# Towards End-to-end Video-based Eye-Tracking
## Supplementary Materials

Seonwook Park, Emre Aksan, Xucong Zhang, and Otmar Hilliges

Department of Computer Science, ETH Zurich
{firstname.lastname}@inf.ethz.ch

**Abstract.** In this supplementary document, we provide further details regarding our proposed EVE dataset and architectures (EyeNet and GazeRefineNet). The EVE dataset is a dataset of videos, including 4 camera views and a video of screen content, as well as corresponding eye gaze position and gaze direction ground-truth (where available). This document provides additional details regarding the demographics captured and the pre-processing routines adopted. Regarding methodology, we provide an in-depth description of our "offset augmentation" procedure, and provide implementation details which are crucial to reproducing our results. In addition, we show additional experiments which demonstrate the effect of screen content, and the robustness of GazeRefineNet to error characteristics from different camera perspective.

The dataset and reference source code are available at:
https://ait.ethz.ch/projects/2020/EVE

**Keywords:** Eye Tracking, Gaze Estimation, Computer Vision Dataset

## 1 The EVE Dataset

Much care was taken in capturing, pre-processing, and analyzing of the EVE dataset. We present a few additional details regarding these steps in this section.

### 1.1 Ethics Approval

The collection of this dataset and the procedure of the study was approved by the Ethics Commission of ETH Zurich (application no. 2019-N-103). Before the beginning of a capture session, we clearly presented the risks (bodily and data-related) to our participants via information sheets and a comprehensive consent form. Participants were recruited via a university job board[1] and after the hour-long session, were paid a fee of 25 Swiss Francs in cash.

[1] https://marktplatz.uzhalumni.ch/

(a) MVC (b) Webcam (Left) (c) Webcam (Center) (d) Webcam (Right)

Fig. 1: Example frames from the EVE dataset, showing the 4 camera views (Machine Vision Camera or MVC from below, and 3 webcams mounted atop the monitor). Note that the outer webcams in particular capture relatively oblique head orientations. The green screen behind the participants should allow for future works to apply background augmentation for training neural networks.

### 1.2 Actual Capture

The quality of eye tracking data can vary greatly depending on specific illumination conditions, ethnicity, gender, and other factors, and as such we placed much care in designing the data collection environment. For example, we used two separate tables placed on top of a carpeted floor: one for holding the eye tracker via a VESA-mount arm, and one for the participants to rest their arms or elbows on (cf. Fig. 2 in the main paper). This was done to minimize the transfer of vibrations due to the participants' movements. We mainly adopted indirect illumination sources for better diffusion of light, and blocked any bright or direct sources of light with black tape or tissue paper. We provide additional samples of collected camera frames in Fig. 1.

Yet, not all nuisance factors can be anticipated and as such an experiment coordinator was present at every data collection session to monitor a live-stream of camera frames and eye movements. We collected a qualitative analysis of gaze data quality in terms of accuracy, precision, and jitter, and provide these alongside the dataset.

### 1.3 Dataset Pre-processing

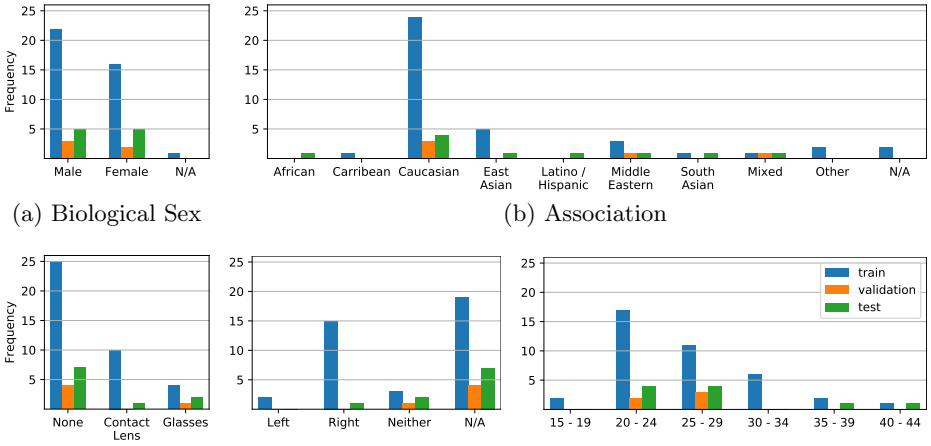
To pre-process the collected data, we first performed camera intrinsics calibration using the OpenCV framework. Extrinsic camera transformation determination was done using a first-surface mirror (to avoid errors due to the refraction occurring in standard mirrors) and code released in [7], with reference points defined by a ChArUco board (flipped as appropriate). Video was collected for every participant while moving the first-surface mirror around each camera such that the reflected ChArUco board was present across the span of the full camera frame with different inclinations.

