Dataset for Eye Tracking on a Virtual Reality Platform

Stephan J. Garbin
University College London
UK

Yiru Shen
Facebook Reality Labs
USA

Immo Schuetz
Facebook Reality Labs
USA

Robert Cavin
Facebook Reality Labs
USA

Gregory Hughes
Google via Adecco
USA

Oleg V. Komogortsev
Facebook Reality Labs
USA

Sachin S. Talathi
Facebook Reality Labs
USA

ABSTRACT

We present a large scale data set of eye-images captured using a virtual-reality (VR) head mounted display mounted with two synchronized eye-facing cameras at a frame rate of 200 Hz under controlled illumination. This dataset is compiled from video capture of the eye-region collected from 152 individual participants and is divided into four subsets: (i) 12,759 images with pixel-level annotations for key eye-regions: iris, pupil and sclera (ii) 252,690 unlabeled eye-images, (iii) 91,200 frames from randomly selected video sequences of 1.5 seconds in duration, and (iv) 143 pairs of left and right point cloud data compiled from corneal topography of eye regions collected from a subset, 143 out of 152, participants in the study. A baseline experiment has been evaluated on the dataset for the task of semantic segmentation of pupil, iris, sclera and background, with the mean intersection-over-union (mIoU) of 98.3 %. We anticipate that this dataset will create opportunities to researchers in the eye tracking community and the broader machine learning and computer vision community to advance the state of eye-tracking for VR applications, which in its turn will have greater implications in Human-Computer Interaction.

CCS CONCEPTS

• **Computing methodologies** → **Image segmentation**; *Reconstruction*; *Neural networks*.

KEYWORDS

eye tracking, appearance-based eye tracking, segmentation, iris, sclera, pupil, virtual reality

ACM Reference Format:
Stephan J. Garbin, Yiru Shen, Immo Schuetz, Robert Cavin, Gregory Hughes, Oleg V. Komogortsev, and Sachin S. Talathi. 2020. Dataset for Eye Tracking on a Virtual Reality Platform. In *ETRA '20: 2020 Symposium on Eye Tracking Research and Applications, June 02–05, 2020, Stuttgart, Germany.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/1122445.1122456

ETRA '20, June 02–05, 2020, Stuttgart, Germany

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Understanding the motion and appearance of the human eye is of great importance to many scientific fields [Holmqvist et al. 2011]. For example, gaze direction can offer information about the focus of a person's attention [Borji and Itti 2013], which in turn can facilitate the research on eye tracking to aid in the study and design of how humans interact with their environment [Smith et al. 2013].

In the context of virtual reality (VR), accurate and precise eye tracking can enable game-changing technological advances, for example, foveated rendering, a technique that exploits the sensitivity profile of the human eye to render only those parts of a virtual scene at full resolution that the user is focused on, can significantly alleviate the computational burden of VR [Patney et al. 2016]. Head-mounted displays (HMDs) that include gaze-contingent variable focus [Kramida 2016] and gaze-driven rendering of perceptually accurate focal blur [Xiao et al. 2018] promise to alleviate vergence-accommodation-conflict and increase visual realism. Finally, gaze-driven interaction schemes could enable novel methods of navigating and interacting with virtual environments [Tanriverdi and Jacob 2000].

The success of data driven machine learning models learned directly from images ([He et al. 2015; Krizhevsky et al. 2012]) is accompanied by the demand for datasets of sufficient size and variety to capture the distribution of natural images sufficiently for the task at hand [Shafaei et al. 2018]. While being able to source vast collections of images from freely available online data has led to the successful creation of datasets such as ImageNet [Krizhevsky et al. 2012] and COCO [Lin et al. 2014], many other research areas require special equipment and expert knowledge for data capture. For example, the creation of the KITTI dataset required synchronization of cameras alongside a laser scanner and localization equipment [Geiger et al. 2013, 2012]. We seek to address this challenge for eye tracking with HMDs.

We opt to capture eye-images using a custom-built VR HMD with two synchronized cameras operating at 200Hz under controlled illumination. Corporate policy precludes us from offering further information on the hardware configuration of the VR HMD used in this study. As outlined below, we also use specialist medical equipment to capture further information about the shape and

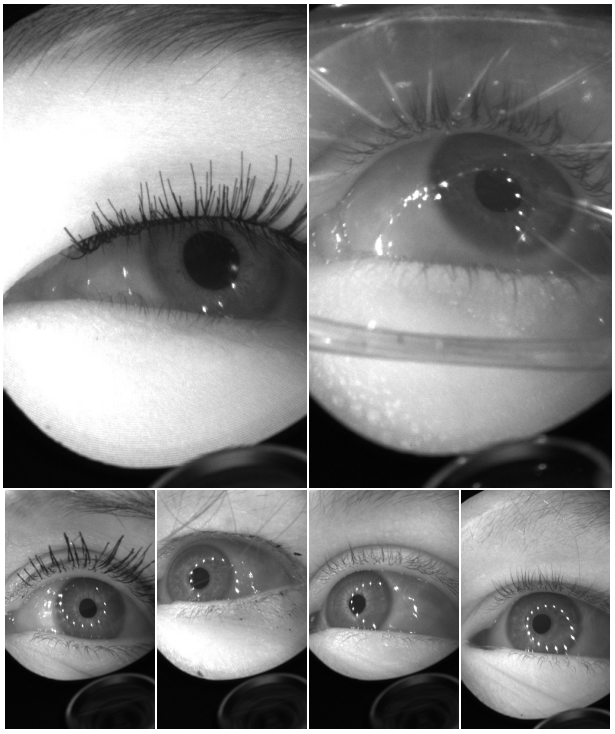

Figure 1: Examples of the acquired HMD images.

