NVGaze: An Anatomically-Informed Dataset for Low-Latency, Near-Eye Gaze Estimation

Joohwan Kim* NVIDIA
Michael Stengel* NVIDIA
Alexander Majercik NVIDIA
Shalini De Mello NVIDIA
David Dunn UNC
Samuli Laine NVIDIA
Morgan McGuire NVIDIA
David Luebke NVIDIA

ABSTRACT

Quality, diversity, and size of training data are critical factors for learning-based gaze estimators. We create two datasets satisfying these criteria for near-eye gaze estimation under infrared illumination: a synthetic dataset using anatomically-informed eye and face models with variations in face shape, gaze direction, pupil and iris, skin tone, and external conditions (2M images at 1280x960), and a real-world dataset collected with 35 subjects (2.5M images at 640x480). Using these datasets we train neural networks performing with sub-millisecond latency. Our gaze estimation network achieves $2.06(\pm 0.44)^\circ$ of accuracy across a wide $30^\circ \times 40^\circ$ field of view on real subjects excluded from training and 0.5° best-case accuracy (across the same FOV) when explicitly trained for one real subject. We also train a pupil localization network which achieves higher robustness than previous methods.

CCS CONCEPTS

• **Computing methodologies** → **Tracking**; *Rendering*; *Neural networks*; • **Human-centered computing**;

KEYWORDS

eye tracking, machine learning, dataset, virtual reality

ACM Reference Format:
Joohwan Kim*, Michael Stengel*, Alexander Majercik, Shalini De Mello, David Dunn, Samuli Laine, Morgan McGuire, and David Luebke. 2019. NVGaze: An Anatomically-Informed Dataset, for Low-Latency, Near-Eye Gaze Estimation. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3290605.3300780

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300780

1 INTRODUCTION

Computer interaction has expanded to include microphone, camera, body tracking-based input and hand-held, windshield, or head-mounted displays–or even systems with no visual display at all. Richer interaction scenarios demand richer input, including a better comprehension of the user through real-time tracking of the user's visual attention for such input and for creating context-aware output.

This new context for interactive computing requires robust gaze estimation and gaze tracking in real time to power applications such as gaze selection [31], attention monitoring [47], gaze communication cues on desktop and in VR [46, 48], active foveated rendering [2, 17, 41], gaze-contingent displays [27], saccadic redirected walking [51], as well as traditional gaze tracking applications for perception research and usability tests in our own human factors community.

Gaze estimation is the process of identifying the line of sight for each eye of a human user at a single instant whereas *gaze tracking* defines the continuous process for following the user's line of sight over time, which typically involves filtering results from individual frames. This paper describes a neural network for gaze estimation that outperforms previous approaches, and presents two novel datasets for training other such networks.

Gaze estimation must run at extremely low latency, in the order of milliseconds, to be useful for real-time interaction [2, 35]. For example, foveated displays and accurate motion blur rendering require the tracking system to return a result faster than the frame duration, or the image can be displayed incorrectly. Ideally, the results should also exhibit less than one degree of error across a wide field of view while being robust to variation in appearance [24]. Commercially available gaze trackers and research systems have recently begun to approach this goal. This work extends previous methods to surpass state-of-the-art results.

As shown by former work, the quality of a neural network-based gaze estimator depends on the combined quality of the training data, training regime, and network structure.

*Joint first authors.
The project page is available at https://sites.google.com/nvidia.com/nvgaze

We enhance the previous state of the art [16, 50, 60, 62] for producing training data by incorporating many additional anatomical features such as pupil constriction shift and line of sight axis correction. We generate a new public dataset of synthetic images that is larger and more realistic than any of the previously available ones. It is also substantially higher in resolution; previous datasets feature images that are typically on the order of 200×200 pixels, whereas ours are 1280×960. We then leverage our dataset under an improved network and training regime to produce an effective gaze estimator when evaluated against real data.

There are two common camera scenarios for gaze estimators: *remote* images captured from a monitor or dashboard-mounted camera, and *near-eye* cameras, which are often intended for use with head-mounted displays. We focus on near-eye image data, an increasingly important use case for augmented and virtual reality headsets. However, we demonstrate the flexibility of our method by successfully training our network on remote image data in a supplemental experiment. Cameras are further divided into *on-axis* and *off-axis* configurations as shown in Fig. 1. We exclusively use on-axis camera configurations in this paper because they are known to provide higher quality data. However, our approach is applicable to any camera configuration. We assume the common head-mounted case of monochrome infrared images under active LED illumination that produces *glints* (corneal reflections) but are not explicitly using the glints for tracking.

We present the following contributions:

- A large, novel dataset of synthetic eye images based on a parametric, anatomically-informed model with variations on face shape, gaze direction, pupil and iris, skin tone, and external conditions. (Sec. 3);
- A large, novel dataset of real eye images matching the on-axis setup of the synthetic ones (Sec. 4);
- An optimized neural network and training regime for gaze and pupil estimation (Sec.5);
- A careful evaluation showing that our estimator achieves higher accuracy and lower latency under real conditions than previous methods (Sec. 5).

Both our real and synthetic images for near-eye gaze tracking with active infrared illumination capture the challenging case of a camera that can slip, transform, or misfocus.

2 RELATED WORK

We focus on recent work directly related to synthetic data and machine learning for gaze estimation. Curious readers can read a detailed up-to-date survey of gaze tracking systems and gaze estimations algorithms found in the work of Kar and Corcoran [24]. Relevant anatomy work with respect to human eyes is covered in Sec. 3.

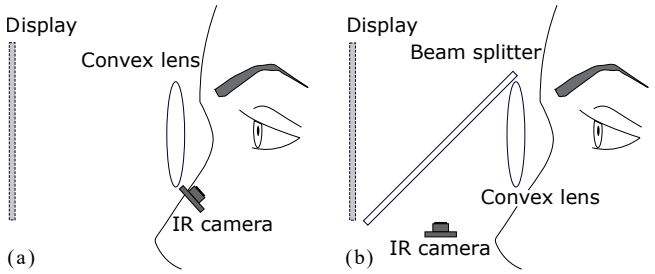

Figure 1: Near-eye display camera configurations. Off- and on-axis placements of gaze tracking cameras inside near-eye displays. (a) The off-axis strategy occupies less space at the cost of accuracy in gaze estimation. (b) The on-axis configuration requires more space but provides frontal view of the eye, which is better for an accurate gaze estimation. Typical locations for display panels in head-mounted displays are denoted by dashed lines.

Eye Rendering and Anatomical Models

Adamo-Villani et al. describe an early simulator for eye motion including eye ball motion and pupil size change [1]. Świrski and Dodgson [52] were the first to apply realistic eye appearance rendering to gaze tracking. They proposed synthetic images for evaluating traditional gaze trackers, whereas the typical approach is to train on synthetic images and validate on real ones. Shrivastava et al. improve the quality of synthetic eye images using a generative adversarial network (GAN) [44]. Our work builds directly on Wood et al.'s SynthesEyes dataset [60], which used a realistic eye model and rendering system for neural network training. We extend their model with additional anatomical detail informed by research on eye glass rendering [30], pupil center shift due to pupil constriction and dilation [58, 59, 63, 64], camera slip/miscalibration, and more sophisticated shading and higher resolution rendering enabled by a modern multi-GPU supercomputer.

Feature-Based Gaze Estimation

Feature-based gaze estimation methods locate the pupil and then map the pupil location to a screen location using user-specific calibration. There are many approaches for locating the pupil. A sampling is discussed in this section.

