’s gaze while removing potential eyeglasses from 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, interpretability, robustness and superior qualitative results on challenging real-world datasets.

I. INTRODUCTION

Driver safety and accident risk are highly determined by a driver’s attention levels, especially visual attention, e.g. a driver’s gaze, for both conventional and self-driving vehicles [1]–[5]. It is crucial to have the ability to monitor the driver’s gaze in real-time and use gaze information to efficiently enhance vehicle safety for intelligent vehicles. Previous work has proven that visual information from imaging sensors and corresponding learning algorithms are effective in determining driver gaze zones [6]–[9] and driver activity in general [10]–[13]. 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 [14], [15], 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 [16]. However, it is difficult to estimate the gaze with a high accuracy using RGB images acquired from dark environments. It is primarily because RGB cameras suffer from a lack of photons captured by imaging sensors, resulting in low signal-to-noise ratios [17].

†Authors contributed equally
Code, datasets & models: https://github.com/arangesh/GPCycleGAN
Video results

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.

In this study, we find that using infrared cameras can mitigate these lighting issues; however, for face images with eye-wear, data pre-processing or algorithmic improvements are necessary to identify the driver’s gaze zones. Since previous works on gaze estimation [8] achieved decent results (under ideal conditions) by using convolutional neural networks (CNNs), we adopt their basic methodology and develop novel pre-processing approaches that can improve gaze estimation in these demanding conditions.

Generative Adversarial Networks (GANs) [18] have shown promising results on a variety of computer vision tasks like image super-resolution [19], realistic face images generation [20], style transfer, image editing [21], [22], etc. There are two sub-models in GAN architecture, 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’s task is to differentiate between real and synthesized images. Both of them are trained together in an adversarial manner until an equilibrium is attained. More recent work on GANs has 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 one such model called the CycleGAN [22], 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...
preserves the gaze and allows for more accurate gaze estimation on images with eyeglasses. Second, unlike previous works [26], [27] on eyeglass removal, there is no need for paired images to train GPCycleGAN. 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 GPCycleGAN model for eyeglass removal specifically optimized for the gaze classification task, c) Experimental analyses to illustrate how the GPCycleGAN model improves the accuracy, interpretability and robustness over a variety of baseline models, and d) A naturalistic driving dataset with labeled gaze zones that includes both IR and RGB images. This work is an extension of our research presented in [30]. In this study, we extend and refine our previous approach with new losses and training schemes, update our previously reported quantitative numbers, and provide new qualitative analysis.

II. RELATED RESEARCH

Recent related studies are listed in Table I with 2 sub-tables that represent two research areas - gaze estimation and eyeglass removal. We mainly focus on vision-based approaches post 2016 for gaze estimation. Please refer to the following studies for earlier work: Vora et al. [8] for gaze estimation; Kar et al. [31] 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. [32]; a review on driver inattention monitoring systems by Dong et al. [33]; and a survey on head pose estimation by Murphy et al. [34]. For the second research area, we were unable to find published studies focused on eyeglass removal for drivers’ face images. We instead provide discussions on general-purpose eyeglass removal papers.

![Table I: Related research](image)

### A. Gaze Estimation

Gaze estimation studies can be categorized in the following ways. First, by the type of model used, they can be divided into convolutional neural networks [8], statistical learning models [35], or geometric approaches [7]. Second, methods can be distinguished by the cues they consider, i.e. studies that only use head pose [6] versus ones that use both head pose and eye information [8], [9], [36], [37]. Third, research can be categorized by the conditions they capture. For instance, studies with limited illumination changes [8] versus research with multiple environments and the variations that come with it [9]. Lastly, studies can be separated by the sensors they use, including RGB cameras [8], IR cameras [38], NIR cameras [23], multiple cameras [39], and wearable sensors [40]. However, only a few studies emphasize the problems associated with driving with eyeglasses or eye-wear. Tawari et al. [39] and Lee et al. [41] 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. [23] combine head pose estimation and pupil detection to determine the eye gaze, but their method only works under ideal capture conditions. Jah and Busso [6] use only head pose in their method. Wang et al. [24] 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 [9], Yoon et al. collect a dataset comprising of images in daytime/nighttime, images with and without eyeglasses using two NIR cameras 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 the training set.