In processing the video of participants, we first undistorted the frames' pixels and detected the face [9] and face-region landmarks [1]. We then performed a 3D morphable model (3DMM) fit to the detected 3D facial landmarks [2] with the purpose of yielding better estimates of gaze ray origins in 3D space. For every participant, we determined a person-specific inter-ocular distance value by exploiting our knowledge of relative camera positions. This inter-ocular distance (defined as the Euclidean distance in millimeters between the outer eye corner landmarks) is then used as a target scale value for scaling every fitted 3DMM. In this way we attempted to further stabilize the yielded eye patches, which were later used as input to our gaze estimation model. The determination of person-specific head-scale was done over 10 randomly sampled frames per participant.

Finally, we applied the "data normalization" procedure for yielding eye patches for gaze estimation [6,10]. The final eye patches are $128 \times 128$ in size and created with the assumption that the virtual camera is located 60cm away from the defined gaze origin, with a focal length of 1800mm. The selected origin of gaze is an average of the 3D eye corner landmarks of the eye in consideration, taken from the fitted 3DMM found in the previous step.

### 1.4 Dataset Characteristics

The final dataset is collected from 54 participants (30 male, 23 female, 1 unknown). The distribution in terms of answers to our demographics questionnaire can be seen in Fig. 2. While there are a few biases in the training data due to the available participant-pool in our local population, the careful selection of our final test set participants (10 participants in total) should allow for conclusions on generalization capabilities to be made. In particular, it can be seen in Fig. 2b that we attempted to sample our 10 test set participants from a variety of ethnicities. More fine-grained per-participant-level information will not be published in order to preserve the participants' privacy.




(a) Biological Sex (b) Association

(c) Visual Aid (d) Dominant Eye (e) Age

Fig. 2: Distribution of biological sex, ethnicity, adopted visual-aid, dominant eye, and age in the training, validation and test subsets of EVE, based on participants' self-reports. "N/A" marks cases where participants either did not know the answer or refused to provide one.

We find that the points-of-gaze (PoGs) in our dataset exhibit a screen-center-bias as previously reported in saliency literature [3] (see Fig. 3). However, this does not indicate that one can naively adjust all estimates of gaze direction to be screen-centered. According experiments are shown in Sec. 4.3 of this document. A notable fact is that the PoG distribution is similar between the training set and the test set, with samples existing in the peripheral regions of the screen.

Measured pupil diameters (as reported by the Tobii Pro SDK and measured by the Tobii Spectrum Eye Tracker) range between 2mm and 4mm (see Fig. 3). While this distribution shifts slightly for the test set participants, we find that the pupil sizes are relatively consistent across the defined subsets. Similarly, distances to the participants as estimated by our pre-processing pipeline (see Sec. 1.3) is consistent across the subsets, and in particular has a mode around the manufacturer recommended distance of 65cm. This demonstrates the care we took in positioning our participants, including a live monitoring of their posture throughout the capture session to avoid large eye tracker errors.

