IEEE Intelligent Vehicles Symposium 2020

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 a 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). This network preserves the driver's gaze while removing potential eyeglasses from infrared face images. 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.). 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.

†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 - use of eyeglasses, harsh illumination, nighttime data etc.

Generative Adversarial Networks (GANs) [8] have shown promising results on a variety of computer vision tasks like generating realistic face images [9], style transfer, image editing [10], [11], etc. In this study, we build on one such model called the CycleGAN [11], which achieved the state-of-the-art on image-to-image translation by using unpaired images from the source and target distributions. 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 [12], [13] 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 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. [19] for eye-gaze estimation systems, algorithms, and performance evaluation methods in consumer platforms; a survey on driver

TABLE I: Related research

(a) Selected research on gaze estimation in vehicular environments

Study	Objective	Sensor	Features	Capture conditions	Methodology
Vora et al. [6]	Gaze zone classification using CNNs	1 RGB camera	HP[1]& gaze	Daytime; w/o eyeglasses	CNN
Martin et al. [5]	Estimating gaze dynamics, glance duration and frequency	1 RGB camera	HP & gaze	Daytime; w/o eyeglasses	CNN & geometry
Naqvi et al. [14]	Gaze zone detection using NIR[2] camera and deep learning	1 NIR camera & NIR LEDs	HP & gaze	Daytime & nighttime; w & w/o eyeglasses	CNN
Yong et al. [7]	Gaze detection using dual NIR cameras and deep residual networks	2 NIR cameras & NIR LEDs	HP & gaze	Daytime & nighttime; w/ & w/o eyeglasses	CNN
Jha & Busso [4]	Gaze region estimation using dense pixelwise predictions	1 RGB camera & 1 headband with AprilTags	HP	Daytime; w/ & w/o eyeglasses	Dense Neural Networks
Wang et al. [15]	Continuous gaze estimation using RGB-D camera	1 RGB-D camera	HP & gaze	Daytime; w/ & w/o eyeglasses	Feature extraction & k-NN

(b) Selected research on eyeglass removal and related topics

Study	Objective	Methodology	Dataset	Advantages	Disadvantages
Zhu et al. [11]	Unpaired image-to-image translation	Cycle-GAN[3]	Unpaired images	Uses unpaired images; performs well on style transfer	Not realistic for eyeglass removal; does not preserve gaze direction
Hu et al. [16]	Eyeglass removal	ER-GAN	Unpaired CelebA & LFW	Good performance on eyeglass removal for frontal faces	Gaze is not preserved; only works for aligned frontal images
Amodio et al. [16]	Unpaired image-to-image translation	TraVeLGAN	Unpaired Images from multiple domains	Good performance on style transfer and eyeglass removal	Eyes are often swapped; gaze is not preserved
Wang et al. [13]	Facial obstruction removal	EC-GAN & LS-GAN	Paired CelebA	Better results than Cycle-GAN; improved face recognition accuracy	Needs an obstruction classifier
Liang et al. [12]	Learn mappings between faces with and without glasses	CNN	Images from web & surveillance cameras	Single step, end-to-end method	Images need to be aligned; needs paired images; uses synthesized eyeglasses
Li et al. [17]	Identity-aware transference of facial attributes	DIAT-GAN	Aligned CelebA	Flexibility to modify different facial attributes	Generated gaze is different; blurry outputs
Shen et al. [18]	Facial attribute manipulation	GAN	CelebA & LFW	Capability to manipulate images with modest pixel modifications	Does not completely remove eyeglasses; the eye region is not preserved

[1] Head Pose
[2] Near Infra-Red
[3] Generative Adversarial Networks

behavior analysis for safe driving by Kaplan et al. [20]; a review on driver inattention monitoring systems by Dong et al. [21]; and a survey on head pose estimation by Murphy et al. [22]. 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 [23], 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], [24], [25]. 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 [26], NIR cameras [14], and wearable sensors [27]. However, only a few studies emphasize the problems associated with driving with eyeglasses or eye-wear. 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

The relevant eyeglass removal research can be found in Table I(b). 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. Methods using deep learning, such as ones proposed by Liang et al. [12] and Wang et al. [13] 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 [11], [16], and [28] do not require paired images, but their models fail to keep gaze information and do not perform well on non-frontal images. 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 640×480. 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

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)	(2768, 4)
daytime; w/ eyeglasses	(43432, 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.

