Neural 3D Gaze: 3D Pupil Localization and Gaze Tracking based on Anatomical Eye Model and Neural Refraction Correction

Conny Lu* UNC Chapel Hill
Praneeth Chakravarthula† Princeton University
Kaihao Liu‡ UNC Chapel Hill
Xixiang Liu§ UNC Chapel Hill
Siyuan Li¶ UNC Chapel Hill
Henry Fuchs‖ UNC Chapel Hill

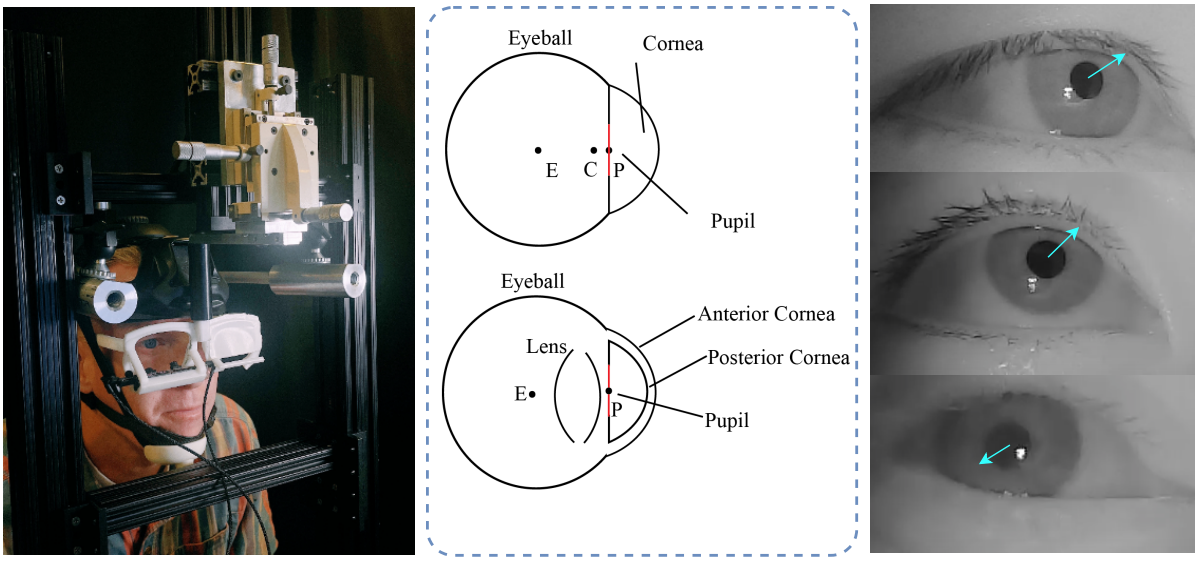

Figure 1: *Left:* The system we built to simultaneously evaluate 3D pupil localization and gaze estimation on a real dataset. *Middle:* We utilize a biologically advanced eye model to replace the commonly used simple eye model and propose a new deep learning-based 2D refraction correction method to achieve better accuracy. *Right:* Gaze estimation results under three different viewing angles, the light blue arrow indicates the estimated gaze direction.

Abstract

Eye tracking has already made its way to current commercial wearable display devices, and is becoming increasingly important for virtual and augmented reality applications. However, the existing model-based eye tracking solutions are not capable of conducting very accurate gaze angle measurements, and may not be sufficient to solve challenging display problems such as pupil steering or eyebox expansion. In this paper, we argue that accurate detection and localization of pupil in 3D space is a necessary intermediate step in model-based eye tracking. Existing methods and datasets either ignore evaluating the accuracy of 3D pupil localization or evaluate it only on synthetic data. To this end, we capture the first 3D pupil-gaze-measurement dataset using a high precision setup with head stabilization and release it as the first benchmark dataset to evaluate both 3D pupil localization and gaze tracking methods. Furthermore, we utilize an advanced eye model to replace the commonly used over-simplified eye model. Leveraging the eye model, we propose a novel 3D pupil localization method with a deep learning-based corneal refraction correction. We demonstrate that our method outperforms the state-of-the-art works by reducing the 3D pupil localization error by 47.5% and the gaze estimation error by 18.7%. Our dataset and codes can be found here: link.

Index Terms: Computing methodologies—Computer graphics—Graphics systems and interfaces—Mixed / augmented reality; Computing methodologies—Artificial intelligence—Computer vision—Computer vision problems

1 Introduction

Eye tracking enables dozens of applications in augmented and virtual reality (AR/VR), physiologically informed rendering such as foveation, gaze-based interaction, diagnosis of conditions like Parkinson's disease, and so on. Typically, existing near-eye tracking solutions employ cameras that are looking at the user's eye to track their gaze. Specifically, in model-based eye tracking, an approximate elliptical model is first fit to the pupil as tracked by the cameras, the 2D pupil is then unprojected into a 3D pupil represented as a circle, and the vector joining the center of the eyeball (or cornea) and the center of the 3D pupil is the estimated user's gaze. While such model-based eye tracking methods have been widely adopted recently in a variety of commercial eye trackers (e.g. Pupil Labs) and near-eye displays (e.g. HTC Vive), they are still not sufficiently accurate, as per performance standards described in prior papers, such as [6].

As a necessary step of model-based gaze estimation, the sig-

*e-mail: connylu@cs.unc.edu
†e-mail: cpk@cs.unc.edu, praneethc@princeton.edu
‡e-mail: kaihao@live.unc.edu
§e-mail: xxl@ad.unc.edu
¶e-mail: lsy0320@ad.unc.edu
‖e-mail: fuchs@cs.unc.edu

nificance of 3D pupil localization is overlooked. In addition to the benefit in gaze estimation, accurate 3D pupil localization itself forms an essential part of the solution to steering the tiny eyebox of (holographic) displays, display dynamic uniformity correction, and display dynamic distortion correction. Moreover, it is essential to also mitigate the pupil swim of world-locked rendering. In this work, we analyze and argue that even minor inaccuracies in the 3D pupil localization result in severe errors in model-based gaze estimation. The state-of-the-art works either did not consider corneal refraction [9, 26] or used oversimplified eye models [4, 5]. More importantly, as far as we know, no work conducted evaluations of 3D pupil localization on real datasets.

We overcome the major challenges of pupil localization in near-eye tracking by 1) specially designing a precision hardware tracker to measure the ground truth of the relative location of the pupil in 3D space, 2) capturing a high accuracy pupil localization and gaze tracking dataset, which is the first to include real 3D pupil movement and the corresponding gaze angle for every individual frame of all the tracking cameras, 3) employing an anatomically-aware advanced eye model, and finally, 4) proposing an image-based refraction correction neural network that can decrease the gaze errors caused by corneal refraction and can be easily added to most of the existed gaze estimation methods using detected pupil contours.