## 2 Offset Augmentation in GazeRefineNet

We provide here a step-by-step explanation of our offset augmentation procedure. This method is introduced to address the large differences in performance in gaze estimation when evaluating a network trained on one set of people, on a new set of people. The person-specific differences are often described as being a consistent offset (also called "angle kappa"), which do not appear in computed

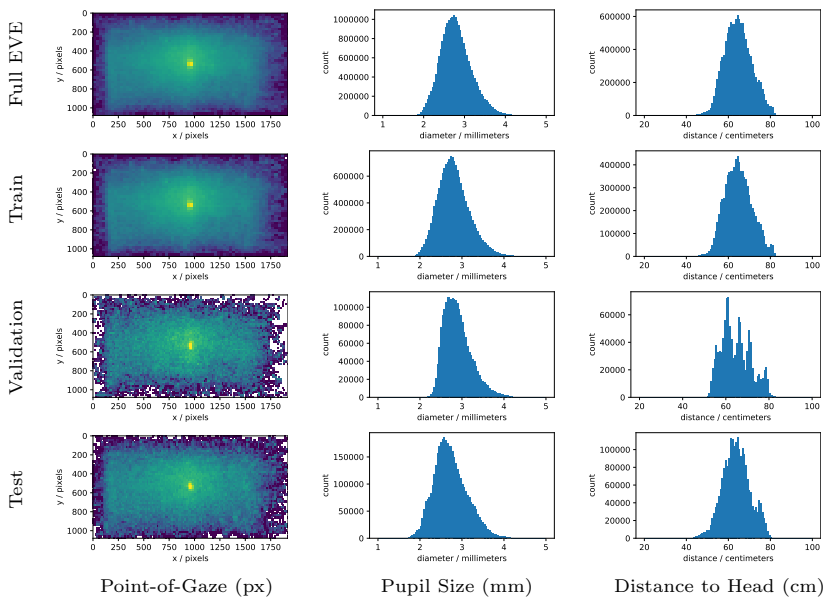



Fig. 3: PoG (on-screen pixels), distance (in cm) and pupil size (in mm) distributions for the defined subsets of the proposed EVE dataset. The number of people involved are 54, 39, 5, 10 respectively for the full dataset, training subset, validation subset, and test subsets. The 2D histograms are coloured with a logarithmic scale, with values normalized by the size of the subset in concern.

training losses, but only in the validation or test losses. We thus implement our augmentation to mimic the effect of this angle kappa.

First, given an estimate for gaze direction $\hat{\mathbf{g}}$, let us assume that this is represented in spherical coordinates representing pitch and yaw angles such that $\theta$ is pitch, and $\phi$ is yaw. Then the unit-vector notation of $\hat{\mathbf{g}} = (\theta, \phi)$ would be calculated with,

$$\hat{\mathbf{v}}_h = \begin{pmatrix} -\cos\theta \sin\phi \\ -\sin\theta \\ -\cos\theta \cos\phi \end{pmatrix}. \tag{1}$$

As the vector was previously defined such that $(\theta, \phi) = (0, 0)$ points towards the camera, we must flip the vector via negation to bring it to the camera-relative coordinate system in which the head model (3DMM) is defined.

Assuming that we know the rotation of the head with respect to the camera (from which the input image was taken from), we then apply the inverse of this known rotation $\mathbf{R}_h$ to calculate the gaze direction relative to the head coordinate system:

$$\hat{\mathbf{v}}_h = \mathbf{R}_h^T \hat{\mathbf{v}}_c. \tag{2}$$

We now return this head-relative gaze direction value to spherical coordinates, with:

$$\hat{\mathbf{g}}_h = \begin{pmatrix} \theta_h \\ \phi_h \end{pmatrix} = \begin{pmatrix} \arcsin -\hat{y}_h \\ \arctan2\left(-\hat{x}_h,\, -\hat{z}_h\right) \end{pmatrix}, \tag{3}$$

where $\hat{\mathbf{v}}_h = (\hat{x}_h,\, \hat{y}_h,\, \hat{z}_h)$. The corresponding rotation matrix is then,

$$\mathbf{R}_h = \begin{pmatrix} \cos\phi_h & 0 & \sin\phi_h \\ 0 & 1 & 0 \\ -\sin\phi_h & 0 & \cos\phi_h \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_h & -\sin\theta_h \\ 0 & \sin\theta_h & \cos\theta_h \end{pmatrix}. \tag{4}$$