The Starburst algorithm [33] iteratively locates the pupil center as the mean of points which exceed a differential luminance threshold along the rays extending from the last best guess. In the SET method [21], the convex hull segments of thresholded regions are fit to sinusoidal components. Swirski et al. [52] and Pupil Labs [25] both start with coarse positioning using Haar features. Swirski et al. then refine by k-means clustering the intensity histogram and a modified RANSAC ellipse fit, while Pupil Labs use ellipse fitting on connected edges. ExCuSe [12], ElSe [14] all use morphological edge

filtering followed by ellipse fitting. ExCuSe and ElSe provide alternative approaches for cases when edge detection is not applicable. Fuhl et al. [11] use circular binary features (CBF) to learn conditional distributions of pupil positions for the datasets on which they test. These distributions are indexed by binary feature vectors and looked up at inference time. This approach is further discussed in Sec. 5.

Machine Learning Gaze Estimation

Balujal et al. [4] and Tew et al. [54] were among the first to research combining near-eye images, neural networks, and synthetic images for gaze tracking. Our work also uses machine learning for gaze estimation as it has been shown to be the most promising approach. The state of the art are mostly based on convolutional neural networks, and include results validated on real images as accurate as 10° [60], 9.44° in seconds[1] [61], 7.9° [62], 4.5° in 38 ms [39, 40], 2.6° in 45 ms for remote images with continuous training and calibration [19], $4.8 \pm 0.8°$ [67], and $6.5 \pm 1.5°$ [50]. The lower error rates tend to be after per-subject calibration during validation, training with a mixture of real and synthetic images that contain the subject, or fine-tuning on real data.

Work on pupil tracking using remote-camera systems is often reported in the metrics of percent-correct inferences within a fixed pixel radius with respect to the screen size or the eye tracking camera frame size instead of angular accuracy. Hence, it is not directly comparable, but it is on roughly the same order: 74% accuracy at 5 pixels in 7 ms [13], 89.2%–98.2% accuracy at pupil diameter radius [16], 1.7–2.5 cm on a mobile phone screen (66 ms) [29], and 0.20 mm median error on a mobile phone screen (2 ms) [37].

Our gaze estimation network is an optimization of previous methods as we operate at lower weight precision, without max pooling, and with fewer layers. These types of networks are derived from the VGG16 network topology [45].

Several broad trends appear from the previous work. More realistic synthetic datasets (both in model and rendering) with more images, as well as higher-resolution data in many cases, appear to improve quality [60]. Improved training quality allows to use simpler and thus faster networks. Near-eye input avoids the problems of head pose and eye-region estimation, and allows use of high-resolution images of the eye. Networks with more layers generally outperform shallower ones, and VGG16 is emerging as consensus topology to be wrapped with preprocessing or context-aware layers [68]. Our datasets and estimation method were designed under these considerations. Our results demonstrate that the improved dataset, network, and training we describe can contribute 2–5x better angular accuracy than the state of the art at throughput that is 10–100x faster, even on an embedded processor.

[1]For all cited methods, we provide runtimes where available.

Remote Gaze Estimation and Multi-Camera Systems

We perform a supplemental experiment on remote images, but otherwise focus exclusively on near-eye images in this paper. The most recent related work on remote images covers training across multiple cameras [65], using the screen as a glint source [20], and machine learning for calibrating trackers [42].

Another interesting multi-camera approach is by Tonsen et al. [55], which employs multiple 25-pixel cameras near the eye and trains a tracker for which they report 1.79° accuracy.

Feit et al. [9] describe strategies for accommodating the error in previous trackers, and sources of error for them; the lighting and camera slip variation in our dataset help address this problem by increasing robustness and accuracy of gaze estimation.

Zhang et al. use full-face images and provide a convolutional network architecture that leverages additional information from different facial regions for gaze estimation [67]. Wood et al. use a morphable eye region model with an analysis-by-synthesis approach to extract facial expression and gaze direction simultaneously [61].

Gaze Datasets

Some key publicly-available labelled gaze datasets are: *EyeChimera* [10] RGB images of 40 subjects at 1920×1080 with manual markers; *Columbia Gaze* [46] 5,880 head images of 56 subjects with 320×240 eye regions; *Świrski and Dodgson* [53] 158 synthetic, near-eye IR passive illumination images at 640×480; *EYEDIAP* [15] 16 subjects with eye images 192×168; *UT Multi-view* [50] 64k near-eye images of 50 subjects and 1.1M synthetic images, both at 60×36; *SynthesEyes* [60] 11.4k synthetic near-eye RGB images with passive illumination at 120×80; *GazeCapture* [29] crowd-sourced 2.5M mobile phone images from 1474 subjects; *LPW* [56] 131k near-eye IR images with active illumination of 22 subjects at 640×480; *MPIIGaze* [68] 214k webcam images of 15 subjects with 60×36 eyes; *PupilNet 2.0* [13] 135k IR near-eye images with 384×288 eyes in varying lighting conditions; *BioID* [23] 1521 images of 23 subjects with 32×20 eyes; *InvisibleEye* [55] 280k images of 17 subjects from four 5×5 pixel cameras; *WebGazer* [38] webcam video of 51 subjects with eye images at 640×480.

We contribute two novel datasets with millions of near-eye, IR, active illumination synthetic (2M images at 1280×960) and real (2.5M images at 640×480) images, with continuous variation in gaze direction, region maps, and gaze labels. This greatly expands the available quantity and quality of public gaze data. We also use the PupilNet 2.0 and MPIIGaze data sets in evaluating our estimators.

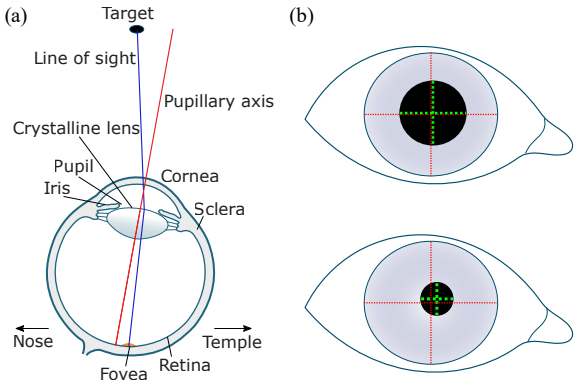

Figure 2: (a) Elements of the eye and axis disparity. (b) Nasal-superior shift under pupil constriction. Red cross-hairs mark iris center; green cross-hairs mark pupil center.

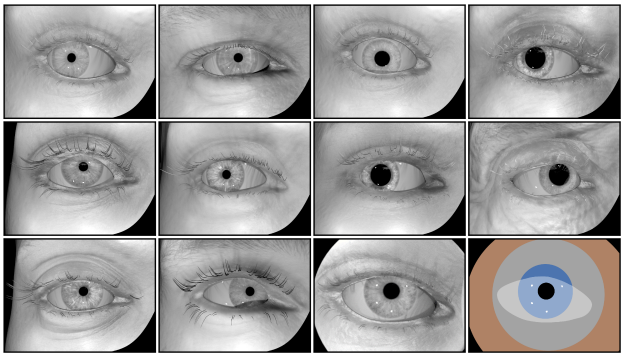

Figure 3: Samples from our synthetic image dataset. The bottom-right image is a composite of the two region maps corresponding to the image on its left, illustrating skin, sclera, visible sclera, iris, pupil, and corneal glints. We augment them in training to vary skin tone, exposure, and environment.

