Towards End-to-end Video-based Eye-Tracking

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

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

Abstract. Estimating eye-gaze from images alone is a challenging task, in large parts due to un-observable person-specific factors. Achieving high accuracy typically requires labeled data from test users which may not be attainable in real applications. We observe that there exists a strong relationship between what users are looking at and the appearance of the user's eyes. In response to this understanding, we propose a novel dataset and accompanying method which aims to explicitly learn these semantic and temporal relationships. Our video dataset consists of time-synchronized screen recordings, user-facing camera views, and eye gaze data, which allows for new benchmarks in temporal gaze tracking as well as label-free refinement of gaze. Importantly, we demonstrate that the fusion of information from visual stimuli as well as eye images can lead towards achieving performance similar to literature-reported figures acquired through supervised personalization. Our final method yields significant performance improvements on our proposed EVE dataset, with up to 28% improvement in Point-of-Gaze estimates (resulting in 2.49° in angular error), paving the path towards high-accuracy screen-based eye tracking purely from webcam sensors. 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 Introduction

The task of gaze estimation from a single low-cost RGB sensor is an important topic in Computer Vision and Machine Learning. It is an essential component in intelligent user interfaces [14,4], user state awareness [21,16], and serves as input modality to Computer Vision problems such as zero-shot learning [25], object referral [2], and human attention estimation [10]. Un-observable person-specific differences inherent in the problem are challenging to tackle and as such high accuracy general purpose gaze estimators are hard to attain. In response, person-specific adaptation techniques [40,33,32] have seen much attention, albeit at the cost of requiring test-user-specific labels. We propose a dataset and accompanying method which holistically combine multiple sources of information explicitly. This novel approach yields large performance improvements without needing ground-truth labels from the final target user. Our large-scale dataset (EVE) and network architecture (GazeRefineNet) effectively showcase the newly proposed task and demonstrate up to 28% in performance improvements.

The human gaze can be seen as a closed-loop feedback system, whereby the appearance of target objects or regions (or visual stimuli) incur particular movements in the eyes. Many works consider this interplay in related but largely separate strands of research, for instance in estimating gaze from images of the user (bottom-up, e.g. [54]) or post-hoc comparison of the eye movements with the visual distribution of the presented stimuli (top-down, e.g. [45]). Furthermore, gaze estimation is often posed as a frame-by-frame estimation problem despite its rich temporal dynamics. In this paper, we suggest that by taking advantage of the interaction between user's eye movements and what they are looking at, significant improvements in gaze estimation accuracy can be attained even in the *absence of labeled samples* from the final target. This can be done without explicit gaze estimator personalization. We are not aware of existing datasets that would allow for the study of these semantic relations and temporal dynamics. Therefore, we introduce a novel dataset designed to facilitate research on the joint contributions of dynamic eye gaze and visual stimuli. We dub this dataset the EVE dataset (**E**nd-to-end **V**ideo-based **E**ye-tracking). EVE is collected from 54 participants and consists of 4 camera views, over 12 million frames and 1327 unique visual stimuli (images, video, text), adding up to approximately 105 hours of video data in total.

Accompanying the proposed EVE dataset, we introduce a novel bottom-up-and-top-down approach to estimating the user's point of gaze. The Point-of-Gaze (PoG) refers to the actual target of a person's gaze as measured on the screen plane in metric units or pixels. In our method, we exploit the fact that more visually salient regions on a screen often coincide with the gaze. Unlike previous methods which adopt and thus depend on pre-trained models of visual saliency [45,46,7], we define our task as that of online and conditional PoG refinement. In this setting a model takes raw screen content and an initial gaze estimate as explicit conditions, to predict the final and refined PoG. Our final architecture yields significant improvements in predicted PoG accuracy on the proposed dataset. We achieve a mean test error of 2.49 degrees in gaze direction or 2.75cm (95.59 pixels) in screen-space Euclidean distance. This is an improvement of up to 28% compared to estimates of gaze from an architecture that does not consider screen content. We thus demonstrate a meaningful step towards the proliferation of screen-based eye tracking technology.

In summary, we propose the following contributions:

- A new task of online point-of-gaze (PoG) refinement, which combines bottom-up (eye appearance) and top-down (screen content) information to allow for a truly end-to-end learning of human gaze,
- EVE, a large-scale video dataset of over 12 million frames from 54 participants consisting of 4 camera views, natural eye movements (as opposed to following specific intructions or smoothly moving targets), pupil size annotations, and screen content video to enable the new task,
- a novel method for eye gaze refinement which exploits the complementary sources of information jointly for improved PoG estimates, in the absence of ground-truth annotations from the user.

Table 1: Comparison of EVE with existing screen-based datasets. EVE is the first to provide natural eye movements (free-viewing, without specific instructions) synchronized with full-frame user-facing video and screen content

Name	Region	# Subjects	# Samples	Temporal Data	Natural Eye Movements	Screen Content Video	Publicly Available
Columbia Gaze [44]	Frame	56	5,800	-	N	N	Y
EYEDIAP [17]	Frame	16	62,500	30Hz	N*	N	Y
UT Multiview [47]	Eyes	50	64,000	-	N	N	Y
MPIIGaze [60]	Eyes	15	213,659	-	N	N	Y
TabletGaze [22]	Frame	51	1,785	-	N	N	Y
GazeCapture [28]	Frame	1,474	2,129,980	-	N	N	Y
Deng and Zhu [12]	Eyes	200	240,000	-	N	N	N
MPIIFaceGaze [61]	Face	15	37,639	-	N	N	Y
DynamicGaze [51]	Eyes	20	645,000	∼30Hz	Y	N	N
EVE (Ours)	Frame	54	12,308,334	30Hz, 60Hz	Y	30Hz	Y

* Only smooth pursuits eye movements are available.

In combination these contributions allow us to demonstrate a gaze estimator performance of 2.49° in angular error, comparing favorably with supervised person-specific model adaptation methods [40,32,8].

2 Related Work

In our work we consider the task of remote gaze estimation from RGB, where a monocular camera is located away from and facing a user. We outline here recent approaches, proposed datasets, and relevant methods for refining gaze estimates.

2.1 Remote Gaze Estimation

Remote gaze estimation from unmodified monocular sensors is challenging due to the lack of reference features such as reflections from near infra-red light sources. Recent methods have increasingly used machine learning methods to tackle this problem [3,35,39] with extensions to allow for greater variations in head pose [34,43,12]. The task of cross-person gaze estimation is defined as one where a model is evaluated on a previously unseen set of participants. Several extensions have been proposed for this challenging task in terms of self-assessed uncertainty [9], novel representations [41,42,57], and Bayesian learning [53,54].