This is the rotation that we can apply on top of a constant sequence-specific and synthetic "offset". Per given training sequence of length $T$ (to maintain consistency with the main paper, Eq.2), we acquire a sequence-specific offset $\kappa_i = (\theta_\kappa,\, \phi_\kappa) \sim \mathcal{N}\left(0,\, 3^\circ\right)$ that is parameterized with pitch and yaw angle values as done in defining $\hat{\mathbf{g}}$. We determined the standard deviation of 3 degrees empirically, and show this in degrees for convenience of understanding. In reality, the sampled values are in radians.

We convert the kappa values to unit vector notation and rotate it by the current gaze direction matrix,

$$\hat{\mathbf{v}}_h^{\text{aug}} = \mathbf{R}_h \hat{\mathbf{v}}_\kappa. \tag{5}$$

This augmented gaze direction is transformed back to the normalized camera coordinates system such that the frontal gaze is defined with $(\theta,\, \phi) = (0,\, 0)$.

## 3 Implementation Details

To facilitate faithful reproduction of our experiments, we provide additional implementation details of our architecture and its training and hyper-parameters.

### 3.1 Validity of Ground-truth Labels

The ground-truth data provided by the EVE dataset often comes from the Tobii Spectrum Pro eye tracker, and associated Tobii Pro SDK. As is often the case with eye trackers, there are cases where tracking fails, such as during eye blinks or when illumination conditions are too poor for features to be tracked. The "validity" of predicted ground-truth is provided by the SDK, and stored alongside all other labels. We apply the validity boolean values to our loss calculation, such that only valid ground-truth labels are used during training.

The collected screen frames and Tobii-origin data do not perfectly coincide in terms of reported timestamps. We perform a manual alignment to ensure consistency between images of the eye-region and the gaze data, and additionally perform bilinear interpolation in PoG given that valid labels exist on both sides (immediately before and immediately after) of the query timestamp. As the eye tracking data is collected at 150Hz (as a reminder, the camera frames have been collected at 30Hz or 60Hz), and by the Nyquist-Shannon sampling theorem, we can assume that the eye tracking data has been reliably handled.

### 3.2 EyeNet

In the main paper (cf. Sec. 4.1), we defined the loss terms for gaze direction as $\mathcal{L}_{\text{gaze}}$ and for pupil size as $\mathcal{L}_{\text{pupil}}$. We define the full loss as:

$$\mathcal{L}_{\text{EyeNet}} = \gamma_{\text{PoG}}\mathcal{L}_{\text{gaze}} + \gamma_{\text{pupil}}\mathcal{L}_{\text{pupil}}, \tag{6}$$

and set $\gamma_{\text{gaze}} = 1.0$ and $\gamma_{\text{pupil}} = 1.0$ empirically. The EyeNet is trained using the Adam optimizer [4] for 8 epochs using a batch size of 16, and $l_2$ parameter decay of 0.005. We apply exponential learning rate decay of factor 0.5 every 1 epoch, beginning from a learning rate of 0.016. The input eye image is resized to be $128 \times 128$ pixels large.

### 3.3 GazeRefineNet

The GazeRefineNet adopts the a mean-squared error loss term for the final PoG (calculated via a soft-argmax layer, cf. Fig. 4b of main paper), and in addition applies a per-pixel cross-entropy loss for guiding the learning of the heatmap. When defining the cross-entropy based loss term as $\mathcal{L}_{\text{XE}}$, we can then define the full loss as:

$$\mathcal{L}_{\text{RefineNet}} = \gamma_{\text{PoG}}\mathcal{L}_{\text{PoG}} + \gamma_{\text{XE}}\mathcal{L}_{\text{XE}}. \tag{7}$$

where we set $\gamma_{\text{PoG}} = 0.001$ and $\gamma_{\text{XE}} = 1.0$ empirically. The GazeRefineNet is trained using the Adam optimizer [4] for 4 epochs using a batch size of 8, and $l_2$ parameter decay of 0.0. We apply exponential learning rate decay of factor 0.5 every 0.5 epochs, beginning from a learning rate of 0.008. The input screen content frame is resized to be $128 \times 72$ pixels large. Please note that during this stage of training, the EyeNet weights are not updated.