Precision Hardware for Pupil Localization To better evaluate the existing 3D pupil localization methods and encourage progress in this direction, we build a precise hardware system to create the first real eye dataset benchmark for 3D pupil localization, which can also measure the accuracy of gaze estimation. This system has robust head and chin stabilization and can evaluate the existing 3D pupil localization methods at a very fine level (0.02mm).

3D Pupil-Gaze Dataset With the precise eye tracking system, the relative 3d pupil center ground truth is measured by moving the optical stage while keeping the viewing angle and head position constant. Our dataset also contains ground truth of 2D pupil and gaze directions. We detect the 2D pupil position by fitting an ellipse using the state-of-the-art method [15] followed by manual adjustment for incorrect detection.

Anatomically-aware Advanced Eye Model Existing solutions use a LeGrand eye model that approximates the eye structure using two intersecting spheres; a large one representing the eyeball and a small one representing the cornea. While this over-simplified eye model accelerates the computation, it also introduces additional errors [4, 25]. Therefore, we use a biologically more accurate eye model to conduct 3D pupil localization where the eye is represented as a precise multi-layer quadratic surface with different refractive indices.

Image-based Refraction Correction As corneal refraction is a non-negligible error source of 3D eye tracking [8], we propose a novel neural network based method to accomplish image-based refraction correction. Compared to previous refraction correction methods, our method does not assume an over-simplified eye model and is also not method-dependent, which can be generalized to most of the existing image-based gaze estimation methods. Compared to the state-of-the-art works, our method achieves higher accuracy of 3D pupil detection and gaze estimation with an increase of 47.5% and 18.7%.

The contributions of our work are presented as follows:

- We build an eye tracking system to conduct more accurate 3D pupil localization and gaze estimation. This system includes a display, eyeglasses with two Pupil Labs cameras attached on an optical stage that can shift the eyeglasses at a very fine level, and a head stabilization device to mitigate the error due to small head movement.
- We create a real dataset for tracking the accurate movement of 3D pupil position, which could be served as a benchmark to evaluate different 3D pupil localization methods.
- We theoretically analyze the significant impact of small error of 3D pupil localization on the accuracy of model-based gaze estimation and propose a novel 3D pupil localization method with deep learning based refraction correction using an advanced eye model.
- We conduct extensive experiments qualitatively and quantitatively, achieving a 18.7% lower gaze angle error in gaze estimation and a 47.5% lower error of 3D pupil location compared to the state-of-the-art works.

2 Related Work

2.1 Model-based gaze estimation

Model-based gaze estimation predicts gaze based on a 3D eye model which can be fitted to features extracted from the eye and/or face images. Yamazoe et al. [31] first built a 3D eye-face model, where gaze direction can be determined by tracking facial features and locating iris centers. Pupil Labs [15] utilized detected pupil contours to infer gaze direction using an average 3D eye model. Wang et al. [29] proposed a 3D eye-face model to enable 3D eye gaze estimation with a single web camera and introduced a unified calibration algorithm to simultaneously reconstruct an individual 3D eye-face model and estimate personal eye parameters. For a more detailed survey of model-based gaze estimation, please refer to the latest survey [16].

2.2 3D Pupil Localization

3D pupil localization methods could be categorized as glint-based (i.e. using LED reflections) and glint-free (i.e. without using LED reflections), where glint-based methods play the dominant role due to their higher accuracy. Before computing the 3D pupil position, most of the glint-based methods estimated the 3D cornea center either by triangulating glints location with coaxial camera and LEDs [3] or by intersecting multiple reflection planes with assumed cornea radius [20, 28, 30]. On the other hand, glint-free methods required simpler hardware configuration and can operate in outside environments. Our method is also glint-free. [7] utilized stereo matching to estimate the 3D pupil center with detected two 2D pupil centers. [26] proposed a temporal approach to conduct glint-free model-based eye tracking using only a single camera. However, both of them computed 3D virtual pupil but not actual pupil, as they did not consider the impact of corneal refraction. [4, 27] provided a detailed analysis of the effects of refraction in glint-free gaze estimation. [4] also introduced an inverse ray-tracing based cost function that accounted for refraction. As follow-up work, [5] accelerated the non-linear optimization process by re-casting model optimization as a least-squares intersection of lines. Despite great results, all their methods are based on the simple Le Grand eye model [21] which has significant approximation errors [4, 25]. Moreover, all the evaluations of the accuracy of 3D pupil detection are based on synthetic datasets, not real datasets.

2.3 Eye dataset

Current real eye datasets mainly contain information on gaze angles [10, 19], 2D pupil [10–13], and 2D eye features represented as segmentation masks [10, 14, 17, 23], but no gold standard 3D labels (e.g. 3D pupil or eyeball center) are provided due to the difficulty of collecting accurate ground truth. To this end, synthetic eye datasets have been used extensively for evaluating the accuracy of eyeball and pupil localization algorithms in 3D. Dierkes et al. [5] generated synthetic images at 640x480 pixels resolution using a ray-tracing pipeline and evaluated the accuracy of their methods for eyeball center detection and gaze estimation based on the synthetic eye images.

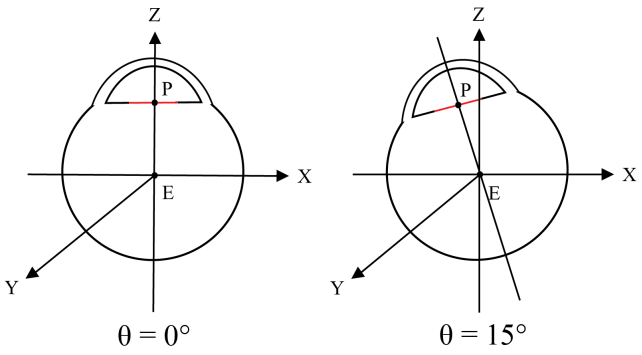

Figure 2: Schematic diagram of eyeball coordinate system and eyeball horizontal rotation. The left image visualizes the eyeball without rotation, while the right image illustrates a 15° horizontal eyeball rotation.

NVGaze [17] is another widely used synthetic dataset, in which the images were labeled with the exact 2D gaze vector, 3D eye location, and 2D pupil location. The 3D pupil location can then be inferred easily.

However, there is still a clear gap between synthetic datasets and real datasets in terms of image quality and realism. Real datasets contain different types of noise that cannot be simulated in a synthetic environment, which leads to a large performance drop when applying the method trained on synthetic datasets to real datasets. Our work aims to provide a real dataset with 2D and 3D pupil ground truth that can be used to evaluate the accuracy of pupil localization and gaze estimation algorithms.