Novel datasets have contributed to the progress of gaze estimation methods and the reporting of their performance, notably in challenging illumination settings [60,61,28], or at greater distances from the user [26,15,17] where image details are lost. Screen-based gaze estimation datasets have had a particular focus [60,61,28,22,12,17,36] due to the practical implications in modern times, with digital devices being used more frequently. Very few existing datasets include videos, and even then often consist of participants gazing at points [22] or following smoothly moving targets only (via smooth pursuits) [17]. While the

RT-GENE dataset includes natural eye movement patterns such as fixations and saccades, it is not designed for the task of screen-based gaze estimation [15]. The recently proposed DynamicGaze dataset [51] includes natural eye movements from 20 participants gazing upon video stimuli. However, it is yet to be publicly released and it is unclear if it will contain screen-content synchronization. We are the first to provide a video dataset with full camera frames and associated eye gaze and pupil size data, in conjunction with screen content. Furthermore, EVE includes a large number of participants (=54) and frames (12.3M) over a large set of visual stimuli (1004 images, 161 videos, and 162 wikipedia pages).

2.2 Temporal Models for Gaze Estimation

Temporal modelling of eye gaze is an emerging research topic. An initial work demonstrates the use of a recurrent neural network (RNN) in conjunction with a convolutional neural network (CNN) for feature extraction [38]. While no improvements are shown for gaze estimates in the screen-space, results on smooth pursuits sequence of the EYEDIAP dataset [17] are encouraging. In [51], a top-down approach for gaze signal filtering is presented, where a probabilistic estimate of state (fixation, saccade, or smooth pursuits) is initially made, and consequently a state-specific linear dynamical system is applied to refine the initially predicted gaze. Improvements in gaze estimation performance are demonstrated on a custom dataset. As one of our evaluations, we re-confirm previous findings that a temporal gaze estimation network can improve on a static gaze estimation network. We demonstrate this on our novel video dataset, which due to its diversity of visual stimuli and large number of participants should allow for future works to benchmark their improvements well.

2.3 Refining Gaze Estimates

While eye gaze direction (and subsequent Point-of-Gaze) can be predicted just from images of the eyes or face of a given user, an initial estimate can be improved with additional data. Accordingly, various methods have been proposed to this end. A primary example is that of using few samples of labeled data - often dubbed "person-specific gaze estimation" - where a pre-trained neural network is fine-tuned or otherwise adapted on very few samples of a target test person's data, to yield performance improvements on the final test data from the same person. Building on top of initial works [28,42], more recent works have demonstrated significant performance improvements with as few as 9 calibration samples or less [33,40,32,56,8]. Although the performance improvements are impressive, all such methods still require labeled samples from the final user.

Alternative approaches to refining gaze estimates in the screen-based setting, consider the predicted visual saliency of the screen content. Given a sufficient time horizon, it is possible to align estimates for PoG so-far, with an estimate for visual saliency [46,48,7,1,52]. However, visual saliency estimation methods can over-fit to presented training data. Hence, methods have been suggested to merge estimates of multiple saliency models [45] or use face positions as likely

(a) Collection Setup (b) Sample frames (each 1080p)

Fig. 1: EVE data collection setup and example of (undistorted) frames collected from the 4 camera views with example eye patches shown as insets.

gaze targets [46]. We propose an alternate and direct approach, which formulates the problem of gaze refinement as one that is conditioned explicitly on screen content and an initial gaze estimate.

3 The EVE Dataset

To study the semantic relations and temporal dynamics between eye gaze and visual content, we identify a need for a new gaze dataset that:

1. allows for the training and evaluation of temporal models on natural eye movements (including fixations, saccades, and smooth pursuits),
2. enables the training of models that can process full camera frame inputs to yield screen-space Point-of-Gaze (PoG) estimates,
3. and provide a community-standard benchmark for a good understanding of the generalization capabilities of upcoming methods.

Furthermore, we consider the fact that the distribution of visual saliency on a computer screen at a given time is indicative of likely gaze positions. In line with this observation, prior work reports difficulty in generalization when considering saliency estimation and gaze estimation as separate components [46,45]. Thus, we define following further requirements for our new dataset:

1. a video of the screen content synchronized with eye gaze data,
2. a sufficiently large set of visual stimuli must be presented to allow for algorithms to generalize better without over-fitting to a few select stimuli,
3. and lastly, gaze data must be collected over time without instructing participants to gaze at specific pin-point targets such that they act naturally, like behaviours in a real-world setting.

We present in this section the methodologies we adopt to construct such a dataset, and briefly describe its characteristics. We call our proposed dataset "EVE", which stands for *"a dataset for enabling progress towards truly **E**nd-to-end **V**ideo-based **E**ye-tracking algorithms"*.

3.1 Captured Data

The minimum requirements for constructing our proposed dataset is the captured video from a webcam, gaze ground truth data from a commercial eye tracker, and screen frames from a given display. Furthermore, we:

- use the Tobii Pro Spectrum eye tracker, which reports high accuracy and precision in predicted gaze[1] even in the presence of natural head movements,
- add a high performance Basler Ace acA1920-150uc machine vision camera with global shutter, running at 60Hz,
- install three Logitech C922 webcams (30Hz) for a wider eventual coverage of head orientations, assuming that the final user will not only be facing the screen in a fully-frontal manner (see Fig. 1b),
- and apply MidOpt BP550 band-pass filters to all webcams and machine vision camera to remove reflections and glints on eyeglass and cornea surfaces due to the powerful near-infra-red LEDs used by the Tobii eye tracker.

All video camera frames are captured at 1920×1080 pixels resolution, but the superior timestamp-reliability and image quality of the Basler camera is expected to yield better estimates of gaze compared to the webcams.

The data captured by the Tobii Pro Spectrum eye tracker can be of very high quality which is subject to participant and environment effects. Hence to ensure data quality and reliability, an experiment coordinator is present during every data collection session to qualitatively assess eye tracking data via a live-stream of camera frames and eye movements. Additional details on our hardware setup and steps we take to ensure the best possible eye tracking calibration and subsequent data quality are described in the supplementary materials.

3.2 Presented Visual Stimuli

A large variety of visual stimuli are presented to our participants. Specifically, we present image, video, and wikipedia page stimuli (shown later in Fig. 4).

For static image stimuli, we select the widely used MIT1003 dataset [24] originally created for the task of image saliency estimation. Most images in the dataset span 1024 pixels in either horizontal or vertical dimensions. We randomly scale the image between 1320 and 1920 pixels in width or 480 to 1080 pixels in height, to be displayed on our 25-inch screen (with a resolution of 1080p).