## 4 Additional Results

Here, we provide additional details with respect to the results shown in Sec. 5 of the main paper, as well as new experiments which further assess our GazeRefineNet architecture. In particular, we experiment with pre-training the gaze estimation network (EyeNet) on an existing in-the-wild dataset, and applying it directly and without modification as part of the GazeRefineNet training. Next, we attempt to understand the inter-play of the proposed offset augmentation and screen content input. We then evaluate the robustness of the GazeRefineNet training to the different error characteristics of the 4 camera views. Lastly, we show the changes in GazeRefineNet performance with varying strength of offset augmentation applied during training.

### 4.1 Evaluation Details

In all experiments, we evaluate on the test split of the EVE dataset consisting of 10 participants. To reduce the data load of both training and evaluation,

Table 1: Experiments where the $\text{EyeNet}_{\text{static}}$ is trained on the GazeCapture dataset [5]. The initial error is high as is typical of eye-patch input gaze estimation networks evaluated in the cross-dataset setting. We see that despite the high initial error, a respectably low error is achieved when training a $\text{GazeRefineNet}_{\text{GRU}}$ atop the predictions from the GazeCapture-trained $\text{EyeNet}_{\text{static}}$. We thus show that our refinement approach can be used in combination with existing gaze estimators to bridge dataset domain gaps

| Model | Gaze Dir. (°) | PoG (cm) | PoG (px) |
|---|---|---|---|
| Baseline (both eyes) | 7.93 | 8.86 | 288.85 |
| $\text{GazeRefineNet}_{\text{GRU}}$ | **3.93** ↓50.57% | **4.33** ↓51.12% | **150.29** ↓47.97% |

we subsample all data such that we take 10 samples per second. A sequence is defined to span 3 seconds of time such that the shortly exposed image stimuli sequences can be trained on as well (exposure time of 3 seconds to participants). Effectively, this means that we sub-sample the number of frames by a factor of $\frac{1}{6}$ and $\frac{1}{3}$ respectively for the machine vision camera and webcams.

For both training and evaluation, we cut all available video data into 3-second-long sequences without gaps or overlaps. This results in $65,116$ sequences in the training sub-set, $7,676$ sequences in the validation sub-set, and $17,660$ sequences in the test sub-set. There are $2,392$ image-stimulus sequences, $10,472$ video-stimulus sequences, and $4,796$ wikipedia-stimulus sequences in the test sub-set.

### 4.2 Training EyeNet on GazeCapture

In order to assess our contribution in the context of existing gaze estimation methods and datasets, we identified that training the gaze estimation part of our architecture ($\text{EyeNet}_{\text{static}}$) and using it without modification to learn the final refinement step, would be the most challenging benchmark. We evaluate this setting by training our $\text{EyeNet}_{\text{static}}$ on the GazeCapture dataset [5] with equivalent pre-processing steps to our data, then train a $\text{GazeRefineNet}_{\text{GRU}}$ while keeping the $\text{EyeNet}_{\text{static}}$ fixed, to finally evaluate performance on the test set of our EVE dataset. We select our own test set as no other publicly available video-based gaze dataset exhibit natural eye movements. The baseline gaze direction error of 7.93° shown in Tab. 1 is typical of network architectures that take single-eye inputs (we perform single-eye gaze estimation to enable binocular gaze estimation in the future - an interesting output for studies on vergence), as shown in recent works [11,8]. We find that a highly significant improvement can be made even with initial errors as large as 27% of the screen height (1080 pixels). This shows that dataset differences can easily be overcome with our GazeRefineNet training, even in the absence of labeled data from test users, and while retaining the errors present in the trained $\text{EyeNet}_{\text{static}}$ (its weights are not changed during $\text{GazeRefineNet}_{\text{GRU}}$ training).