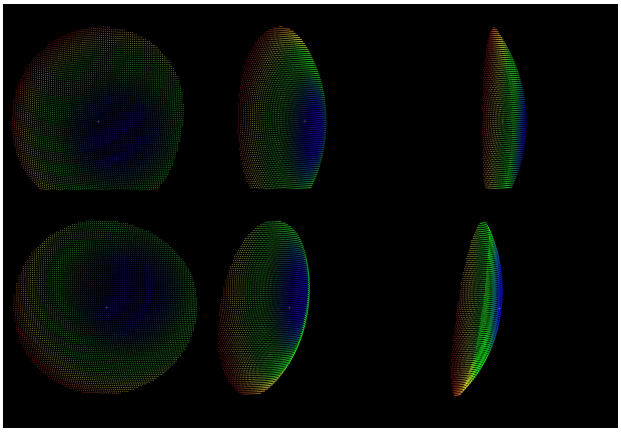

Figure 2: Examples of corneal topography, represented as point clouds. Row from top to bottom: left eye, right eye. Column from left to right: rotations along Y-axis. Color variation is along Z-axis. Better viewed in color.

optical properties of each participant's eyes contained in the dataset. It is the unique combination of advanced data capture, usually only performed in a clinical setting, with high resolution eye images and corresponding annotation masks for key eye-regions that sets this dataset apart from comparable datasets. We hope that this dataset bridges the gap between the computer vision and eye tracking communities and provides novel opportunities for research in the domain of eye tracking for HMDs.

Our contributions are summarized as follows:

- A large scale dataset captured using an HMD with two synchronized cameras under controlled illumination and high frame rates;
- A large scale dataset of annotation masks for key eye-regions: the iris, the sclera and the pupil;
- point cloud data from corneal topography captures of eye regions.

We want to however note that the dataset presented in this work was collected with a single type of VR head mounted device. It is quite possible that the results of any modeling based on this dataset may not be generalizable to use with other eye trackers that are available in the public domain.

2 RELATED WORK

2.1 Eye Tracking Datasets

Due to the difficulty of capturing binocular eye data especially in the VR context, there exists only a limited number of large-scale high resolution human image datasets in this domain. A comparison of our dataset to existing datasets of similar image modality can be found in Table 1.

The most similar dataset in terms of domain and image specifications is the recently published NVGaze dataset [Kim et al. 2019], consisting of 2.5 million infrared images recorded from 30 participants using an HMD (640x480 at 30 Hz). NVGaze includes annotation masks for key eye-regions for an additional dataset of 2 million synthetic eye images but does not provide segmentation annotations for the human eye image set at this point. The LPW dataset [Tonsen et al. 2015] includes a number of images recorded from 22 participants wearing a head mounted camera. Images are from indoor and outdoor recordings with varying lighting conditions and thus very different from the controlled lighting conditions in a VR HMD.

Some eye focused image sets are aimed at gaze prediction and released with gaze direction information, but do not include annotation masks, such as the Point of Gaze (PoG) dataset [McMurrough et al. 2012]. Other large-scale eye image datasets were captured in the context of appearance-based gaze estimation and record the entire face using RGB cameras as opposed to the eye region [Funes et al. 2014; Huang et al. 2015; Zhang et al. 2015]. For example, Gaze Capture [Krafka et al. 2016] consists of over 2.5 million images at various resolutions, recorded through crowd-sourcing on mobile devices, and its images are not specifically focused on the eye but contain a large portion of the surrounding face. In all these datasets, the focus is not solely on captures of eye-images, making them less suitable for the specific computer vision and machine learning challenges in the VR context. Finally, a different category of dataset, such as the UBIRIS [Hugo et al. 2010] and UBIRIS v2 [Proenca et al. 2010], were conceived with iris recognition in mind and therefore contain only limited annotation mask information.

Dataset	#Images	#Participants	Resolution	Framerate	Controlled Light	Sync. Left/Right	Optometric Data
STARE [Hoover et al. 2000]	-	-	-	-	No	No	Yes
PoG [McMurrough et al. 2012]	-	20	-	30 Hz	Yes	No	No
MASD [Das et al. 2017]	2,624	82	-	-	No	No	No
Ubiris v2 [Hugo et al. 2010]	11,102	261	400×300	-	No	-	No
LPW [Tonsen et al. 2015]	130,856	22	480×640	95 Hz	No	No	No
NVGaze [Kim et al. 2019]	2,500,000	30	480×640	30 Hz	Yes	No	No
Gaze Capture [Krafka et al. 2016]	2,500,000	1,450	Various	-	-	No	No
ExCuSe [Fuhl et al. 2015]	39,001	7	384×288	-	No	No	No
ElSe [Fuhl et al. 2016b]	55712	17	384×288	-	No	No	No
PupilNet [Fuhl et al. 2016a]	41,217	5	384×288	-	No	No	No
Closed-Eyes [Santini et al. 2018]	49,790	41	384×288	-	No	No	No
[Fuhl et al. 2019]	501,230	20	384×288	25 Hz	No	No	No
Swirski [Świrski et al. 2012]	600	2	620×460	-	No	No	No
Ours	356,649	152	400×640	200 Hz	Yes	Yes	Yes

Table 1: Publicly available datasets in the field of eye tracking.

2.2 Eye Segmentation

Segmentation of periocular regions, such as the pupil, iris, sclera is critical to subsequent estimation of a gaze location and thus will have a direct impact on subsequent classification and characteristic estimation of such eye movements as saccade, fixation and gaze estimation [Venkateswarlu 2003]. A large amount of studies have been investigated on segmenting a single trait (e.g., only the iris, sclera or eye region) [Das et al. 2017; Lucio et al. 2018; Radu et al. 2015; Sankowski et al. 2010; Thoma 2016]. A detailed survey of iris and sclera segmentation is presented in [Adegoke et al. 2013; Das et al. 2013]. We note that, although there are several advantages to having segmentation information on all key eye-regions simultaneously, a very limited amount of studies have been done on multi-class eye segmentation [Luo et al. 2019; Rot et al. 2018]. However, study in [Rot et al. 2018] trained a convolutional encoder-decoder neural network on a small data set of 120 images from 30 participants. A study by [Luo et al. 2019] trained a convolutional neural network coupled with conditional random field for post-processing on a data set of 3,161 low resolution images to segment only two classes: iris and sclera. In this paper, we try to address the gap for multi-class eye segmentation including pupil, iris, sclera and background, in a large data set of images in high resolution of 400×640.

2.3 Eye Rendering

Generating eye appearances under various environmental conditions including facial expressions, color of the skin and the illumination settings, play an important role in gaze estimation and tracking [Kim et al. 2019]. Two approaches have been studied in eye rendering: graphics-based approaches to generate eye images using a 3D eye model usually with a rendering framework to provide geometric variations such as gaze direction or head orientation [Wood et al. 2016, 2015]. Another is machine learning based approach. Study in [Shrivastava et al. 2017] used a generative adversarial network to train models with synthetic eye images while testing on realistic images. Study in [Wang et al. 2018] used a conditional bidirectional generative adversarial network to synthesize eye images consistent with the given eye gaze. However, these studies are focused on rendering synthetic eye images. We are interested in rendering realistic eye images that conform to the captured distribution of an individual and anticipate this dataset will encourage researchers to use the large corpus of eye-images and the annotation masks to develop solutions that can render realistic eye images. Realistic synthetic eye images are critical for the development and validation of novel and accurate eye tracking algorithms because they provide the data necessary to train and validate corresponding machine learning approaches for gaze estimation.