3 SYNTHETIC DATASET

We rendered 2M infrared images of eyes at 1280×960 resolution under active illumination (with 4 simulated IR LEDs) from the view of a virtual, axis-aligned, near-eye gaze tracking camera. Each image is labeled with the exact 2D gaze vector, 3D eye location, 2D pupil location, and a segmentation of pupil, iris, and sclera, skin and glints (corneal reflections) allowing novel training strategies. This is the highest resolution and most diverse such dataset available. Publishing it is one of our main contributions.

Wood et al. [60] previously developed a good synthetic model for pupil tracking under daylight conditions, which includes face shape variation, eye lashes, pupil diameter animation, eyelid motion, and eyeball rotation. To produce even more realistic images with further variation (e.g., Fig. 3), we extended their parametric model with additional anatomical accuracy and detail for infrared lighting conditions as described in this section. Our results of higher accuracy than Wood et al. [62] give evidence that these improvements reduce error during training, as described in Sec. 5.

Due to the level of detail of the model and resolution of the images, each image took about 30 seconds to ray trace on a single GPU with shadows, subsurface scattering, reflection, refraction, and anti-aliasing. It took the equivalent of 3.8 years of single-machine processing time to produce the dataset, using a supercomputer continuously for a week.

Geometry and Animation

We begin with ten geometric models of real human faces (5 females and 5 males of various ages and ethnic groups) generated by 3D scans[2] with manual retouching by Wood et al. [60] to represent a variety of face shapes. We rescaled each head to accommodate a human-average 24 mm-diameter eyeball, giving a more realistic fit than the original work. We inserted the average eyeball, modeling a 7.8 mm radius of curvature at the apex of cornea and approximately 10 mm radius at the boundary with the sclera [34, 54].

For each sample in our synthetic dataset, we displaced the head including the eye by a small, random offset to model the slippage of a head-mounted camera during use. This kind of shift after calibration is a significant and common source of error in gaze trackers [7, 63]. We then chose a randomly selected point of regard on a fixed screen at 1 m from the virtual head. This defined the *line of sight*, which passes through the geometrical center of the eye.

For the selected gaze direction, we modeled the ~5° disparity between the line of sight and pupillary axis of the eye [26, p.74] by rotating the virtual eyeball in the temporal direction (side of the head). We randomly selected eyelid positions ranging from fully open to roughly two-thirds closed [60]. For each position, the top lid covers approximately 4x more eye surface area than the lower lid in order to simulate physically correct eye appearance during a blink [18, 49].

We selected the pupil size from the useful range of 2 mm to 8 mm and modeled the nasal-superior (i.e., towards the forehead above the nose) shift of the pupil under constriction due to illumination [26, p.511]. We used key frames of about 0.1 mm, nasal and superior, for a dilated 8 mm pupil in dim light, about 0.2 mm nasal, 0.1 mm superior for a typical 4 mm pupil, and 0.25 mm nasal, 0.1 mm superior for a constricted 2 mm pupil in bright light (Fig. 2). We allowed the iris texture to randomly rotate around the center of the pupil to provide additional variation of the eye appearance.

[2]Purchased from http://www.3dscanstore.com/.

Materials

The original textures were designed for visible light. We modified the skin and iris textures in both pattern and intensity to match the observed properties of those surfaces under monochromatic (λ = 950 nm) infrared imaging. Accordingly, we modeled air with a unit refractive index and the cornea with an index of refraction $n = 1.38$ [43], yielding the realistic highly-reflective corneal surface on which LED glints appear.

The ten face models provide different skin textures. Although there is much less tonal and pigment variation between individuals in near-infrared wavelengths than in visible light [3, 69], we recommend varying the skin tone using our provided skin region masks to amplify the effective data size as commonly done for neural network training.

Region Maps and Labels

We provide the 2D gaze vector (point of regard on a screen, described as horizontal and vertical gaze angle from a constant reference eye position), head position, eye lid states, and pupil size used to generate each image and the 2D iris center and pupil center (for comparison to older work) in the image. For each sample we produce two exact region maps. The first one identifies skin, pupil, iris, sclera, and LED glints on the cornea. In the second region map, we render the non-skin structures with the face geometry removed, so that pixel-accurate data is provided for the remaining features even when parts of the eye are occluded by eyelids or rest of the face (see Fig. 3, bottom right).

4 REAL-WORLD DATASET

We captured a novel binocular dataset consisting of 1M labelled frames from two high-speed (120 Hz) on-axis near-eye infrared cameras of the eyes of real humans at 640×480 resolution per eye (Fig. 4). The resolution is lower than in our synthetic data due to the limitations of near-eye gaze tracking cameras. This is still a significantly higher resolution than the eye images of previous pupil estimation datasets [13] and two orders of magnitude more pixels per image than previous gaze estimation datasets [68]. This is also the first binocular gaze dataset captured during an acuity task to increase precision.

Environment and Subjects

We captured images from 30 subjects with variation in gender, ethnicity, age, eye shape, and face shape. We induced incidental factors of eyeliner, eyeshadow, mascara, eyeglasses, and contact lenses. These data have comparable active infrared LED characteristics and camera parameters to the synthetic set. For each subject, the data includes varying gaze direction, pupil size (due to ambient visible illumination changes), and infrared illumination (Fig. 4).

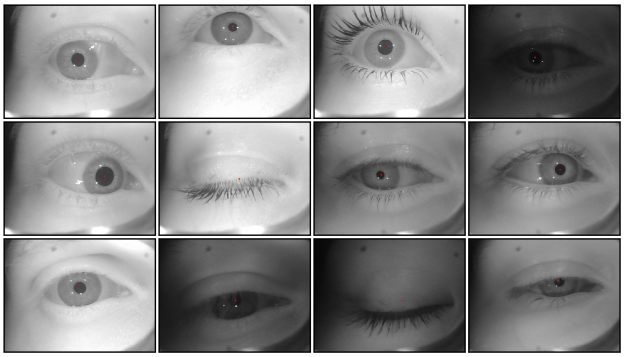

Figure 4: Samples from our real image dataset containing varying pupil size and lighting. The pupil locations estimated by our pupil estimation network are red pixels. The soft dots present in the upper regions of each frame are camera aberrations.

Two hardware setups were used. The first setup emulates the use case of virtual reality headsets with a constant infrared illumination, where we gathered data from 10 subjects. The second setup emulates a more general use case such as augmented reality with changing infrared illumination to cover uncalibrated lighting conditions, where we gathered data from 20 subjects. We randomly decide for constant lighting or vary the infrared LED intensity by using pulse-width modulation and oscillating the intensity between defined min/max values with a sine wave of 1 Hz frequency.

Task and Stimulus

To ensure precise gaze direction labels for the captured images, subjects performed an acuity task during capture, which requires accurate fixation and can reduce occurrence of microsaccades [5, 28]. We placed the subject in a quiet and dimmed office environment, wearing a VR headset with integrated infrared cameras or looking at a computer monitor (27” inches at 53cm distance) with a face stabilizer and mounted cameras.

For each trial, we displayed at a random location on screen a capital letter ‘E’ that subtended 5 arcmin on its long axis, which by definition is the smallest size discernible to a viewer with 20/20 vision, and rotated it to a random multiple of 90° orientation. The subject attempted to identify the orientation, which requires looking directly at the target, and then (without looking) selected an appropriate button. When the subject gave an incorrect response, we rejected the image and ran an additional randomized trial.

When the subject responded correctly, a video recording of 2 seconds duration window began and we instructed the

subject to remain fixated until the target disappeared, providing frames that differ in blink and micro-saccades. Three hundred milliseconds after the video began, we induced further variation in pupil center shift and diameter by ramping screen background intensity (including ambient reflection) from 400 lumens ("white") to 2 lumens ("black").