All video stimuli are displayed in 1080p resolution (to span the full display), and taken from the DIEM [37], VAGBA [30], and Kurzhals et al. [29] datasets. These datasets consist of 720p, 1080p, and 1080p videos respectively, and thus are of high-resolution compared to other video-based saliency datasets. DIEM consists of various videos sampled from public repositories such as trailers and documentaries. VAGBA includes human movement or interactions in everyday scenes, and Kurzhals et al. contain purposefully designed video sequences with intentionally-salient regions. To increase the variety of the final set of video stimuli further, we select 23 videos from Wikimedia (at 1080p resolution).

[1] See https://www.tobiipro.com/pop-ups/accuracy-and-precision-test-report-spectrum/?v=1.1

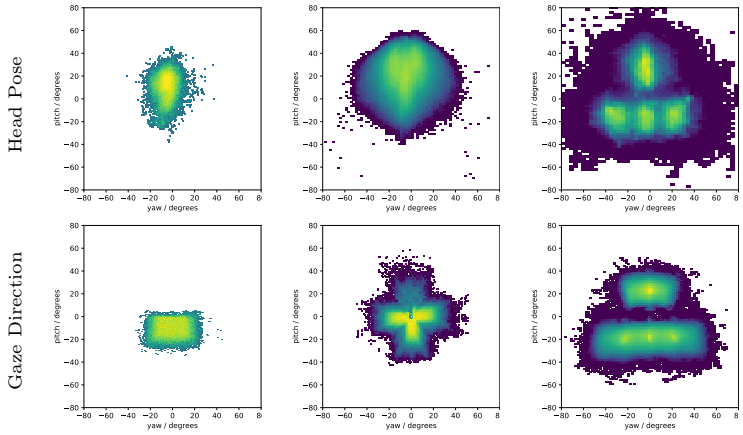


Fig. 2: Head orientation and gaze direction distributions are compared with existing screen-based gaze datasets [61,28]. We capture a larger range of parameter space due to a multi-view camera setup and 25-inch display. 2D histogram plot values are normalized and colored with log-scaling.

Wikipedia pages are randomly selected on-the-fly by opening the following link in a web browser: `https://en.m.wikipedia.org/wiki/Special:Random#/random` and participants are then asked to freely view and navigate the page, as well as to click on links. Links leading to pages outside of Wikipedia are automatically removed using the GreaseMonkey web browser extension.

In our data collection study, we randomly sample the image and video stimuli from the mentioned datasets. We ensure that each participant observes 60 image stimuli (for three seconds each), at least 12 minutes of video stimuli, and six minutes of wikipedia stimulus (three 2-minute sessions). At the conclusion of data collection, we found that each image stimulus has been observed 3.35 times ($SD = 0.73$), and each video stimulus has been observed 9.36 times ($SD = 1.28$).

3.3 Dataset Characteristics

The final dataset is collected from 54 participants (30 male, 23 female, 1 unspecified). The details of responses to our demographics questionnaire can be found in our supplementary materials along with how we pre-process the dataset. We ensure that the subjects in both training and test sets exhibit diverse gender, age, and ethnicity, some with and some without glasses.

In terms of gaze direction and head orientation distributions, EVE compares favorably to popular screen-based datasets such as MPIIFaceGaze [61] and GazeCapture [28]. Figure 2 shows that we cover a larger set of gaze directions and head poses. This is likely due to the 4 camera views that we adopt, together with a large screen size of 25 inches (compared to the other datasets).

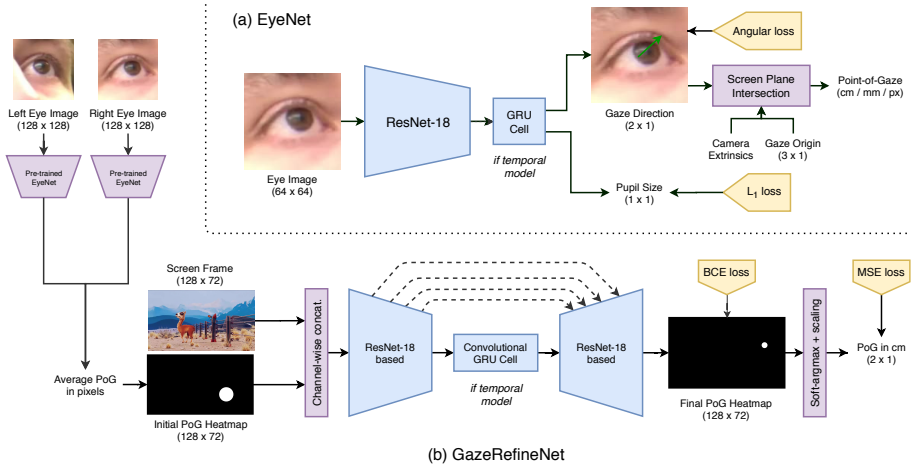


Fig. 3: We adopt (a) a simple EyeNet architecture for gaze direction and pupil size estimation with an optional recurrent component, and propose (b) a novel GazeRefineNet architecture for label-free PoG refinement using screen content.

4 Method

We now discuss a novel architecture designed to exploit the various sources of information in datasets and to serve as baseline for follow-up work. We first introduce a simple eye gaze estimation network ($\text{EyeNet}_{\text{static}}$) and its recurrent counterparts ($\text{EyeNet}_{\text{RNN}}$, $\text{EyeNet}_{\text{LSTM}}$, $\text{EyeNet}_{\text{GRU}}$) for the task of per-frame or temporal gaze and pupil size estimation (see Fig. 3a). As the EVE dataset contains synchronized visual stimuli, we propose a novel technique to process these initial eye-gaze predictions further by taking the raw screen content directly into consideration. To this end, we propose the GazeRefineNet architecture (Fig. 3b), and describe its details in the second part of this section.

4.1 EyeNet Architecture

Learning-based eye gaze estimation models typically output their predictions as a unit direction vector or in Euler angles in the spherical coordinate system. The common metric to evaluate the goodness of predicted gaze directions is via an angular distance error metric in degrees. Assuming that the predicted gaze direction is represented by a 3-dimensional unit vector $\hat{\mathbf{g}}$, the calculation of the angular error loss when given ground-truth $\mathbf{g}$ is then:

$$\mathcal{L}_{\text{gaze}}\left(\mathbf{g}, \hat{\mathbf{g}}\right) = \frac{1}{NT} \sum^{N} \sum^{T} \frac{180}{\pi} \arccos\left(\frac{\mathbf{g} \cdot \hat{\mathbf{g}}}{\|\mathbf{g}\| \|\hat{\mathbf{g}}\|}\right) \tag{1}$$

where a mini-batch consists of N sequences each of length T.