Table 2: Ablation study to further understand the effect in the absence of any screen content input. Each row adds a factor (such that the last row includes all changes). The refinement network without screen content simply refines a given heatmap, and thus could be considered a method of screen-center-bias enforcement, a form of gaze position prior.

| Model | Gaze Dir. (°) | PoG (cm) | PoG (px) |
|---|---|---|---|
| Baseline (both eyes) | 3.48 | 3.85 | 132.56 |
| + Refinement Network* | 3.41 ↓2.18% | 3.77 ↓2.17% | 130.78 ↓1.34% |
| + Offset Augmentation | 3.00 ↓13.84% | 3.31 ↓13.90% | 115.10 ↓13.17% |
| + Screen Content | **2.49** ↓28.43% | **2.75** ↓28.49% | **95.59** ↓27.89% |

* with GRU and skip connections between encoder and decoder.

### 4.3 Offset augmentation without screen content

To better understand the effect of the screen content input, we performed an ablative study of our contributions in the absence of screen content. What this means is that no appearance-based context is given to the task of gaze estimate refinement, except for the dimensions of the heatmap with which PoG is represented. More specifically, the GazeRefineNet could be conjectured to be performing a center-bias application. We find in Tab. 2 that this assumption is only partly true, and that applying the GazeRefineNet alone without screen input nor offset augmentation results in comparable results to the baseline. This means that the center-bias present in the data is not useful in further improving gaze estimates. We do find however, that the offset augmentation still works relatively well in the absence of screen content. With screen content input we can reach the final best reported performance.

### 4.4 Cross-Camera Evaluation

To assess the sensitivity of our GazeRefineNet approach, we evaluate performance changes when training on predicted gazes from different camera views in Table 3, where gaze estimates are still provided from a pre-trained $\text{EyeNet}_{\text{GRU}}$ but the $\text{GazeRefineNet}_{\text{GRU}}$ is trained from gaze data only from the source camera view, and tested on frames from the target camera view. We find that in general, the best performances can be seen when evaluating on the machine vision camera frames, as image quality and detail are expectedly higher. Nonetheless in general, improvements can be seen across the board, showing that the GazeRefineNet is not sensitive to changes in camera view (and the consequent change in the errors of initial PoG predictions).

### 4.5 Effect of offset augmentation strength

The amount of offset to apply to initial gaze direction predictions is an important hyperparameter. For example, a $\text{GazeRefineNet}_{\text{GRU}}$ trained with weak offset augmentation may not handle high test-time offsets whereas a $\text{GazeRefineNet}_{\text{GRU}}$

Table 3: Final refined gaze direction errors (in degrees, lower is better) for cross-camera evaluations. While testing on the high-quality machine vision camera frames yield the best results, it can be seen that the refinement step is mostly agnostic to where the gaze data comes from and can generalize to gaze data from new views, despite differences in characteristics of the error

| Source \ Target | Webcam (Left) | Webcam (Center) | Webcam (Right) | MVC |
|---|---|---|---|---|
| Webcam (Left) | **3.03** ↓ 21.62% | **2.55** ↓ 23.47% | 3.12 ↓ 22.79% | **2.24** ↓ 16.90% |
| Webcam (Center) | **3.04** ↓ 21.53% | **2.55** ↓ 23.53% | 3.11 ↓ 22.96% | 2.26 ↓ 16.31% |
| Webcam (Right) | 3.07 ↓ 20.52% | 2.58 ↓ 22.66% | 3.14 ↓ 22.17% | 2.29 ↓ 15.32% |
| MVC | **3.03** ↓ 21.69% | **2.54** ↓ 23.70% | **3.09** ↓ 23.36% | **2.23** ↓ 17.50% |

Note: MVC stands for "Machine Vision Camera".
Improvements are with respect to initial PoG estimates from $\text{EyeNet}_{\text{GRU}}$.