B. Eyeglass Removal

In real-world cabin environment, large variations, including head pose, eyeglass type, environment conditions, and the presence of reflection artifacts, make eyeglass removal is a complex task. Existing 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 do not require high computational power but have limitations on eyeglasses, environmental conditions, and head poses [42], [43]. On the other hand, methods using deep learning, such as ones proposed by Liang et al. [27], Wang et al. [26], Din et al. [44], and Lee et al. [45] modified GANs for better results, but they need paired and aligned images for training. Such datasets are expensive and tedious to collect. Different GAN architectures proposed in [22], [25], and [46] 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. [28] and Shen et al. [29] are capable of changing multiple facial attributes but do not produce satisfying results on eyeglass removal. All the above methods on eyeglass removal have general or different purposes. However, none of these methods are constructed to preserve 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 mirror. To ensure a good compromise between larger fields-of-view and faster processing speeds, we settled on a capture mirror. To ensure a good compromise between larger fields-of-view and faster processing speeds, we settled on a capture mirror.

Table II: Dataset size (# of images, # of subjects) across different splits and capture conditions. There is no overlap of subjects between different splits to ensure cross-subject validation and testing.

| Capture conditions | Dataset split | Training | Validation | Testing |
|--------------------|---------------|----------|------------|---------|
| daytime; w/o eyeglasses | (67151, 9) | (9908, 1) | (2758, 4) |
| nighttime; w/o eyeglasses | (59352, 9) | (8510, 1) | (2608, 4) |
| daytime; w/ eyeglasses | (43332, 5) | (9062, 1) | (3294, 4) |
| nighttime; w/ eyeglasses | (33189, 5) | (8103, 1) | (2897, 4) |
| Total (all conditions) | (203124, 9) | (35583, 1) | (11717, 4) |

Fig. 2: Example images from our dataset under different capture conditions.

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 [8]), each trained on data captured under different conditions 1 - 9. From the Table, we glean that model 5 validated on data 1, 6, and 2 produce similar accuracies, and model.
This study, we propose to do the latter by training an eyeglass removal network. Between images with and without eyeglasses. In this study, we propose to do the latter by training an eyeglass

The accuracy of model 1

The accuracy of model 6

This demonstrates that the gaze classifier models work well on daytime and nighttime data when trained on data containing both conditions. Thus, problems related to lighting can be effectively solved by training using IR images, appropriate normalization, and histogram equalization. We also learn that nighttime data is easier to model and results in better accuracy in general. Next, to analyze the models’ performance on daytime and nighttime data when trained on data containing all conditions.

The two domains for the eyeglass removal problem are eye images with and without eyeglasses, $X$ and $Y$. This implies that simply training on data with and without eyeglasses is not sufficient to ensure good generalization across both conditions. Thus, we need to explicitly handle eyeglasses through modeling, or reduce the domain gap between images with and without eyeglasses. In this study, we propose to do the latter by training an eyeglass removal network.

### B. Vanilla CycleGAN-based Approach

The two domains for the eyeglass removal problem are eye image crops without eyeglasses, $X$, and eye image crops with eyeglasses, $Y (X,Y \subset \mathbb{R}^{H \times W \times C}, H = 256, W = 256$, and $C = 1$ for IR images, $C = 3$ for RGB images). We denote $X (X \in X)$ and $Y (Y \in Y)$ as sample images from domains $X$ and $Y$ respectively.

First, we consider a baseline gaze classifier model (Step 1 in Fig. 4(b)), for which we use input images from domain $X$. We use the SqueezeNet architecture as before and denote its output gaze zone probabilities as $\{p_i\}_i$. Standard cross-entropy loss ($L_{CE}$) is used for training the gaze classifier. $L_{CE}$ is defined as:

$$L_{CE} = - \sum_i z_i \log p_k,$$

where $i$ is the class index and $z$ is the one-hot encoding of the ground truth class.