To calculate PoG, the predicted gaze direction must first be combined with the 3D gaze origin position $\mathbf{o}$ (determined during data pre-processing), yielding a gaze ray with 6 degrees of freedom. We can then intersect this ray with the screen plane to calculate the PoG by using the camera transformation with respect to the screen plane. Pixel dimensions (our 1920×1080 screen is 553mm wide and 311mm tall) can be used to convert the PoG to pixel units for an alternative interpretation. We denote the predicted PoG in centimeters as $\hat{\mathbf{s}}$.

Assuming that the pupil size can be estimated, we denote it as $\hat{\mathbf{p}}$ and define an ℓ_1 loss given ground-truth $\mathbf{p}$ as:

$$\mathcal{L}_{\text{pupil}}\left(\mathbf{p},\, \hat{\mathbf{p}}\right) = \frac{1}{NT} \sum^{N} \sum^{T} \left\|\mathbf{p} - \hat{\mathbf{p}}\right\|_1 \tag{2}$$

The two values of gaze direction and pupil size are predicted by a ResNet-18 architecture [18]. To make the network recurrent, we optionally incorporate a RNN [49], LSTM [19], or GRU [11] cell.

4.2 GazeRefineNet Architecture

Given the left and right eye images $\mathbf{x}_l$ and $\mathbf{x}_r$ of a person, we hypothesize that incorporating the corresponding screen content can improve the initial PoG estimate. Provided that an initial estimate of PoG $\tilde{\mathbf{s}} = f\left(\mathbf{x}\right)$ can be made for the left and right eyes $\tilde{\mathbf{s}}_l$ and $\tilde{\mathbf{s}}_r$ respectively, we first take the average of the predicted PoG values with $\tilde{\mathbf{s}} = \frac{1}{2}\left(\tilde{\mathbf{s}}_l + \tilde{\mathbf{s}}_r\right)$ to yield a single estimate of gaze. Here f denotes the previously described EyeNet. We define and learn a new function, $\mathbf{s} = g\left(\mathbf{x}_S, \tilde{\mathbf{s}}\right)$, to refine the EyeNet predictions by incorporating the screen content and temporal information. The function g is parameterized by a fully convolutional neural network (FCN) to best preserve spatial information. Following the same line of reasoning, we represent our initial PoG estimate $\tilde{\mathbf{s}}$ as a confidence map. More specifically, we use an isotropic 2D Gaussian function centered at the estimated gaze position on the screen. The inputs to the FCN are concatenated channel-wise.

To allow the model to better exploit the temporal information, we use an RNN cell in the bottleneck. Inspired by prior work in video-based saliency estimation, we adopt a convolutional recurrent cell [31] and evaluate RNN [49], LSTM [19], and GRU [11] variants.

The network optionally incorporate concatenative skip connections between the encoder and decoder layers, as this is shown to be helpful in FCNs. We train the GazeRefineNet by using pixel-wise binary cross-entropy loss on the output heatmap and MSE loss on the final numerical estimate of the PoG. It is calculated in a differentiable manner via a soft-argmax layer [6,20]. The PoG is converted to centimeters to keep the loss term from exploding (due to its magnitude). Please refer to Fig. 3b for the full architecture diagram, and our supplementary materials for implementation details.

Offset augmentation In the task of cross-person gaze estimation, it is common to observe high discrepancies between the training and validation objectives. This is not necessarily due to overfitting or non-ideal hyperparameter selections but rather due to the inherent nature of the problem. Specifically, every human has a person-specific offset between their optical and visual axes in each eye, often denoted by a so-called Kappa parameter. While the optical axis can be observed by the appearance of the iris, the visual axis cannot be observed at all as it is defined by the position of the fovea at the back of the eyeball.

During training, this offset is absorbed into the neural network's parameters, limiting generalization to unseen people. Hence, prior work typically incur a large error increase in cross-person evaluations ($\sim 5°$) in comparison to person-specific evaluations ($\sim 3°$). Our insight is that we are now posing a gaze refinement problem, where an initially incorrect assessment of offset could actually be corrected by additional signals such as that of screen content. This is in contrast with the conventional setting, where no such corrective signal is made available. Therefore, the network should be able to learn to overcome this offset when provided with randomly sampled offsets to a given person's gaze.

This randomization approach can intuitively be understood as learning to undo all possible inter-personal differences rather than learning the corrective parameters for a specific user, as would be the case in traditional supervised personalization (e.g., [40]). We dub our training data augmentation approach as an *"offset augmentation"*, and provide further details of its implementation in our supplementary materials.

5 Results

In this section, we evaluate the variants of EyeNet and find that temporal modelling can aid in gaze estimation. Based on a pre-trained $\text{EyeNet}_{\text{GRU}}$, we then evaluate the effects of our contributions in refining an initial estimate of PoG using variants of GazeRefineNet. We demonstrate large and consistent performance improvements even across camera views and visual stimulus types.

5.1 Eye Gaze Estimation

We first consider the task of eye gaze estimation purely from a single eye image patch. Tab. 2 shows the performance of the static $\text{EyeNet}_{\text{static}}$ and its temporal variants ($\text{EyeNet}_{\text{RNN}}$, $\text{EyeNet}_{\text{LSTM}}$, $\text{EyeNet}_{\text{GRU}}$) on predicting gaze direction, PoG, and pupil size. The networks are trained on the training split of EVE. Generally, we find our gaze direction error values to be in line with prior works in estimating gaze from single eye images [60], and see that the addition of recurrent cells improve gaze estimation performance modestly. This makes a case for training gaze estimators on temporal data, using temporally-aware models, and corroborates observations from a prior learning-based gaze estimation approach on natural eye movements [51].

Table 2: Cross-person gaze estimation and pupil size errors of EyeNet variants, evaluated on the test set of EVE. The GRU variant performs best in terms of both gaze and pupil size estimates

Model	Left Eye				Right Eye			
	Gaze Dir. (°)	PoG (cm)	PoG (px)	Pupil Size (mm)	Gaze Dir. (°)	PoG (cm)	PoG (px)	Pupil Size (mm)
$\text{EyeNet}_{\text{static}}$	4.54	5.10	172.7	**0.29**	4.75	5.29	181.0	**0.29**
$\text{EyeNet}_{\text{RNN}}$	4.33	4.86	166.7	**0.29**	4.91	5.48	186.5	**0.28**
$\text{EyeNet}_{\text{LSTM}}$	4.17	4.66	161.0	0.32	**4.71**	**5.25**	**180.5**	0.33
$\text{EyeNet}_{\text{GRU}}$	**4.11**	**4.60**	**158.5**	**0.28**	4.80	5.33	183.9	**0.29**

Table 3: An ablation study of our contributions in GazeRefineNet, where a frozen and pre-trained $\text{EyeNet}_{\text{GRU}}$ is used for initial gaze predictions. Temporal modelling and our novel offset augmentation both yield large gains in performance.

