ETH-XGaze: A Large Scale Dataset for Gaze Estimation under Extreme Head Pose and Gaze Variation

Xucong Zhang[1], Seonwook Park[1], Thabo Beeler[2], Derek Bradley, Siyu Tang[1] and Otmar Hilliges[1]

[1] Department of Computer Science, ETH Zurich
{xucong.zhang, spark, siyu.tang, otmar.hilliges}@inf.ethz.ch
[2] Google Inc.
tbeeler@google.com

Abstract. Gaze estimation is a fundamental task in many applications of computer vision, human computer interaction and robotics. Many state-of-the-art methods are trained and tested on custom datasets, making comparison across methods challenging. Furthermore, existing gaze estimation datasets have limited head pose and gaze variations, and the evaluations are conducted using different protocols and metrics. In this paper, we propose a new gaze estimation dataset called ETH-XGaze, consisting of over one million high-resolution images of varying gaze under extreme head poses. We collect this dataset from 110 participants with a custom hardware setup including 18 digital SLR cameras and adjustable illumination conditions, and a calibrated system to record ground truth gaze targets. We show that our dataset can significantly improve the robustness of gaze estimation methods across different head poses and gaze angles. Additionally, we define a standardized experimental protocol and evaluation metric on ETH-XGaze, to better unify gaze estimation research going forward. The dataset and benchmark website are available at https://ait.ethz.ch/projects/2020/ETH-XGaze

1 Introduction

Estimating eye-gaze from monocular images alone has recently received significant interest in computer vision [35,38,9] due to its significance in many application domains ranging from the cognitive sciences and HCI to robotics and semi-autonomous driving [7,23,32]. Many arising computing paradigms such as smart-home appliances, autonomous cars and robots, as well as body-worn cameras will rely on understanding the attention and intent of humans without directly interacting with the observed person. We argue that in order to be more robust to a larger variety of environmental conditions, future methods should be able to accurately estimate the gaze of humans in a broader range of settings, including variation of viewpoint, extreme gaze angles, lighting variation, input image resolutions, and in the presence of occluders such as glasses.

Unfortunately, existing gaze datasets do not cater to such use-cases and are mostly limited to the frontal setting, covering a relatively narrow range of head poses and gaze directions. These are typically collected via laptops [42], mobile devices [19,13] or in stationary settings [10]. Recent work has moved towards more unconstrained environmental conditions in particular with respect to lighting but the coverage of head pose and gaze direction ranges remains limited [16,9,40].

In this paper we detail a new dataset, dubbed ETH-XGaze, to facilitate research into robust gaze estimation methods. The dataset exhaustively samples large variations in head poses, up to the limit of where both eyes are still visible (maximum $\pm 70^\circ$ from directly facing the camera) as well as comprehensive gaze directions (maximum $\pm 50^\circ$ in the head coordinate system) [27]. The dataset will allow for the development of new methods that can robustly estimate gaze direction without requiring a quasi-frontal camera placement. We show experimentally that i) the data distribution of ETH-XGaze is more comprehensive than other datasets (e.g., our dataset broadens the scope for eye-gaze research), and ii) that training on our dataset significantly improves robustness towards head pose and gaze direction variations. Beyond extending the gaze and head-pose ranges, the proposed dataset allocates considerably more pixels to the periocular region compared to existing datasets (e.g. refer to Fig. 3). This allows to train gaze estimators that can take advantage of the high-resolution imagery of modern camera hardware to improve gaze prediction. We collect data from 110 participants with different ethnicity, age, and gender – some with glasses and some without – in order to provide a rich and diverse dataset. For each of the participants we capture over 500 gaze directions with full-on illumination, plus an additional 90 samples under 15 different illumination conditions. This results in a total of over 1 million labeled samples. For all samples, the ground-truth gaze direction is known since the gaze is guided by stimuli displayed on a large screen in front of the participant, ensuring good label quality even under extreme view angles. The capture setup is depicted in Fig. 1 (left).