3 Theoretical Analysis

Model-based glint-free gaze estimation methods compute the optical axis by connecting the 3D pupil center and eyeball center or obtaining the normal of the 3D pupil. In both cases, computing the 3D pupil center position is a necessary step. We conduct a theoretical analysis to investigate the impact of the estimation error of 3D pupil center on the accuracy of gaze angle estimation. We argue that even a small error of 3D pupil localization worsens the accuracy of model-based gaze estimation significantly.

As shown in Fig. 2, we set the center of the eyeball as the origin, the eye direction of the eyeball axial length as z axis, the up direction as y axis, and the axis parallel to them while intersecting at the origin as x axis. To compute the correlation between the 3D pupil center error and gaze angle error, we first fix a gaze angle θ and express the ground truth gaze vector as $g = (d\sin\theta, 0, d\cos\theta)$, where d is the distance between the eyeball center and 3D pupil center. Then we express the erroneous gaze vector $\hat{g}$ after adding a small error δ to 3D pupil center estimation as $(d\sin\theta+\delta, 0, d\cos\theta)$, $(d\sin\theta, \delta, d\cos\theta)$, and $(d\sin\theta, 0, d\cos\theta+\delta)$ in x, y, and z axis, respectively. In the experiment, we set θ to 0°, 10°, 20°, 30°, d to 10.39cm [5].

Finally, we compute the angle error between the ground truth gaze vector g and the erroneous gaze vector $\hat{g}$ using Equation 1:

$$E_{gaze} = arccos(\frac{g \cdot \hat{g}}{||g||||\hat{g}||}) \quad (1)$$

Note we only consider the horizontal movement of the eye in the analysis, since the vertical movement of the eye gives the same result.

Fig. 3 illustrates the change of gaze angle error with 3D pupil position error when the gaze angle θ is 0°, 10°, 20°, 30°, respectively.

We observed that, when $\theta = 0°$, the 3D pupil center errors in x and y axis have the same effect on gaze angle error, making angle error increases linearly as the pupil error increases. while the 3D pupil center error in z axis has no effect on gaze angle error. As θ increases, the error curve in y axis keeps the same, the error curve in x axis shows an increasing asymmetry in the positive and negative directions, while the error curve in z axis illustrates a more significant impact of 3D pupil error to the gaze angle error, but the impact is still smaller than other two axes.

Most importantly, we found that, at any angle, a 0.5mm(1mm) error of 3D pupil estimation in x or y axis induces a 3°(5°) error of gaze angle, which emphasizes the importance of accurate assessment of existing 3D pupil localization works. In later sections, we will also demonstrate that on our benchmark, current methods are still far from accurate, and developing a more accurate 3D pupil localization method is necessary.

4 Dataset

In this section, we describe our dataset building process in three steps. First, we present our hardware setup used to capture real eye data and build the benchmark. Second, we introduce the specific experimental procedures to obtain all necessary eye videos and raw data. Finally, we illustrate the methods of computing the ground truth of 2D pupil, 3D pupil as well as the gaze direction from captured images and raw data.

4.1 Setup building

To acquire high quality eye datasets, we carefully design our setup with two high resolution Pupil Labs cameras attached to eyeglasses, a head stabilization frame, and a display for visualizing target points. The whole setup can be seen in Fig. 4.

For image capturing, we utilize two high quality near-eye Pupil Labs cameras [15] attached to the eyeglasses. We adjust the camera positions to enable the eye to occupy a large space in the whole image, while allowing the eye to move freely without exceeding the edge of the image. Each camera is placed where clear and unobstructed eye images can be captured at any horizontal and vertical rotation angles within +-30°, avoiding the heavy eyelashes occlusion when the eyes look down.

In our experiment, an essential source of error is the head movement of subjects during the capture. Traditional eye tracking setup normally constrains the head movement using a chinrest, however, our experiment has a significantly higher requirement for head stabilization, as even a 0.5mm error leads to a large 3°gaze angle error. Using a chinrest only, subjects can still easily move their heads at least 1cm left and right. To this end, we design a head stabilization setup in consist of a helmet and a chinrest. The helmet with metal horns can be placed in the corresponding slots on a fixed frame to tightly control the forehead movement, while the chinrest controls the chin movement as in prior studies.

One thing worth noting is that we do not directly measure the *absolute* ground truth of 3D pupil center, as 3D pupil center is covered by the cornea and anterior chamber and cannot be measured directly. Current medical instruments can only measure eye shape parameters (e.g. cornea thickness) to a very fine level, but not the position relative to outside cameras. Thus, our dataset captures high-precision *relative* ground truth that represents the relative change of 3D pupil center. To obtain the relative ground truth, we attach the eyeglasses to an optical stage that can shift the whole eyeglasses with cameras in a very small amount of 0.02 mm.

Finally, we place a display in front of the user to visualize dots in known positions from the user's point of view, which enables data collection under a variety of gaze angles. We calibrated the position of the display in the eye camera coordinate using an external camera.

4.2 Dataset collection procedure

Before the data capture process, we first calibrate and fix each subject's head on our head stabilization setup. We used two thin boards

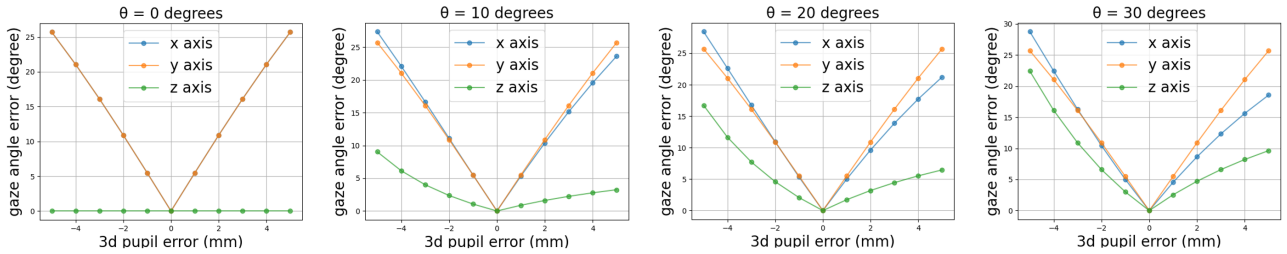

Figure 3: The correlation between 3D pupil center error and gaze angle error at 0°, 10°, 20°, and 30° horizontal gaze angle.