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 [29], (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 and validated on ⓐ - ⓘ. From the Table, we glean that model ⑤ validated on data ⓐ, ⓑ, and ⓔ produce similar accuracies. This demonstrates that the gaze classifier models work well on daytime and nighttime data when

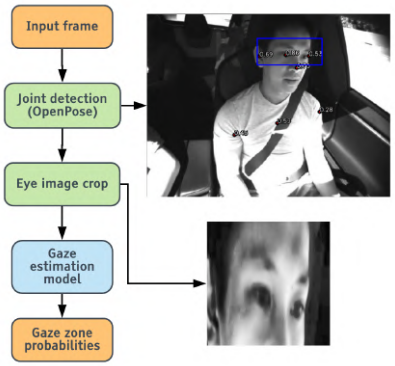

Fig. 3: Overall processing pipeline for driver gaze zone estimation.

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 accuracies in general. Next, to analyze the models' performance on data with and without glasses, we observe that the accuracy of model ⑤ validated on data ⓔ (87.0276%) is much higher than the accuracy of model ⑥ validated on data ⓕ (73.2966%). Similarly, the accuracy of model ① is better than model ③, and model ② is better than model ④ when their validation sets comprise of data from their training distribution. The model trained with all the data (⑨) has better accuracies when validated on data without glasses (ⓐ, ⓑ, and ⓔ), than that with glasses (ⓒ, ⓓ, and ⓕ). 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 modelling, 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. Formulation

The two domains for the eyeglass removal problem are eye image crops without eyeglasses, X, and eye image crops with eyeglasses, Y ($X, Y \subset \mathbb{R}^{\mathcal{H} \times \mathcal{W} \times 1}$, $\mathcal{H} = 256$, and $\mathcal{W} = 256$). 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, for which we use input images from both domains. We use the SqueezeNet architecture as before and denote its output gaze zone probabilities as $\{p_i\}_i$. Standard cross-entropy loss ($\mathcal{L}_{CE}$) is used for training the gaze classifier. $\mathcal{L}_{CE}$ is defined as:

$$\mathcal{L}_{\text{CE}} = -\sum_i y_i \log p_i, \tag{1}$$

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

Next, we consider a CycleGAN-based eyeglass removal network (denoted as model Ⓐ), followed by a SqueezeNet-based gaze classifier. 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, $G_{w/o}$

TABLE III: Validation accuracies for gaze models trained on data with different capture conditions

Model ⓗ & training data used \ Validation data	ⓐ daytime; w/o eyeglasses	ⓑ nighttime; w/o eyeglasses	ⓒ daytime; w/ eyeglasses	ⓓ nighttime; w/ eyeglasses	ⓔ w/o eyeglasses	ⓕ w/ eyeglasses	ⓖ daytime	ⓗ nighttime	ⓘ all conditions
① daytime; w/o eyeglasses	81.0774	62.4657	45.4588	11.0680	72.5677	29.7526	64.1875	38.1209	52.2757
② nighttime; w/o eyeglasses	73.5272	87.2226	29.3782	25.6684	79.7891	27.6839	52.5923	58.0671	55.0941
③ daytime; w/ eyeglasses	60.4746	42.7325	70.6334	39.7423	52.3625	56.5256	65.2918	41.3162	54.3356
④ nighttime; w/ eyeglasses	51.4649	76.7639	45.3773	64.8151	63.0322	54.2544	48.5782	71.1043	58.8720
⑤ w/o eyeglasses	85.5508	88.7809	43.9101	20.6954	87.0276	33.3080	65.8053	56.5317	61.5675
⑥ w/ eyeglasses	71.3746	82.2239	69.5738	77.7254	76.3351	73.2966	70.5207	80.0932	74.8951
⑦ daytime	77.2131	74.7569	73.5095	39.9224	76.0901	58.1704	75.4569	58.2573	67.5971
⑧ nighttime	55.1192	83.7696	49.3596	61.3520	68.2189	54.8365	52.3881	73.1514	61.8763
⑨ all conditions	81.6969	83.3084	56.7885	66.4358	82.4337	61.1944	69.8857	75.3166	72.3675