3 DATA COLLECTION

To ensure eye-safety during our data-collection process, which is important because collecting the data with a wearable VR headset exposes users' eyes to infrared light, the exposure levels were controlled to be well below the maximum permissible exposure as laid out in IEC 62471:2006, as well as the American National Standard for the Safe use of Lasers (ANSI Z136.1-2000, (IEC) 60825).

The dataset was collected from voluntary participants of ages between 19 and 65 using the aforementioned VR headset with light blocking facial interfaces. The illumination profile was controlled to be similar for all participants. Participants provided written informed consent to release their eye images before taking part in the study. There was no selection bias in our selection of participants for the data collection study, except that we required subjects to have corrected visual acuity above legal blindness, have a working knowledge of the English language, and not be pregnant. Participants were paid for their participation per session, and were given the choice to withdraw from the study at any point.

In addition to the image data captured from the HMD, the dataset released as part of this work also comes with an anonymized set of metadata per participant:

- Age (19-65), sex (male/female), usage of glasses (yes/no);
- Corneal topography.

Other than a pre-capture questionnaire for every participant, the rest of the data is captured in an approximately hour long session which is further split into sub-sessions as follows: All optometric examination are conducted in the first 10 minutes, followed by a 5 minute break. The break is taken to prevent any effect of optometric examinations on data captured with the HMD. Finally, two 20 minute capture sessions using the HMD are used to capture the eye-data during which the users were asked to perform several tasks i.e., a combination of explicit gaze target fixation task, smooth pursuit task, and free-viewing task in VR, separated by further breaks. The images present in the initial release of the dataset are taken from these 20 minute capture sessions.

4 ANNOTATION

We generated annotation masks for key eye-regions from a total of 12,759 eye-images as follows: a) eyelid, (using human-annotated key points) b) Iris (using an ellipse), c) Iris (using human-annotated points on the boundary), d) Pupil (using an ellipse) and e) Pupil (using human-annotated key points). The key-point dataset was used to generate annotation masks, which are released as part of this dataset. Below, we provide more details into the annotation protocol that was followed by the annotators in generating the key-points and the ellipse. The annotators were hired through a contracting company that specializes in the aforementioned annotation tasks. In total, 9 annotators worked on the annotation task for a period of one month. We also want to acknowledge the fact that that any polygonal annotations have sharp edges and corners, which may not reflect the actual semantic boundaries that were generated to produce the semantic mask released with the dataset.

4.1 Iris & Pupil Annotation with Ellipses

In order to produce the ellipse annotations for the Iris and the Pupil regions, the annotators performed the following three steps:

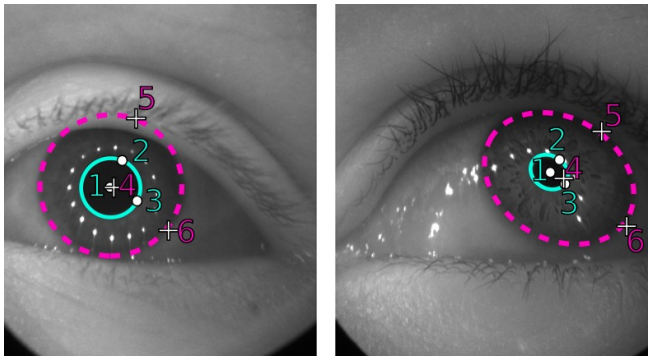

Figure 3: Ellipse Annotations. Points 1, 2 and 3 describe the Pupil, whereas points 4, 5 and 6 describe the Iris. Note that the points 1 and 4 are the center points. Best viewed digitally in color at high-resolution zoomed in.

- Position the center point;
- Position the second control point to adjust the shape and rotation; and
- Position the third control point to constraint the remaining degree of freedom for ellipse fitting.

All points are colored and numbered appropriately in the annotation tool to avoid confusion. Figure 3 offers an illustration of this process.

Since the ellipses do not provide any information about occlusion due to viewpoint and eyelid closure, we also obtain a more detailed annotation for which the annotators are asked to place a larger number for control points, which allows us to extract complex polygonal regions with high accuracy. We instruct the annotators to skip the ellipse-based annotation if the iris and pupil are not visible, but allow ellipses if they can be inferred with reasonable certainty. Some of the difficult cases are shown in Figure 4. Some of the images do not contain any useful information because the eyelids are completely closed or there is severe occlusion of the iris and pupil. In these situations annotators are asked to skip labeling images. In particular, we do not annotate eye-images if:

- The eye is closed;
- Eyelashes occlude the eye to the point where it is unreasonable to make an estimation;
- The eye is out of view, which can happen e.g. if the HMD is misaligned in a particular frame. Typically, this happens when the user is either putting on the headset or adjusting it to their head.

4.2 Iris & Pupil Annotation with Polygons

For the pupil and iris, we use 10 points each to create a polygon annotation. In both cases, the process starts by placing two points at the top and bottom of the feature of interest (labelled with numbers) and spacing a further 4 points equally between them on the boundary (labelled with letters). The top and bottom points do not have to be placed, and, if both cannot be identified, we fall back to labelling just with the points for the left and right side. With this process, the annotators are instructed to proceed with the left side

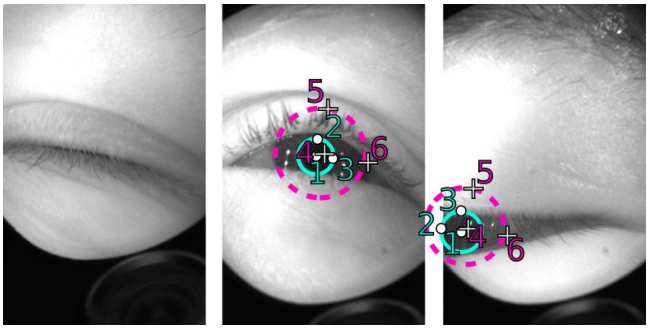

Figure 4: Ellipse annotations for difficult cases. For images where iris and pupil are not visible, no annotations are provided *(left)*. However, if positions can be inferred, we obtain ellipses for each class *(middle, right)* Please note that the images are padded for annotation so that ellipses can extend beyond the image boundaries *(right)*. Best viewed digitally in color and zoomed in.