Labels

Gaze direction was labeled as defined in Sec. 3. For the benefit of future work, we also computed the pupil position and blink labels using the pupil estimator described in Sec. 5 and provide those as additional labels.

5 EXPERIMENTAL RESULTS

We trained neural networks with the proposed dataset and evaluated their performance for practical applications such as gaze estimation and pupil detection. The network architecture that we used was a convolutional neural network motivated by Laine et al. [32], which was optimized for speed and accuracy in performing gaze estimation (see details in the supplemental material).

Evaluation of Proposed Synthetic Dataset

We conducted an ablation study to assess the contribution of our extensions to the original SynthesEyes model of Wood et al. [60] for the case of near-eye gaze estimation under IR lighting. We created 5 synthetic datasets as below. The first two datasets directly compared our dataset and the original SynthesEyes model. For the additional three datasets, we individually removed one of the following features from our model: geometrical correction of eye model, texture adjustment for infrared lighting, and pupil center shift. To evaluate how well a trained network generalizes on a novel subject, we defined *generalization error* as the absolute error between the test labels and inferred values transformed according to a per-subject affine calibration transform, computed between the set of inferred values and the set of test labels. We rendered 16K images across 10 synthetic subjects for each condition, trained gaze networks for them and evaluated them on real data from 9 subjects. We repeated training for each condition 10 times and performed a two-way ANOVA to identify the statistically significant effects.

For the main effects, we observed statistically significant differences between the various training data sets ($p<0.05$), but not between the testing subjects. No interaction between training data sets and testing subjects was found. Furthermore, pairwise comparisons between the different training sets (after Bonferroni correction) revealed that our proposed dataset (with and without the pupil constriction shift, rows 1 and 5 in Table 1) resulted in a significant improvement ($p<0.05$) over the original SynthesEyes model (row 2). Additionally, both our eye model and infrared textures (rows 3 and 4 in Table 1 vs. 1) showed a trend towards improving accuracy with the latter being more significant ($p<0.1$). While further experimentation with more data would help to understand the individual effects more clearly, it is clear that all factors together lead to the improved gaze performance of our synthetic model versus the existing SynthesEyes model in the near-eye infrared setting.

	Dataset	Generalization Error (°)
1	Our model	3.51
2	SynthesEyes model	3.87
3	Our model without geometrical correction of eye	3.62
4	Our model without texture adjustment for IR	3.82
5	Our model without pupil-center shift	3.50

Table 1: Ablation study to assess benefit of proposed synthetic dataset. When trained with our synthetic dataset, the neural network could estimate gaze of unseen, real subjects more accurately. The ablation study suggests that most of the advantage of our synthetic model comes from geometrical correction of the eye model and texture adjustment for the IR lighting condition.

Near-Eye Gaze Estimation

Using our synthetic and real-world VR headset datasets, we evaluated the gaze estimation accuracy of our neural network architecture with 6 convolutional layers, input resolution of 127x127, and 8 channels in the first layer. We chose this network architecture as it resulted in a reasonable compromise between accuracy and computational cost (see supplemental material for more details). We evaluated three training methods: 1) training specifically on data from one real subject and testing on the same subject, 2) training exclusively on data consisting of synthetic images and testing on real subjects, and 3) training on data consisting of both synthetic and real images and testing on real subjects. We achieve remarkable accuracy in all three scenarios.

Training and Testing on One Real Subject. For each subject, a training set consisted of about 5,000 to 7,400 images collected for 45 to 50 gaze directions and varying pupil sizes. The test set consisted of about 1,400 to 1,900 images taken for 11 to 13 gaze directions, which were not present in the training set. The details of the training procedure are in the supplemental material. On average, across all subjects, our network achieved an absolute estimation error of 0.84° with the best-case accuracy being 0.50°.

Training on Synthetic Data and Testing on Real Subjects. The training set consisted of 240k images rendered using 10 synthetic subjects. To effectively increase the size of data, we augmented the training inputs by using region maps; we applied random amounts of blur, intensity modulation, and

contrast modulation to the iris, sclera, and skin regions independently. The test set was all the images acquired from 7 real subjects. We achieved 3.1° generalization error on average across all subjects with the best-case accuracy being 2.3°.

Training on Synthetic and Real Data and Testing on Real Subjects. The training set consisted of all previously used synthetic images and real images from 3 real subjects. We tested on the remaining 7 real subjects (same as in the previous test). We achieved 2.1° generalization error on average, the best accuracy being 1.7°.

Remote Gaze Estimation

We also evaluated the efficacy of our proposed neural network architecture for remote gaze tracking. Note that this is a harder task than near-eye gaze estimation, as low-resolution eye images are captured with a remote camera placed 0.5–1 meters away from the subject, under highly variable ambient lighting conditions and with the presence of the full range of motion of the subject's head. Recently, Park et al. [40] proposed a top-performing method, containing several hourglass networks, for unconstrained eye landmark detection and gaze estimation. They used millions of synthetic eye images generated by the UnityEyes model [62], of size 90×150, to train their network and reported an error of 8.3° on the real-world benchmark MPIIGaze dataset [66] when no images from the MPIIGaze data were used for training or calibration (confirmed via personal communication with the author).

To directly compare the performance of our CNN against their approach, we trained it with one million synthetic images generated from the UnityEyes model and evaluated its performance on all the 45K images from the MPIIGaze dataset. Our network, for this task, was identical to the one that we used in the previous experiment for near-eye gaze estimation, with the exception that we normalized the activations of the first four convolutional layers via instance normalization [8, 57] and used leaky ReLU [36] with $\alpha = 0.1$ instead of ReLU as the non-linearity. We empirically determined these to be useful for stabilizing training and convergence. For training, we used the Adam optimizer with a learning rate of 10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, batch size of 64, and trained for 300 epochs. We also used all the data augmentation steps employed previously by Park et al. [40], except for random image rotations during training. Our network resulted in an error of 8.4°, which is equivalent to that of Park et al., but our network was 100x faster (2000Hz vs. 26Hz theirs'). Considering accuracy and latency together, our network is superior for remote gaze tracking.

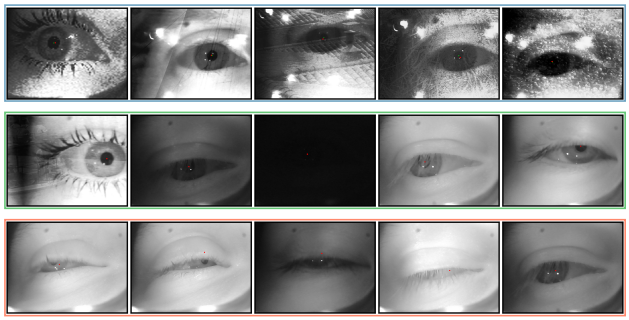

Figure 5: Samples for pupil estimation network. The first row shows augmented images during training. Our network performs well even for challenging samples including bad lighting conditions, dark eye lashes and reflections (second row). Unsuccessful cases due to strong pupil occlusion are shown in the third row.

Pupil Location Estimation

Most existing high-quality video-based gaze tracking systems initially perform pupil estimation in the eye tracking camera frame followed by mapping the pupil position to a screen location with a polynomial calibration function [24]. To compare against such approaches, we trained our network to estimate the pupil center from infrared eye images.

As input we use a subset of 16,000 images of our synthetic dataset containing 1,600 of each head model (Fig. 3) combined with 7,128 images from 3 real subjects from our second real-world dataset (Fig. 4), yielding a synthetic to real image ratio of approximately 2:1. Labels for pupil location are given for our synthetic images whereas initial labels for the real-world data set are computed using the PupilLabs pupil tracker [25] and validated by manual inspection.