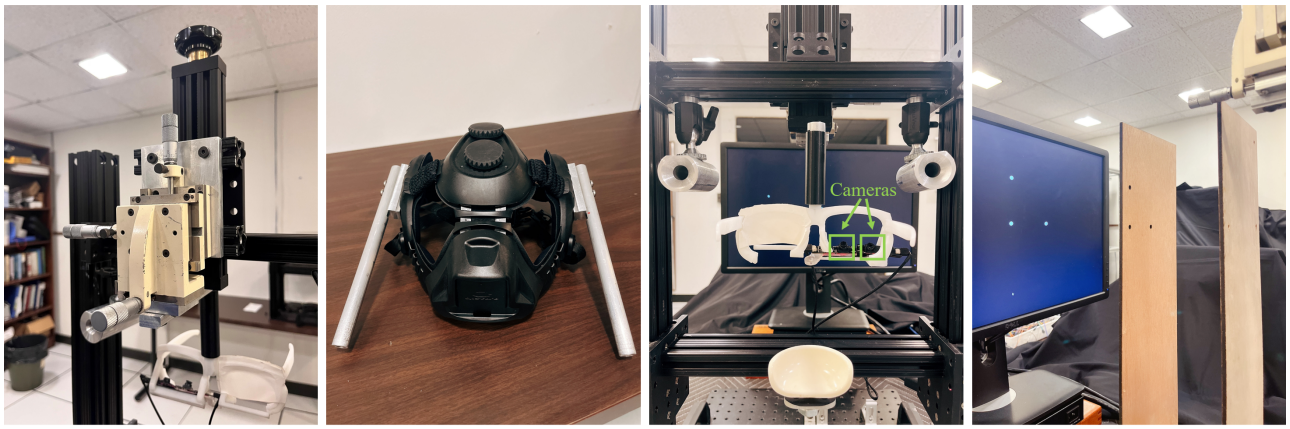

Figure 4: The hardware setup we used to capture our real dataset. *First column*: optical stage for moving cameras precisely; *Second column*: helmet with metal horns that can be placed in the corresponding slots for fixing the subject's head position, used with the chinrest; *Third column*: eyeglasses with two eye cameras that capture eye images; *Fourth column*: screen for showing the dots enabling calibration, calibrating boards for ensuring a zero gaze direction.

with three small holes that form a right-angle shape to calibrate the zero gaze angle. Two aligned thin boards are vertically placed in front of the display with a 20cm distance. At a sight angle of zero, the subject is expected to see three complete dots at pre-calculated positions on the screen through holes in both boards. By adjusting the chin rest in all three axes, all users' eyes can then be calibrated in the same 3D position. To synchronize the Pupil Labs cameras and the display, we match their timestamps using the system clock.

After all preparations are completed, we turn on the cameras and run the experiment program by visualizing 25 dots on the display that are arranged in 5 rows and 5 columns and span a space of +-12 vertical degrees and +-15 horizontal degrees. Following [22], users are required to click the dot using a mouse to ensure their fixation on the dot. To sharpen the user's focus, the dot is set to shrink after clicking and fully disappears in another 0.5 seconds.

The 25 dots focus experiment is repeated 9 times with different eyeglasses (as well as camera) positions, where both cameras move 1.27mm, 2.54mm, and 3.81mm in each of the three axes together. Since the user's head does not move during the experiment, the gaze direction of focusing on the same dot is always the same no matter how the cameras move. The whole experiment procedure is illustrated in Algorithm 1 and lasts about 10 minutes for each subject.

4.3 Ground truth generation

With the raw videos captured by two Pupil Labs cameras and the corresponding dot positions on the display, we create a dataset with accurate ground truth of 2D pupil, 3D pupil, and gaze angle. Our dataset is the only one that includes data on 2D pupil, 3D pupil, and gaze angle.

2D pupil We first compute the 2D pupil position in the eye images based on the state-of-the-art 2D pupil detection method [15], which approximates the 2D pupil as an ellipse and estimates the 2D pupil center, axes, and radius. This method can already achieve ground truth level accuracy in general, but still contains detection errors in some cases, such as eyelid occlusion and rapid eye movement. We manually check every image and relabel all the detection errors.

3D pupil As we only move the cameras, not the user's head, the head position is fixed relative to the display. Thus, when the eye looks at the same target on the display, the eye rotation is always the same in the world coordinate, no matter how the camera moves. Given the same eye rotation, the relative distance of the 3D pupil center in two frames can then be represented as the fine-grained distance the camera moves. For example, at the same gaze angle, assuming two frames are captured before and after the 1.27mm movement of the optical stage along the x-axis, the ground truth distance of 3D pupil centers in two frames should also be 1.27mm.

Gaze direction To obtain the accurate gaze direction in the eye image, we extract the 2D dot position on the display and then transfer it to a 3D point in millimeters with the knowledge of the transformation between pixels and millimeters and the transformation between the eye and display coordinate. The gaze is represented as a 3D vector (g_x, g_y, g_z) pointing from the eye rotation center to the target on the screen.

Algorithm 1: Dataset collection process

Cameras Activation: Adjust the positions of two Pupil Labs cameras to see appropriate eye images. *Note we only need to do this once for all subjects*;
Subject-specific Calibration: Calibrate the eye origin position with zero eye rotation for a specific user using a thin board with holes;
Head Stabilization: Stabilize user head using chin rest and the helmet with metal horns that can be placed in the corresponding slots;
Data Capture:
for $a \in \{x, y, z\}$ **do**
  for $t \leftarrow 0$ **to** 0.15 *inch* $(3.81mm)$ **by** 0.05 *inch* $(1.27mm)$ **do**
    Move all cameras t *inch* along a axis while setting two other axes to 0 ;
    for $vert \leftarrow -12°$ **to** $12°$ **by** $6°$ **do**
      for $hori \leftarrow -15°$ **to** $15°$ **by** $7.5°$ **do**
        Visualize a dot on the display at $(vert, hori)$ degree;
        Shrink the dot 0.5 seconds after the user clicks on it;
        Record the dot position and timestamp;
      end
    end
  end
end

5 3D Pupil Localization

As we have shown in Sect. 3, the accuracy of model-based gaze estimation is significantly subject to the estimation error of 3D pupil localization. To achieve better 3D pupil localization, we propose a novel method utilizing a biologically more accurate eye model [1] and developing a deep learning-based refraction correction method to mitigate the error introduced by corneal refraction, as shown in Fig. 7. Results are evaluated in our captured real dataset.