Next, we consider a CycleGAN-based eyeglass removal network, followed by a SqueezeNet-based gaze classifier (Step 2 in Fig. 4(b)). The CycleGAN-based eyeglass removal model learns the mapping $G_{w/o}$ from $Y$ to $X$ (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 $\{p_i\}_i$. In a similar manner, the mapping for adding eyeglasses to $X \in X$ is $G_{w/o}: X \rightarrow Y$, 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 [22] (Eq. 2) that encourages cycle consistency between the real images and the reconstructed images; adversarial losses [18] (Eq. 3) for making the generated images indistinguishable from real images; and identity losses [48] (Eq. 4) to ensure identity mapping when the input belongs to the target domain.

$$L_{cycle}(G_{w/o}, G_{w/o}) = E_{X \sim p(x)} \left[ \|G_{w/o}(G_{w/o}(X)) - X\|_1 \right]$$

$$+ E_{Y \sim p(y)} \left[ \|G_{w/o}(G_{w/o}(Y)) - Y\|_1 \right].$$

$$L_{adv}(G_{w/o}, G_{w/o}, D_{w/o}, D_{w/o}) = E_{Y \sim p(y)} \left[ \log D_{w/o}(Y) \right]$$

$$+ E_{X \sim p(x)} \left[ \log (1 - D_{w/o}(G_{w/o}(X))) \right]$$

$$+ E_{X \sim p(x)} \left[ \log D_{w/o}(X) \right]$$

$$+ E_{Y \sim p(y)} \left[ \log (1 - D_{w/o}(G_{w/o}(Y))) \right].$$

| Model \(\#\) & training data used | Validation data \(\#\) daytime; w/o eyeglasses | Validation data \(\#\) nighttime; w/o eyeglasses | Validation data \(\#\) daytime; w/ eyeglasses | Validation data \(\#\) nighttime; w/ eyeglasses | Validation data \(\#\) w/o eyeglasses | Validation data \(\#\) daytime | Validation data \(\#\) nighttime | Validation data \(\#\) all conditions |
|----------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| 1 | daytime; w/o eyeglasses | 81.0774 | 62.4657 | 45.4588 | 11.0680 | 75.2607 | 29.7562 | 64.1875 | 38.1269 |
| 2 | nighttime; w/o eyeglasses | 73.5272 | 87.2236 | 29.3782 | 25.6684 | 79.7801 | 27.6839 | 52.5923 | 58.0671 |
| 3 | daytime; w/ eyeglasses | 60.4746 | 42.7325 | 70.6334 | 39.7423 | 52.3625 | 49.3596 | 48.5782 | 54.3356 |
| 4 | nighttime; w/ eyeglasses | 51.4649 | 76.7639 | 43.9101 | 64.8151 | 63.0322 | 33.3080 | 48.5782 | 58.8720 |
| 5 | w/o eyeglasses | 82.5508 | 88.7809 | 99.3596 | 20.6954 | 87.0276 | 33.3080 | 65.8053 | 61.5675 |
| 6 | w/ eyeglasses | 71.3746 | 82.2239 | 69.5738 | 77.7254 | 73.3561 | 75.4569 | 56.5317 | 61.5675 |
| 7 | daytime | 77.2131 | 74.7569 | 73.5095 | 39.9224 | 76.0901 | 58.1704 | 75.4569 | 67.5971 |
| 8 | nighttime | 55.1192 | 83.7696 | 49.3596 | 61.3520 | 68.2189 | 53.8881 | 67.5971 | 61.5675 |
| 9 | all conditions | 81.6969 | 83.3084 | 56.7885 | 66.4358 | 75.3166 | 72.3675 | 75.3166 | 72.3675 |
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.

As a final refinement step, we retrain the gaze classifier from step 1 using real images from $\mathcal{X}$ and fake images from $G_{w/o}(\mathcal{Y})$ (Step 3 in Fig. 4(b)). Unlike step 1, we use a selective cross-entropy loss ($L_{\text{sel-CE}}$) to avoid overfitting to errors resulting from $G_{w/o}(\cdot)$. $L_{\text{sel-CE}}$ is defined as:

$$L_{\text{sel-CE}} = \left\{ \begin{array}{ll} -\sum_i z_i \log p_i, & \text{if } \arg \max_i p_i = \arg \max_i z_i \\ 0, & \text{otherwise,} \end{array} \right.$$  

where only the cross-entropy loss from correctly classified samples are backpropagated. This ensures that the gaze classifier is not trained to incorrectly classify fake images with a potentially different gaze zone than the original. After the fine-tuning step, the generator $G_{w/o}$ and the gaze classifier are deployed sequentially to estimate gaze zones in the presence of eyeglasses (Fig. 5(b)).

