Towards Eyeglasses Refraction in Appearance-based Gaze Estimation

Junfeng Lyu* Feng Xu†

School of Software and BNRist, Tsinghua University

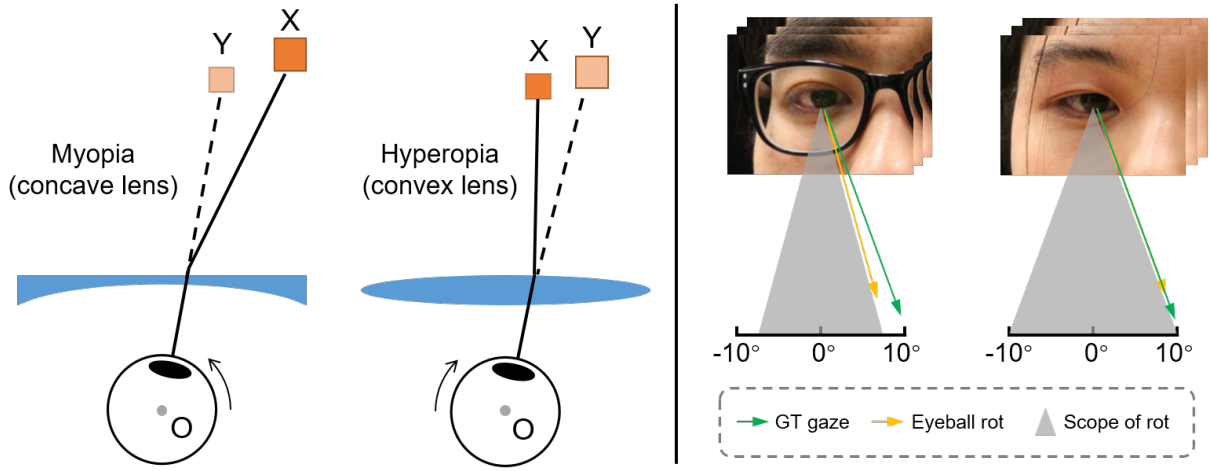

Figure 1: Diagram of eyeglasses refraction. Left: eyeglasses influence the eyeball rotation via refracting the light path. O: the eyeball center. X: the actual position of the object. Y: the position perceived by subjects. Right: compared with naked eyes, myopia eyeglasses result in a more narrow rotation scope for the same gaze targets.

Abstract

For myopia and hyperopia subjects, eyeglasses would change the position of objects in their views, leading to different eyeball rotations for the same gaze target (Fig. 1). Existing appearance-based gaze estimation methods ignore this effect, while this paper investigates it and proposes an effective method to consider it in gaze estimation, achieving noticeable improvements. Specifically, we discover that the appearance-gaze mapping differs for spectacled and unspectacled conditions, and the deviations are nearly consistent with the physical laws of the ideal lens. Based on this discovery, we propose a novel multi-task training strategy that encourages networks to regress gaze and classify the wearing conditions simultaneously. We apply the proposed strategy to some popular methods, including supervised and unsupervised ones, and evaluate them on different datasets with various backbones. The results show that the multi-task training strategy could be used on the existing methods to improve the performance of gaze estimation. To the best of our knowledge, we are the first to clearly reveal and explicitly consider eyeglasses refraction in appearance-based gaze estimation. Data and code are available at `https://github.com/StoryMY/RefractionGaze`.

Index Terms: Computing methodologies—Gaze estimation; Computing methodologies—Eyeglasses refraction; Computing methodologies—Multi-task learning

1 Introduction

Gaze is an important cue to reflect human intentions and mental activities. Appearance-based gaze estimation predicts gaze directions or gaze points from 2D images. As it is noninvasive and easy to use, it has been applied in many applications like human-machine interaction [1,46], driver monitoring [4,20,35], VR/AR rendering [5,7,24], avatar animation [43,60] and even medical analysis [8,51,55]. Recently, with the development of deep learning, this technique has been vastly improved and has drawn much more attention from both academia and industry.

Although large-scale datasets and deep learning techniques greatly improve the performance of appearance-based gaze estimation, there are still many challenges for accurate and robust gaze estimation. Some of them have been considered by previous works. For example, the variation of the angle kappa, which refers to the angle between the optical and visual axis of the human eyeball, has been handled via differential networks [30], calibrated parameters [29], and meta-learning [38]. Meanwhile, contrastive learning has been applied to handle domain gaps, like backgrounds and illuminations, in the recorded images [19,53]. Despite these efforts, the ever-growing demand for accurate gaze estimation motivates further exploration of other factors that may affect accuracy. In this paper, we reveal eyeglasses refraction as a factor that negatively affects gaze estimation, which has not been explicitly considered in previous works. Given the increasing prevalence of refractive errors among the population [13], it is necessary to conduct more in-depth research to better understand the relationship between gaze and eyeglasses.

In theory, eyeglasses would cause eyeball rotation deviation for a subject looking at the same gaze target. As shown in Fig. 1, eyeglasses would change the position of objects in the subject's view by refracting light paths. For myopia conditions, the observed appearance of the eyeball would indicate a smaller eyeball rotation angle when gazing at the same object compared with the case without eyeglasses. This effect would consequently lead to different appearance-gaze mappings for spectacled and unspectacled conditions. Previous methods do not consider this effect and usually treat the appearance-gaze mapping of spectacled and unspectacled

*e-mail: ljf19@mails.tsinghua.edu.cn
†e-mail: feng-xu@tsinghua.edu.cn

data equally, which introduces ambiguity for training the gaze estimation networks. Therefore, we aim to explicitly consider this effect in appearance-based gaze estimation methods to improve the estimation accuracy.

To validate the existence of this refraction-related effect, we design experiments to demonstrate that a naked-eye based gaze estimation network cannot work well for spectacled data, and we further prove that a correction model can be deduced to correct the errors. To be specific, we first collect a dataset that contains both spectacled and unspectacled data for each subject. All the spectacled samples in our dataset have ground-truth diopter labels. Then we train a network to learn the appearance-gaze mapping of naked eyes and use it to measure the mapping deviation between spectacled and unspectacled data, which is significant to our experiments. Notice that to exclude the influence of different colors between spectacled and unspectacled data (For example, eyeglasses themselves may also lead to poor results as the model trained with unspectacled data has never seen eyeglasses.), in the network training, we propose and use a gaze representation called Relative Coordinate Representation (RCR) which represents an eye region image by the 2D relative positions of the iris and the eye. Finally, based on the physical laws of ideal lenses, we deduce a correction model which can transform the results of the naked-eye model to the corrected gaze results given the ground-truth diopter values. From the experiments, this correction model can correct the deviation, which indicates the existence of the refraction effect.

As the correction model requires ground-truth diopter labels that are not easy to obtain in real cases, it is difficult to directly apply this model to existing gaze estimation methods, which motivates us to find a new way to consider eyeglasses refraction. We believe the key to handling eyeglasses refraction is to train the network to learn the correlation between gaze and eyeglasses. To achieve this, we propose a multi-task training strategy to encourage the networks to regress gaze and classify wearing conditions simultaneously. This helps the network to solve the multiple mappings without inputting ground-truth diopters. Experiments show that the proposed multi-task training strategy can improve the performance of supervised and unsupervised methods across different backbones.

In summary, our main contributions are listed as follows:

- To the best of our knowledge, we are the first to reveal and model the influence of eyeglasses refraction (i.e., the eyeball rotation deviation) in appearance-based gaze estimation.
- We demonstrate the existence of eyeglasses refraction by using a unique dataset and a special gaze representation, and we propose a correction model based on the ideal lens assumption for gaze correction.
- We propose a multi-task training strategy to help networks consider eyeglasses refraction and improve the performance of both supervised and unsupervised methods.

2 Related Works

Appearance-based Gaze Estimation. Early works [2, 32, 49] use linear functions or small neural networks to estimate gaze in the constrained environment. Thanks to the development of deep learning, convolutional neural networks (CNNs) show outstanding performance on visual tasks. Zhang et al. [63] are the first to use CNN to regress the gaze from eye images, which outperforms the traditional methods. Cheng et al. [10] propose an asymmetric architecture and leverage the information of both eyes to achieve better performance. Instead of eye images, Park et al. estimate gaze on the semantic representation [39] as well as landmark heatmaps [40]. The methods based on eye region information usually require an eye detection module for cropping and an additional head pose as input. To address this, Zhang et al. [64] propose an end-to-end approach based on the full face. Deng et al. [12] combine the information of the full face and eye region by a transform layer. As the optical axis and visual axis of the human eyeball are not coincident, the general models suffer from the difference between identities. To solve the personalization problem, Liu et al. [30] propose a differential approach to estimate the relative change rather than the absolute gaze. Lindén et al. [29] predict additional calibration parameters. Yu et al. [58] use few labeled samples to finetune the general model. Park et al. [38] take the problem as a transfer task and apply meta-learning to address it. In addition to identity differences, domain gaps introduced by backgrounds and illuminations are also challenging for appearance-based methods. Recently, methods based on contrastive learning have been proposed to address domain gaps [19, 53]. Bao et al. [3] propose a novel approach based on rotation consistency to generalize the gaze estimation. In this paper, we reveal eyeglasses refraction as a factor that also challenges the performance of appearance-based gaze estimation.