Model	Screen Content	Offset Augmen.	Skip Conn.	Gaze Dir. (°)	PoG (cm)	PoG (px)
Baseline ($\text{EyeNet}_{\text{GRU}}$)				3.48	3.85	132.56
	o			3.33	3.67	127.59
	o	o		2.80	3.09	107.42
$\text{GazeRefineNet}_{\text{static}}$	o	o	o	2.87	3.16	109.85
	o	o		2.67	2.95	102.36
$\text{GazeRefineNet}_{\text{RNN}}$	o	o	o	2.57	2.83	98.38
	o	o		**2.49**	**2.75**	**95.43**
$\text{GazeRefineNet}_{\text{LSTM}}$	o	o	o	2.53	2.79	96.97
	o	o		2.51	2.77	96.24
$\text{GazeRefineNet}_{\text{GRU}}$	o	o	o	**2.49**	**2.75**	**95.59**

Pupil size errors are presented in terms of mean absolute error. Considering that the size of pupils in our dataset vary from 2mm to 4mm, the presented errors of 0.3mm should allow for meaningful insights to be made in fields such as the cognitive sciences. We select the GRU variant ($\text{EyeNet}_{\text{GRU}}$) for the next steps as it shows consistently good performance for both eyes.

5.2 Screen Content based Refinement of PoG

GazeRefineNet consists of a fully-convolutional architecture which takes as input a screen content frame, and an offset augmentation procedure at training time. Our baseline performance for this experiment is different to Tab. 2 as gaze errors are improved when averaging the PoG from the left and right eyes, with according adjustments to the label (averaged in screen space). Even with the new competitive baseline from PoG averaging, we find in Tab. 3 that each of our additional contributions yield large performance improvements, amounting to a

Table 4: Improvement in PoG prediction (in px) of our method in comparison with two saliency-based alignment methods, as evaluated on the EVE dataset.

Method \ Stimulus Type	Image	Video	Wikipedia
Saliency-based (scale + bias)	78.4 ↓36.3%	116.7 ↓12.0%	198.3 ↑43.6%
Saliency-based (kappa)	75.0 ↓39.2%	110.9 ↓17.0%	258.0 ↑84.4%
GazeRefineNet$_{\text{GRU}}$ (Ours)	**48.7** ↓60.4%	**96.7** ↓27.1%	**116.3** ↓15.8%

Table 5: Final gaze direction errors (in degrees, lower is better) from the output of GazeRefineNet$_{\text{GRU}}$, evaluated on the EVE test set in cross-stimuli settings. Indicated improvements are with respect to initial PoG predictions (mean of left+right) from EyeNet$_{\text{GRU}}$ trained on specified source stimuli types.

Source \ Target	Images	Videos	Wikipedia
Images	**1.30** ↓60.55%	3.60 ↑4.10%	4.74 ↑30.13%
Videos	1.97 ↓40.09%	**2.60** ↓24.88%	3.71 ↑1.94%
Wikipedia	2.12 ↓35.75%	3.32 ↓3.84%	**3.04** ↓16.62%

28% improvement in gaze direction error, reducing it to 2.49°. While not directly comparable due to differences in setting, this value is lower even than recently reported performances of supervised few-shot adaptation approaches on in-the-wild datasets [40,32]. Specifically, we find that the offset augmentation procedure yields the greatest performance improvements, with temporal modeling further improving performance. Skip connections between the encoder and decoder do not necessarily help (except in the case of GazeRefineNet$_{\text{RNN}}$), presumably because the output relies mostly on information processed at the bottleneck. We present additional experiments of GazeRefineNet in the following paragraphs, and describe their setup details in our supplementary materials.

Comparison to Saliency-based Methods. In order to assess how our GazeRefineNet approach compares with existing saliency-based methods, we implement two up-to-date methods loosely based on [1] and [52]. First, we use the state-of-the-art UNISAL approach [13] to attain high quality visual saliency predictions. We accumulate these predictions over time for the full exposure duration of each visual stimulus in EVE (up to 2 minutes), which should provide the best context for alignment (as opposed to our online approach, which is limited to 3 seconds of history). Standard back propagation is then used to optimize for either scale and bias in screen-space (similar to [1]) or the visual-optical axis offset, kappa (similar to [52]) using a KL-divergence objective between accumulated visual saliency predictions and accumulated heatmaps of refined gaze estimates in the screen space. Tab. 4 shows that while both saliency-based base-

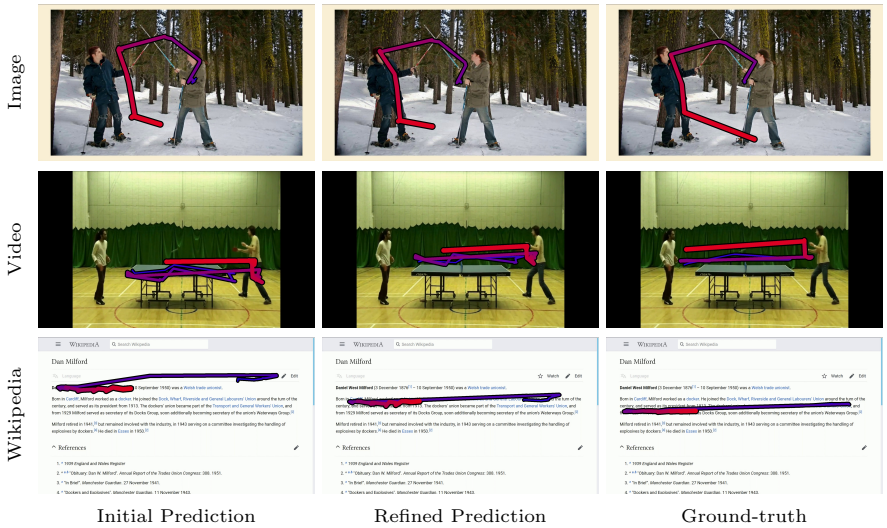


Fig. 4: Qualitative results of our gaze refinement method on our test set, where PoG over time are colored from blue-to-red (old-to-new). It can be seen that GazeRefineNet corrects offsets between the initial prediction and ground-truth.

lines perform respectably on the well-studied image and video stimuli, they fail completely on wikipedia stimuli despite the fact that the saliency estimation model was provided with full 1080p frames (as opposed to the 128×72 input used by GazeRefineNet$_{\text{GRU}}$). Furthermore, our direct approach takes raw screen pixels and gaze estimations up to the current time-step as explicit conditions and thus is a simpler yet explicit solution for live gaze refinement that can be learned end-to-end. Both the training of our approach and its large-scale evaluation is made possible by the EVE, which should allow for insightful comparisons in the future.