To ensure fair and systematic comparisons between future methods that leverage this new large-scale dataset, we also propose a standardized evaluation protocol. Unlike other fields in computer vision that have benefited from such benchmark frameworks (i.e. image classification [28], face recognition [24], full-body [15], hand pose estimation [43] and multiview stereo reconstruction [29]), the gaze estimation community has so far relied on a heterogeneous environment where many papers employ custom data pre-processing and evaluation protocols, rendering direct comparisons challenging. Motivated by the benchmarking approaches in adjacent areas we create a website open to the public to submit, evaluate and compare gaze estimation methods based on ETH-XGaze.

Finally, in order to provide initial insights into the value of our dataset, we provide results from a simple gaze estimation method that can serve as a baseline. Our estimation approach leverages a standard CNN architecture (i.e., ResNet-50 [11]), trained with the task of estimating gaze from a monocular face patch. We present the estimation results as well as an ablation study of training

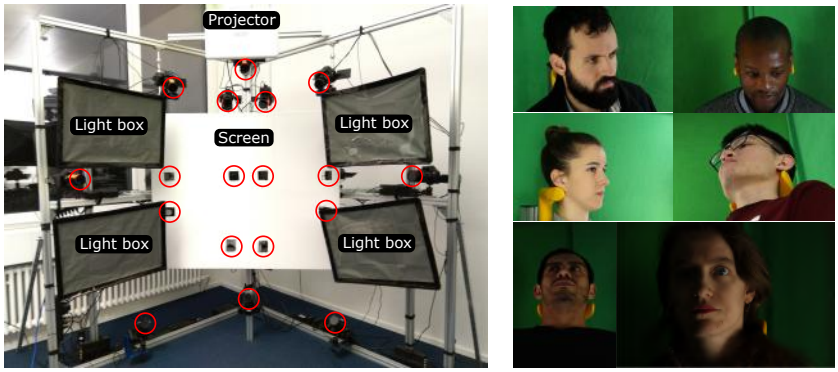

Fig. 1: Our data collection device includes 18 high-resolution Canon 250D digital SLR cameras (marked with red circles), a projector to project the stimuli on the screen, and four Walimex Daylight 250 light boxes. A chair with a head rest is positioned approximately one meter away from the screen. Captured samples under different head poses and lighting conditions are shown on the right.

on different subsets of our dataset, indicating the importance of all sampled dimensions (e.g. head pose and gaze angles, number of subjects, lighting conditions and input image resolution). We hope this baseline method and evaluations will inspire future research in gaze estimation using our ETH-XGaze dataset.
In summary, our contribution is three-fold:

- A large scale dataset (over 1 Mio samples) for gaze estimation covering a large head pose and gaze range from 110 participants of different age, gender and ethnicity with consistent label quality and high-resolution images.
- Standardized experimental protocol and evaluation metrics including a new robustness evaluation.
- Detailed analysis on different factors for gaze estimation training.

2 Related Work

2.1 Gaze Estimation Algorithms

Initial learning-based gaze estimation methods often assume a static head pose [1,21], with later works allowing for gradually more head pose freedom [22,33]. In parallel, gaze estimation errors on public datasets have improved rapidly in recent years, through the use of domain adaptation [30], Bayesian networks [34], adversarial approaches [35], coarse-to-fine [6], and multi-region CNNs [19,9]. Recent development in the person-specific adaptation of gaze estimators [25,38,20] are quickly reducing error metrics on public datasets even further. However, gaze-estimation is studied mostly in the frontal setting which does not apply to many emerging application domains. There is hence a need for a systematic method to understanding the robustness of a model with regards to gaze direction and head orientation ranges. We thus propose our gaze estimation dataset to cover these factors and propose concrete tasks for their evaluation.