C. Proposed GPCycleGAN-based Approach

Our proposed GPCycleGAN model with the gaze classifier (Fig. 4(c)) builds on the previous model, with the addition of a gaze consistency loss ($L_{\text{gaze}}$) in step 2. 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 consistency loss is defined as follows:

$$L_{\text{gaze}}(G_{w/}, G_{w/o}) = \frac{1}{N} \sum_{i=1}^N \| T(A_{w/\alpha; \text{real}}^i) - T(A_{w/\alpha; \text{rec}}^i) \|_2,$$

where $A_{w/\alpha; \text{real}}^i$ and $A_{w/\alpha; \text{rec}}^i$ are the Class Activation Maps for the real and reconstructed images, respectively.
where \( N \) stands for the total number of classes (gaze zones), \( i \) is the class index, and \( \{ A^i \}_{i=1}^N \) 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 \( G_{w/o}(G_{w/\cdot}) \) using the same eye image crop. The transformation \( T(\cdot) \) applied to the CAMs is a simple sigmoid function with temperature \( \tau \): 

\[
T(A^i) = \frac{1}{1 + \exp(-\tau \cdot A^i)}.
\] (8)

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 directly.

The full objective for the proposed GPCycleGAN model is:

\[
L_{\text{total}} = L_{\text{adv}} + \lambda_1 L_{\text{cyc}} + \lambda_2 L_{\text{identity}} + \lambda_3 L_{\text{gaze}},
\] (9)

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

D. Implementation Details

The training and inference architectures are the same for the baseline gaze classification models 1⃣-9⃣, but different for models A⃣ and B⃣. Models 1⃣-9⃣ are adopted from Vora et al. [8], and differ in the data they were trained on (see Table III). Models A⃣ and B⃣ 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 B⃣). Both models consist of generators with 9 residual blocks and 70×70 PatchGANs [22] as discriminators. As illustrated in Fig. 4, training models A⃣ and B⃣ 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 for fine-tune the gaze classifier from step 1 using generated images \( G_{w/o}(x, y) \). The inference for models 1⃣ - 9⃣ (Fig. 5(a)) is a simple forward propagation through the gaze classifier to obtain the gaze zone probabilities. Models A⃣ and B⃣ have the same inference setup (Fig. 5(b)), where eye image crops 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 \( \tau = 0.01 \) in Eq. 8, and \( \lambda_1 = 10, \lambda_2 = 5, \lambda_3 = 1 \) in Eq. 5 and Eq. 9. We use a learning rate of 0.0004 with an Adam optimizer for training all SqueezeNet classifiers, a learning rate of 0.0002 for CycleGAN/GPCycleGAN in step 2, and a reduced learning rate of 0.0001 for the final finetuning step. The gaze classifiers are trained for a total of 50 epochs, while the GANs are trained for 15 epochs - both with early stopping. The total runtime for inference using our approach is 27 ms when using an RTX 2080 graphics card (OpenPose: 22 ms, eyeglass removal network: 3 ms, and gaze classifier network: 2 ms) - indicating real-time performance.
A. Quantitative Metrics & Results

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:

\[
\text{Macro-average accuracy} = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{\text{True positives}}{\text{Total population}} \right)_i, \quad (10)
\]

\[
\text{Micro-average accuracy} = \frac{\sum_{i=1}^{N} \left( \frac{\text{True positives}}{\text{Total population}} \right)_i}{\sum_{i=1}^{N} \text{Total population}}_i, \quad (11)
\]

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. So, if the data is balanced, which means the number of images for every class is the same, micro-average accuracy will equal macro-average accuracy. In our case, there are more data in higher accuracy classes, so the micro-average accuracies are higher than macro-average accuracies. Thus, we mainly compare the macro-average accuracy since the dataset is not balanced.

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 modeled. 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 produces slightly worse accuracies than model 9. This can be attributed to the fact that the gaze classifier has only been trained on half the training set i.e. images without eyeglasses. However, after fine-tuning using the entire training set (Step 3 in Fig. 4), the accuracies for both models A and B increase considerably. 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. 6. The error modes of our best performing model (Fig. 6f) 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.

B. Qualitative Results

To carry out a qualitative comparison between different GAN variants, we also show 10 examples of eyeglass removal on real images using CycleGAN and GPCycleGAN in Fig. 7. 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 the glare from glasses is a relatively common occurrence in the real world, and often occludes the eyes, making it harder for 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
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 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.