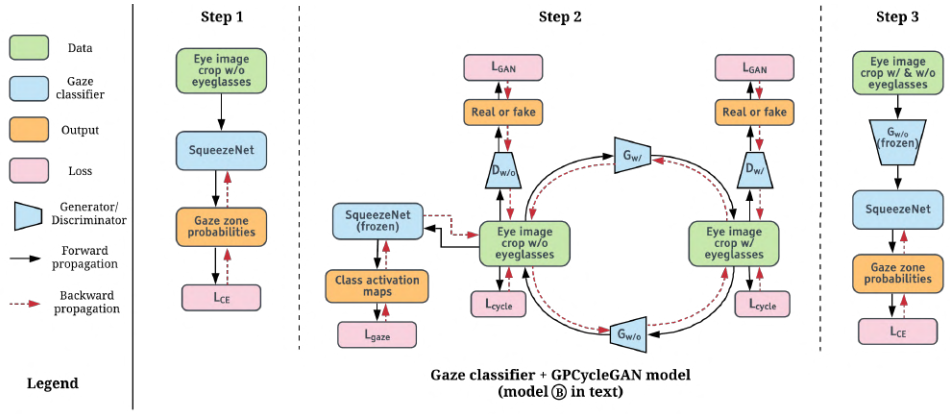

Fig. 4: Training setup and architectures for different gaze zone estimation models.

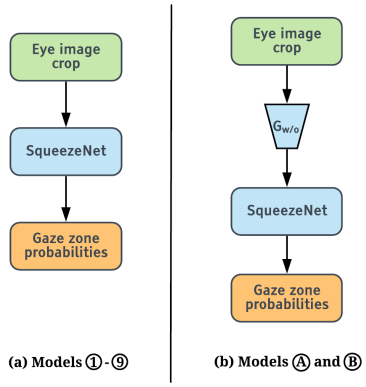

Fig. 5: Inference setup and architectures for different gaze zone estimation models.

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 $\{p_i\}_i$. 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 $G_{w/}(x)$, 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 [11] (Eq. 2) that encourages cycle consistency between the real images and the reconstructed images; adversarial losses [8] (Eq. 3) for making the generated images indistinguishable from real images; and identity losses [30] (Eq. 4) to ensure identity mapping when the input belongs to the target domain.

$$\begin{aligned}\mathcal{L}_{\text{cyc}}(G_{w/}, G_{w/o}) = \mathbb{E}_{x\sim P(x)}\left[\|G_{w/o}(G_{w/}(x)) - x\|_1\right] \\ + \mathbb{E}_{y\sim P(y)}\left[\|G_{w/}(G_{w/o}(y)) - y\|_1\right],\end{aligned} \tag{2}$$

$$\begin{aligned}\mathcal{L}_{\text{adv}}(G_{w/}, G_{w/o}, D_{w/}, D_{w/o}) \\ = \mathbb{E}_{y\sim P(y)}\left[\log D_{w/}(y)\right] \\ + \mathbb{E}_{x\sim P(x)}\left[\log\left(1 - D_{w/}\left(G_{w/}\left(x\right)\right)\right)\right] \\ + \mathbb{E}_{x\sim P(x)}\left[\log D_{w/o}(x)\right] \\ + \mathbb{E}_{y\sim P(y)}\left[\log\left(1 - D_{w/o}\left(G_{w/o}\left(y\right)\right)\right)\right],\end{aligned} \tag{3}$$

and,

$$\begin{aligned}\mathcal{L}_{\text{identity}}\left(G_{w/}, G_{w/o}\right) = \mathbb{E}_{y\sim P(y)}\left[\|G_{w/}(y) - y\|_1\right] \\ + \mathbb{E}_{x\sim P(x)}\left[\|G_{w/o}(x) - x\|_1\right],\end{aligned} \tag{4}$$