2.2 Gaze Datasets

Newly introduced datasets in any area of research tend to push the limits of the data distribution represented in existing datasets. Multi-view cameras have been used to cover head poses in previous works. However, there are limited range of head poses [31], or limited effective resolution on face region using machine vision cameras [33] or wide-angle cameras [40]. The Columbia dataset uses five high-resolution camera while only 5,880 samples with discrete gaze directions are recorded [31]. UT Multi-view (UTMV) [33] is recorded with eight machine version cameras, and HUMBI is recorded with multiple wide-angle cameras, however, their resolution of eye region is small in the captured image. Capturing different head poses with a single camera can be achieved by asking participants to explicitly move their head during recording as in EYEDIAP [10], moving the camera and gaze target around the participant as in Gaze360 [16], or both as in RT-GENE [9]. Some of these approaches do result in lower resolution images, and as such are not informative in the development of generative models [30] or gaze redirection methods [12,38]. Therefore, these methods had to revert to the synthetic data from UnityEye [37] or the relatively small Columbia datasets [31]. In addition, it is more challenging to aim for the acquisition of a balanced dataset in terms of head pose and gaze estimation ranges when capturing in the wild (cf. [42,19,16]), as is later shown in this paper in parameter range comparisons between our proposed dataset and existing ones. Our high resolution dataset tackles the mentioned challenge of limited head pose and gaze direction ranges in existing datasets, taking meaningful steps towards constructing a balanced set of training data for learning high performance and robust gaze estimation models. Furthermore, we see potential in leveraging the high quality imagery to enable future work in areas adjacent to gaze-estimation such as generative modeling of the eye-region, Computer Graphics and facial reconstruction.

A comprehensive summary of current gaze estimation datasets in relationship to ours is shown in Tab. 1.

2.3 Evaluation Protocols

Having public benchmark frameworks for evaluation of popular algorithms is common for many computer vision tasks such as image classification [28], face recognition [17], pedestrian detection [8] and hand pose estimation [43]. Unfortunately, there is neither a unified evaluation protocol for gaze estimation nor an existing dataset that can serve as a general evaluation platform. Despite existing best practices, most previous work relies on their own data pre-processing and sometimes uses different training-test splits for evaluation. To provide a platform for gaze estimation evaluation, we share our dataset ETH-XGaze and define a set of clearly defined evaluation procedures. Furthermore, an online evaluation system and public leader-board are released along with the dataset).

	# Peo.	Maximum Head Pose	Maximum Gaze	# Data	Resolution
Columbia [31]	56	0°, ±30°	±15°, ±10°	5,880	5184×3456
UTMV [33]	50	±36°, ±36°	±50°, ±36°	64,000	1280 × 1024
EYEDIAP [10]	16	±15°, 30°	±25°, 20°	237 min	HD & VGA
MPIIGaze [42]	15	±15°, 30°	±20°, ±20°	213,659	1280 × 720
GazeCapture [19]	1,474	±30°, 40°	±20°, ±20°	2,445,504	640 × 480
RT-GENE [9]	15	±40°, ±40°	±40°, −40°	122,531	1920 × 1080
Gaze360 [16]	238	±90°, unknown	±140°, −50°	172,000	4096 × 3382
ETH-XGaze	110	**±80°, ±80°**	**±120°, ±70°**	1,083,492	**6000 × 4000**

Table 1: Overview of popular gaze estimation datasets showing the number of participants, the maximum head poses and gaze in horizontal (around yaw axis) and vertical (around pitch axis) directions in the camera coordinate system, amount of data (number of images or duration of video), and image resolution.

3 ETH-XGaze Dataset

There are several parameters that define a comprehensive gaze estimation dataset, including: head pose, gaze direction, subject appearance, illumination condition, and image resolution. We design the ETH-XGaze data collection procedure with the main objective to maximize the parameter range along each of those dimensions as much as possible.