To test the viability of our approach on different data types, we repeated our prescribed training procedure using an RGB dataset collected for the same task. Fig. 8 depicts qualitative results for eyeglass removal using our RGB model. For this RGB dataset, we notice that the resulting images after eyeglass removal tend to be sharper, more realistic, and preserve more details. We believe this is because of the higher resolution and better signal-to-noise ratio of the RGB sensor used to capture the dataset. We provide links to download both IR and RGB datasets in our repository.

C. Comparison with the Other Methods on Different Datasets

To further demonstrate the generality of our approach, we train and test it on two popular datasets - the Columbia gaze dataset [49], and the MeGlass dataset [50].

Results on the Columbia Gaze Dataset. We split the dataset such that the train and test splits contain 45 and 11 subjects respectively. Of the 21 subjects who wore prescription glasses, 6 were included in the test set and the remaining in the train set. The Columbia dataset has very few images to train a gaze classifier (5880 images for 21 gaze classes). So we instead used our pre-trained gaze classifier from the RGB dataset (Section V-B) to provide the gaze consistency loss for the GPCycleGAN. The results from this model on the test split are shown in Fig. 9. Despite using a gaze classifier trained on a different dataset, we can see that the resulting images preserve the gaze direction in the Columbia gaze dataset.

Results on the MeGlass Dataset. For this experiment, we split the dataset based on person identities. Of all 1710 identities in the dataset, we assign 1197 to the train set and 513 to the test set. This ensures cross-subject testing in our evaluation. Similar to the Columbia dataset, as the MeGlass dataset does not have labelled gaze directions, we simply use our pre-trained gaze classifier to train the GPCycleGAN model. Since there is no prior work on eyeglass removal for gaze estimation, we compare our approach with the recently proposed eyeglasses removal model ERGAN [51] (Fig. 10). To aid comparison, we overlay our eye crop images after eyeglass removal on the original face image. From this comparison, we can see that ERGAN fails to preserve the gaze direction as it does not enforce any constraints on the gaze. On the other hand, our approach can preserve gaze despite using a gaze classifier trained on a completely different dataset. This also illustrates the benefits of applying eyeglass removal models to eye crop images rather than entire face images, especially if the downstream task is gaze estimation. We believe this is a more efficient use of the generator’s modelling capacity.

D. Analyzing the Class Activation Maps

Observing the Class Activation Maps (CAMs) produced by classifier models is a simple and effective way of interpreting CNNs and assessing their robustness. In this spirit, we visualize and contrast the CAMs from our proposed approach (model B) and the baseline solution (model 9). This visualization is presented in Fig. 11, where eye image crops from the test set are overlaid with the corresponding ground truth CAMs i.e. the CAMs corresponding to the correct gaze zone. In particular, we compare CAMs generated when raw eye image crops with eyeglasses are directly fed to a gaze classifier (as in model 9) with CAMs that are obtained by first removing eyeglasses from the eye crop images using a pre-processing network, then feeding this processed image to a fine-tuned gaze classifier (as in model B). The CAMs clearly indicate that aside from being superior in terms of quantitative metrics, our proposed model exhibits better interpretability, robustness, and more accurately localizes the eye regions. The baseline model - despite being trained on images with eyeglasses - fails to learn the true nature of the task, and instead overfits to unrelated visual cues. This portends better generalization for our model in comparison to the baseline.

VI. CONCLUDING REMARKS

Fig. 9: Example real RGB images with eyeglasses from the Columbia gaze dataset, and corresponding generated images after eyeglass removal.

Fig. 10: Example real RGB images with eyeglasses from the MeGlass dataset, and corresponding generated images after eyeglass removal. We compare the proposed GPCycleGAN approach with ERGAN.
Fig. 11: IR eye image crops from the test set overlaid with ground truth CAMs produced by the gaze classifier. **Top row:** raw eye image crops with ground truth CAMs produced by model 9. **Bottom row:** same eye crops after eyeglass removal and updated ground truth CAMs produced by model 8.

VII. ACKNOWLEDGMENTS

We would like to thank the Toyota Collaborative Safety Research Center (CSRC) for their generous and continued support. We would also like to thank our colleagues for their useful inputs and help in collecting and labeling the dataset. Finally, we greatly appreciate all comments and suggestions from the reviewers, associate editor and editor-in-chief, which have improved the overall quality and readability of this work.

REFERENCES

[1] V. Beanland, M. Fitzharris, K. Young, and M. Lenné, “Erratum: Driver inattention and driver distraction in serious casualty crashes: Data from the Australian national crash in-depth study (accident analysis and prevention 54c (2013) 99-107),” Accident Analysis and Prevention, vol. 59, p. 626, 2013.