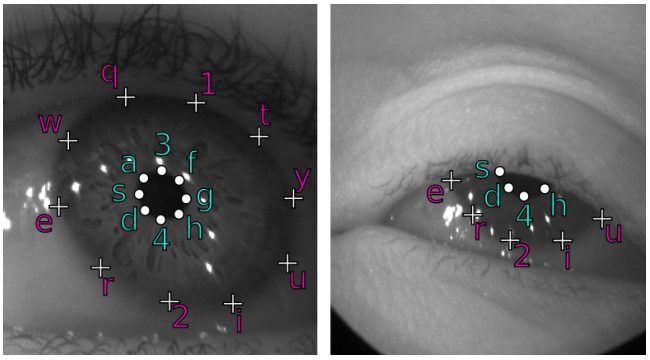

Figure 5: Iris and pupil annotations with dots (up to 10 points per feature). Top / bottom points are labelled with numbers, and further boundary points are denoted with letters. We note that even if one or the other of the top and bottom points is not visible, we still obtain annotations.

first, and then the right (see Figure 5). Again, the annotators are asked to skip difficult-to-label images, as explained in Section 4.1.

4.3 Eyelid Annotation

18 points are used for annotating the upper and lower eyelid. Similar to the iris and pupil case, the annotators are instructed to place these points equally spaced along the eyelid boundaries. Since the eyelid is much bigger than the other two cases (Iris and Pupil), in order to help with equal spacing we give the instruction to split the line recursively while adding points. Examples of this process are given in Figure 6. Figure 7 shows completed annotations for a variety of cases.

5 GENERAL STATISTICS

The dataset has a total of 12,759 images with annotation masks, 252,690 further images, 91,200 frames from contiguous 1.5 second video snippets, and 286 point cloud datasets for corneal topography.

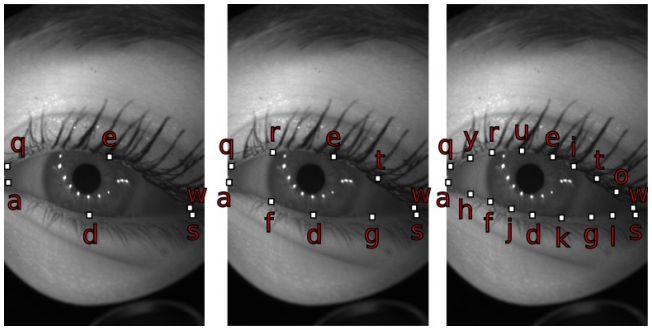

Figure 6: Eyelid annotation process. Note how the annotators proceed by splitting the annotation boundary recursively *(left to right)*.

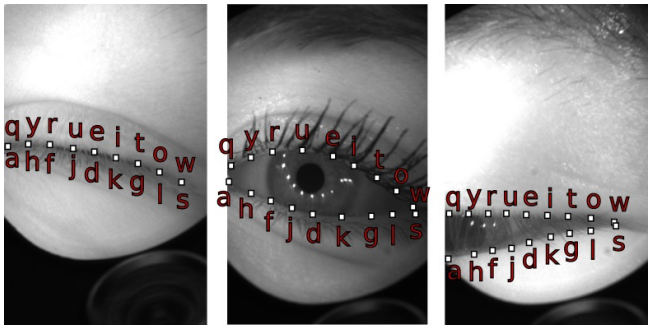

Figure 7: Eyelid annotation examples. Note how in case of the eye fully closed, we require the eyelid annotation points to overlap.

The latter image and image sequence sets are not accompanied by semantic segmentation annotations. Table 2 shows the statistics w.r.t. demographics (sex, age group, wearing glasses or not), and the amount of images and point clouds provided.

5.1 Corneal Topography

We provide corneal topography point cloud data that captures the surface curvature of the cornea for both left and right eyes. Corneal topography of each participant was measured via Scheimpflug imaging using an OCULUS® Pentacam® HR corneal imaging system, where corneal elevation maps were exported and converted to a point cloud. Figure 2 shows an example. We note that there is at most one point cloud estimate per participant.

6 STATISTICAL ANALYSIS

As outlined above, we choose to split the dataset by identity of the study participants as we found this to be both intuitive, and an easy setting to assess and avoid bias. When selecting the validation and test sets, we resample (or alternatively reweight for evaluation) the data to account for factors such as age and sex. Resampling or weighting of this kind is motivated by the fact that under-sampled modes of the true data distribution are the hardest to accurately capture by data-driven approaches. To avoid bias arising from this,

2*	Sex		Age					Glasses		#Images			2*#CT.
	Female	Male	18-23	24-30	31-40	41-50	51-65	No	Yes	SeSeg.	IS.	Seq.	
Train	51	44	3	15	39	26	12	92	3	8,916	193,882	57,000	178
Val.	13	15	1	5	5	8	9	27	1	2,403	57,337	16,800	52
Test	18	11	4	3	7	6	9	27	2	1,440	1,471	17,400	56
Total	82	70	8	23	51	40	30	146	6	12,759	252,690	91,200	286

Table 2: Statistics of data used for train, validation and test. SeSeg.: Images with semantic segmentation annotations. IS.: Images without annotations. Seq.: Image sequence set. CT: corneal topography. We report #identities regarding demographics (sex, age), wearing glasses or not. We also report #images and #corneal topography (represented as point clouds).

we ensure that the reweighting and selection of the test set penalizes approaches that do not take this into consideration.

An example of this is the age distribution. Figure 9 shows histograms characterizing the age distribution of the training vs test and validation images (with a bin-width of 5 years). Note how our choice of dataset splits already removes a significant amount of bias.