3.1 Acquisition Setup

The setup used for data collection is shown in the left of Fig. 1. We capture the subject with 18 Canon 250D digital SLR cameras from different viewpoints to cover a large range of head poses. There are five paired cameras for geometry capture and eight cameras for texture acquisition, such as to enable 3D face reconstruction in the future. The resolution of the captured images is very high (6000 × 4000 pixels). All cameras are connected via ESPER trigger boxes[3] to a Raspberry Pi, and a wireless mouse is used to send the triggering signal to the Raspberry Pi. The delay between mouse click and triggering the camera is below 0.05 seconds. A large screen (120 × 100 cm) is placed in the center of the cameras to show the stimuli controlled by the Raspberry Pi and projected by a projector. Since some cameras are placed behind the screen, we create cutout holes for their lenses. There are four light boxes (Walimex Daylight 250) surrounding the screen and each of them is equipped with a light bulb that emits ∼4500lm. The Raspberry Pi can turn each of the light boxes on or off to simulate different illumination conditions. We mount polarization filters in front of both the light box and camera and carefully adjust the filter angle to attenuate specular reflection off the face of the participants. During recording, the participants are sitting at approximately one meter distance in front of the screen, with the head placed in a head rest to reduce unintentional head motion.

[3] https://www.esperhq.com

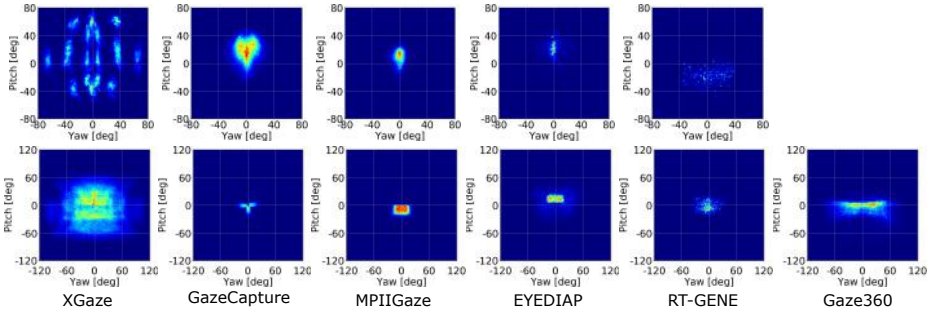

Fig. 2: Head pose (top row) and gaze direction (bottom row) distributions of different datasets. The head pose of Gaze360 is not shown here since it is not provided by the dataset.

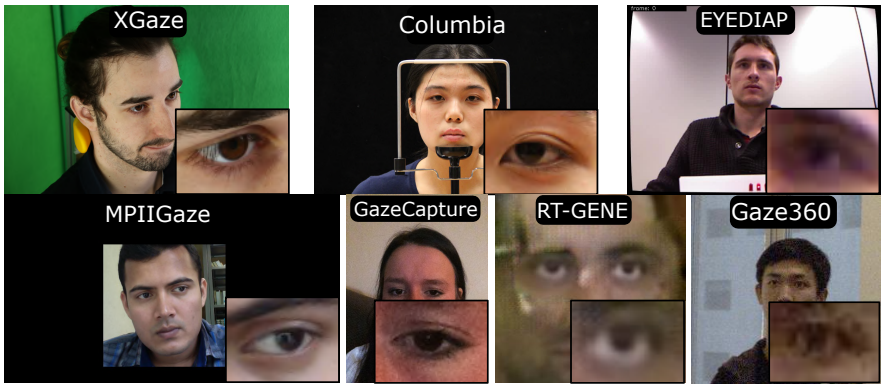

Fig. 3: Data examples and corresponding cropped eye images from different gaze estimation datasets. ETH-XGaze has the highest image resolution and quality.

3.2 Collection Procedure

During data collection, the participant focuses on a shrinking circle and clicks the mouse when the circle becomes a dot, providing the gaze point. The position of gaze points are randomly distributed on the screen. We have three methods to ensure the participant is looking at the dot when clicking the mouse. First, the participant has a short time window of 0.5 second to click the mouse to successfully collect one sample. Second, the shrinking time of the circle is random such that the participant has to focus on the shrinking circle to avoid missing the triggering time window. Third, the participant is told to collect a fixed amount of samples and any missing mouse click will increase the collection time. For most of the data collection, all four light boxes are fully on, in order to provide the maximum brightness, but we additionally simulate 15 illumination conditions by switching on and off the four light boxes.