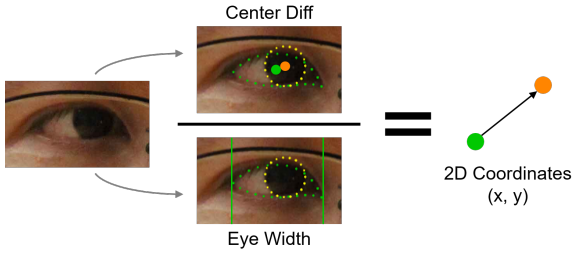

Figure 2: Diagram of relative coordinate representation (RCR). Green: eye landmarks and eye center. Yellow/Orange: iris landmarks and iris center. The center difference is normalized by eye width.

Gaze Datasets. Existing gaze datasets roughly fall into two categories: i) dataset in the controlled environment and ii) dataset in the wild. ColumbiaGaze [47] is a dataset in the controlled environment that contains 6K images of 56 subjects. UTMultiview [48] uses a screen-based capture system with eight attached cameras to collect the eye images and generate synthetic images, which finally results in 1.1M images of 50 participants. EYEDIAP [16] is a video dataset collected by an RGB-D camera. It contains 94 videos of 16 participants, and the length of each video is about 2-3 minutes. RT-GENE [14] uses a mobile eye-tracker to acquire 123K images of 15 subjects. ETH-XGaze [61] is the most recent dataset collected in the laboratory, which includes 1.1M high-resolution images of 110 subjects. It also contains extreme head pose and 16 illuminations, providing various conditions for robust gaze estimation. GazeCapture [25] is collected by mobile devices in daily life, which contains 2.4M images of 1,474 participants with 2D annotations. It is re-annotated by [38] with 3D gaze labels. Zhang et al. propose MPIIGaze [65], which is collected by laptops with free head pose and illuminations. It contains 213K eye images of 15 subjects and also has a full-face version called MPIIFaceGaze [64]. Gaze360 [21] is recorded with 238 participants in both indoor and outdoor environments, containing 172K images in total with a large range of head poses. However, all the datasets do not have diopter labels, and the subjects either wear eyeglasses or not, which is unsuitable for validating the existence of eyeglasses refraction. Instead, we collect a unique dataset that contains both spectacled and unspectacled samples for each subject with ground-truth diopter labels.

Multi-task Learning. Multi-task learning is successfully used in many applications of machine learning [6, 45]. Based on the implicit data augmentation and regularization, it can improve the performance of the main task, including natural language processing [11] and computer vision [17]. Face-related works [41, 42, 52, 56, 68] usually leverage multi-task frameworks based on the high correlation between face attributes. There are also works that can generalize

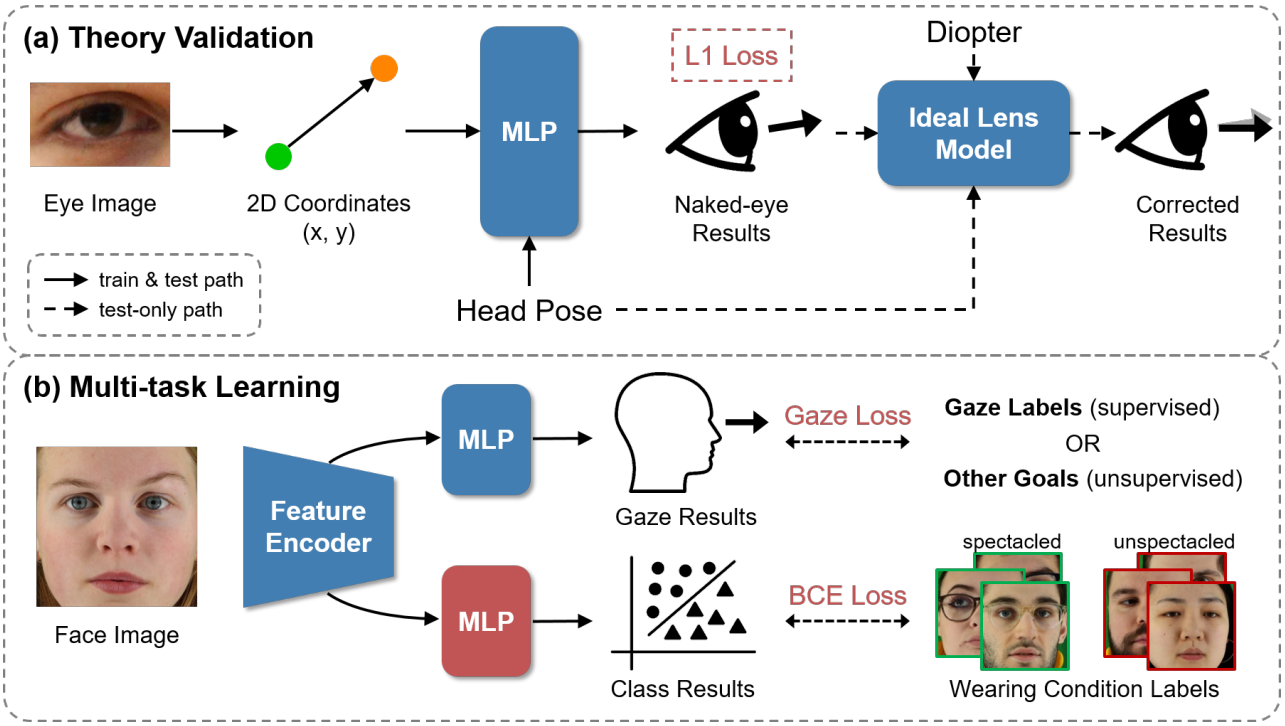

Figure 3: Overview of validation pipeline and proposed multi-task training strategy. (a) For each subject, we train a person-specific multilayer perceptron (MLP) based on the 2D coordinates representation of unspectacled data. Then, the learned appearance-gaze mapping is tested on both spectacled and unspectacled data to get naked-eye gaze results. Finally, we use the proposed model based on the ideal lens assumption to get corrected gaze results. (b) Our multi-task training strategy employs an additional MLP on top of supervised or unsupervised methods to encourage the network to estimate gaze and classify wearing conditions simultaneously.

on other tasks [36], dynamically modify the network branches [33], and adaptively adjust the loss weights [9, 18, 22, 31]. In addition, multi-task learning is also beneficial for gaze estimation. Yu et al. [57] propose a Constrained Landmark-Gaze Model to handle the joint variation of the eye landmarks and gaze directions in an explicit way. Lian et al. [27] propose a multi-task multi-view architecture to deal with the 2D gaze points and 3D gaze estimation simultaneously. They also propose another multi-task approach to regress the gaze point with the help of depth information [28]. Wu et al. [54] construct a multi-task CNN by combining eye segmentation, glint detection, pupil and cornea center estimation. Unlike these works, we propose a novel multi-task training strategy that encourages networks to learn gaze estimation and wearing classification simultaneously.

Few works study the effect of eyeglasses refraction on gaze estimation. Kübler et al. [26] share most similarity with ours, which show gaze mappings vary with diopters. However, as they are in synthetic context, the refraction effect in their work is totally different from that in ours. In their work, the eyeglasses effect refers to the influence on pupil-glint features of correctly rotated eyeballs, while the effect in ours refers to the biased eyeball rotations.

3 Theory Background

As an optical instrument, eyeglasses are usually made into concave lenses for myopia and convex lenses for hyperopia. Both of them can refract the light path and lead to eyeball rotation deviation.

From the perspective of subjects, the lens refracts the light path from objects, changing their positions in the subject's view. As shown in Fig. 1, the ground-truth gaze direction is from O to X, but the direction indicated by appearance is from O to Y. For qualitative analysis, when gazing at the same object, eyeball rotation angles would decrease for the concave lens and increase for the convex lens. Furthermore, the deviation is related to the lens diopter and the incidence angle of the light path. Generally, the larger diopter and incidence angle result in a larger deviation. In terms of 0-diopter eyeglasses or looking straight ahead, the spectacled and unspectacled conditions would share a similar appearance-gaze mapping.