Apart from the information we can directly gain from the collected metadata, we further investigate the dataset for forms of sampling bias. To do this, we take an imagenet [Deng et al. 2009] pretrained 150-layer ResNet [He et al. 2015], as provided by the PyTorch library [Paszke et al. 2017], and encode a subset of 60,000 images sampled uniformly among the study participants to 4096-dimensional feature space (the output of the second-to-last fully connected layer). We subsequently use the k-means implementation of FAISS [Johnson et al. 2017] to cluster the encodings. Analysis of the resulting partitioning of the images reveals that most clusters predominantly contain images of either the left or right eye of *one* identity. This supports the intuition that identities are distinctive, and that it is easy to tell apart the left and right side of the face. We are however more concerned with clusters that exhibit a similar proportion of left *and* right eyes, or contain more than one identity. These clusters, if not corrected for in the process of sampling the identities used for each subset of the data splits, are evidence of a potentially significant difference of the distribution for the training, test and validation sets. The existence of such out of distribution cases could cast doubt over the validity of the test and validation process [Liang et al. 2017], and we therefore correct for it as far as possible.

Apart from identifying invalid images which can be trivially discarded (e.g. due to occlusion of the cameras or misalignment of the HMD), we identify the following difficult and under-sampled cases from the clustering process that are *not* correlated with the previously identified factors of variation such as sex, age, and left/right differences:

(1) Identities with glasses,
(2) Images of almost closed eyes, and
(3) Images of completely closed eyes.

In order to make sure that identities with glasses are represented across all data splits, we provide additional per-user annotations of the glasses case and make sure that at least one such case is contained in the test and validation sets.

Ensuring that nearly and fully closed eyes are represented in the dataset is a more challenging problem. This is due to the volume of data and the fact that these cases cannot be easily delineated. For example, deciding what counts as 'nearly'closed is not well defined. We address this by building a simple heuristic from a small (< 10000) number of cases for *open*, *nearly closed*, and *closed* eyes selected from the clusters described above. Example images given by this heuristic are shown in Figure 8.

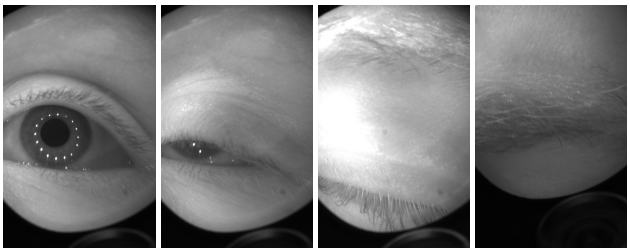

Figure 8: Representative images classified by our heuristic as (open, almost closed, closed and misaligned)

Given a small number of manual annotations, we train a 50-layers ResNet to classify a subset of 2.5 million images from the raw data as open, almost closed, closed or misaligned. By raw data we refer to images obtained during the capture which we do not use for the dataset release (this data is not annotated or curated). We extend the small initial dataset by additional annotations for images with high uncertainty, and then proceed to retrain the classifier. After repeating this process 4 times, we achieve more than 96% accuracy on the heuristic. We find that this heuristic finds approximately twice as many 'nearly'than fully closed eyes (on a subset of 12.6 million images, we estimate the percentages for these cases to be 2.21% and 4.31%, respectively). This is expected as the eye is near-fully closed just before and after blinking, thus providing evidence for the usefulness of our heuristic.

We make sure that all three identified eye states are present in the data when selecting images for the dataset but do not absolutely guarantee that any particular ratios are preserved.

6.1 Identity-Centric Dataset Splits

We partition the dataset in several ways that allow for the principled evaluation of a variety of machine learning problems, as shown in Table 2. Reasoning from an identity-centric perspective, let U be the set of all participants. We require semantic segmentation training, test and validation sets

$$ss_{train} \subset U, ss_{val} \subset U, ss_{test} \subset U, \quad (1)$$

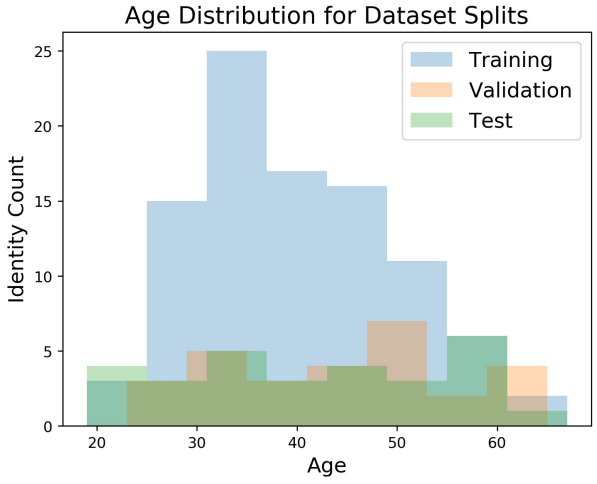

Figure 9: Age distribution of our proposed dataset splits. Note how we correct for bias in the collected data as far as possible.

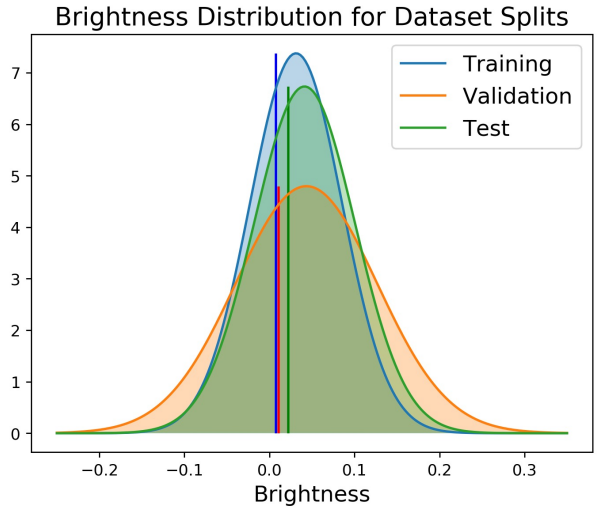

Figure 10: Per split distribution of the the mean image luminance. The median is denoted by vertical lines. Note that we do not apply any transformations to equalize these.

where $ss_{train} \cap ss_{val} \cap ss_{test} = \emptyset$.

The set of users from which we uniformly select images for the additional image and sequence datasets without annotation, $E \subset U$, can contain images from users contained in ss_{train}, ss_{test} and ss_{val}. The point of additionally providing E is to encourage the use of unlabelled data to improve the performance of supervised learning approaches.

For the case of semantic segmentation, we roughly follow the ratio of popular datasets such as MSCOCO [Lin et al. 2014]. We thus have, from a user-centric perspective, a $ss_{train}/ss_{val}/ss_{test}$ split of $\frac{95}{152}/\frac{28}{152}/\frac{29}{152}$.