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

The full objective for the CycleGAN model is the following:

$$\mathcal{L}_{total} = \mathcal{L}_{\text{adv}} + \lambda_1 \mathcal{L}_{\text{cyc}} + \lambda_2 \mathcal{L}_{\text{identity}}, \tag{5}$$

where λ_1 is the weight of the cycle loss, and λ_2 is the weight of the identity loss.

Finally, our proposed GPCycleGAN model (denoted as model Ⓑ) with the gaze classifier (Fig. 4) builds on the previous model (model Ⓐ), with the addition of a gaze identity loss ($\mathcal{L}_{gaze}$). 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:

$$\mathcal{L}_{\text{gaze}}(G_{w/}, G_{w/o}) = \frac{1}{N}\sum_{i=1}^{N}(A^i_{w/o;real} - A^i_{w/o;rec})^2, \quad (6)$$

where N is total the number of classes, i is the class index, and A^is 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 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:

$$\mathcal{L}_{total} = \mathcal{L}_{\text{adv}} + \lambda_1\mathcal{L}_{\text{cyc}} + \lambda_2\mathcal{L}_{\text{identity}} + \lambda_3\mathcal{L}_{\text{gaze}}, \quad (7)$$

where λ_3 is the weight for the gaze identity loss.

C. Implementation Details

The training and inference architectures are the same for the baseline gaze classification models ①-⑨, but different for models Ⓐ and Ⓑ. Models ①-⑨ are adopted from Vora et al. [6], and differ in the data they were trained on (see Table III). Models Ⓐ and Ⓑ 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 Ⓑ). Both models consist of generators with 9 residual blocks and 70×70 PatchGANs [11] as discriminators. As illustrated in Fig. 4, training models Ⓐ and Ⓑ 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 $G_{w/o}(\cdot)$. 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 Ⓐ and Ⓑ 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:

$$\text{Macro-average accuracy} = \frac{1}{N}\sum_{i=1}^{N}\frac{(\text{True positives})_i}{(\text{Total population})_i}, \quad (8)$$

TABLE IV: Test set metrics for different models

Model	Micro-average accuracy(%)	Macro-average accuracy(%)
⑤	59.83	56.71
⑨	73.45	72.57
Ⓐ	72.82	72.14
Ⓐ with fine-tuning	79.01	72.14
Ⓑ	74.92	73.00
Ⓑ with fine-tuning	**80.49**	**79.00**

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

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 Ⓐ and Ⓑ improves the accuracies over the baseline model ⑨. Nonetheless, the improvement is meager, with model Ⓑ producing accuracies slightly higher than the baseline and model Ⓐ. However, after fine-tuning using the entire training set, the model Ⓑ accuracies increase considerably, while that of model Ⓐ remain relatively fixed. In conclusion, our proposed model Ⓑ demonstrates significant improvement over both the baseline model ⑨ and the vanilla CycleGAN-based model Ⓐ 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.

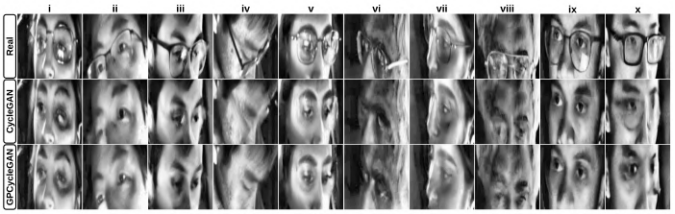

Fig. 6: Example real images with eyeglasses, and corresponding generated images after eyeglass removal.