5.1 Advanced eye model

Currently, all the state-of-the-art glint-free 3D pupil localization methods are based on a simple two-sphere eye model [21], i.e. LeGrand model. As shown in Fig. 5, the LeGrand eye model approximates the eye structure using two intersecting spheres, the large one representing the eyeball, and the small one representing the cornea. Pupil and iris are modeled as two concentric circles in the intersecting plane perpendicular to the optical axis. Using spheres to approximate the eyeball and cornea simplifies and accelerates the computation process, especially when computing the reflection and refraction of light. However, this over-simplified eye model also introduces non-negligible errors. To achieve more accurate 3D pupil localization results, we argue that using an advanced eye model can improve the 3D pupil localization results by a large margin.

The biologically more accurate eye model we use is based on [1, 2]. In this eye model, the eye is represented as more precise multi-layer quadratic surfaces with different refractive indices that are set for the wavelength domain used to illuminate the eye. The cornea consists of the anterior and posterior surface that differ by 0.55mm, where the anterior surface of the cornea is modeled as a tri-axial ellipsoid. Another important feature of this eye model is that it also incorporates dynamic eye properties. The simple eye model assumes the eyeball sphere center as the fixed rotation center, but experiments revealed that the center of rotation for horizontal eye movements is deeper than that for vertical eye movements [2]. Therefore, the advanced eye model also parameterized the depth of the rotation center for horizontal and vertical eye movements.

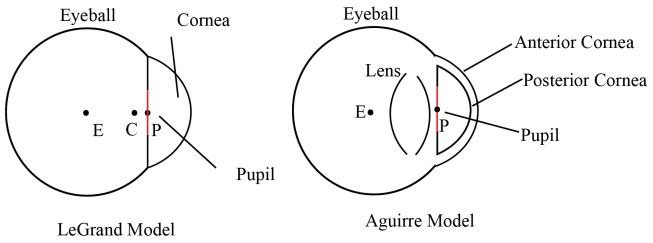

Figure 5: Eye model comparison. *Left:* A simple two-sphere eye model that is commonly used, *Right:* The more advanced eye model of [2].

5.2 Deep learning-based refraction correction

With the advanced eye model, we now develop a novel deep learning based refraction correction method to reduce 3D pupil errors caused by corneal refraction.

3D pupil localization normally starts with 2D pupil detection by fitting an ellipse in the image. It is worth noting that this ellipse is distorted and magnified by cornea surfaces with a refractive index bigger than 1, especially when the viewing angle is large. [8] showed that as viewing angle increases, the virtual pupil moves forward, tilts, and curves towards the observer's direction. Fig. 6 is a schematic figure illustrating the pupil location with and without considering refraction. We can observe an apparent change of size and position when comparing the real pupil and virtual pupil, which results in a significant decrease in the accuracy of 3D pupil localization. Glint-based methods [28] utilize glints (i.e.first purkinje image or anterior corneal reflection) to estimate cornea center and shape that benefit refraction correction, but the glint-free methods can only rely on pupil information. [4] and [5] are two recent glint-free methods that incorporated refraction correction into eyeball center estimation and 3D pupil localization. [4] proposed an optimization-based method based on the unprojection of 2D ellipse in 3D space, while [5] presenting a faster and simpler method for estimating two empirical correction functions of eyeball center and gaze direction to account for corneal refraction effects. However, their empirical correction functions assume an over-simplified eye model as we mentioned in Sect. 5.1 and correct method-dependent errors, so that the system correction cannot be generalized to other methods. In contrast, we propose a novel refraction correction method that operates on the general 2D space. Instead of assuming no refraction during the computation and then adding the refraction effects, we directly de-refract the detected 2D pupil, which can be modeled as a neural network. Our method is not tied to a specific 3D algorithm and can be easily applied to other 3D pupil localization methods, where 2D pupil detection is a necessary step.

To this end, we do not explicitly model parameters that affect the appearance of 2D pupil (including eye structure, eye dynamic movement, the transformation between eye and camera coordinate, and camera projection), but directly input the 2D pupil into a small 3-layer neural network and output the corresponding de-refracted pupil. We represent the pupil using an ellipse with five parameters (c_x, c_y, a, b, γ), where (c_x, c_y), (a, b), and γ are ellipse center, axes and tilt angle, respectively. Since our input and output dimensions are both low, a 3-layer neural network can be trained fast and also mitigate overfitting. During the training, we optimize a L1 loss function that minimizes the distance between the normalized ground truth ellipse vector and the predicted ellipse vector. Note that de-refraction is also affected by other user-specific parameters (e.g. cornea shape), this network can probably infer more accurate results if we consider feeding cornea parameters as well. However, it

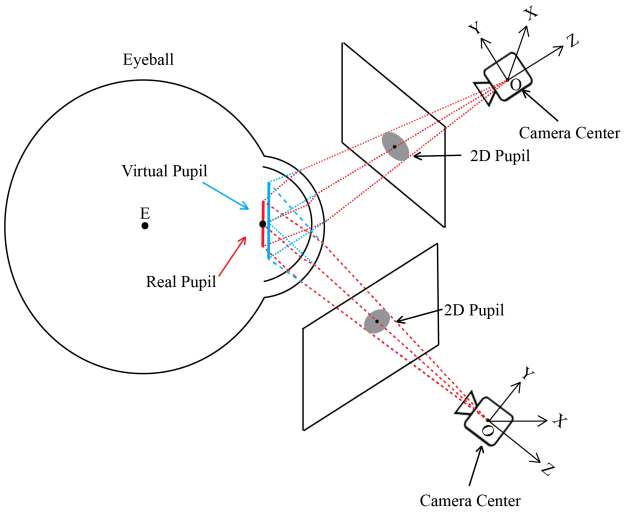

Figure 6: Schematic figure illustrating the 3D pupil location with (real pupil) and without (virtual pupil) considering refraction. With refraction, two rays in red cast by two cameras are bent when passing through the anterior and posterior cornea surface. Without refraction, the rays in blue are not bent and form a virtual pupil in front of the real pupil.

is hard to obtain precise cornea-related parameters without using medical-grade equipment. The intuition of our method is that ellipse position, axes, and tilt angle indicate useful information about the eyeball and cornea position in the camera coordinate, which leads to better results compared to previous studies. One challenge with this approach is that it is difficult to obtain the ground truth of the de-refracted 2D ellipse in real data. We thus generate a synthetic dataset including original and de-refracted 2D ellipses, built on the advanced eye model. We express the 2D ellipse without corneal refraction by setting the refractive index of the cornea and anterior chamber to 1. Note in general, there is a domain gap between synthetic and real data, it is mitigated in our case due to the low-dimensional simple parameter space and the usage of the biologically accurate eye model.