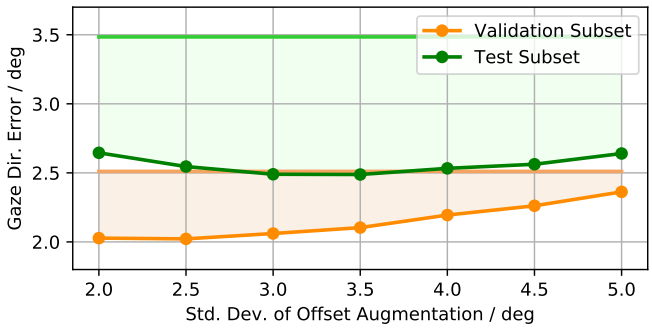



Fig. 4: Varying $\sigma_\kappa$ results in differing performance improvements on the validation subset of the EVE dataset, compared to that on the test subset. Specifically, the test subset is significantly more challenging and thus a stronger amount of offset augmentation is required than in the case of the validation subset.

trained with strong offset augmentation may perform overly aggressive corrections. We show this trade-off in Fig. 4 where we see that the relatively easier validation requires lower amounts of offset augmentation at training time compared to the test set.

A more comprehensive study of discrepancies between learned models' predictions of gaze direction should be performed in the future, in relation to the differences in demographics in various gaze datasets. Furthermore, these offsets are most certainly not due to textbook anatomical differences only (between optical and visual axes in each eyeball). For instance, the determination of 3D gaze origin is always done in an approximate manner and may vary greatly depending on (a) how the head pose was determined, and (b) how the head-pose-relative gaze origin was determined. In pre-processing the EVE dataset, we apply a

3DMM fitting approach with interocular-distance-based scale-normalization to alleviate these issues.

## 5 Ethical Considerations

In this work we effectively demonstrate that it is possible to improve predictions of PoG given the screen content, even without prompting the user (ground-truth label acquisition or gaze estimator calibration). We are certain that the field will progress quickly, and will soon be reporting methods and architectures which yield higher accuracy and robustness for screen-based eye tracking based on our initial insights and the EVE dataset.

We are aware of the ethical implications of further developments to our approach in the context of data privacy. Specifically, a malicious agent could attempt to elicit information regarding a user's habits or preferences without their awareness.

To eliminate such efforts, we hope that operating system developers can build secure sandbox environments where front-facing camera usage is increasingly restricted. Furthermore, we recommend that the Computer Vision community work on: (a) allowing for light-weight model architectures through knowledge distillation or weight quantization to quickly enable edge-only prediction of eye gaze such as to restrict the transfer of original front-facing camera frames, and (b) development of eye movement descriptors which need not expose fine-grained person-specific traits yet assist in intelligent interactive systems such as user-state-aware interfaces (e.g. changing of layout or appearance based on perceived stress or cognitive load).

## References


1. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In: ICCV. pp. 1021–1030 (2017)
2. Huber, P., Hu, G., Tena, R., Mortazavian, P., Koppen, P., Christmas, W.J., Ratsch, M., Kittler, J.: A multiresolution 3d morphable face model and fitting framework. In: VISIGRAPP (2016)
3. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: ICCV. pp. 2106–2113. IEEE
4. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
5. Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., Torralba, A.: Eye Tracking for Everyone. In: CVPR (2016)
6. Sugano, Y., Matsushita, Y., Sato, Y.: Learning-by-Synthesis for Appearance-based 3D Gaze Estimation. In: CVPR (2014)
7. Takahashi, K., Nobuhara, S., Matsuyama, T.: Mirror-based camera pose estimation using an orthogonality constraint. IPSJ Transactions on Computer Vision and Applications **8**, 11–19 (2016)
8. Wood, E., Baltrušaitis, T., Morency, L.P., Robinson, P., Bulling, A.: A 3d morphable eye region model for gaze estimation. In: ECCV. pp. 297–313. Springer (2016)
9. Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., Li, S.Z.: S3fd: Single shot scale-invariant face detector. In: ICCV. pp. 192–201 (2017)
10. Zhang, X., Sugano, Y., Bulling, A.: Revisiting data normalization for appearance-based gaze estimation. In: ETRA (2018)
11. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. TPAMI (2019), `https://perceptual.mpi-inf.mpg.de/files/2017/11/zhang17_pami.pdf`