Our research focuses on the eyeball rotation deviation caused by eyeglasses refraction. As it fundamentally changes the eyeball rotation, this influence cannot be simply addressed in image space. To some extent, it is similar to the problem caused by angle kappa between optical and visual axes, which yields a "lying" appearance to confuse networks. However, unlike the angle kappa fixed on a particular identity, the deviation varies with different naked-eye gaze directions on a particular diopter, which is much more complicated and difficult.

4 Theory Validation

Appearance-based gaze estimation aims to learn a mapping between the appearance and the gaze. To validate the existence of eyeglasses refraction, we plan to demonstrate that the mapping for unspectacled data is different from that for spectacled data. A straightforward way to achieve this is to train a network on the unspectacled data and test it on both. Unfortunately, appearance-based methods usually suffer from other factors like identities, illuminations as well as the distortion and reflection introduced by eyeglasses. To eliminate their interference, we collect a unique dataset and leverage a special gaze representation.

4.1 Data Acquisition

Almost all existing datasets mix the identity difference with the eyeglasses difference, which means that each subject in the dataset either wears eyeglasses or not. In order to exclude the identity influence, we collect a new dataset in which each subject has both

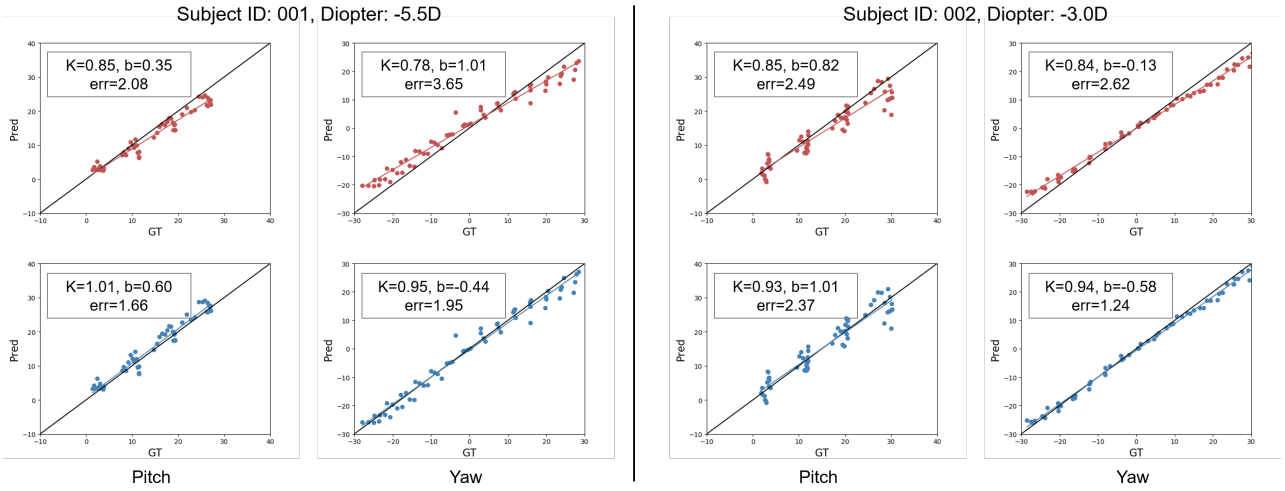

Figure 4: Scatter plot of the predicted gaze (y-axis) and ground-truth gaze (x-axis) for subjects with myopia eyeglasses. The first row shows the naked-eye results, and the second row shows the results corrected by the ideal lens model.

spectacled and unspectacled samples with the ground-truth diopter labels.

A capture system similar to [16, 48] is built for data collection. In detail, the system contains a 27-inch monitor (Philips Brilliance 272P) and an RGB-D camera (Azure Kinect) under the screen. The calibration between the screen and the camera is achieved by a mirror-based method [44]. During collection, the target dots are displayed sequentially on the screen, and the participant sitting in front of the monitor is asked to follow the dots by rotating their eyes but keeping their head still. The images and gaze directions will be recorded when the participant focuses on the dot and press the keyboard. Note that to mitigate distortion in recorded images, we only collect data on the frontal head pose. More details of the capture system can be found in the supplementary material.

Each participant repeats this procedure multiple times for spectacled and unspectacled data collection. For spectacled samples, the diopters of eyeglasses would be recorded. For unspectacled samples, the diopter would be taken as 0. Note that all the data are pre-processed and normalized using the method described in [62] to mitigate the influence of the capture distance and the roll angle of head pose.

4.2 Relative Coordinate Representation

Most appearance-based methods estimate gaze from images. Although the image representation includes rich gaze-related information, it also contains irrelevant information like image distortion and specular reflection. Inspired by [39] and [40], we think the key gaze-related information is the relative position between the iris and the eye. Based on this idea, we simplify the representation a step further by converting an eye region image into a 2D vector that depicts the position of the iris relative to the eye. This representation can remove almost all irrelevant information while retaining the key information for gaze estimation.

As shown in Fig. 2, given an eye region image, we first extract the eye and iris landmarks by the commercial software of SenseTime[1]. Then, their center coordinates are represented as follows,

$$\boldsymbol{p}^c_{eye} = (\boldsymbol{p}^{lm}_{eye} + \boldsymbol{p}^{rm}_{eye})/2 \tag{1}$$

$$\boldsymbol{p}^c_{iris} = \frac{1}{N_{iris}} \sum_i \boldsymbol{p}^i_{iris} \tag{2}$$

[1] https://www.sensetime.com

Table 1: Quantitative results of MLPs trained on unspectacled data and tested on both. The results are gaze errors in degrees.

ID	Testing Diopter	Naked-eye	Corrected
001	0	2.36	2.36
001	-5.50D	4.48	2.87
002	0	2.59	2.59
002	-3.00D	4.00	2.88
003	0	2.57	2.57
003	-5.25D	5.52	2.81
004	0	2.83	2.83
004	-3.25D	4.75	3.15
005	0	2.85	2.85
005	-3.50D	4.58	3.23
006	0	3.58	3.58
006	-7.00D	4.99	3.78

where N_{iris} is the number of iris landmarks. $\boldsymbol{p}^c$ is the center coordinates. $\boldsymbol{p}^{lm}$ and $\boldsymbol{p}^{rm}$ are the leftmost and rightmost landmarks, respectively. $\boldsymbol{p}^i$ is the i-th landmark.

The relative position is represented by the difference of the two center coordinates. In order to reduce the influence of image distortion, the difference between the two center coordinates is normalized by the eye width. Finally, the RCR is formulated as follows,

$$\boldsymbol{p} = (\boldsymbol{p}^c_{iris} - \boldsymbol{p}^c_{eye})/w_{eye} \tag{3}$$

where w_{eye} is the distance between the leftmost and rightmost landmarks on the horizontal axis.

4.3 Preliminary Validation

The data acquisition and gaze representation supplement each other. Although the RCR is prone to eye shapes, it can still work well on our dataset which only contains data on the frontal head pose and guarantees the training/testing data for each experiment group are of the same subject. Conversely, as a low-dimension representation, the RCR significantly decreases the demand for data. Compared with images, the mapping between the RCR and the gaze can be learned from little data, preventing the highly expensive cost of collecting various spectacled subjects with diopter labels.

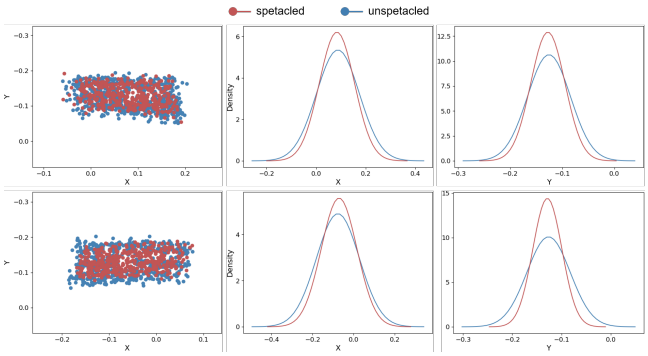

Figure 5: Distribution of the 2D coordinates on ColumbiaGaze. The first and the second row are visualization of the left and right eye, respectively. Left: scatter plot of the 2D coordinates. Middle: probability density of the x-axis. Right: probability density of the y-axis.