One characteristic of our image domain we leave intentionally unaltered for the dataset release is the image brightness distribution. To estimate it, we compute the mean luminance for each image, and then estimate the mean, median and standard deviation across the dataset splits. As shown in Figure 10, this can vary strongly on a per-identity basis. The reason for this are individual-specific reflectance properties of human skin and eye, the fit of the HMD (which can vary depending on the shape of an individual's face), and whether or not makeup is applied. We provide this information to encourage future analysis to focus on generalizing to brightness variations of this kind.

6.2 Dataset Generation

Generating the data splits requires us to solve a constrained discrete optimization problem, subject to the constraints of having a balance of the sex, age group, wearing glasses, and eye-state attributes outlined in this section. We simplify this process by greedily selecting ss_{test}, ss_{val} and ss_{train}, in that order. We obtain the individual balanced selections using 10 million random configuration samples each, and picking the most balanced configuration among those samples. In other words, we simply select images from the raw data a large number of times, compute how well balanced the image sets are, and then use the best selection. We found that this provides well-balanced splits. The results of this procedure can be seen in Figure 9, where the age distribution is significantly more well-balanced in the test and validation as opposed to the training set.

	Train	Validation	Test
Train	0.0	0.003932	0.00487
Validation	0.003932	0.0	0.00489
Test	0.00487	0.00489	0.0

Table 3: Maximum Mean Disrepancy between the training, test and validation sets (Evaluated on a random set of 1024 images per split, averaged over 10 runs)

As a final sanity check, we compute the Maximum Mean Discrepancy (MMD) ([Sriperumbudur et al. 2010]) between the training, validation and test sets to ensure the identity-based splits of the data produce sufficiently correlated subsets. The standard deviation of the Gaussian RBF kernel ($k(x, x') = \exp(-\frac{\|x-x'\|_2^2}{2\sigma^2})$) used in computing the MMD is set based on the median Euclidian norm between all image pairs in the dataset [Liang et al. 2017]. In our case, this amounts to setting $\frac{1}{2\sigma^2} = 0.0087719$. Evaluating the metric for random sets of 1024 images across the data splits gives the results shown in Table 3, which strongly indicates that the three derived image sets are drawn from the same underlying data distribution.

7 INVESTIGATIONS INTO NEURAL NETWORK MODELS TO SEGMENT KEY EYE-REGIONS

There is a growing interest in segmentation of the ocular biometric traits of key eye regions such as the pupil, the iris and the sclera as these can provide useful information to model eye movements in various eye states such as as saccades and fixation [Venkateswarlu

2003]. As such, a large amount of works have focused on segmenting single eye region, i.e. only iris or only pupil, [Das et al. 2017; Lucio et al. 2018; Radu et al. 2015; Sankowski et al. 2010; Thoma 2016]. On the other hand, a very limited amount of investigations have focused on multi-class eye segmentation. Studies in [Rot et al. 2018] and [Luo et al. 2019] are the most recent examples for works focusing on multi-class eye segmentation. The study by [Rot et al. 2018] trained a neural network to segment iris and sclera on a very small dataset of 120 images drawn from 30 participants. Along the similar lines, the study by [Luo et al. 2019] trained a neural network coupled with a graphical model, the conditional random fields, [?] to segment iris and sclera on a data set of 3161 low resolution eye images.

In this Section, we attempt to address the problem of multi-class segmentation of eye regions, using the present dataset, which is comprised of 12,759 high resolution eye images and the corresponding ground truth segmentation masks generated from a pool of 152 participants. Specifically we adopt a popular framework of semantic segmentation (SS), also known as pixel level classification, to develop a neural network model that maps a given input image into a semantically labeled output mask. In particular, we investigate neural network models derived from one of the popular deep learning model for semantic segmentation, the SegNet model, [Badrinarayanan et al. 2017].

SegNet is comprised of an encoder neural network that maps the input image to a compressed latent representation, followed by a decoder neural network that uses pooling indices computed in the max-pooling step of the corresponding layer of the encoder network to perform non-linear upsampling. The final layer is a pixel wise classification layer. The entire network is trainable in an end-to-end fashion producing semantically segmented regions of interest for a given input image. For our investigations, we consider the following 3 modifications: (a) limit the number convolutional layers in the SegNet encoder and decoder sub-networks to 4 (b) add a boundary refinement (BR) layer in between the decoder network and the final pixel level classification layer. BR layer have been shown to improve pixel localization performance near object boundaries [Peng et al. 2017], and (c) replace the convolutional layers with separable convolution (SC) layers. SC factorizes convolution operation on multi-channel image into depth-wise 1×1 convolution along the dimension of the channels, which significantly reduces the computational cost for model inference [Howard et al. 2017]. The final modification to the SegNet model involves replacing the additive skip connection from the first layer of encoder network to last layer of decoder network with a multiplicative skip connection. The motivation for multiplicative skip connection is to suppress low probability pixel level class predictions. The SegNet model with 4 convolutional layer coupled with multiplicative skip connection is referred to as the modified-SegNet (mSegNet) model. The two additional variants that we investigate are: (a) mSegNet coupled with a BC layer and (b) mSegNet with SC.

We trained each network for 200 epochs on a NVIDIA RTX 2080 GPU using PyTorch [Paszke et al. 2017] with the ADAM optimizer [Kingma and Ba 2014]. The training parameters are: initial learning rate 0.001, mini-batch size 8, and regularization weight of 1e−5. No data augmentation was performed. The training was performed to minimize the pixel-wise multi-class cross-entropy loss as in the SegNet paper [Badrinarayanan et al. 2017].

Model	Pixel acc.	Mean acc.	Mean F1	Mean IoU	#Param. (M)
mSegNet	98.0	96.8	97.9	90.7	3.5
mSegNet w/ BR	98.3	97.5	98.3	91.4	3.5
mSegNet w/ SC	97.6	96.6	97.4	89.5	0.4

Table 4: Semantic Segmentation results. #Params: the number of learnable parameters, where "M" stands for million. *mSegnet*: 4-layer segnet with multiplicative skip connection between the last layer of the encoder and the decoder network.

The models are evaluated using the four common metrics used to evaluate semantic segmentation, following the procedure described in [Long et al. 2015]. In addition, we also report the model complexity measured in terms of the learnable model parameters. Table 4 shows the results of our experiments. In terms of accuracy, mSegNet with BR achieved the best performance with a mean IoU of 91.4%. In terms of model complexity, unsurprisingly, mSegNet with SC is the smallest model size with only 400,000 trainable parameters.