3.3 Data Characteristics

In total, we collect data from 110 participants (47 female and 63 male), aged between 19 and 41 years. 17 of them wore contact lenses and 17 of them wore eye glasses during recording. The ethnicities of the participants includes Caucasian, Middle Eastern, East Asian, South Asian and African. Each participant collected 525 gaze points under the full-lighting condition, and 90 gaze points under the varying lighting conditions - six gaze points for each of the 15 lighting conditions. For each gaze point, a total of 18 images was collected by the 18 different cameras. We manually removed samples for which the participant was not looking at the ground-truth point-of-regard due to blinking, motion blur etc. This results in total 1,083,492 images samples for whole ETH-XGaze dataset.

A comparison between the proposed and existing datasets can be found in Tab. 1. Our dataset surpasses existing datasets regarding the following aspects.

Head pose. ETH-XGaze has the largest range of head poses compared to existing datasets, as shown in the first row of Fig. 2. Examples from ETH-XGaze with different head poses are shown in Fig. 5. In [16], it is stated that the effective head pose range of Gaze360 is $\pm 90^\circ$ in horizontal direction and limited head poses in vertical direction. However, head pose annotations are not provided in their dataset and hence we cannot visualize it here.

Gaze direction. ETH-XGaze has the largest range of gaze directions compared to existing datasets. The second row of Fig. 2 compares the gaze direction distributions. Although Gaze360 reports $\pm 140^\circ$ coverage on the horizontal gaze direction, the dataset contains only very few samples beyond $\pm 70^\circ$. ETH-XGaze is evenly sampled across a large range of horizontal and vertical gaze directions.

Image resolution. ETH-XGaze has the highest image resolution compared to existing datasets, especially the effective resolution on the face region. We show some examples and corresponding cropped eye images from different datasets in Fig 3. The Columbia dataset also has high image resolution, however, the dataset is comprised of only 5,880 samples. While EYEDIAP, MPIIGaze, RT-GENE and Gaze360 have fairly high resolution imagery as well, the participant is far away from the camera which results in low effective eye region resolution.

Controlled illumination conditions. ETH-XGaze provides a set of controlled illumination conditions. Although uncontrolled in-the-wild illumination conditions are important for gaze estimation [42,19], controlled illumination conditions provide complementary information to better understand illumination impact and enable light synthesis. As shown in Fig. 4, we record 16 different illumination conditions.

3.4 ETH-XGaze Utility

ETH-XGaze makes it possible to *train* gaze estimators that cover large ranges of head poses and gaze directions. This allows to better estimate gaze from oblique viewpoints, such as overhead cameras. ETH-XGaze can also be used to *evaluate* the robustness of a gaze estimation method with respect to these factors. In our dataset the head pose remains fixed and thus does not follow the traditional

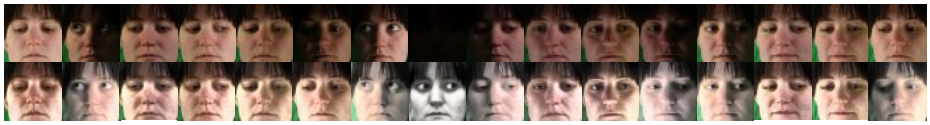

Fig. 4: Samples of the 16 illumination conditions created by switching on and off the four light boxes. The first row are the original samples, and the second row employs histogram equalization. The first column is the full-lighting setting.

head-pose-following-gaze pattern. However, by imaging from 18 viewpoints we densely sample all natural pose-gaze combinations with respect to the camera, suitable for varied applications like gaze estimation from a personal laptop or attention measurement inside a smart home.

Our dataset allows future gaze prediction methods to train on high-resolution imagery, which is critical for generative methods [12,25,38,34]. Since the generated image quality highly depends on the training image quality, Columbia and the synthetic UnityEYE dataset have been used during training in the past. Our ETH-XGaze provides high-resolution images (6000×4000 pixels), and more importantly the face region occupies a large portion of the image.