Network Architecture and Training. The network architecture is equivalent to the previous experiment, except that we use 7 convolutional layers. To compensate for significant noise in the tested images representing challenging augmented reality conditions, we increased the kernel sizes of the first 4 convolutional layers to 9,7,5,5 and added one additional layer with respect to our baseline architecture in order to increase robustness against image noise, such as reflections and bad lighting conditions. This slightly enlarged network can be still evaluated very quickly on the GPU (see next section). In comparison to the 6-layer network used for gaze estimation, the slightly bigger 7-layer network performs more robust, particularly in the case of strong reflections covering the eye.

We again use 2×2 stride at each convolutional layer, add dropout layers after each convolutional layer, and apply no padding or pooling. The input image size is 293×293 pixels. Because we are not estimating the line of sight with this network, no per-subject post-process transformation is applied after the fully-connected final layer. During training

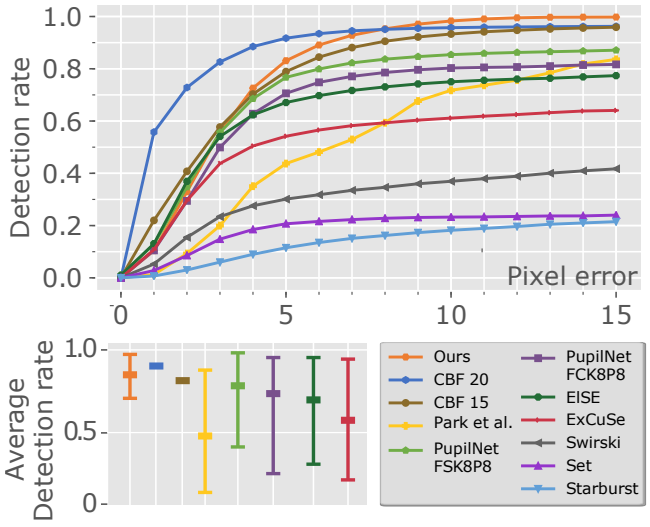

Figure 6: Average pupil estimation error on PupilNet datasets. Top: We compare the average detection rate of our pupil estimation network against Starburst [33], Set [21], Swirski et al. [52], ExCuSe [12], ElSe [14], PupilNet [13], Park et al. [40] and CBF [11]. Bottom: The 5-pixel error is averaged across individual PupilNet datasets (bold marker) and bounded by best and worst error values for all datasets (upper and lower markers). Our approach reaches highest robustness. Note that for CBF only the average detection rates over all datasets were published, not the detection rate for individual datsets [11].

we always rescale the image to the network input resolution using bicubic filtering. We then perform various augmentation steps during training as we did for the gaze estimation network, making a subset of our synthetic data sufficient for convergence in training of the network. Specifically, we randomize image samples using affine image transformation, pixel-wise intensity noise, global intensity offset, Gaussian filtering, image shrinking followed by upscaling, histogram equalization, and normalization with mean shift. We also simulate environment reflections in the eye by randomly overlaying the image with images out of a dataset of 326 natural photographs [22]. For details about the training procedure we refer the reader to the supplementary material. Augmented images are shown in Fig. 5 (first row).

Pupil Estimation Accuracy. Accuracy of pupil estimation is usually given in form of a probability of estimating the pupil location with a maximum distance of 5 pixels from the ground truth pupil location (”5-pixel error” or ”detection rate”) [13]. When training on 10 synthetic and 3 real subjects we reach very high pupil estimation accuracy across all remaining subjects of our second real-world dataset. Fig. 5 shows that our network is able to estimate the pupil center even for very challenging cases such as bad lighting conditions, dark eye lashes, partly occluded pupils, and reflections. Pupil estimation during blinks and in case of other strong occlusions are ill-posed problems and typically result in higher estimation error. However, since we include pupil labels even for occluded pupils in our synthetic data we effectively limit the offset from the ground truth pupil location during a blink (Fig. 5, third row). Note that our pupil localization network is trained on on-axis images and performs well for this camera configuration (Fig. 1, on-axis). For other camera configuration, respective images must be included in training or the network will perform suboptimal.

Recent pupil estimation methods [12–14, 21, 33, 40, 52] have been evaluated on the PupilNet dataset containing 29 individual datasets with 135,000 frames each with different challenges such as different geometric configurations, strong environment reflections, camera noise, difficult lighting, and even incomplete frames [13]. Following Fuhl et al., we trained with images from the PupilNet datasets in addition to our synthetic dataset excluding images from the PupilNet dataset that we use for validation. For a fair comparison to values given in previous papers we compute results with respect to the native dataset resolution of 384x288. This means our 293x293 network has to deliver sub-5-pixel accuracy. For Park et al. we test on 180x108 cropped images centered on the pupil location given by the label, allowing their network to effectively work on the full native resolution while making sure that the pupil is still contained.

We reach a 5-pixel error of 83.1% which is significantly superior to other CNN-based approaches such as PupilNet.v2 with 76.7% and Park et al. 43.7% and higher than ExCuse 67.1%, ElSe 54.2% (see Fig. 6, top). We reason that our network architecture in combination with image augmentation helps significantly to increase robustness against noise visible in challenging real-world images. In Fig. 6 (bottom) we plot the best case and worst case value over all individual datasets of PupilNet. The graph shows that our network reaches consistently high robustness with low variance from dataset to dataset (69.1% worst case, 96.3% best case). Using the network of Park et al. overall shows a much higher variance and lower performance across the PupilNet dataset. These results are worse in comparison to numbers reported in their paper when testing on the MPIIGaze real-world dataset [40]. We reason that Park et al. do not augment with random reflections during training which significantly lowers the detection rate on the PupilNet dataset.

CBF-20 exceeds our performance on the PupilNet dataset below 8 pixels of error and CBF-15 below 3 pixels. However, our trained model requires only 8 MB of memory whereas the CBF models consumes 3 orders of magnitude more (3 GB and 9.5 GB respectively). Therefore, CBF may be used in the case where best case accuracy with high memory consumption is an acceptable tradeoff over robust worst-case performance and low memory footprint.

Real-Time Performance

We implemented our trained network in cuDNN [6], a framework of optimized GPU kernels for deep learning built on NVIDIA CUDA. We tested inference times for different networks on desktop and mobile class GPUs as shown in Table 2. The Near-Eye Gaze Estimation and Pupil Localization networks refer to the networks described earlier in this section. The times reported are averages for a single frame over 100,000 inferences using 16 bit floating point (half) precision.[3]

On NVIDIA Titan V, our networks run at well over 1,000 Hz. On Jetson TX2, our Gaze Estimation network again runs at over 1,000 Hz, while our slightly larger Pupil Location network achieves over 260 Hz.

Network	Titan V	Jetson TX2
Gaze Estimation	0.496 ms	0.659 ms
Pupil Localization	0.914 ms	3.781 ms

Table 2: Inference performance on tested hardware.

6 DISCUSSION

Accurate synthetic data is essential for training machine learning systems within practical resource limits. Our novel synthetic dataset is accurate and comprehensive for the implemented model components. It enables training for information which is hard to obtain and control in the real world, and our novel real image dataset improves accuracy and provides real-world validation. We demonstrated the effectiveness of these datasets in contributing to one of the best-performing gaze estimation networks, and have shown that adaptation to new hardware configurations is simple, fast, and robust.