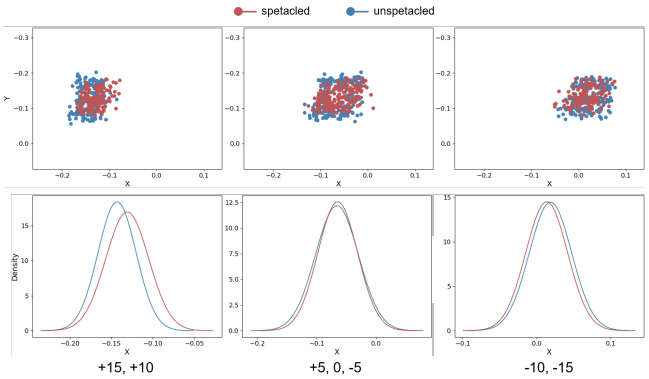

Figure 6: Detailed distribution of the 2D coordinates on ColumbiaGaze (right eye). Left: samples with $+15°$ and $+10°$ yaw. Middle: samples with $+5°$, $0°$, $-5°$ yaw. Right: samples with $-10°$ and $-15°$ yaw.

Based on our proposed dataset and the RCR, we train several person-specific gaze estimation networks for the corresponding subjects on their unspectacled training data. Although our dataset is not involved with multiple head poses, the head pose is still involved to solve unconscious movements. Specifically, we learn the mapping between the RCR and the gaze as follows,

$$\boldsymbol{g} = M(\boldsymbol{p}) + \boldsymbol{h} \tag{4}$$

where $M(\cdot)$ is a small multilayer perceptron (MLP) that consists of 5 fully connected layers (dim from input to output: 2-64-32-16-8-2) and 3 ReLU activation functions. $\boldsymbol{g}$ is the pitch and yaw of gaze in the camera coordinate system. $\boldsymbol{h}$ is the pitch and yaw of head pose in the camera coordinate system. The network is trained with L1 loss between the predicted and the ground-truth gaze.

After training, each person-specific network is tested on both spectacled and unspectacled data of the same subject. Experiment results in Sect. 6.1 show that the predicted gaze is accurate on the unspectacled data while having an obvious deviation on the spectacled data. This indicates that the appearance-gaze mapping for spectacled data differs from that for unspectacled data. Note that we have validated that our MLP is not over-paramterized, thus the deviation is irrelevant with network modeling. Please refer to the supplementary material for details if necessary.

4.4 Further Validation

Showing difference alone is insufficient to prove the existence of eyeglasses refraction. For further validation, we model the eyeglasses as ideal lenses and deduce a model to correct the deviation physically and mathematically. The correction is based on naked-eye gaze results and diopter labels without additional training.

The detail of the derivation process can be found in the supplementary material (Appendix A). As a result, we come to two equations for myopia and hyperopia gaze correction, respectively.

$$\tan(\hat{\theta}) = (1 + |D| \cdot d) \cdot \tan(\theta) \tag{5}$$

$$\tan(\hat{\theta}) = (1 - |D| \cdot d) \cdot \tan(\theta) \tag{6}$$

where θ and $\hat{\theta}$ are the naked-eye eyeball rotation angle and the corrected eyeball rotation angle, respectively. $|D|$ is the absolute value of eyeglasses diopter. d is the distance between lens center and eyeball center. As the diopter is usually negative for myopia and positive for hyperopia, the Equation 5 and Equation 6 can be merged as follows,

$$\tan(\hat{\theta}) = (1 - D \cdot d) \cdot \tan(\theta) \tag{7}$$

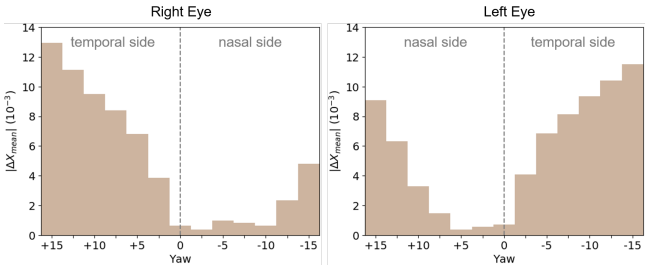

Figure 7: Trend of x-axis difference changing with yaw. The difference measures the x-axis distance between the centers of spectacled and unspectacled samples.

As the gaze is usually composed of head pose and eyeball rotation, the gaze correction can be achieved by the following equation,

$$\hat{\theta}_g = \arctan\left((1 - D \cdot d) \cdot \tan(\theta_g - \theta_h)\right) + \theta_h \tag{8}$$

where θ_g and θ_h are the angles of gaze and head pose, respectively. Corrected results of spectacled data in Sect. 6.1 show similar accuracy with the unspectacled data, which indicates the deviation is caused by the refraction of eyeglasses.

5 Multi-task Learning

Our goal is to explicitly consider eyeglasses refraction in appearance-based gaze estimation methods and improve the estimation accuracy. Although the correction model can achieve this in controlled setups, it has many limitations for real-world applications. For example, the correction based on Equation 7 requires diopters as an additional input. Furthermore, as a two-step method, it also presupposes that naked-eye gaze results are accurate in unspectacled conditions. An ideal approach is expected to enhance the performance without additional input and assumption, which implies that the network should autonomously identify the correlation between eyeglasses and gaze. To achieve this, we propose a multi-task learning framework to extend the theory to image representation and benefit the existing appearance-based methods.

The architecture of the proposed training strategy is illustrated in Fig. 3. Appearance-based methods usually leverage an encoder to extract image features and use an MLP to regress the gaze. Based on this framework, we additionally apply an MLP to classify whether the subject in the input image wears eyeglasses or not, which shares the same encoder with the gaze MLP. The proposed framework is designed based on the key idea that appearance-gaze mapping

Table 2: Within-dataset and cross-dataset performance of supervised methods with and without classification loss.

Name	Backbone	Dataset	Aux. Task	ETH-XGaze	ColumbiaGaze	MPIIFaceGaze	Our Data	Acc. (Wearing)
XGaze	ResNet-18	ETH-XGaze	None	4.60	8.06	7.99	5.30	0.6690
XGaze	ResNet-18	ETH-XGaze	Wearing	**4.59**	**6.92**	**7.56**	**4.56**	0.9336
XGaze	ResNet-18	ETH-XGaze	Gender	4.67	7.66	8.54	4.62	-
XGaze	ResNet-34	ETH-XGaze	None	4.56	6.08	7.14	4.08	0.6887
XGaze	ResNet-34	ETH-XGaze	Wearing	**4.55**	**5.59**	**6.64**	**3.96**	0.9173
XGaze	ResNet-34	ETH-XGaze	Gender	4.57	6.61	6.75	4.23	-
XGaze	ResNet-50	ETH-XGaze	None	4.50	5.36	**5.58**	3.50	0.7785
XGaze	ResNet-50	ETH-XGaze	Wearing	**4.49**	**5.04**	5.64	**3.36**	0.9315
XGaze	ResNet-50	ETH-XGaze	Gender	4.54	5.29	5.84	3.88	-

Table 3: Within-dataset and cross-dataset performance of unsupervised methods with and without classification loss.

Name	Backbone	Dataset	Aux. Task	MiniEVE	ColumbiaGaze	MPIIFaceGaze	Our Data	Acc. (Wearing)
GazeCLR	ResNet-18	EVE	None	4.97	7.90	7.53	11.78	0.7051
GazeCLR	ResNet-18	EVE	Wearing	**3.88**	**7.45**	**7.24**	**8.47**	0.9190
GazeCLR	ResNet-18	EVE	Gender	4.13	7.97	7.43	9.80	-
GazeCLR	ResNet-34	EVE	None	4.56	8.39	7.75	10.39	0.6879
GazeCLR	ResNet-34	EVE	Wearing	**4.17**	**7.53**	**7.51**	**9.95**	0.8901
GazeCLR	ResNet-34	EVE	Gender	4.62	7.63	7.78	10.94	-
GazeCLR	ResNet-50	EVE	None	5.26	7.46	7.59	10.94	0.7541
GazeCLR	ResNet-50	EVE	Wearing	**4.30**	**7.28**	**7.43**	**9.63**	0.9034
GazeCLR	ResNet-50	EVE	Gender	4.51	7.52	7.70	11.34	-

differs for spectacled and unspectacled conditions. On this premise, if a network can handle the multiple mappings, it should be able to recognize the eyeglasses in input images at least.

We learn the classification by Binary Cross Entropy (BCE) loss as follows,

$$\mathcal{L}_{class} = -c \log \hat{c} - (1 - c) \log(1 - \hat{c}) \tag{9}$$

where c and $\hat{c}$ represent the class labels and classification results, respectively. For both supervised and unsupervised methods, the total training loss for the proposed framework is formulated as

$$\mathcal{L} = \mathcal{L}_{gaze} + \lambda \mathcal{L}_{class} \tag{10}$$

where $\mathcal{L}_{gaze}$ is the loss for gaze estimation task, which is usually L1 or L2 loss for supervised methods and other goals for unsupervised ones. λ is the weight for the classification task.