Since the data in ETH-XGaze has been captured to allow for 3D geometry reconstruction using multi-view photogrammetry methods (i.e. [2]), it provides the potential of synthesizing high-quality gaze estimation data in the future. Parametric eye models [36,3] can be fit to the data to build a controllable rig of the eye [4]. Such a rig can then be used to re-render novel images of different lighting conditions, gaze directions, and head poses with state-of-the-art rendering techniques, providing additional training data for gaze estimation task.

3.5 Data Pre-processing

We crop the face patch out of the original image as input for gaze estimation model training. For each input image sample, we first perform face and facial landmark detection using a state-of-the-art method [5]. We then fit a 3D morphable model of the face to the detected landmarks to estimate the 3D head pose [14]. The 3D head pose along with camera calibration information is used to perform data normalization [41]. In a nutshell, the data normalization method maps the input image to a normalized space where a virtual camera is used to warp the face patch out of the original input image according to 3D head pose. It rotates a virtual camera to cancel the head rotation around the row axis, and moves the virtual camera to a fixed distance from the face center to warp the face patch of fixed size. More details can be found in the original paper [41]. During data normalization, we define the face center as the center of the four eye corners and two nose corners, we set the focal length of the virtual camera to be 960 mm, the normalized distance to be 300 mm, and the cropped face image is 448×448 pixels. Examples of face patches after data normalization are shown in Fig. 5. The processed data along with original imagery are released to public.

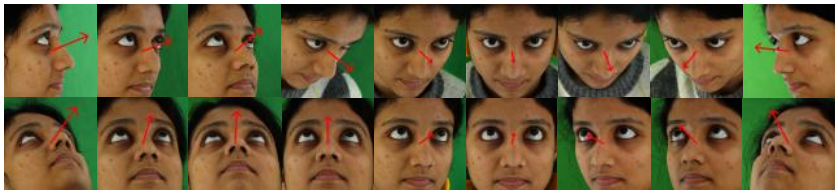

Fig. 5: Data examples captured by 18 different camera views. The red arrow is the gaze direction. The face patch images shown are after data normalization.

4 Evaluation Protocol

One goal of this paper is to establish a benchmark to evaluate gaze estimation algorithms. For this purpose, we define four evaluations on ETH-XGaze. The first three evaluations - cross-dataset, within-dataset, and person-specific evaluations - are popular evaluations found in the current gaze estimation literature. In addition, we propose to also assess robustness over head poses and gaze directions as a fourth evaluation criteria, which is made possible by ETH-XGaze.

4.1 Baseline Method

We provide a baseline gaze estimation method using an off-the-shelf ResNet-50 network [11]. This baseline takes the full-face patch covering 224×224 pixels as input and outputs the horizontal and vertical gaze angles. We used the ADAM [18] optimizer with an initial learning rate of 0.0001, and the batch size is set to be 50. We trained the baseline model for 25 epochs and decay the learning rate by a factor of 0.1 every 10 epochs.

4.2 Dataset Preparation

We split ETH-XGaze into three parts: a training set $\mathbb{TR}$ comprised of 80 participants, a test set for within-dataset evaluation $\mathbb{TE}$ containing 15 participants, and a test set for person-specific evaluation $\mathbb{TES}$ consisting of another 15 participants. Splitting the test data into two disjoint sets allows us to release ground truth gaze required for the person-specific evaluation (Sec. 4.5). We ensured that the subjects in both training and test sets exhibit diverse gender, age, and ethnicity, some with and some without glasses. While we release both ground-truth gaze and imagery for the training set, we withhold the ground-truth gaze for the test sets. Authors are encouraged to submit gaze predictions on test samples to the benchmark website, and the performance will be evaluated and reported. This enables future research to compare to existing methods on neutral grounds.