Our eye model does not include eyeball elongation common in myopic eyes, the complicated optical elements behind the pupil such as the crystalline lens, rotational movements of the eyeball according to Listing's Law, or the movement of fluid within the eye during gaze changes. We did not model these because we hypothesized that they have milder impact on gaze and pupil estimation compared to what we incorporated in our dataset. Having addressed the challenging larger sources of error by our improved anatomical model and rendering shaders, the previously mentioned smaller sources of error are now good candidates for follow-up study.

We include the region maps to promote future research with region-wise augmentation of our synthetic dataset. For example, extended iris texture, eye lash variation, alternative environment reflections, additional physiological structures in the sclera, alternative camera lens distortion and vignetting properties, and more diverse makeup application can now be explored as 2D imaging operations during training without the immense computing power required to path trace millions of high-resolution images from 3D models.

Including head slippage in training data is essential for robust and accurate gaze estimation. Our approach was to randomize head positions, thereby covering a space encompassing typical head positions encountered when using a headset. This strategy is simple to implement but can possibly include head positions never encountered in real use cases. While this approach may generalize better for unseen subjects, a more realistic slippage modeling based on measurements could enhance accuracy even further for specific real-world scenarios. We hope to explore this approach in future work.

Robust, accurate gaze tracking enables novel gaze-based HCI methodologies, particularly in VR. The methodologies that have been explored for VR are limited by the accuracy of gaze trackers, leading to approximations of both the VR headset[2][3] and the gaze tracker[4]. Even those setups using a real VR headset with a real gaze tracker cite the accuracy of the tracker as a confounding factor in their experiments[6][7]. Methods that overcome gaze tracker inaccuracy by "snapping to locations" or other approximations are confounded by nearly-overlapping objects, limiting test scene complexity, and requiring heuristics to separate items[5]. Clearly, gaze tracker accuracy and robustness is a limitation in HCI research on gaze-based interaction.

Our presented network, training technique, and datasets compose a method for training a robust, accurate gaze tracker for arbitrary head-mounted setups. Previously, experiments were limited to the accuracy achievable by off-the-shelf trackers which, though capable of high accuracy in the ideal case, do not achieve robust accuracy for all experimental setups, let alone all experiment participants. Our result provides the best of both worlds: experimenters can use our robust, pretrained gaze tracking network, or follow our method to train their own using the datasets and eye model that we publish with the paper. Based on our experience constructing our network and dataset, we make the following recommendations to experimenters in the HCI community seeking to implement our technique:

- Despite their simplicity, stacked convolutional network architectures are very accurate once trained to convergence, provide low-latency inference on modern GPUs, and are easy to implement. They should be preferred models for VR gaze tracking setups. We provide analysis of training parameters (number of convolutional layers, feature counts, etc.) in the supplementary material to assist researchers in creating stacked convolutional networks that fit their experimental setups.

[3]For both networks, we verified that inference accuracy is identical (to 1/1,000th of a degree/pixel) for 32 and 16 bit floating point precision.

- The most important physical properties of synthetic eye models are accurate representation of anatomical structure and reflectance in the given lighting condition, affecting size and brightness of features in images. Experimenters should take this into account both when using synthetic data and when using real data/training on live participants.
- The most important hardware setup properties to simulate are lighting condition and camera parameters (view, sensor properties, exposure, noise). Experimenters should take this into account when synthesizing images and when designing experiments.

Finally, though many gaze-based interaction methodologies are enabled by our approach, we find blink-based interactions to be of particular interest [10]. A robust blink detection technique combined with our gaze tracking network would enable robust and precise blink interaction, allowing researchers to differentiate blink events (e.g. voluntary vs. involutary blinks), thus enlarging the space of possible interaction techniques.

7 CONCLUSION

We have presented 1) a robust, accurate gaze estimation network, 2) a general method for training image-based gaze estimators from custom hardware setups, and 3) the NVGaze datasets containing millions of real and synthetic images of high quality augmented with an eye model and rendering pipeline to create highly realistic eye images. Our network achieves state-of-the-art accuracy for gaze estimation and pupil detection and is more robust (in terms of worst case performance) than all previous methods. Our approach is easily adaptable to arbitrary hardware configurations, and we include recommendations for training based on our experience implementing our presented network. We share dataset, our eye model, and rendering and animation code with the community to allow researchers to easily render synthetic data specific to their hardware setup. Our network, method, and datasets constitute a significant advance in the state of the art for image-based head-mounted gaze tracking, enabling numerous opportunities for research in gaze-based rendering and HCI. With this work we hope to make an impetus for novel research topics covering gaze interaction, visual perception, gaze-contingent displays and gaze-based rendering.

8 ACKNOWLEDGEMENTS

We thank all reviewers for their valuable feedback. We also thank Eric Whitmire for improving the animation scripts of our models. We thank Peter Shirley and Pavlo Molchanov for their insightful comments and suggestions. We thank Erroll Wood for providing the Blender eye model from the SynthesEyes dataset [60].

REFERENCES

[1] Nicoletta Adamo-Villani, Gerardo Beni, and Jeremy White. 2005. EMOES: Eye Motion and Ocular Expression Simulator. *International Journal of Information Technology* 2, 3 (2005), 170–176.

[2] Rachel Albert, Anjul Patney, David Luebke, and Joohwan Kim. 2017. Latency Requirements for Foveated Rendering in Virtual Reality. *ACM Trans. Appl. Percept.* 14, 4, Article 25 (Sept. 2017), 13 pages. https://doi.org/10.1145/3127589

[3] Elli Angelopoulou. 1999. *The Reflectance Spectrum of Human Skin*. Technical Report MS-CIS-99-29. University of Pennsylvania. 16 pages.

[4] Shumeet Baluja and Dean Pomerleau. 1994. Non-Intrusive Gaze Tracking Using Artificial Neural Networks. In *Advances in Neural Information Processing Systems 6*, J. D. Cowan, G. Tesauro, and J. Alspector (Eds.). Morgan-Kaufmann, San Francisco, CA, USA, 753–760.

[5] Bruce Bridgeman and Joseph Palca. 1980. The role of microsaccades in high acuity observational tasks. *Vision Research* 20, 9 (1980), 813–817.

[6] Sharan Chetlur, Cliff Woolley, Philippe Vandermersch, Jonathan Cohen, John Tran, Bryan Catanzaro, and Evan Shelhamer. 2014. cuDNN: Efficient primitives for deep learning. *CoRR* abs/1410.0759 (2014).

[7] Kyoung Whan Choe, Randolph Blake, and Sang-Hun Lee. 2016. Pupil size dynamics during fixation impact the accuracy and precision of video-based gaze estimation. *Vision Research* 118 (2016), 48–59.

[8] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. 2016. A Learned Representation For Artistic Style. *CoRR* abs/1610.07629. arXiv:1610.07629 http://arxiv.org/abs/1610.07629

[9] Anna Maria Feit, Shane Williams, Arturo Toledo, Ann Paradiso, Harish Kulkarni, Shaun Kane, and Meredith Ringel Morris. 2017. Toward Everyday Gaze Input: Accuracy and Precision of Eye Tracking and Implications for Design. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 1118–1130. https://doi.org/10.1145/3025453.3025599

[10] Laura Florea, Corneliu Florea, Ruxandra Vranceanu, and Constantin Vertan. 2013. Can Your Eyes Tell Me How You Think? A Gaze Directed Estimation of the Mental Activity. In *BMVC 2013 - Electronic Proceedings of the British Machine Vision Conference 2013*. BMVA Press, 60.1–60.11.