To build the synthetic dataset that can simulate the real situations as much as possible, we randomly select a variety of eye and scene-related parameters. Firstly, we sample the eye rotation in a large area, -40 to 40 degrees horizontally and -10 to 50 degrees vertically. As the rotation is represented in the camera coordinate, we sample more upward-looking vertical angles based on the assumption that current near-eye cameras are normally placed in front of or under the eyes to avoid the interference of eyelashes and eyelids. Likewise, the camera position is also bound to a constraint space to get closer to the real distribution. Specifically, the camera translation is drawn randomly from a uniform distribution within a box with a width from -20 to 20cm, a height from -20 to 0cm, and a length from 20 to 40cm. To incorporate more pupil size changes, we randomly draw it from a uniform distribution between 1mm to 4mm. We finally save the 5-dimensional parameters of 2D ellipses before and after refraction.

With the synthetic dataset, we are now capable of training the 2D de-refraction neural network using synthetic ground truth. After obtaining the 2D de-refracted pupil, we compute 3D pupil location using one or multiple cameras. We also develop a stereo camera-based 3D pupil localization method using triangulation. We first detect 2D pupil in the image and then feed it into the de-refraction neural network. After getting the de-refracted 2D pupil in both cameras, we triangulate the 2D pupil center to infer the final 3D pupil center.

5.2.1 Network architecture

Our neural network is a 3-layer network with 2 hidden layers, the first and second hidden layer contains 1024 and 512 nodes respectively. We used ReLU as our activation function.

5.2.2 Training details

We use PyTorch [24] to train our 3-layer neural network. For the training, we use the Adam solver [18] with a learning rate of 0.003 and betas of (0.5, 0.999). To accelerate the training, We also set a learning rate scheduler to lower the learning rate by 0.3 at epoch 50, 100, and 150. The whole training with 200 epochs on a dataset with 50000 data takes less than 10 minutes on a GTX 2080 Ti.

6 Experiments

In this section, we evaluate our results of 3D pupil center localization and gaze estimation on our proposed benchmark. We first demonstrate and analyze our results qualitatively and quantitatively, and then compare our results with state-of-the-art works and show that our method outperforms existing methods by 47.5% on 3D pupil localization and 18.7% on gaze estimation. Finally, we conduct ablation studies to demonstrate the superiority of using the biologically advanced eye model and of our deep learning based refraction correction.

6.1 Experiment settings

6.1.1 Dataset description

Our real dataset contains 5 subjects of different ages (15-50), genders (3 females and 2 males), and races (1 Black, 3 Asian, and 1 White), capturing gaze angles spanning in a +-15°space in the horizontal direction and a +-12°space in the vertical direction. For each subject, the dataset contains two 640×480 eye videos captured by two Pupil Lab cameras and provides the corresponding .csv files recording the target positions on the display, the time interval when the eye fixates on each target, and the exact distance the camera has moved.

6.1.2 Evaluation metrics

3D pupil localization As we capture the relative ground truth, namely ground truth that represents changes in data, it is not possible to compute the absolute error between the estimated 3D pupil position and the ground truth one. Thus, to evaluate the accuracy of various 3D pupil center localization methods on our benchmark, we propose a novel evaluation metric fitted well to our *relative* ground truth.

First, we assume that eye rotation is always the same when the subject stares at the same target, as the user's head does not move relative to the display during the whole data capture process, thanks to our carefully designed head stabilization setup. In the experiment, the cameras are moved 8 times and have a total of 9 positions. The data with the largest movement in z axis is removed due to the heavy occlusion of eyeglasses in the images. We call each position of the camera a spot. At each spot, the user looks at 25 targets on the screen, in total 225 targets. For each target, we compute the distance between two point sets by computing the euclidean distance between the center of each point set. Note we don't choose the classical Hausdorff distance due to its high sensitivity to even a single outlier. As we know how far the camera has moved relative to the first spot (reference spot), we can obtain the exact movement distance of 3D pupil center at all other spots when staring at the same target. We expect to see the estimated position change of 3D pupil center between two spots is the same at different gaze angles. We also expect that the estimated position change is as close as possible to the pre-measured spot distance. We design our evaluation criteria based on these two expectations.

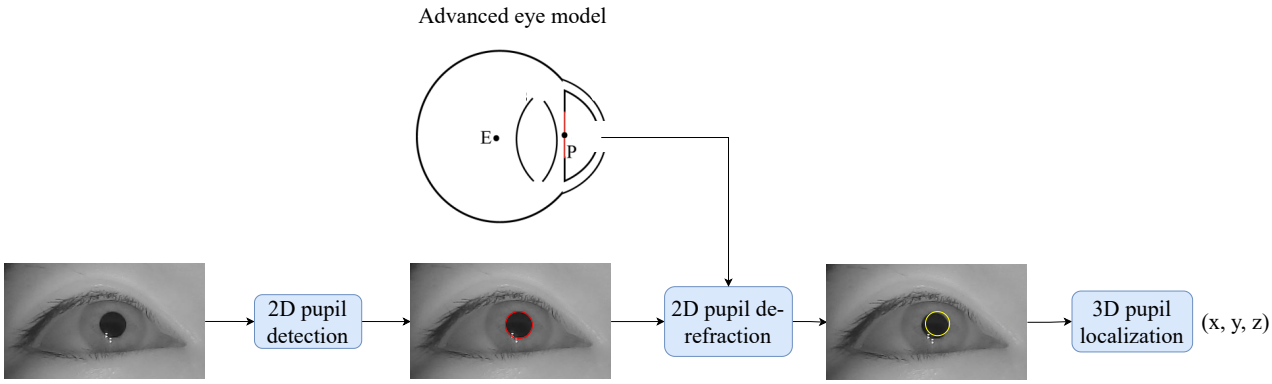

Figure 7: Our pipeline of 3D pupil localization. We first detect 2D pupil in the image highlighted in red and then de-refract the 2D pupil using a neural network trained on a biologically accurate eye model expressed in a yellow circle that is finally used to localize the 3D pupil.

Table 1: Comparison of $Pupil_{mean}$ error with the state-of-the-art 3D pupil localization methods, measured in (mm).

Method	S1	S2	S3	S4	S5	Mean	Std
Swirski et al. [26]	3.60	11.14	1.35	4.34	1.45	4.38	3.58
Dierkes et al. [5]	2.93	9.39	1.29	3.46	1.79	3.77	2.91
Ours (single)	**1.47**	8.64	**1.09**	4.18	**1.07**	3.29	2.91
Ours (stereo)	2.44	**2.02**	1.35	**2.45**	1.63	**1.98**	**0.43**

Table 2: Comparison of $Pupil_{std}$ error with the state-of-the-art 3D pupil localization methods, measured in (mm).