Aside from the proposed ETH-XGaze dataset, we also evaluated other existing datasets with our baseline method. These datasets were pre-processed as we described in Sec. 3.5. For the *EYEDIAP* dataset, we used both screen sequence and floating target sequences and sampled the video sequences every 15 frames. For the *GazeCapture* dataset, we used the pre-processing pipeline from [25] to obtain 3D head poses since the dataset does not provide camera parameters. For

Train \ Test	MPIIGaze	EYEDIAP	Gaze Capture	RT -GENE	Gaze360	ETH-XGaze	Ave. Rank
MPIIGaze	-	17.9	**6.3**	14.9	31.7	34.9	2.6
EYEDIAP	16.9	-	14.2	15.6	33.7	41.7	4.2
GazeCapture	**4.5**	13.7	-	**14.7**	30.2	29.4	**1.8**
RT-GENE	12.0	21.2	13.2	-	34.7	42.6	4.6
Gaze360	10.3	11.3	12.9	26.6	-	**17.0**	2.8
ETH-XGaze	7.5	**11.0**	10.5	31.2	**27.3**	-	2.0

Table 2: Gaze estimation errors in degrees on cross-dataset evaluations. The last column shows the average ranking on each test sets, and all other numbers are gaze estimation error in degrees.

the *Gaze360* dataset, we used the face bounding box provided by the dataset to crop the face patch, alongside the 3D gaze ground-truth. We will ask authors of these datasets for permission to release the processed data such that the community can use it for evaluations on ETH-XGaze.

4.3 Cross-dataset Evaluation

Cross-dataset evaluation has gained popularity since it indicates the generalization capabilities of a gaze estimation method. We define the cross-dataset evaluation as training the model on ETH-XGaze and testing on other datasets, as well as training on other datasets and testing on ETH-XGaze.

We conducted the pair-wise cross-dataset evaluations on different datasets and show results achieved by the baseline in Tab. 2. The results exhibit rather large gaze estimation errors when testing on our ETH-XGaze, indicating that there is a big domain gap between ETH-XGaze and previous datasets. This stems from the fact that ETH-XGaze exhibits much larger variation in head pose and gaze direction compared to other datasets. Therefore, the gaze estimator has to extrapolate to those unseen head poses and gaze directions which is known to be a difficult machine learning task.

Training on GazeCapture achieves the best overall ranking since it contains similar head pose and gaze ranges compared to MPIIGaze, RT-GENE and EYEDIAP. However, it performs poorly on test datasets that exhibit large variation in head pose and gaze direction such as Gaze360 and our ETH-XGaze. In contrast, ETH-XGaze enables thorough benchmarking of generalization capabilities of future gaze estimation approaches.

The model trained on Gaze360 achieves the best cross-dataset performance on ETH-XGaze since they contain similar head pose and gaze direction ranges. However, Gaze360 has been collected "in the wild" setting and can suffer from low-quality images and gaze labels (see Fig. 6). Our dataset, despite the lab setting, still allows for good performance (the best on EYEDIAP and Gaze360) without any data augmentation.

Fig. 6: Test samples from different datasets. We show results from training on Gaze360 and testing on ETH-XGaze (left), and training on ETH-XGaze and testing on Gaze360 (middle) and RT-GENE (right). The green arrow denotes ground truth and the red arrow is the prediction. The numbers give the respective gaze estimation errors in degrees.

	ETH-XGaze	MPIIGaze	EYEDIAP	GazeCapture	RT-GENE
[25]	-	5.2	-	3.5	-
[9]	-	4.8	-	-	**8.7**
[26]	-	**4.5**	10.3	-	-
[39]	-	-	6.8	-	-
Baseline	4.5	4.8	**6.5**	**3.3**	12.0

Table 3: Comparison of the baseline with current state-of-the-art on within dataset evaluations. Numbers are gaze estimation errors in degrees.