[11] Wolfgang Fuhl, David Geisler, Thiago Santini, Tobias Appel, Wolfgang Rosenstiel, and Enkelejda Kasneci. 2018. CBF: Circular Binary Features for Robust and Real-time Pupil Center Detection. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications (ETRA '18)*. ACM, New York, NY, USA, Article 8, 6 pages. https://doi.org/10.1145/3204493.3204559

[12] Wolfgang Fuhl, Thomas Kübler, Katrin Sippel, Wolfgang Rosenstiel, and Enkelejda Kasneci. 2015. ExCuSe: Robust Pupil Detection in Real-World Scenarios. In *Computer Analysis of Images and Patterns*, George Azzopardi and Nicolai Petkov (Eds.). Springer International Publishing, Cham, 39–51.

[13] Wolfgang Fuhl, Thiago Santini, Gjergji Kasneci, Wolfgang Rosenstiel, and Enkelejda Kasneci. 2017. PupilNet v2.0: Convolutional Neural Networks for CPU based real time Robust Pupil Detection. *CoRR* abs/1711.00112 (2017). http://arxiv.org/abs/1711.00112

[14] Wolfgang Fuhl, Thiago C. Santini, Thomas Kübler, and Enkelejda Kasneci. 2016. ElSe: Ellipse Selection for Robust Pupil Detection in Real-world Environments. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications (ETRA '16)*. ACM, New York, NY, USA, 123–130. https://doi.org/10.1145/2857491.2857505

[15] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. 2014. 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 (ETRA '14)*. ACM, New York, NY, USA, 255–258. https://doi.org/10.1145/2578153.2578190

[16] Chao Gou, Y. Wu, Kang Wang, Fei-Yue Wang, and Q. Ji. 2016. Learning-by-synthesis for accurate eye detection. *2016 23rd International Conference on Pattern Recognition (ICPR)* 1, 1 (Dec 2016), 3362–3367. https://doi.org/10.1109/ICPR.2016.7900153

[17] Brian Guenter, Mark Finch, Steven Drucker, Desney Tan, and John Snyder. 2012. Foveated 3D Graphics. *ACM Trans. Graph.* 31, 6, Article 164 (Nov. 2012), 10 pages. https://doi.org/10.1145/2366145.2366183

[18] Michael J Hawes and Richard K Dortzbach. 1982. The microscopic anatomy of the lower eyelid retractors. *Archives of ophthalmology* 100, 8 (1982), 1313–1318.

[19] Michael Xuelin Huang, Tiffany C.K. Kwok, Grace Ngai, Stephen C.F. Chan, and Hong Va Leong. 2016. Building a Personalized, Auto-Calibrating Eye Tracker from User Interactions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 5169–5179. https://doi.org/10.1145/2858036.2858404

[20] Michael Xuelin Huang, Jiajia Li, Grace Ngai, and Hong Va Leong. 2017. ScreenGlint: Practical, In-situ Gaze Estimation on Smartphones. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 2546–2557. https://doi.org/10.1145/3025453.3025794

[21] Amir-Homayoun Javadi, Zahra Hakimi, Morteza Barati, Vincent Walsh, and Lili Tcheang. 2015. SET: A Pupil Detection Method using Sinusoidal Approximation. *Frontiers in Neuroengineering* 8 (2015), 4. https://doi.org/10.3389/fneng.2015.00004

[22] Herve Jegou, Matthijs Douze, and Cordelia Schmid. 2008. Hamming embedding and weak geometric consistency for large scale image search. In *European Conference on Computer Vision (ECCV)*. Springer, Springer International Publishing, 304–317.

[23] Oliver Jesorsky, Klaus J. Kirchberg, and Robert Frischholz. 2001. Robust Face Detection Using the Hausdorff Distance. In *Proceedings of the Third International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA '01)*. Springer International Publishing, Berlin, Heidelberg, 90–95. https://www.bioid.com/facedb/

[24] Anuradha Kar and Peter Corcoran. 2017. A Review and Analysis of Eye-Gaze Estimation Systems, Algorithms and Performance Evaluation Methods in Consumer Platforms. *CoRR* abs/1708.01817 (2017). arXiv:1708.01817 http://arxiv.org/abs/1708.01817

[25] Moritz Kassner, William Patera, and Andreas Bulling. 2014. 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 (UbiComp '14 Adjunct)*. ACM, New York, NY, USA, 1151–1160. https://doi.org/10.1145/2638728.2641695

[26] Paul L Kaufman, Leonard A Levin, Francis Heed Adler, and Albert Alm. 2011. *Adler's Physiology of the Eye*. Elsevier Health Sciences, St. Louis, MO.

[27] Robert Konrad, Nitish Padmanaban, Keenan Molner, Emily A. Cooper, and Gordon Wetzstein. 2017. Accommodation-invariant Computational Near-eye Displays. *ACM Trans. Graph.* 36, 4, Article 88 (July 2017), 12 pages. https://doi.org/10.1145/3072959.3073594

[28] Elleen Kowler and Robert M Steinman. 1979. Miniature saccades: eye movements that do not count. *Vision Research* 19, 1 (1979), 105–108.

[29] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba. 2016. Eye Tracking for Everyone. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 1. 2176–2184. https://doi.org/10.1109/CVPR.2016.239

[30] Thomas C. Kübler, Tobias Rittig, Enkelejda Kasneci, Judith Ungewiss, and Christina Krauss. 2016. Rendering Refraction and Reflection of Eyeglasses for Synthetic Eye Tracker Images. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications (ETRA '16)*. ACM, New York, NY, USA, 143–146. https://doi.org/10.1145/2857491.2857494

[31] Mikko Kytö, Barrett Ens, Thammathip Piumsomboon, Gun A. Lee, and Mark Billinghurst. 2018. Pinpointing: Precise Head- and Eye-Based Target Selection for Augmented Reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 81, 14 pages. https://doi.org/10.1145/3173574.3173655

[32] Samuli Laine, Tero Karras, Timo Aila, Antti Herva, Shunsuke Saito, Ronald Yu, Hao Li, and Jaakko Lehtinen. 2017. Production-Level Facial Performance Capture Using Deep Convolutional Neural Networks. *Proc. Symposium on Computer Animation (SCA)*.

[33] Dongheng Li, D. Winfield, and D. J. Parkhurst. 2005. Starburst: A hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, Vol. 1. 79–79. https://doi.org/10.1109/CVPR.2005.531

[34] Hwey-Lan Liou and Noel A Brennan. 1997. Anatomically accurate, finite model eye for optical modeling. *Journal of the Optical Society of America* 14, 8 (1997), 1684–1695.

[35] Lester C Loschky and George W McConkie. 2002. Investigating spatial vision and dynamic attentional selection using a gaze-contingent multiresolutional display. *Journal of Experimental Psychology: Applied* 8, 2 (2002), 99.

[36] Andrew L Maas, Awni Y Hannun, and Andrew Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. International Conference on Machine Learning (ICML)*, Vol. 30.

[37] Alex Mariakakis, Jacob Baudin, Eric Whitmire, Vardhman Mehta, Megan A Banks, Anthony Law, Lynn Mcgrath, and Shwetak N Patel. 2017. PupilScreen: Using Smartphones to Assess Traumatic Brain Injury. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 81.

[38] Alexandra Papoutsaki, Aaron Gokaslan, James Tompkin, Yuze He, and Jeff Huang. 2018. The Eye of the Typer: A Benchmark and Analysis of Gaze Behavior During Typing. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications (ETRA '18)*. ACM, New York, NY, USA, Article 16, 9 pages. https://doi.org/10.1145/3204493.3204552

[39] Seonwook Park, Adrian Spurr, and Otmar Hilliges. 2018. Deep Pictorial Gaze Estimation. *European Conference on Computer Vision (ECCV)* 16, 1 (2018), 741–757.