6 Experiments

In this section, we first describe the datasets and the implementation details. Then, we discuss the theory validation based on the RCR. Finally, we evaluate the proposed multi-task training strategy on existing methods, including supervised and unsupervised ones, which are trained on public datasets with various settings.

Datasets. For theory validation, we use the procedure described in Sect. 4.1 to collect data. The raw resolution of the collected images is 1080p. Then the images are converted into 2D coordinates using the approach described in Sect. 4.2. For each image, we only process one single eye of subjects (we select right eye in our experiment). In total, we recruit six participants (male Asians aged from 20 to 30). Each participant repeats the procedure three times to collect three 60-sample sets: i) unspectacled for training, ii) unspectacled for testing, and iii) spectacled for testing. For multi-task learning evaluation, we take ETH-XGaze [61] and EVE [37] as the training datasets for supervised and unsupervised methods, respectively. ColumbiaGaze [47], MPIIFaceGaze [64], and our data (full-face version) are used for cross-dataset evaluation.

Implementation Details. For theory validation, we use Adam optimizer [23] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The learning rate is 0.001, and the batch size is 60. For multi-task learning evaluation, we take XGaze [61] and GazeCLR [19] as the baselines for supervised and unsupervised settings, respectively. The training configuration mostly follows the original settings in their released code. We additionally apply the classification loss with $\lambda = 0.1$. All the code is implemented with PyTorch.

6.1 Eyeglasses Refraction Validation

We aim to demonstrate that the spectacled and unspectacled data share different appearance-gaze mappings. To achieve this, we train a person-specific MLP on one of the unspectacled sets for each subject to obtain the mapping between 2D coordinates and gazes. Subsequently, we apply the learned mapping to both the other unspectacled set and the spectacled set to acquire gaze results for measurement. Quantitative results, as shown in Table 1, reveal that the gaze errors of the spectacled data are universally greater than those of the unspectacled data. Moreover, the scatter plot illustrated in Fig. 4 indicates that the deviation displays certain directionality, further confirming the divergence of appearance-gaze mappings between spectacled and unspectacled data.

To verify that the deviation results from eyeglasses refraction, we model eyeglasses as ideal lenses and apply the ideal lens model described in Sect. 4.4 to correct the deviation. Apart from the naked-eye rotation angle and ground-truth diopters, the model also requires the distance between the lens center and the eyeball center as input. We empirically set $d = 0.04m$ for all subjects and correct the pitch and yaw separately. Results in Table 1 show that the gaze errors of the spectacled data decrease to the same level as the unspectacled data after correction. Our linear fitting analysis, presented in Fig. 4, further illustrates the details of the correction process. Before the

correction, the slope of the fitting lines is less than 1.0, indicating that the predicted angle is generally smaller than the ground truth. This observation is consistent with our theory that myopia eyeglasses (concave lens) would decrease the rotation angle when gazing at the same target. After the correction, the slope of the fitting lines is close to 1.0, which indicates that the deviation is primarily caused by eyeglasses refraction.

In addition to the mapping, the distribution of the RCR itself also contains clues that indicate the existence of eyeglasses refraction. For example, the 2D coordinates of the same gaze targets are probably located in different positions for spectacled and unspectacled conditions. To validate this, we choose ColumbiaGaze to visualize the distribution of the 2D coordinates because ColumbiaGaze is collected with fixed gaze targets and contains a considerable proportion of spectacled subjects (22 spectacled subjects vs. 34 unspectacled subjects), which suits this setting best. Like our dataset, we only process the samples with the frontal head pose. However, the 2D coordinates still suffer from the different eye shapes of subjects, leading to different semantics for the same coordinates. To normalize the 2D coordinates, we define a canonical coordinate by the mean of zero-gaze (0 pitch and yaw) samples. For samples of each subject, we translate them by making the zero-gaze sample aligned with the canonical coordinate. The visualization of the translated 2D coordinates is shown in Fig. 5 (left). It clearly shows that the spectacled samples are covered by the unspectacled samples. This is because almost all the spectacled subjects in ColumbiaGaze wear myopia eyeglasses, which decrease the rotation angles for the same gaze targets and result in a more narrow distribution of the 2D coordinates. We also fit two gaussian distributions for depicting the probability density of the x-axis and y-axis, respectively. The results in Fig. 5 show that the 2D coordinates of the spectacled samples are more probably located in the inner positions, while those of the unspectacled samples are more probably located in the outer ones.

The experiment can be extended by splitting the 2D coordinates according to different gaze targets. Fig. 6 displays the separation of right eye samples into three groups based on yaw angles, along with the probability density of the x-axis for each group. The outcomes show that the distributions of the middle group are almost coincident. This is because small angles indicate the ray is close to the lens center, and the lens has minimal impact on the light path close to its center. In this case, the spectacled and unspectacled samples share similar appearance-gaze mapping, resulting in the similar distributions. The difference mainly occurs in the groups with large yaw angles. As shown in Fig. 6, the left group shows the greatest distinction between the two distributions, with the distribution of spectacled samples shifting notably towards the inner side on the x-axis. The spectacled samples in the right group also shift towards the inner side, but the difference is less significant than that of the left group. Since the visualization is based on the right eye of the subjects, negative yaw refers to eyeball rotation towards the nasal side. The minor difference in the right group is likely due to the small appearance difference of the right eye when subjects look at left targets. A similar phenomenon can be found in Fig. 7, where samples of the left eye with positive yaw exhibit a smaller x-axis difference than those with negative yaw, which is almost symmetric with the right eye.

6.2 Multi-task Learning Evaluation

In this subsection, we evaluate the proposed multi-task training strategy on top of existing methods using public datasets. We apply our strategy to both supervised and unsupervised methods and compare the results with the baselines. Specifically, we choose XGaze [61] as the representative of supervised methods because it is the latest supervised baseline. As unsupervised settings are multifarious, it is difficult to make evaluations on all unsupervised methods. Instead, we choose GazeCLR [19] as the representative of unsupervised

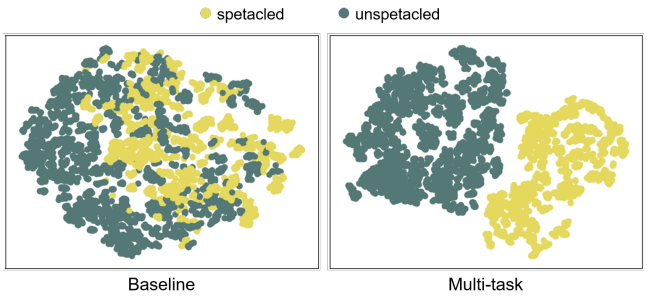

Figure 8: Visualization of ColumbiaGaze feature using t-SNE.

methods because it is based on a typical unsupervised method with released code. For classification supervision, we manually label wearing conditions for each subject in the public datasets.

Within-dataset Evaluation. We train XGaze on ETH-XGaze dataset with and without the classification loss. The original version of XGaze only takes ResNet-50 as backbone. For universal evaluation, we additionally train the version based on ResNet-18 and ResNet-34 backbones. After training, all the trained networks are tested on the testing set of ETH-XGaze. However, as the baseline only improves by 0.1 degree from ResNet-18 to Resnet-50 (Table 2), it indicates that within-dataset testing is too simple to reflect the gap between different supervised settings, which explains the minor improvement of introducing wearing classification.

For the unsupervised setting, we train GazeCLR on EVE dataset with and without the classification loss, and we formulate a similar protocol as [19] to test gaze performance. In detail, we take five subjects out of 39 as testing subjects and take the others as training subjects. Each subject contains 3,000 samples randomly selected from the whole EVE dataset. We freeze the pre-trained encoder and train an MLP for gaze regression on the training subjects with gaze labels. The gaze performance on the testing subjects is reported in Table 3, labeled as "MiniEVE". Like the evaluation on XGaze, we extend the original version (ResNet-18) to other backbone settings, i.e., ResNet-34 and ResNet-50. Note that the training configuration of the additional settings obeys the original configuration of ResNet-18. As unsupervised methods are usually sensitive to training settings, the configuration of the small backbone might not be the best for large backbones and may cause performance degradation. Therefore, the gaze errors of ResNet-34 and ResNet-50 are even larger than those of ResNet-18. However, the comparisons between settings with the same backbone are still reasonable and meaningful. Considering the same backbone, the settings with wearing classification have better gaze performance than the baselines.