Cross-Stimuli Evaluation. We study if our method generalizes to novel stimuli types, as this has previously been raised as in issue for saliency-based gaze alignment methods (such as in [45]). In Tab. 5, we confirm that indeed training and testing on the same stimulus type yields the greatest improvements in gaze direction estimation (shown in diagonal of table). We find in general that large improvements can be observed even when training solely on video or wikipedia stimuli types. One assumes that this is the case due to the existence of text in our video stimuli and the existence of small amounts of images in the wikipedia stimulus. In contrast, we can see that training a model on static images only does not lead to good generalization on the stimuli types.

Qualitative Results. We visualize our results qualitatively in Fig. 4. Specifically, we can see that when provided with initial estimates of PoG over time from EyeNet$_{\text{GRU}}$ (far-left column), our GazeRefineNet$_{\text{GRU}}$ can nicely recover person-

specific offsets at test time to yield improved estimates of PoG (center column). When viewed in comparison with the ground-truth (far-right column), the success of GazeRefineNet$_{\text{GRU}}$ in these example cases is clear. In addition, note that the final operation is not one of pure offset-correction, but that the gaze signal is more aligned with the visual layout of the screen content post-refinement.

6 Conclusion

In this paper, we introduced several effective steps towards increasing screen-based eye-tracking performance even in the absence of labeled samples or eye-tracker calibration from the final target user. Specifically, we identified that eye movements and the change in visual stimulus have a complex interplay which previous literature have considered in a disconnected manner. Subsequently, we proposed a novel dataset (EVE) for evaluating temporal gaze estimation models and for enabling a novel online PoG-refinement task based on raw screen content. Our GazeRefineNet architecture performs this task effectively, and demonstrates large performance improvements of up to 28%. The final reported angular gaze error of 2.49° is achieved without labeled samples from the test set.

The EVE dataset is made publicly available[2], with a public web server implemented for consistent test metric calculations. We provide the dataset and accompanying training and evaluation code in hopes of further progress in the field of remote webcam-based gaze estimation. Comprehensive additional information regarding the capture, pre-processing, and characteristics of the dataset is made available in our supplementary materials.

Acknowledgements

We thank the participants of our dataset for their contributions, our reviewers for helping us improve the paper, and Jan Wezel for helping with the hardware setup. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme grant agreement No. StG-2016-717054.

[2] https://ait.ethz.ch/projects/2020/EVE

References

1. Alnajar, F., Gevers, T., Valenti, R., Ghebreab, S.: Calibration-free gaze estimation using human gaze patterns. In: ICCV (December 2013)
2. Balajee Vasudevan, A., Dai, D., Van Gool, L.: Object referring in videos with language and human gaze. In: CVPR. pp. 4129–4138 (2018)
3. Baluja, S., Pomerleau, D.: Non-Intrusive Gaze Tracking Using Artificial Neural Networks. In: NeurIPS. pp. 753–760 (1993)
4. Biedert, R., Buscher, G., Schwarz, S., Hees, J., Dengel, A.: Text 2.0. In: ACM CHI EA (2010)
5. 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)
6. Chapelle, O., Wu, M.: Gradient descent optimization of smoothed information retrieval metrics. Information retrieval **13**(3), 216–235 (2010)
7. Chen, J., Ji, Q.: Probabilistic gaze estimation without active personal calibration. In: CVPR. pp. 609–616 (2011)
8. Chen, Z., Shi, B.: Offset calibration for appearance-based gaze estimation via gaze decomposition. In: WACV (March 2020)
9. Cheng, Y., Lu, F., Zhang, X.: Appearance-based gaze estimation via evaluation-guided asymmetric regression. In: ECCV (2018)
10. Chong, E., Ruiz, N., Wang, Y., Zhang, Y., Rozga, A., Rehg, J.M.: Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In: ECCV (2018)
11. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. In: NeurIPS Workshop on Deep Learning (2014)
12. Deng, H., Zhu, W.: Monocular free-head 3d gaze tracking with deep learning and geometry constraints. In: ICCV. pp. 3143–3152 (2017)
13. Droste, R., Jiao, J., Noble, J.A.: Unified Image and Video Saliency Modeling. In: ECCV (2020)
14. Feit, A.M., Williams, S., Toledo, A., Paradiso, A., Kulkarni, H., Kane, S.K., Morris, M.R.: Toward everyday gaze input: Accuracy and precision of eye tracking and implications for design. In: ACM CHI. pp. 1118–1130 (2017)
15. Fischer, T., Chang, H.J., Demiris, Y.: RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments. In: ECCV (2018)
16. Fridman, L., Reimer, B., Mehler, B., Freeman, W.T.: Cognitive load estimation in the wild. In: ACM CHI (2018)
17. Funes Mora, K.A., Monay, F., Odobez, J.M.: Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In: ACM ETRA. ACM (Mar 2014)
18. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: ICCV (2015)
19. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)
20. Honari, S., Molchanov, P., Tyree, S., Vincent, P., Pal, C., Kautz, J.: Improving landmark localization with semi-supervised learning. In: CVPR (2018)
21. Huang, M.X., Kwok, T.C., Ngai, G., Chan, S.C., Leong, H.V.: Building a personalized, auto-calibrating eye tracker from user interactions. In: ACM CHI. p. 5169–5179. New York, NY, USA (2016)