[2] A. Rangesh, N. Deo, K. Yuen, K. Pirozhenko, P. Gunaratne, H. Toyoda, and M. Trivedi, “Exploring the situational awareness of humans inside autonomous vehicles,” in IEEE International Conference on Intelligent Transportation Systems, 2018.

[3] A. Tawari, S. Sivaraman, M. Trivedi, T. Shannon, and M. Tippelhofer, “Looking-in and looking-out vision for urban intelligent assistance: Estimation of driver attentive state and dynamic surround for safe merging and braking,” in IEEE Intelligent Vehicles Symposium, 2014.

[4] M. M. Trivedi et al., “Attention monitoring and hazard assessment with bio-sensing and vision: Empirical analysis utilizing cnns on the kiti dataset,” in 2019 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2019, pp. 1673–1678.

[5] M. Daily, S. Medasani, R. Behringer, and M. Trivedi, “Self-driving cars,” Computer, vol. 50, no. 12, pp. 18–23, 2017.

[6] S. Jha and C. Busso, “Probabilistic estimation of the gaze region of the driver using dense classification,” in 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Nov 2018, pp. 697–702.

[7] S. Martin, S. Vora, K. Yuen, and M. M. Trivedi, “Dynamics of driver’s gaze: Explorations in behavior modeling and maneuver prediction,” IEEE Transactions on Intelligent Vehicles, vol. 3, no. 2, p. 141–150, Jun 2018. [Online]. Available: http://dx.doi.org/10.1109/TIV.2018.2804160

[8] 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, p. 254–265, Sep 2018. [Online]. Available: http://dx.doi.org/10.1109/TIV.2018.2843120

[9] H. S. Yoon, N. R. Baek, N. Q. Truong, and K. R. Park, “Driver gaze detection based on deep residual networks using the combined single image of dual near-infrared cameras,” IEEE Access, vol. 7, pp. 93448–93461, 2019.

[10] N. Deo and M. M. Trivedi, “Looking at the driver/ rider in autonomous vehicles to predict take-over readiness,” IEEE Transactions on Intelligent Vehicles, vol. 5, no. 1, pp. 41–52, 2019.

[11] A. Roitberg, C. Ma, M. Haurilet, and R. Stiefelhagen, “Open set driver activity recognition,” in Intelligent Vehicles Symposium (IV). IEEE, 2020.

[12] M. Martin, A. Roitberg, M. Haurilet, M. Horne, S. Reibl, M. Voit, and R. Stiefelhagen, “Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles,” in The IEEE International Conference on Computer Vision (ICCV), Oct 2019.

[13] A. Rangesh, N. Deo, R. Greer, P. Gunaratne, and M. M. Trivedi, “Autonomous vehicles that alert humans to take-over controls: Modeling with real-world data,” arXiv preprint arXiv:2104.11489, 2021.

[14] “Vision impact institute releases study on corrective lens wearers in the u.s.” [Online]. Available: https://www.essilorusa.com/newsroom/vision-impact-institute-releases-study-on-corrective-lens-wearers-in-the-u-s

[15] “Vision impact institute stats.” [Online]. Available: https://visionimpactinstitute.org/download-stats/

[16] “The most dangerous time to drive.” [Online]. Available: https://www.nsc.org/road-safety/safety-topics/night-driving

[17] M. Gamadia, N. Kehrtarnavaz, and K. Roberts-Hoffman, “Low-light auto-focus enhancement for digital and cell-phone camera image pipelines,” IEEE Transactions on Consumer Electronics, vol. 53, no. 2, pp. 249–257, May 2007.

[18] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” 2014.

[19] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and et al., “Photo-realistic single image super-resolution using a generative adversarial network,” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul 2017. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2017.19

[20] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Jun 2019. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2019.00453

[21] Y. Heng, M.-Y. Liu, S. Belongie, and J. Kautz, “Multimodal unsupervised image-to-image translation,” Lecture Notes in Computer Science, p. 179–196, 2018. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-01219-9_11

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 include 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 outperforms both the baseline approach and the vanilla CycleGAN + SqueezeNet model considerably. We also provide example results and illustrations to demonstrate the superiority of our model in terms of interpretability, robustness, and generality. Future work entails improving on the architectures of different components like generator, discriminator, and gaze classifier.