Cross-dataset Evaluation. In order to better show the gap between different settings, we perform a cross-dataset evaluation using a similar protocol as [59]. Specifically, we use 8-fold, 15-fold, and 6-fold evaluations for Columbia, MPIIFaceGaze, and our data. respectively. In each fold, we freeze the feature encoder and train a new MLP based on 100 randomly selected samples with annotations on top of that. The mean gaze errors are reported in Table 2. Thanks to the classification loss, the feature encoder of the multi-task setting can distinguish spectacled and unspectacled input (Fig. 8), which helps the MLP learn different mappings for different wearing conditions. Note that the classification loss benefits not only spectacled but also unspectacled conditions. We show the detailed gaze errors in Fig. 9 by grouping wearing conditions. It demonstrates that the networks with the multi-task setting have better performance on both spectacled and unspectacled samples. In terms of improvement, the small backbone (ResNet-18) shows a more significant decrease in gaze errors. For the large backbone (ResNet-50), the improvement descends, and the performance may be similar with the baseline on MPIIFaceGaze. There are two possible reasons for this phenomenon.

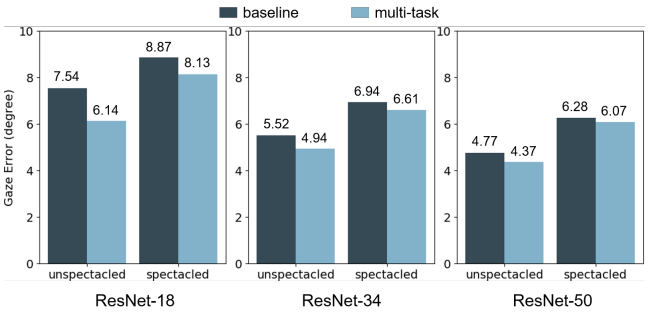

Figure 9: Detailed performance of XGaze on ColumbiaGaze.

First, as the parameter number grows, the network becomes more powerful and can implicitly handle multiple tasks. One clue shown in Table 2 is that the ResNet-50 baseline outperforms other baselines on the wearing classification task. The values represent the classification accuracy on ColumbiaGaze using 10% data to finetune an MLP. Second, MPIIFaceGaze only contains five spectacled subjects with ten unspectacled subjects, while the ratio of ColumbiaGaze is 22:34. It may not fully reflect the ability to handle multiple mappings when testing on the dataset with a lower proportion of spectacled subjects, as the multiple appearance-gaze mappings can be more easily approximated by a single one in this condition.

In addition to the supervised method, we also perform a cross-dataset evaluation on the unsupervised method (GazeCLR) using the same protocol and report the performance in Table 3. Similar to the within-dataset evaluation, some results of the large backbones are worse than ResNet-18 due to the cross-backbone configuration. However, with regard to the same backbone, the settings with wearing classification generally outperform the baselines. We also find the unsupervised method needs more data to learn good gaze mappings, which leads to the performance degradation on our data that has small data scale.

Improvement Validation. The multi-task learning is expected to work well only if the auxiliary tasks are closely related to the main task [50]. However, it is still potential for auxiliary tasks to achieve a robust representation and make improvements. To further validate that considering eyeglasses refraction benefits the improvement, we additionally take gender classification, which is also a binary classification like ours, as the auxiliary task with the same configuration. Results in Table 2 and Table 3 show that it performs worse than ours and even worse than the baseline in some settings. This indicates that our multi-task method is benefited from considering eyeglasses refraction.

7 Discussion

Our key idea is to consider eyeglasses refraction in appearance-based gaze estimation. As it has been ignored before in the literature, we first validate the theory of eyeball rotation deviation in controlled setups and then propose the multi-task training strategy as a practical solution based on the theory. They are highly-relevant with each other and both important to this paper. Although the multi-task training strategy seems simple in hindsight, we believe it still has novelty and the motivation of considering eyeglasses refraction provides insights for the community. Unlike previous multi-task gaze estimation that usually uses landmark detection and 2D gaze points regression as auxiliary tasks, our approach constructs the multi-task setting by classifying wearing conditions, which is from a new perspective. Considering the booming demand for gaze estimation in mobile and wearable devices, we believe it is good to see that such a simple adjustment can improve the performance of gaze estimation, especially on small networks that are highly demanded by mobile and wearable devices.

Position among Literature. Different from previous works that aim to make appearance-based gaze estimation robust to known factors like identities and head poses, this paper introduces the eyeglasses refraction as a new factor that challenges the appearance-based gaze estimation. Note that we are not to claim that the eyeglasses refraction is more important than other known factors. In fact, they are orthogonal to each other, and all of them are significant to achieving better gaze estimation.

Applications on VR/AR. The proposed multi-task method is a remote eye tracking method that can be directly used in 360-Degree VR Domes (e.g., Fulldome, Pacific Domes) and 3D video communication systems [66, 67] to achieve better gaze interactions. Meanwhile, wearable eye tracking is also useful for HMD-based VR/AR settings. In this situation, as the head pose relative to the camera is fixed, and the specular reflection of the lens is excluded, the proposed correction model (Sect. 4.4) has the potential to be directly used in wearable devices after individual calibration.

Despite the aforementioned success, our approach is still subject to some limitations. The method in the theory validation part succeeds in validating the existence of eyeglasses refraction, but it is not a perfect model. For example, eyeglasses are assumed to be ideal lenses, which is not true in practice. The distortion in recorded images is complicated and may not be fully addressed by using frontal head pose and normalized 2D coordinates. Astigmatism and other refractive errors are not modeled either. In addition, almost all public datasets and our dataset contain few subjects with hyperopia eyeglasses. As hyperopia normally presents during infancy and early childhood [15], hyperopia participants are difficult to recruit. Therefore, although the proposed model can handle hyperopia theoretically, the practical performance is still unknown. To some extent, this can partly explain why the proposed multi-task training strategy can work on the public dataset. As the spectacled samples almost all wear myopia eyeglasses, the appearance-gaze mappings can be roughly divided into two clusters rather than three. Strictly speaking, the binary classification is insufficient for totally solving eyeglasses refraction as the spectacled mappings differ for different diopters, which explains the less improvements of spectacled samples in Fig. 9. Future works can be made to build a conditional network to learn the mappings based on diopter labels or explicitly embed the physical laws in the network architecture.

8 Conclusion

In this paper, we validate the eyeball rotation deviation caused by eyeglasses refraction and propose a multi-task framework to leverage this for appearance-based gaze estimation. With the help of Relative Coordinate Representation (RCR), we successfully validate the existence of eyeglasses refraction on both our proposed and public datasets qualitatively and quantitatively. Based on this discovery, the proposed multi-task framework leverages a classification loss to encourage the networks to predict gaze and classify wearing conditions simultaneously, which helps the networks to learn the multiple appearance-gaze mappings. Experiments demonstrate that the proposed framework can improve gaze performance on supervised and some unsupervised methods across various backbones. To the best of our knowledge, we are the first to clearly reveal and explicitly consider eyeglasses refraction in appearance-based gaze estimation, which could provide valuable insights for future research.

Acknowledgments

This work was supported by the National Key R&D Program of China (2018YFA0704000), Beijing Natural Science Foundation (M22024), the NSFC (No.62021002), and the Key Research and Development Project of Tibet Autonomous Region (XZ202101ZY0019G). This work was also supported by THUIBCS, Tsinghua University, and BLBCI, Beijing Municipal Education Commission. Feng Xu is the corresponding author.

REFERENCES

[1] S. Andrist, X. Z. Tan, M. Gleicher, and B. Mutlu. Conversational gaze aversion for humanlike robots. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pp. 25–32, 2014.

[2] S. Baluja and D. Pomerleau. Non-intrusive gaze tracking using artificial neural networks. *Advances in Neural Information Processing Systems*, 6, 1993.

[3] Y. Bao, Y. Liu, H. Wang, and F. Lu. Generalizing gaze estimation with rotation consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4207–4216, 2022.

[4] T. Bär, J. F. Reuter, and J. M. Zöllner. Driver head pose and gaze estimation based on multi-template icp 3-d point cloud alignment. In *2012 15th International IEEE Conference on Intelligent Transportation Systems*, pp. 1797–1802. IEEE, 2012.

[5] A. Burova, J. Mäkelä, J. Hakulinen, T. Keskinen, H. Heinonen, S. Siltanen, and M. Turunen. Utilizing vr and gaze tracking to develop ar solutions for industrial maintenance. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2020.