22. Huang, Q., Veeraraghavan, A., Sabharwal, A.: Tabletgaze: Dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. Machine Vision Applications **28**(5-6), 445–461 (Aug 2017)
23. 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)
24. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: ICCV. pp. 2106–2113. IEEE
25. Karessli, N., Akata, Z., Schiele, B., Bulling, A.: Gaze embeddings for zero-shot image classification. In: CVPR (2017)
26. Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., Torralba, A.: Gaze360: Physically unconstrained gaze estimation in the wild. In: ICCV (October 2019)
27. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
28. Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., Torralba, A.: Eye Tracking for Everyone. In: CVPR (2016)
29. Kurzhals, K., Bopp, C.F., Bässler, J., Ebinger, F., Weiskopf, D.: Benchmark data for evaluating visualization and analysis techniques for eye tracking for video stimuli. In: Proceedings of the fifth workshop on beyond time and errors: novel evaluation methods for visualization. pp. 54–60 (2014)
30. Li, Z., Qin, S., Itti, L.: Visual attention guided bit allocation in video compression. Image and Vision Computing **29**(1), 1–14 (2011)
31. Linardos, P., Mohedano, E., Nieto, J.J., O'Connor, N.E., Giro-i Nieto, X., McGuinness, K.: Simple vs complex temporal recurrences for video saliency prediction. In: BMVC (2019)
32. Lindén, E., Sjostrand, J., Proutiere, A.: Learning to personalize in appearance-based gaze tracking. In: ICCVW. pp. 0–0 (2019)
33. Liu, G., Yu, Y., Mora, K.A.F., Odobez, J.: A differential approach for gaze estimation with calibration. In: BMVC (2018)
34. Lu, F., Okabe, T., Sugano, Y., Sato, Y.: A head pose-free approach for appearance-based gaze estimation. In: BMVC (2011)
35. Lu, F., Sugano, Y., Okabe, T., Sato, Y.: Inferring human gaze from appearance via adaptive linear regression. In: ICCV (2011)
36. Martinikorena, I., Cabeza, R., Villanueva, A., Porta, S.: Introducing i2head database. In: PETMEI. pp. 1–7 (2018)
37. Mital, P.K., Smith, T.J., Hill, R.L., Henderson, J.M.: Clustering of gaze during dynamic scene viewing is predicted by motion. Cognitive computation **3**(1), 5–24 (2011)
38. Palmero, C., Selva, J., Bagheri, M.A., Escalera, S.: Recurrent cnn for 3d gaze estimation using appearance and shape cues. In: BMVC (2018)
39. Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., Hays, J.: Webgazer: Scalable webcam eye tracking using user interactions. In: IJCAI. pp. 3839–3845 (2016)
40. Park, S., Mello, S.D., Molchanov, P., Iqbal, U., Hilliges, O., Kautz, J.: Few-shot adaptive gaze estimation. In: ICCV (2019)
41. Park, S., Spurr, A., Hilliges, O.: Deep Pictorial Gaze Estimation. In: ECCV (2018)
42. Park, S., Zhang, X., Bulling, A., Hilliges, O.: Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In: ACM ETRA (2018)
43. Ranjan, R., Mello, S.D., Kautz, J.: Light-weight head pose invariant gaze tracking. In: CVPRW (2018)

44. Smith, B., Yin, Q., Feiner, S., Nayar, S.: Gaze Locking: Passive Eye Contact Detection for Human-Object Interaction. In: ACM UIST. pp. 271–280 (Oct 2013)
45. Sugano, Y., Bulling, A.: Self-calibrating head-mounted eye trackers using egocentric visual saliency. In: ACM UIST. p. 363–372. New York, NY, USA (2015)
46. Sugano, Y., Matsushita, Y., Sato, Y.: Calibration-free gaze sensing using saliency maps. In: CVPR. pp. 2667–2674 (2010)
47. Sugano, Y., Matsushita, Y., Sato, Y.: Learning-by-Synthesis for Appearance-based 3D Gaze Estimation. In: CVPR (2014)
48. Sugano, Y., Matsushita, Y., Sato, Y., Koike, H.: An incremental learning method for unconstrained gaze estimation. In: ECCV. pp. 656–667. Springer (2008)
49. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: NeurIPS. pp. 3104–3112 (2014)
50. 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)
51. Wang, K., Su, H., Ji, Q.: Neuro-inspired eye tracking with eye movement dynamics. In: CVPR. pp. 9831–9840 (2019)
52. Wang, K., Wang, S., Ji, Q.: Deep eye fixation map learning for calibration-free eye gaze tracking. In: ACM ETRA. p. 47–55. New York, NY, USA (2016)
53. Wang, K., Zhao, R., Ji, Q.: A hierarchical generative model for eye image synthesis and eye gaze estimation. In: CVPR (2018)
54. Wang, K., Zhao, R., Su, H., Ji, Q.: Generalizing eye tracking with bayesian adversarial learning. In: CVPR. pp. 11907–11916 (2019)
55. 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)
56. Yu, Y., Liu, G., Odobez, J.M.: Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In: CVPR. pp. 11937–11946 (2019)
57. Yu, Y., Odobez, J.M.: Unsupervised representation learning for gaze estimation. In: CVPR (June 2020)
58. 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)
59. Zhang, X., Sugano, Y., Bulling, A.: Revisiting data normalization for appearance-based gaze estimation. In: ETRA (2018)
60. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: CVPR (2015)
61. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: It's written all over your face: Full-face appearance-based gaze estimation. In: CVPRW (2017)
62. 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`

Appendix

A 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.

A.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[3] and after the hour-long session, were paid a fee of 25 Swiss Francs in cash.

A.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. 5.

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.

A.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 [50], 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.

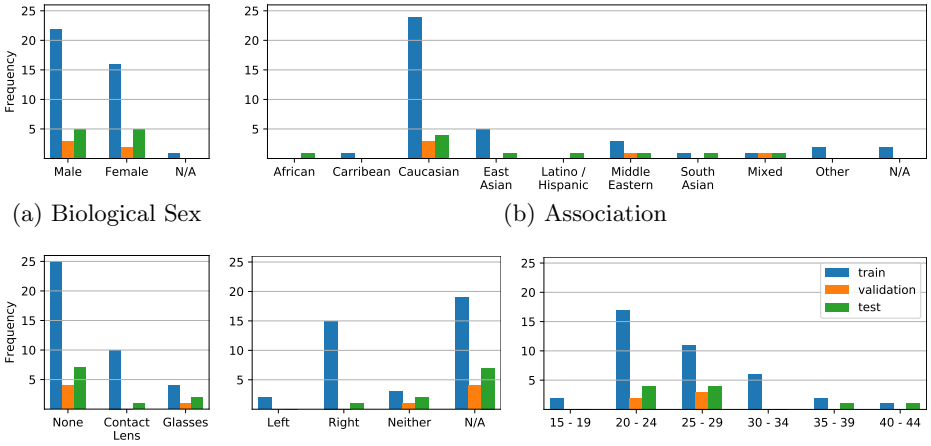
[3] `https://marktplatz.uzhalumni.ch/`

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

Fig. 5: 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.