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 models to learn discriminative gaze features. In columns ii, iv, vi, and viii, 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

In this study, we improved the robustness and generalization of gaze estimation on real-world data captured under extreme conditions. 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 1.9% 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

[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*, 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] S. Jha and C. Busso, "Probabilistic estimation of the gaze region of the driver using dense classification," in *International Conference on Intelligent Transportation Systems*, 2018.

[5] 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*, Jun 2018.

[6] 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*, 2018.

[7] 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*, 2019.

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

[9] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *IEEE/Conference on Computer Vision and Pattern Recognition*, Jun 2019.

[10] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," *Lecture Notes in Computer Science*, 2018.

[11] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[12] M. W. Liang, Y. Xue, K. Xue, and A. Yang, "Deep convolution neural networks for automatic eyeglasses removal," 2017.

[13] Y. Wang, X. Ou, L. Tu, and L. Liu, "Effective facial obstructions removal with enhanced cycle-consistent generative adversarial networks," in *Artificial Intelligence and Mobile Services*. Springer International Publishing, 2018.

[14] "Deep learning-based gaze detection system for automobile drivers using a nir camera sensor," *Sensors*, vol. 18, no. 2, p. 456, Feb 2018.

[15] Y. Wang, G. Yuan, Z. Mi, J. Peng, X. Ding, Z. Liang, and X. Fu, "Continuous driver's gaze zone estimation using rgb-d camera," *Sensors*, vol. 19, no. 6, p. 1287, Mar 2019.

[16] B. Hu, W. Yang, and M. Ren, "Unsupervised eyeglasses removal in the wild," in *arXiv preprint arXiv:1909.06989*, 2019.

[17] M. Li, W. Zuo, and D. Zhang, "Deep identity-aware transfer of facial attributes," in *arXiv preprint arXiv:1610.05586*, 2016.

[18] W. Shen and R. Liu, "Learning residual images for face attribute manipulation," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[19] A. Kar and P. Corcoran, "A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms," *IEEE Access*, vol. 5, p. 16495–16519, 2017.

[20] S. Kaplan, M. A. Guvensan, A. G. Yavuz, and Y. Karalurt, "Driver behavior analysis for safe driving: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 6, pp. 3017–3032, Dec 2015.

[21] Y. Dong, Z. Hu, K. Uchimura, and N. Murayama, "Driver inattention monitoring system for intelligent vehicles: A review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 2, pp. 596–614, June 2011.

[22] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, p. 607–626, Apr. 2009.

[23] L. Fridman, P. Langhans, J. Lee, and B. Reimer, "Driver gaze region estimation without use of eye movement," *IEEE Intelligent Systems*, vol. 31, no. 3, p. 49–56, May 2016.

[24] A. Tawari, K. Chen, and M. Trivedi, "Where is the driver looking: Analysis of head, eye and iris for robust gaze zone estimation," in *IEEE International Conference on Intelligent Transportation Systems*, 2014.

[25] E. Ohn-Bar, S. Martin, A. Tawari, and M. Trivedi, "Head, eye, and hand patterns for driver activity recognition," in *International Conference on Pattern Recognition*, 2014.

[26] W. Chinsatit and T. Saitoh, "Cnn-based pupil center detection for wearable gaze estimation system," *Applied Computational Intelligence and Soft Computing*, vol. 2017, 2017.

[27] A. Tsukada, M. Shino, M. Devyver, and T. Kanade, "Illumination-free gaze estimation method for first-person vision wearable device," in *IEEE International Conference on Computer Vision Workshops*, 2011.

[28] M. Amodio and S. Krishnaswamy, "Travelgan: Image-to-image translation by transformation vector learning," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019.

[29] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields," in *arXiv preprint arXiv:1812.08008*, 2018.

[30] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," in *arXiv preprint arXiv:1611.02200*, 2016.