Method	S1	S2	S3	S4	S5	Mean	Std
Swirski et al. [26]	2.82	1.34	1.06	2.88	0.89	1.80	0.87
Dierkes et al. [5]	2.28	**1.15**	**0.54**	1.72	0.76	1.29	0.64
Ours (single)	1.09	1.66	0.68	1.69	**0.55**	1.13	0.47
Ours (stereo)	**1.05**	1.82	1.06	**0.73**	0.76	**1.02**	**0.19**

According to the first expectation we have:

$$E_{pupil_{std}} = \frac{1}{8}\sum_{i=1}^{8}(std(\{X_{i,j} - X_{0,j} | j = 1,2,...25\})) \quad (2)$$

where $X_{0,j}$ and $X_{i,j}$ are the 3D pupil positions while staring at j-th target at the reference spot and i-th spot. The second expectation leads to the equation below:

$$E_{pupil_{mean}} = \frac{1}{200}\sum_{i=1}^{8}((\sum_{j=1}^{25}||X_{i,j} - X_{0,j}| - d_i|)) \quad (3)$$

where d_i is the ground truth distance between i-th spot and the reference spot.

Gaze estimation To evaluate the gaze direction, we compute the cosine angle between the estimated and real gaze angle using Equation 1.

6.2 3D pupil localization

We compare our 3D pupil localization method with state-of-the-art works. For a fair comparison, we only compare with glint-free model-based methods as glint-based methods use additional glint information. We don't compare with [9] as they do neither directly provide the 3D pupil center nor the indirect 3D eyeball center. We also utilize the same 2D pupil detection method [15] to get the initial 2D ellipse for all the methods. As for the compared methods, we used the implementation they provided. Table 1 and Table 2 illustrate the advantages of our method in both criteria. We refer to our method using triangulation with stereo cameras as Ours(stereo), the method replacing the original refraction correction in [5] with our 2D refraction correction function as Ours(single). We found that our stereo method can achieve much more stable results (i.e. smaller std) than all other methods with a 47.5% improvement. Our single camera method also performs best in all of the single camera-based methods.

We also list the two types of 3D pupil errors of each subject in our dataset, our results are basically the lowest for most of the users, showing the generalizability and robustness of our method. We found that data of one subject (S2) raises an unusually high error in most of the methods. After checking the data, we observed clear nystagmus during the experiment, worsening the accuracy of the ground truth and leading to bad results in most of the methods.

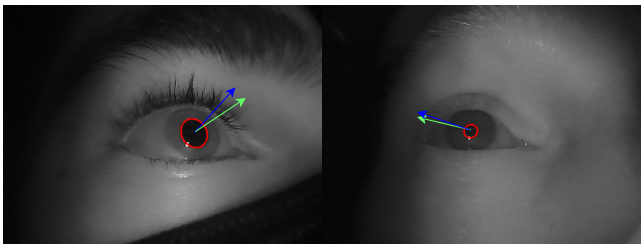

Figure 8: Qualitative results of gaze estimation. The blue arrow represents the estimated gaze direction, while the green arrow represents the ground truth gaze direction. We also highlight the detected 2D pupil contour using a red ellipse.

6.3 Gaze estimation

Qualitative results We first visualize two examples of our gaze estimation results in Fig. 8. The blue and green arrow represents the estimated and the ground truth gaze direction, respectively. We also highlight the 2D pupil using a red ellipse. The left image visualizes an angle error of 3.64°, while the right image describes an angle error of 0.75°.

Quantitative results We also compare our gaze estimation method with the state-of-the-art methods on our proposed dataset as shown in Fig. 9. Compared with other methods, our result achieves a lowest mean error of 4.38°, 18.7% lower than [5], 23.0% lower than [9], and 44.0% lower than [26]. This result also indicates the

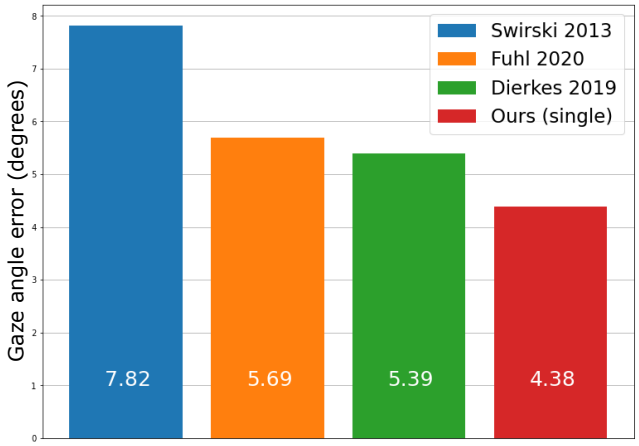

Figure 9: Gaze angular error of different methods for gaze estimation on our dataset in degrees.

Table 3: Comparison of using simple eye model and advanced eye model

Used eye model	Pupil$_{mean}(mm)$	Pupil$_{std}(mm)$	Angle error (°)
Simple	3.51	1.69	5.16
Advanced	3.29	1.13	4.38

correlation between the 3D pupil localization and gaze estimation. Methods that have a lower error on the 3D pupil localization task also perform better in the gaze estimation task.

6.4 Ablation studies

Advanced eye model and simple eye model We conduct additional ablation studies to analyze the effect of the advanced eye model in terms of two 3D pupil evaluation criteria and the gaze angle error as shown in Table 3. In the experiments, we replace the advanced eye model with the simple eye model when generating the synthetic dataset for training the de-refraction method. In all three indicators, the advanced eye model performs better than the simple eye model.

With and without refraction correction Eye refraction is an important factor while doing 3D pupil localization. Here we evaluate the effect of our refraction function added to different methods. Compared to [5], our 2D neural network-based refraction correction has a lower error in the pupil mean criteria and achieves the same accuracy in pupil std criteria, indicating a more accurate 3D pupil localization. Compared to [9] that did not consider refraction correction, with our de-refraction step, their method is able to achieve a smaller gaze angle error, which again illustrates the generality of our 2D-based refraction correction method.

Table 4: Comparison of methods with and without neural network based refraction correction. *w. RC* and *w/o RC* represents with and without our refraction correction, respectively. *w. 3DRC** indicates the use of refraction correction proposed in [5].