Given the focus of this paper, we did not iterate training to achieve best in class model performance. As such, all models failed to produce high fidelity eye region segments for eye-images impacted by occlusions from eye-glasses or heavy mascara or non-conforming pupil orientation. Figure 11 shows examples of these failure cases. We hope that our findings will encourage researchers from the community to leverage this dataset to develop improved semantic segmentation models to classify eye regions of interest.

8 CONCLUSION

We have presented a new dataset of images and optometric data of the human eye. Initial algorithmic analysis of the provided labels demonstrates the usefulness of the data for semantic segmentation. We made the dataset open to the public and we encourage the community to improve upon the reported results. We strongly believe that the foundation of this work will improve eye tracking robustness and efficiency in the future which has strong potential in enabling a variety of applications in VR environments, including but not limited to gaze interaction, foveated rendering, usability evaluation and others. We also believe that in addition to the eye tracking applications this dataset will be useful for researchers that study the variability of periocular regions of human eyes.

REFERENCES

B.O. Adegoke, E.O. Omidiora, S.A. Falohun, and J.A. Ojo. 2013. Iris Segmentation: a survey. *International Journal of Modern Engineering Research (IJMER)* 3, 4 (2013), 1885–1889.

V. Badrinarayanan, A. Kendall, and R. Cipolla. 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence* 39, 12 (2017), 2481–2495.

A. Borji and L. Itti. 2013. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence* 35, 1 (2013), 185–207.

Abhijit Das, Umapada Pal, Michael Blumenstein, and Miguel Angel Ferrer Ballester. 2013. Sclera recognition-a survey. In *2013 2nd IAPR Asian Conference on Pattern Recognition*. IEEE, 917–921.

A. Das, U. Pal, M.A. Ferrer, M.Blumenstein, D. Štepec, P. Rot, Z. Emeršič, P. Peer, V. Štruc, and S.V. Kumar. 2017. SSERBC 2017: Sclera segmentation and eye recognition

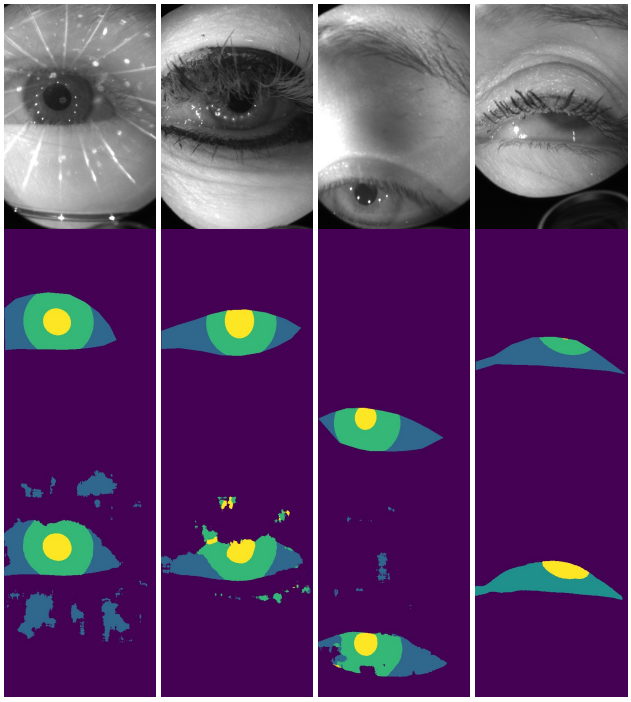

Figure 11: Examples of challenging samples on test data set of semantic segmentation. Rows from top to bottom: images, ground truth, predictions from SegNet w/ BR. Columns from left to right: eyeglasses, heavy mascara, dim light, varying pupil size. Better viewed in color.

benchmarking competition. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 742–747.

A. Das, U. Pal, M. A. Ferrer, M. Blumenstein, D. Štepec, P. Rot, Z. Emersic, P. Peer, V. Štruc, S. V. A. Kumar, and B. S. Harish. 2017. SSERBC 2017: Sclera segmentation and eye recognition benchmarking competition. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*. 742–747. https://doi.org/10.1109/BTAS.2017.8272764

J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

W. Fuhl, T. Kübler, K. Sippel, W. Rosenstiel, and E. Kasneci. 2015. Excuse: Robust pupil detection in real-world scenarios. In *International Conference on Computer Analysis of Images and Patterns*. Springer, 39–51.

W. Fuhl, W. Rosenstiel, and E. Kasneci. 2019. *500,000 Images Closer to Eyelid and Pupil Segmentation*. 336–347. https://doi.org/10.1007/978-3-030-29888-3_27

W. Fuhl, T. Santini, G. Kasneci, and E. Kasneci. 2016a. PupilNet: Convolutional Neural Networks for Robust Pupil Detection. *CoRR* abs/1601.04902 (2016). arXiv:1601.04902 http://arxiv.org/abs/1601.04902

W. Fuhl, T. Santini, T. Kübler, and E. Kasneci. 2016b. 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

M. Funes, A. Kenneth, F. Monay, and J. 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*. ACM, 255–258.

A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. 2013. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research* 32, 11 (2013), 1231–1237.

A. Geiger, P. Lenz, and R. Urtasun. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3354–3361.

K. He, X. Zhang, S. Ren, and J. Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015). arXiv:1512.03385 http://arxiv.org/abs/1512.03385

K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. van de Weijer. 2011. Eye Tracking: A Comprehensive Guide To Methods And Measures. (01 2011).

A. D. Hoover, V. Kouznetsova, and M. Goldbaum. 2000. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical Imaging* 19, 3 (March 2000), 203–210. https://doi.org/10.1109/42.845178

A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).

Q. Huang, A.Veeraraghavan, and A. Sabharwal. 2015. TabletGaze: A Dataset and Baseline Algorithms for Unconstrained Appearance-based Gaze Estimation in Mobile Tablets. *CoRR* abs/1508.01244 (2015). arXiv:1508.01244 http://arxiv.org/abs/1508.01244

P. Hugo, F. Silvio, S. Ricardo, O. Joao, and A. Luis A. 2010. The ubiris. v2: A database of visible wavelength iris images captured on-the-move and at-a-distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 8 (2010), 1529–1535.