[40] Seonwook Park, Xucong Zhang, Andreas Bulling, and Otmar Hilliges. 2018. Learning to Find Eye Region Landmarks for Remote Gaze Estimation in Unconstrained Settings. *ACM Symposium on Eye Tracking Research and Applications (ETRA)* (2018).

[41] Anjul Patney, Marco Salvi, Joohwan Kim, Anton Kaplanyan, Chris Wyman, Nir Benty, David Luebke, and Aaron Lefohn. 2016. Towards Foveated Rendering for Gaze-tracked Virtual Reality. *ACM Trans. Graph.* 35, 6, Article 179 (Nov. 2016), 12 pages. https://doi.org/10.1145/2980179.2980246

[42] Thiago Santini, Wolfgang Fuhl, and Enkelejda Kasneci. 2017. CalibMe: Fast and Unsupervised Eye Tracker Calibration for Gaze-Based Pervasive Human-Computer Interaction. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 2594–2605. https://doi.org/10.1145/3025453.3025950

[43] Dhiraj K Sardar, Guang-Yin Swanland, Raylon M Yow, Robert J Thomas, and Andrew TC Tsin. 2007. Optical properties of ocular tissues in the near infrared region. *Lasers in medical science* 22, 1 (2007), 46–52.

[44] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. 2017. Learning from Simulated and Unsupervised Images through Adversarial Training. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 2. 2242–2251. https://doi.org/10.1109/CVPR.2017.241

[45] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014). arXiv:1409.1556 http://arxiv.org/abs/1409.1556

[46] Brian A. Smith, Qi Yin, Steven K. Feiner, and Shree K. Nayar. 2013. Gaze Locking: Passive Eye Contact Detection for Human-object Interaction. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology (UIST '13)*. ACM, New York, NY, USA, 271–280. https://doi.org/10.1145/2501988.2501994

[47] P. Smith, M. Shah, and N. da Vitoria Lobo. 2003. Determining Driver Visual Attention with One Camera. *Trans. Intell. Transport. Sys.* 4, 4 (Dec. 2003), 205–218. https://doi.org/10.1109/TITS.2003.821342

[48] Michael Stengel, Steve Grogorick, Martin Eisemann, Elmar Eisemann, and Marcus A. Magnor. 2015. An Affordable Solution for Binocular Eye Tracking and Calibration in Head-mounted Displays. In *Proceedings of the 23rd ACM International Conference on Multimedia (MM '15)*. ACM, New York, NY, USA, 15–24. https://doi.org/10.1145/2733373.2806265

[49] William Steptoe, Oyewole Oyekoya, and Anthony Steed. 2010. Eyelid Kinematics for Virtual Characters. *Computer Animation and Virtual Worlds* 21, 3-4 (2010), 161–171.

[50] Y. Sugano, Y. Matsushita, and Y. Sato. 2014. Learning-by-Synthesis for Appearance-Based 3D Gaze Estimation. *2014 IEEE Conference on Computer Vision and Pattern Recognition* 1, 1 (June 2014), 1821–1828. https://doi.org/10.1109/CVPR.2014.235

[51] Qi Sun, Anjul Patney, Li-Yi Wei, Omer Shapira, Jingwan Lu, Paul Asente, Suwen Zhu, Morgan Mcguire, David Luebke, and Arie Kaufman. 2018. Towards Virtual Reality Infinite Walking: Dynamic Saccadic Redirection. *ACM Trans. Graph.* 37, 4, Article 67 (July 2018), 13 pages. https://doi.org/10.1145/3197517.3201294

[52] Lech Świrski and Neil Dodgson. 2014. Rendering Synthetic Ground Truth Images for Eye Tracker Evaluation. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '14)*. ACM, New York, NY, USA, 219–222. https://doi.org/10.1145/2578153.2578188

[53] Lech Świrski and Neil A. Dodgson. 2013. A fully-automatic, temporal approach to single camera, glint-free 3D eye model fitting. In *Proceedings of ECEM 2013*. http://www.cl.cam.ac.uk/research/rainbow/projects/eyemodelfit/

[54] A. I. Tew. 1997. Simulation results for an innovative point-of-regard sensor using neural networks. *Neural Computing & Applications* 5, 4 (01 Dec 1997), 230–237. https://doi.org/10.1007/BF01424228

[55] Marc Tonsen, Julian Steil, Yusuke Sugano, and Andreas Bulling. 2017. InvisibleEye: Mobile Eye Tracking Using Multiple Low-Resolution Cameras and Learning-Based Gaze Estimation. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 106 (Sept. 2017), 21 pages. https://doi.org/10.1145/3130971

[56] Marc Tonsen, Xucong Zhang, Yusuke Sugano, and Andreas Bulling. 2016. Labelled Pupils in the Wild: A Dataset for Studying Pupil Detection in Unconstrained Environments. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications (ETRA '16)*. ACM, New York, NY, USA, 139–142. https://doi.org/10.1145/2857491.2857520

[57] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. 2017. Improved Texture Networks: Maximizing Quality and Diversity in Feed-Forward Stylization and Texture Synthesis. *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 4105–4113. https://doi.org/10.1109/CVPR.2017.437

[58] Glyn Walsh. 1988. The effect of mydriasis on pupillary centration of the human eye. *Ophthalmic Physiol Opt.* 8 (02 1988), 178–82.

[59] Ulrich Wildenmann and Frank Schaeffel. 2013. Variations of pupil centration and their effects on video eye tracking. *Ophthalmic and Physiological Optics* 34, 1 (09 2013). https://doi.org/10.1111/opo.12102

[60] Erroll Wood, Tadas Baltruaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling. 2015. Rendering of Eyes for Eye-Shape Registration and Gaze Estimation. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV) (ICCV '15)*. IEEE Computer Society, Washington, DC, USA, 3756–3764. https://doi.org/10.1109/ICCV.2015.428

[61] Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. 2016. A 3D Morphable Eye Region Model for Gaze Estimation. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 297–313.

[62] Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. 2016. Learning an Appearance-based Gaze Estimator from One Million Synthesised Images. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications (ETRA '16)*. ACM, New York, NY, USA, 131–138. https://doi.org/10.1145/2857491.2857492

[63] Harry J. Wyatt. 2010. The human pupil and the use of video-based eyetrackers. *Vision Research* 50, 10 (2010), 1982–1988. https://doi.org/10.1016/j.visres.2010.07.008

[64] Yabo Yang, Keith Thompson, and Stephen Burns. 2002. Pupil Location under Mesopic, Photopic, and Pharmacologically Dilated Conditions. *Invest Ophthalmol Vis Sci.* 43 (08 2002), 2508–12.

[65] Xucong Zhang, Michael Xuelin Huang, Yusuke Sugano, and Andreas Bulling. 2018. Training Person-Specific Gaze Estimators from User Interactions with Multiple Devices. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 624, 12 pages. https://doi.org/10.1145/3173574.3174198

[66] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2015. Appearance-based gaze estimation in the wild. 1, 1 (June 2015), 4511–4520. https://doi.org/10.1109/CVPR.2015.7299081

[67] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2016. It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation. *CoRR* abs/1611.08860. arXiv:1611.08860 http://arxiv.org/abs/1611.08860

[68] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2017. MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation. *CoRR* abs/1711.09017 (2017). arXiv:1711.09017 http://arxiv.org/abs/1711.09017

[69] George Zonios and Aikaterini Dimou. 2009. Light scattering spectroscopy of human skin in vivo. *Opt. Express* 17, 3 (Feb 2009), 1256–1267. https://doi.org/10.1364/OE.17.001256