[6] R. Caruana. *Multitask learning*. Springer, 1998.

[7] M. Chen, Y. Jin, T. Goodall, X. Yu, and A. C. Bovik. Study of 3d virtual reality picture quality. *IEEE Journal of Selected Topics in Signal Processing*, 14(1):89–102, 2019.

[8] W. Chen, R. Li, Q. Yu, A. Xu, Y. Feng, R. Wang, L. Zhao, Z. Lin, Y. Yang, D. Lin, et al. Early detection of visual impairment in young children using a smartphone-based deep learning system. *Nature Medicine*, 29(2):493–503, 2023.

[9] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pp. 794–803. PMLR, 2018.

[10] Y. Cheng, F. Lu, and X. Zhang. Appearance-based gaze estimation via evaluation-guided asymmetric regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 100–115, 2018.

[11] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pp. 160–167, 2008.

[12] H. Deng and W. Zhu. Monocular free-head 3d gaze tracking with deep learning and geometry constraints. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3143–3152, 2017.

[13] E. Dolgin. The myopia boom. *Nature*, 519(7543):276, 2015.

[14] T. Fischer, H. J. Chang, and Y. Demiris. RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments. In *European Conference on Computer Vision*, pp. 339–357, September 2018.

[15] N. J. Friedman. *The Massachusetts Eye and Ear Infirmary illustrated manual of ophthalmology*. Saunders/Elsevier, Philadelphia, Pa, 3rd ed. ed., 2009.

[16] K. A. Funes Mora, F. Monay, and J.-M. Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proceedings of the symposium on eye tracking research and applications*, pp. 255–258, 2014.

[17] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[18] M. Guo, A. Haque, D.-A. Huang, S. Yeung, and L. Fei-Fei. Dynamic task prioritization for multitask learning. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 270–287, 2018.

[19] S. Jindal and R. Manduchi. Contrastive representation learning for gaze estimation. In *Annual Conference on Neural Information Processing Systems*, pp. 37–49. PMLR, 2023.

[20] I. Kasahara, S. Stent, and H. S. Park. Look both ways: Self-supervising driver gaze estimation and road scene saliency. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pp. 126–142. Springer, 2022.

[21] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6912–6921, 2019.

[22] A. Kendall, Y. Gal, and R. Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7482–7491, 2018.

[23] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[24] R. Konrad, A. Angelopoulos, and G. Wetzstein. Gaze-contingent ocular parallax rendering for virtual reality. *ACM Transactions on Graphics (TOG)*, 39(2):1–12, 2020.

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

[26] T. C. Kübler, T. Rittig, E. Kasneci, J. Ungewiss, and C. Krauss. Rendering refraction and reflection of eyeglasses for synthetic eye tracker images. ETRA '16, p. 143–146. Association for Computing Machinery, 2016.

[27] D. Lian, L. Hu, W. Luo, Y. Xu, L. Duan, J. Yu, and S. Gao. Multiview multitask gaze estimation with deep convolutional neural networks. *IEEE transactions on neural networks and learning systems*, 30(10):3010–3023, 2018.

[28] D. Lian, Z. Zhang, W. Luo, L. Hu, M. Wu, Z. Li, J. Yu, and S. Gao. Rgbd based gaze estimation via multi-task cnn. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 2488–2495, 2019.

[29] E. Lindén, J. Sjostrand, and A. Proutiere. Learning to personalize in appearance-based gaze tracking. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pp. 0–0, 2019.

[30] G. Liu, Y. Yu, K. A. F. Mora, and J.-M. Odobez. A differential approach for gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):1092–1099, 2019.

[31] S. Liu, E. Johns, and A. J. Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1871–1880, 2019.

[32] F. Lu, Y. Sugano, T. Okabe, and Y. Sato. Adaptive linear regression for appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 36(10):2033–2046, 2014.

[33] Y. Lu, A. Kumar, S. Zhai, Y. Cheng, T. Javidi, and R. Feris. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5334–5343, 2017.

[34] J. Martschinke, J. Martschinke, M. Stamminger, and F. Bauer. Gaze-dependent distortion correction for thick lenses in hmds. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 1848–1851. IEEE, 2019.

[35] A. G. Mavely, J. Judith, P. Sahal, and S. A. Kuruvilla. Eye gaze tracking based driver monitoring system. In *2017 IEEE international conference on circuits and systems (ICCS)*, pp. 364–367. IEEE, 2017.

[36] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3994–4003, 2016.

[37] S. Park, E. Aksan, X. Zhang, and O. Hilliges. Towards end-to-end video-based eye-tracking. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pp. 747–763. Springer, 2020.

[38] S. Park, S. D. Mello, P. Molchanov, U. Iqbal, O. Hilliges, and J. Kautz. Few-shot adaptive gaze estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9368–9377, 2019.

[39] S. Park, A. Spurr, and O. Hilliges. Deep pictorial gaze estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 721–738, 2018.

[40] S. Park, X. Zhang, A. Bulling, and O. Hilliges. Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In *Proceedings of the 2018 ACM symposium on eye tracking research & applications*, pp. 1–10, 2018.

[41] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):121–135, 2017.

[42] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa. An all-in-one convolutional neural network for face analysis. In *2017 12th IEEE international conference on automatic face & gesture recognition*

(FG 2017), pp. 17–24. IEEE, 2017.

[43] A. Richard, C. Lea, S. Ma, J. Gall, F. De la Torre, and Y. Sheikh. Audio- and gaze-driven facial animation of codec avatars. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 41–50, 2021.

[44] R. Rodrigues, J. P. Barreto, and U. Nunes. Camera pose estimation using images of planar mirror reflections. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pp. 382–395. Springer, 2010.

[45] S. Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

[46] S. Sheikhi and J.-M. Odobez. Combining dynamic head pose–gaze mapping with the robot conversational state for attention recognition in human–robot interactions. *Pattern Recognition Letters*, 66:81–90, 2015.

[47] B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar. Gaze locking: passive eye contact detection for human-object interaction. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, pp. 271–280, 2013.

[48] Y. Sugano, Y. Matsushita, and Y. Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1821–1828, 2014.

[49] K.-H. Tan, D. J. Kriegman, and N. Ahuja. Appearance-based eye gaze estimation. In *Sixth IEEE Workshop on Applications of Computer Vision, 2002.(WACV 2002). Proceedings.*, pp. 191–195. IEEE, 2002.

[50] P. Vafaeikia, K. Namdar, and F. Khalvati. A brief review of deep multi-task learning and auxiliary task learning. *arXiv preprint arXiv:2007.01126*, 2020.

[51] M. Vidal, J. Turner, A. Bulling, and H. Gellersen. Wearable eye tracking for mental health monitoring. *Computer Communications*, 35(11):1306–1311, 2012.

[52] F. Wang, H. Han, S. Shan, and X. Chen. Deep multi-task learning for joint prediction of heterogeneous face attributes. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pp. 173–179. IEEE, 2017.

[53] Y. Wang, Y. Jiang, J. Li, B. Ni, W. Dai, C. Li, H. Xiong, and T. Li. Contrastive regression for domain adaptation on gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19376–19385, 2022.

[54] Z. Wu, S. Rajendran, T. Van As, V. Badrinarayanan, and A. Rabinovich. Eyenet: A multi-task deep network for off-axis eye gaze estimation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 3683–3687. IEEE, 2019.

[55] Y. Yang, J. Lyu, R. Wang, Q. Wen, L. Zhao, W. Chen, S. Bi, J. Meng, K. Mao, Y. Xiao, et al. A digital mask to safeguard patient privacy. *Nature Medicine*, 28(9):1883–1892, 2022.

[56] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim. Rotating your face using multi-task deep neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 676–684, 2015.

[57] Y. Yu, G. Liu, and J.-M. Odobez. Deep multitask gaze estimation with a constrained landmark-gaze model. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pp. 0–0, 2018.

[58] Y. Yu, G. Liu, and J.-M. Odobez. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11937–11946, 2019.

[59] Y. Yu and J.-M. Odobez. Unsupervised representation learning for gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7314–7324, 2020.

[60] J. Zhang, J. Chen, H. Tang, W. Wang, Y. Yan, E. Sangineto, and N. Sebe. Dual in-painting model for unsupervised gaze correction and animation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1588–1596, 2020.

[61] X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang, and O. Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pp. 365–381. Springer, 2020.

[62] X. Zhang, Y. Sugano, and A. Bulling. Revisiting data normalization for appearance-based gaze estimation. In *Proceedings of the 2018 ACM symposium on eye tracking research & applications*, pp. 1–9, 2018.