J. Johnson, M. Douze, and H. Jégou. 2017. Billion-scale similarity search with GPUs. *CoRR* abs/1702.08734 (2017). arXiv:1702.08734 http://arxiv.org/abs/1702.08734

J. Kim, M. Stengel, A. Majercik, S. De Mello, D. Dunn, S. Laine, M. McGuire, and D. Luebke. 2019. NVGaze: An Anatomically-Informed Dataset for Low-Latency, Near-Eye Gaze Estimation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 10. https://doi.org/10.1145/3290605.3300780

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba. 2016. Eye Tracking for Everyone. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

G. Kramida. 2016. Resolving the vergence-accommodation conflict in head-mounted displays. *IEEE transactions on visualization and computer graphics* 22, 7 (2016), 1912–1931.

A. Krizhevsky, I. Sutskever, and G.E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

S. Liang, Y. Li, and R. Srikant. 2017. Principled Detection of Out-of-Distribution Examples in Neural Networks. *CoRR* abs/1706.02690 (2017). arXiv:1706.02690 http://arxiv.org/abs/1706.02690

T. Y. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. 2014. Microsoft COCO: Common Objects in Context. *CoRR* abs/1405.0312 (2014). arXiv:1405.0312 http://arxiv.org/abs/1405.0312

J. Long, E. Shelhamer, and T. Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.

D.R. Lucio, R. Laroca, E. Severo, A. Britto Jr, and D. Menotti. 2018. Fully convolutional networks and generative adversarial networks applied to sclera segmentation. *CoRR, vol. abs/1806.08722* (2018).

B. Luo, J. Shen, Y. Wang, and M. Pantic. 2019. The iBUG Eye Segmentation Dataset. In *2018 Imperial College Computing Student Workshop (ICCSW 2018) (OpenAccess Series in Informatics (OASIcs))*, Edoardo Pirovano and Eva Graversen (Eds.), Vol. 66. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 7:1–7:9. https://doi.org/10.4230/OASIcs.ICCSW.2018.7

C. D. McMurrough, V. Metsis, J. Rich, and F. Makedon. 2012. An Eye Tracking Dataset for Point of Gaze Detection. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '12)*. ACM, New York, NY, USA, 305–308. https://doi.org/10.1145/2168556.2168622

A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. 2017. Automatic differentiation in PyTorch. (2017).

A. Patney, J. Kim, M. Salvi, A. Kaplanyan, C. Wyman, N. Benty, A. Lefohn, and D. Luebke. 2016. Perceptually-based Foveated Virtual Reality. In *ACM SIGGRAPH 2016 Emerging Technologies (SIGGRAPH '16)*. ACM, New York, NY, USA, Article 17, 2 pages. https://doi.org/10.1145/2929464.2929472

Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. 2017. Large Kernel Matters–Improve Semantic Segmentation by Global Convolutional Network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4353–4361.

H. Proenca, S. Filipe, R. Santos, J. Oliveira, and L. A. Alexandre. 2010. The UBIRIS.v2: A Database of Visible Wavelength Iris Images Captured On-the-Move and At-a-Distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 8 (Aug 2010), 1529–1535. https://doi.org/10.1109/TPAMI.2009.66

P. Radu, J. Ferryman, and P. Wild. 2015. A robust sclera segmentation algorithm. In *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 1–6.

P. Rot, Z. Emeršič, V. Struc, and P. Peer. 2018. Deep multi-class eye segmentation for ocular biometrics. In *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*. IEEE, 1–8.

W. Sankowski, K. Grabowski, M. Napieralska, M. Zubert, and A. Napieralski. 2010. Reliable algorithm for iris segmentation in eye image. *Image and vision computing* 28, 2 (2010), 231–237.

T. Santini, W. Fuhl, and E. Kasneci. 2018. PuRe: Robust pupil detection for real-time pervasive eye tracking. *Computer Vision and Image Understanding* 170 (2018), 40 –

50. https://doi.org/10.1016/j.cviu.2018.02.002

A. Shafaei, M. Schmidt, and J. J. Little. 2018. Does Your Model Know the Digit 6 Is Not a Cat? A Less Biased Evaluation of "Outlier" Detectors. *CoRR* abs/1809.04729 (2018). arXiv:1809.04729 http://arxiv.org/abs/1809.04729

A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. 2017. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2107–2116.

B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar. 2013. Gaze locking: passive eye contact detection for human-object interaction. In *UIST*.

B.K. Sriperumbudur, A. Gretton, Kenji K.F., B. Schölkopf, and G. Lanckriet. 2010. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research* 11, Apr (2010), 1517–1561.

Lech Świrski, Andreas Bulling, and Neil Dodgson. 2012. Robust real-time pupil tracking in highly off-axis images. In *Proceedings of the Symposium on Eye Tracking Research and Applications*. ACM, 173–176.

V. Tanriverdi and R. Jacob. 2000. Interacting with eye movements in virtual environments. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 265–272.

Martin Thoma. 2016. A survey of semantic segmentation. *arXiv preprint arXiv:1602.06541* (2016).

M. Tonsen, X. Zhang, Y. Sugano, and A. Bulling. 2015. Labeled pupils in the wild: A dataset for studying pupil detection in unconstrained environments. *CoRR* abs/1511.05768 (2015). arXiv:1511.05768 http://arxiv.org/abs/1511.05768

R. Venkateswarlu. 2003. Eye gaze estimation from a single image of one eye. In *Proceedings Ninth IEEE International Conference on Computer Vision*. IEEE, 136–143.

K. Wang, R. Zhao, and Q. Ji. 2018. A hierarchical generative model for eye image synthesis and eye gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 440–448.

E. Wood, T. Baltrušaitis, LP. Morency, P. Robinson, and A. 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*. ACM, 131–138.

E. Wood, T. Baltrusaitis, X. Zhang, Y. Sugano, P. Robinson, and A. Bulling. 2015. Rendering of eyes for eye-shape registration and gaze estimation. In *Proceedings of the IEEE International Conference on Computer Vision*. 3756–3764.

L. Xiao, A. Kaplanyan, A. Fix, M. Chapman, and D. Lanman. 2018. DeepFocus: Learned Image Synthesis for Computational Displays. *ACM Trans. Graph.* 37, 6, Article 200 (Dec. 2018), 13 pages. https://doi.org/10.1145/3272127.3275032

X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. 2015. Appearance-based gaze estimation in the wild. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4511–4520. https://doi.org/10.1109/CVPR.2015.7299081