In processing the video of participants, we first undistorted the frames' pixels and detected the face [58] and face-region landmarks [5]. We then performed a 3D morphable model (3DMM) fit to the detected 3D facial landmarks [23] 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 [47,59]. The final eye patches are 128×128 in size and cre-



(a) Biological Sex (b) Association

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

Fig. 6: 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.

ated 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.

A.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. 6. 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. 6b 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.

We find that the points-of-gaze (PoGs) in our dataset exhibit a screen-center-bias as previously reported in saliency literature [24] (see Fig. 7). 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. D.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. 7).

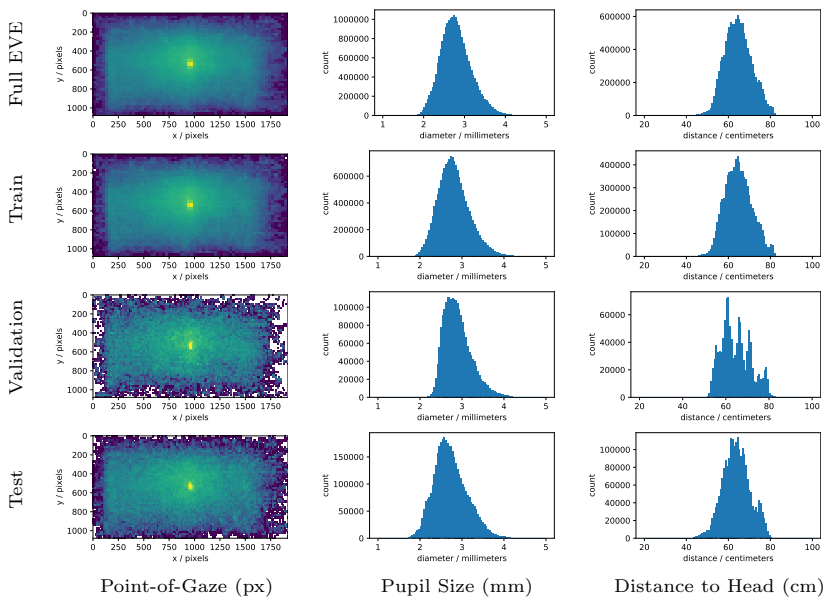


Fig. 7: PoG (on-screen pixels), distance (in cm) and pupil size (in mm) distributions for the defined subsets of the proposed EVE dataset. The number of people involved are 54, 39, 5, 10 respectively for the full dataset, training subset, validation subset, and test subsets. The 2D histograms are coloured with a logarithmic scale, with values normalized by the size of the subset in concern.

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. A.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.

B 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 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 θ

is pitch, and ϕ is yaw. Then the unit-vector notation of $\hat{\mathbf{g}} = (\theta, \phi)$ would be calculated with,

$$\hat{\mathbf{v}}_h = \begin{pmatrix} -\cos\theta \sin\phi \\ -\sin\theta \\ -\cos\theta\cos\phi \end{pmatrix}. \tag{3}$$

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:

$$\hat{\mathbf{v}}_h = \mathbf{R}_h^T \hat{\mathbf{v}}_c. \tag{4}$$

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

$$\hat{\mathbf{g}}_h = \begin{pmatrix} \theta_h \\ \phi_h \end{pmatrix} = \begin{pmatrix} \arcsin -\hat{y}_h \\ \arctan2\left(-\hat{x}_h, -\hat{z}_h\right) \end{pmatrix}, \tag{5}$$

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

$$\mathbf{R}_h = \begin{pmatrix} \cos\phi_h & 0 & \sin\phi_h \\ 0 & 1 & 0 \\ -\sin\phi_h & 0 & \cos\phi_h \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_h & -\sin\theta_h \\ 0 & \sin\theta_h & \cos\theta_h \end{pmatrix}. \tag{6}$$

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 $\kappa_i = (\theta_\kappa, \phi_\kappa) \sim \mathcal{N}(0, 3^\circ)$ 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,

$$\hat{\mathbf{v}}_h^{\text{aug}} = \mathbf{R}_h \hat{\mathbf{v}}_\kappa. \tag{7}$$

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

C Implementation Details

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

C.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.

C.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:

$$\mathcal{L}_{\text{EyeNet}} = \gamma_{\text{PoG}}\mathcal{L}_{\text{gaze}} + \gamma_{\text{pupil}}\mathcal{L}_{\text{pupil}}, \tag{8}$$

and set $\gamma_{\text{gaze}} = 1.0$ and $\gamma_{\text{pupil}} = 1.0$ empirically. The EyeNet is trained using the Adam optimizer [27] 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 128×128 pixels large.

C.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:

$$\mathcal{L}_{\text{RefineNet}} = \gamma_{\text{PoG}}\mathcal{L}_{\text{PoG}} + \gamma_{\text{XE}}\mathcal{L}_{\text{XE}}. \tag{9}$$

where we set $\gamma_{\text{PoG}} = 0.001$ and $\gamma_{\text{XE}} = 1.0$ empirically. The GazeRefineNet is trained using the Adam optimizer [27] 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 128×72 pixels large. Please note that during this stage of training, the EyeNet weights are not updated.

Table 6: Experiments where the $\text{EyeNet}_{\text{static}}$ is trained on the GazeCapture dataset [28]. 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%

D 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.

D.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, 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.

D.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

Table 7: 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.

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 [28] 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. 6 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 [62,55]. 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).

D.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. 7 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.

Table 8: 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 $\text{EyeNet}_{\text{GRU}}$.

D.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 8, 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).

D.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 $\text{GazeRefineNet}_{\text{GRU}}$ trained with strong offset augmentation may perform overly aggressive corrections. We show this trade-off in Fig. 8 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

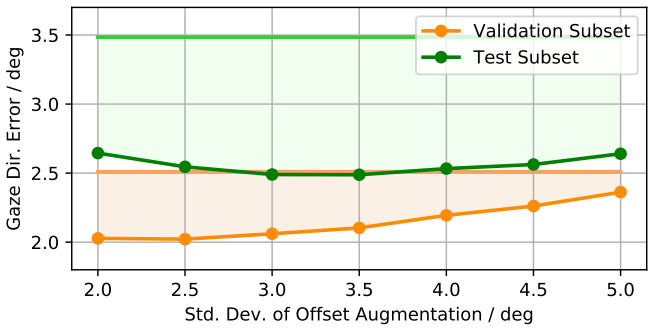


Fig. 8: Varying σ_κ results in differing performance improvements on the validation subset of the EVE dataset, compared to that on the test subset. Specifically, the test subset is significantly more challenging and thus a stronger amount of offset augmentation is required than in the case of the validation subset.

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.

E 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).