[63] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4511–4520, 2015.

[64] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 51–60, 2017.

[65] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):162–175, 2017.

[66] Y. Zhang, Z. Li, S. Xu, C. Li, J. Yang, X. Tong, and B. Guo. Remote-touch: Enhancing immersive 3d video communication with hand touch. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 1–10. IEEE, 2023.

[67] Y. Zhang, J. Yang, Z. Liu, R. Wang, G. Chen, X. Tong, and B. Guo. Virtualcube: An immersive 3d video communication system. *IEEE Transactions on Visualization and Computer Graphics*, 28(5):2146–2156, 2022.

[68] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pp. 94–108. Springer, 2014.

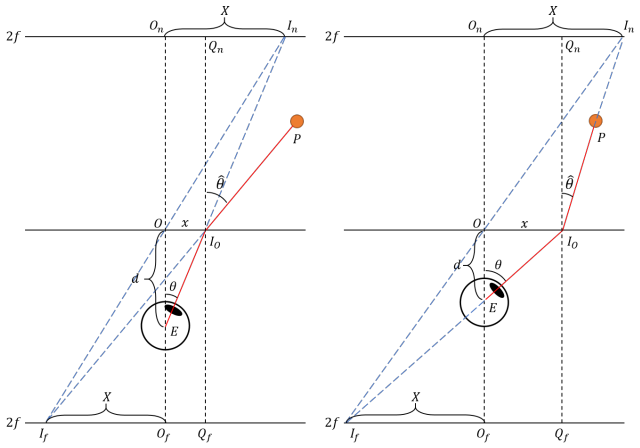

Figure 10: Illustration of the ideal lens refraction. Left: myopia. Right: hyperopia.

A Derivation of Equation

We model eyeglasses as ideal lenses with focal length f. As shown in Fig. 10, from top to bottom, there are three parallel lines representing $2f$-plane (near), lens plane and $2f$-plane (far), respectively. O represents the lens center. E represents the eyeball center. P represents the object being watched.

For myopia, the light emits from the object and hits the lens plane at I_O. According to the physical laws of the ideal lens, the light path after refraction can be determined by the following steps:

- Extend the ray $\overrightarrow{PI_O}$ and intersect the far plane at I_f.
- Extend the ray $\overrightarrow{I_fO}$ and intersect the near plane at I_n.
- Ray $\overrightarrow{I_nI_O}$ is the light path after refraction.

The ground-truth gaze direction is $\overrightarrow{EP}$ which can be described by rotation angle $\angle PEO$. However, this angle is difficult to calculate as it is related to the distance between the object and the lens plane. Instead, we use $\angle PI_OQ_n$ to approximate $\angle PEO$ as the distance between the object and the lens is usually much greater than the distance between the eyeball and the lens. After approximation, our goal is to find the relationship between $\angle PI_OQ_n$ (labeled as $\hat{\theta}$) and $\angle I_OEO$ (labeled as θ).

For $\triangle OEI_O$, we have the following equation,

$$x = d \cdot \tan(\theta) \tag{11}$$

For $\triangle O_nEI_n$, we have the following equation,

$$X = (2f + d) \cdot \tan(\theta) \tag{12}$$

For $\triangle I_OI_fQ_f$, we have the following equation,

$$X + x = 2f \cdot \tan(\hat{\theta}) \tag{13}$$

After simplification, we have the following equation,

$$\tan(\hat{\theta}) = (1 + \frac{d}{f}) \cdot \tan(\theta) \tag{14}$$

As diopter D and focal length f have the following relationship,

$$|D| = \frac{1}{f} \tag{15}$$

we can get the final equation as follows,

$$\tan(\hat{\theta}) = (1 + |D| \cdot d) \cdot \tan(\theta) \tag{16}$$

For hyperopia, the light path after refraction can be determined by the following steps:

- Extend the ray $\overrightarrow{I_OP}$ and intersect the near plane at I_n.
- Extend the ray $\overrightarrow{I_nO}$ and intersect the far plane at I_f.
- Ray $\overrightarrow{I_OI_f}$ is the light path after refraction.

Similar to myopia, we use $\angle PI_OQ_n$ to approximate $\angle PEO$ as the distance between the object and the lens is usually much greater than the distance between the eyeball and the lens.

For $\triangle OEI_O$, we have the following equation,

$$x = d \cdot \tan(\theta) \tag{17}$$

For $\triangle O_fEI_f$, we have the following equation,

$$X = (2f - d) \cdot \tan(\theta) \tag{18}$$

For $\triangle Q_nI_OI_n$, we have the following equation,

$$X - x = 2f \cdot \tan(\hat{\theta}) \tag{19}$$

After simplification, we have the following equation,

$$\tan(\hat{\theta}) = (1 - \frac{d}{f}) \cdot \tan(\theta) \tag{20}$$

Finally, based on the relationship between diopter D and focal length f, we have the following equation,

$$\tan(\hat{\theta}) = (1 - |D| \cdot d) \cdot \tan(\theta) \tag{21}$$

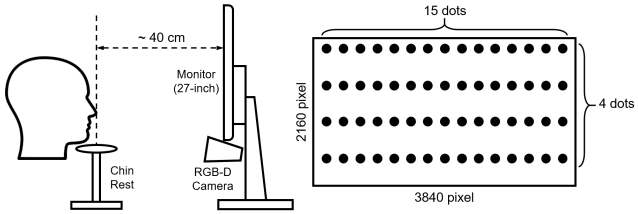

Figure 11: Illustration of capture system and the distribution of dots.

B Details of Capture System

As shown in Fig. 11, our capture system consists of a 4K monitor (27-inch), a RGB-D camera, and a chin rest. The chin rest is put in front of the monitor, and the distance between them is about 40 cm. The camera is set under the monitor, and is calibrated by a mirror-based method [44] to locate the screen in the camera coordinate system (CCS). We pre-set 60 dots (4 rows with 15 dots per row) on the screen. During data collection, the dots are displayed sequentially, and participants are asked to follow the dots without changing their head poses. When the participants focus on the dots, they can press the keyboard that controls the camera to record the RGB-D image.

Given a RGB-D frame, we can get the 3D position of the eye in the CCS. With the dot index and pre-calibrated transformation, we can also get the 3D dot position in the CCS. The gaze direction in the CCS is defined by the direction from the 3D eye position to the 3D dot position. Finally, we convert the direction vector of gaze to the pitch and yaw angles. As for head pose, we use the same definition as [48], where the direction of head pose is defined by the normal of the triangle (two eyes and the mouth). We also convert the direction vector of head pose to the pitch and yaw angles.

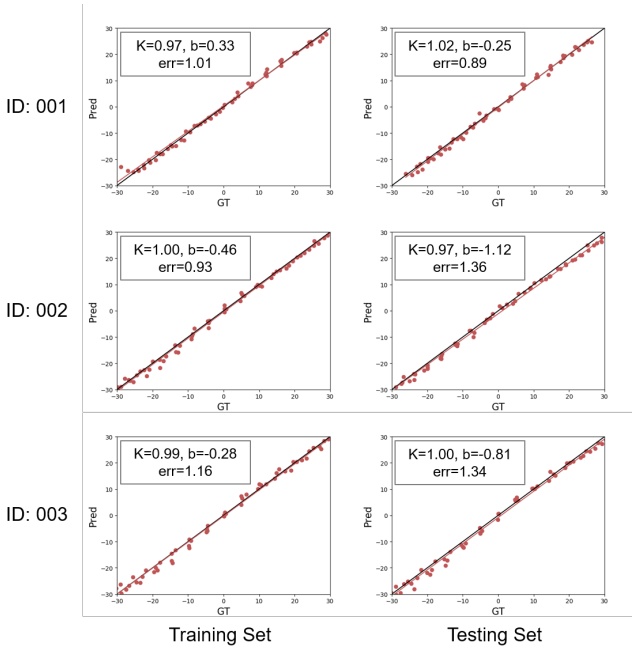

Figure 12: Scatter plots of yaw results on unspectacled sets for training and testing.

C Parameter Number of Multilayer Perceptron

In the theory validation, we train an MLP using one unspectacled set (i.e., unspectacled for training) to learn the mapping between the RCR and gaze, and then test it on the other unspectacled set (i.e., unspectacled for testing) and the spectacled set. To ensure that our MLP is not over-parameterized, we compare the performance of the trained MLP on both the training and testing unspectacled set. We show yaw results of three subjects as examples in Fig. 12, where the trained MLPs have similar errors on the training and testing set. This indicates that our MLP is not over-parameterized, and the deviation shown in the main text is irrelevant with network modeling.