Method	Pupil$_{mean}(mm)$	Pupil$_{std}(mm)$	Angle error (°)
[5] w. 3DRC*	3.77	2.91	5.39
[5] w. RC	3.29	2.91	4.38
[9] w/o RC	-	-	6.83
[9] w. RC	-	-	5.69

7 Conclusion

In this paper, we theoretically analyzed the importance of 3D pupil localization on model-based gaze estimation, where 0.5mm 3D pupil error may lead to a 3°angle error. We also proposed the first real dataset including ground truth of gaze direction and relative 3D pupil location, serving as the benchmark of current 3D pupil localization and model-based gaze tracking methods. To obtain the gold standard ground truth of the data, we built a setup with high stability. We also introduced a novel 3D pupil localization method using an advanced eye model and deep learning-based refraction correction. Extensive experiments show that our method achieves a 47.5% higher accuracy in the 3D pupil localization task than the state-of-the-art work and a 18.7% higher accuracy in gaze estimation. We also found that our image-based refraction correction can also improve methods that ignore corneal refraction. We hope that our work can attract more attention to 3D pupil localization and help solve challenges in the field of display technologies. In future research, we plan to capture a bigger dataset with more subjects to evaluate 3D pupil localization results in a more varied environment. Moreover, as our dataset precisely measures the movement of the head relative to the camera, future research could be devoted to the development and evaluation of slippage detection. We will also conduct experiments on holographic displays with a tiny eye box to explore the practical value of our 3D pupil localization method.

References

[1] G. K. Aguirre. A model of the entrance pupil of the human eye. *Scientific reports*, 9(1):1–10, 2019.

[2] G. K. Aguirre. A model of the appearance of the moving human eye. *bioRxiv*, 2021.

[3] J. Chen, Y. Tong, W. Gray, and Q. Ji. A robust 3d eye gaze tracking system using noise reduction. In *Proceedings of the 2008 symposium on Eye tracking research & applications*, pp. 189–196, 2008.

[4] K. Dierkes, M. Kassner, and A. Bulling. A novel approach to single camera, glint-free 3d eye model fitting including corneal refraction. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pp. 1–9, 2018.

[5] K. Dierkes, M. Kassner, and A. Bulling. A fast approach to refraction-aware eye-model fitting and gaze prediction. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, pp. 1–9, 2019.

[6] D. Dunn. Required accuracy of gaze tracking for varifocal displays. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 1838–1842. IEEE, 2019.

[7] Y. Ebisawa. Realtime 3d position detection of human pupil. In *2004 IEEE Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems, 2004.(VCIMS).*, pp. 8–12. IEEE, 2004.

[8] C. Fedtke, F. Manns, and A. Ho. The entrance pupil of the human eye: a three-dimensional model as a function of viewing angle. *Optics express*, 18(21):22364–22376, 2010.

[9] W. Fuhl, H. Gao, and E. Kasneci. Neural networks for optical vector and eye ball parameter estimation. In *ACM Symposium on Eye Tracking Research and Applications*, pp. 1–5, 2020.

[10] W. Fuhl, G. Kasneci, and E. Kasneci. Teyed: Over 20 million real-world eye images with pupil, eyelid, and iris 2d and 3d segmentations, 2d and 3d landmarks, 3d eyeball, gaze vector, and eye movement types. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 367–375. IEEE, 2021.

[11] W. Fuhl, T. Kübler, K. Sippel, W. Rosenstiel, and E. Kasneci. Excuse: Robust pupil detection in real-world scenarios. In *International conference on computer analysis of images and patterns*, pp. 39–51. Springer, 2015.

[12] W. Fuhl, T. Santini, G. Kasneci, and E. Kasneci. Pupilnet: Convolutional neural networks for robust pupil detection. *arXiv preprint arXiv:1601.04902*, 2016.

[13] W. Fuhl, T. C. Santini, T. Kübler, and E. Kasneci. Else: Ellipse selection for robust pupil detection in real-world environments. In *Proceedings*

of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications, pp. 123–130, 2016.

[14] S. J. Garbin, Y. Shen, I. Schuetz, R. Cavin, G. Hughes, and S. S. Talathi. Openeds: Open eye dataset. *arXiv preprint arXiv:1905.03702*, 2019.

[15] M. Kassner, W. Patera, and A. Bulling. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: Adjunct publication*, pp. 1151–1160, 2014.

[16] H. Kaur, S. Jindal, and R. Manduchi. Rethinking model-based gaze estimation. 2022.

[17] J. Kim, M. Stengel, A. Majercik, S. De Mello, D. Dunn, S. Laine, M. McGuire, and D. Luebke. Nvgaze: An anatomically-informed dataset for low-latency, near-eye gaze estimation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–12, 2019.

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

[19] R. Kothari, Z. Yang, C. Kanan, R. Bailey, J. B. Pelz, and G. J. Diaz. Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities. *Scientific reports*, 10(1):1–18, 2020.

[20] C.-C. Lai, S.-W. Shih, and Y.-P. Hung. Hybrid method for 3-d gaze tracking using glint and contour features. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(1):24–37, 2014.

[21] Y. Le Grand. *Light, colour and vision*. Chapman & Hall, 1968.

[22] C. Lu, P. Chakravarthula, Y. Tao, S. Chen, and H. Fuchs. Improved vergence and accommodation via purkinje image tracking with multiple cameras for ar glasses. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 320–331. IEEE, 2020.

[23] B. Luo, J. Shen, Y. Wang, and M. Pantic. The ibug eye segmentation dataset. In *2018 Imperial College Computing Student Workshop (ICCSW 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

[24] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.

[25] B. Petersch and K. Dierkes. Gaze-angle dependency of pupil-size measurements in head-mounted eye tracking. *Behavior Research Methods*, 54(2):763–779, 2022.

[26] L. Swirski and N. Dodgson. A fully-automatic, temporal approach to single camera, glint-free 3d eye model fitting. *Proc. PETMEI*, pp. 1–11, 2013.

[27] A. Villanueva, R. Cabeza, et al. Evaluation of corneal refraction in a model of a gaze tracking system. *IEEE Transactions on Biomedical Engineering*, 55(12):2812–2822, 2008.

[28] A. Villanueva, J. J. Cerrolaza, and R. Cabeza. Geometry issues of gaze estimation. In *Advances in Human Computer Interaction*. IntechOpen, 2008.

[29] K. Wang and Q. Ji. Real time eye gaze tracking with 3d deformable eye-face model. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1003–1011, 2017.

[30] Z. Wu, S. Rajendran, T. van As, J. Zimmermann, V. Badrinarayanan, and A. Rabinovich. Magiceyes: A large scale eye gaze estimation dataset for mixed reality. *arXiv preprint arXiv:2003.08806*, 2020.

[31] H. Yamazoe, A. Utsumi, T. Yonezawa, and S. Abe. Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions. In *Proceedings of the 2008 symposium on Eye tracking research & applications*, pp. 245–250, 2008.