4.4 Within-dataset Evaluation

Within-dataset evaluation is another popular means of evaluating gaze estimation methods. Here the method is trained on $\mathbb{TR}$ and evaluated on $\mathbb{TE}$. Tab. 3 shows performances of the baseline alongside comparisons to recent state-of-the-art methods. The baseline achieves an error of 4.7 degrees on average on ETH-XGaze, which is reasonably low given the large ranges of head poses and gaze directions. On other datasets, the baseline exhibits an accuracy comparable to current state-of-the-art methods, indicating that it is a strong baseline. The results of the other methods are taken from the respective publications.

4.5 Person-specific Evaluation

Person-specific gaze estimation has gained a lot of attention in recent years [25,38,20] due to the huge improvements that can be achieved from even just a few personal calibration samples. We randomly selected 200 samples from each participant in $\mathbb{TES}$ as the personal calibration samples. The protocol is to train the model with $\mathbb{TR}$ and up to 200 personal calibration samples, and to test on the remaining samples of $\mathbb{TES}$ – for each of the 15 test subjects. We pre-trained the model on $\mathbb{TR}$ and then fine-tune it using the 200 samples with 25 epochs.

Results from the baseline in Fig. 7 show that personal calibration improves the gaze estimates by a large margin. The goal of this evaluation is not only to achieve good results but also to rely on as few calibration samples as possible.

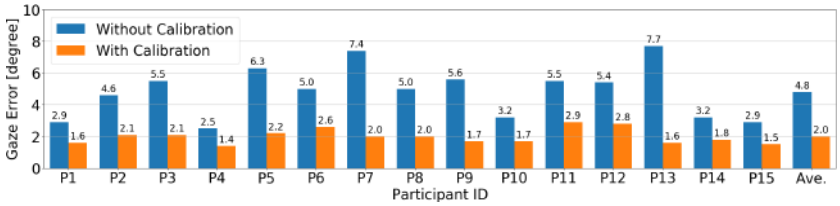

Fig. 7: Gaze estimation errors for person-specific evaluation of our baseline. We show the gaze estimation errors with and without training with 200 calibration samples. The number above each bar is the gaze estimation error in degrees.

4.6 Robustness Evaluation

Previous gaze estimation works usually only report the mean gaze estimation errors without detailed analysis across head poses and gaze directions. This is partly due to the lack of sufficient data samples to cover a wide range. Knowing the performance of an algorithm with respect to these factors is important, since a method with a higher overall error might have lower error within a specific range of interest. Hence we introduce a detailed evaluation to show the robustness across head poses and gaze directions. Fig. 8 shows the performance of the baseline on $\mathbb{TE}$ over horizontal and vertical axes of the head pose and gaze direction. The different colors represent the different training sets. While these plots evaluate the performance of the different training sets, the benchmark will compare different algorithms instead. A flat curve across the entire graph, as in the case of training on ETH-XGaze, indicates robustness to head pose and gaze direction variation.

5 Demonstration of ETH-XGaze

In this section, we evaluate the importance of different factors during training. Previous gaze estimation datasets cannot serve as the evaluation set for an ablation study of different factors such as head poses, gaze directions and illumination conditions due to the limited coverage. In contrast, the proposed ETH-XGaze is an ideal dataset for these evaluations.

Head Pose and Gaze Direction. We created several training subsets from $\mathbb{TR}$ by constraining the head poses and/or gaze directions angle ranges to be ± 80, ± 60, ± 40, and ± 20 in both horizontal and vertical directions. To keep the same amount of training samples for each subsets, we randomly re-sampled each training subset to have the same amount of samples as the minimal training set, i.e. the training set of ± 20 in both head poses and gaze directions. The results of testing on $\mathbb{TE}$ are shown in the left of Fig. 9. As we can see from the figure, constraining the head pose and gaze direction results in worse performance in general, especially when we constrain both head pose and gaze direction ranges. Constraining the gaze directions achieves worse results than constraining head poses, which indicates gaze directions have more impact than the head poses. Specifically, when we constrained the angle range to be ± 40 degrees, the performance decrease caused by constraining head poses is 34.6%, constraining gaze

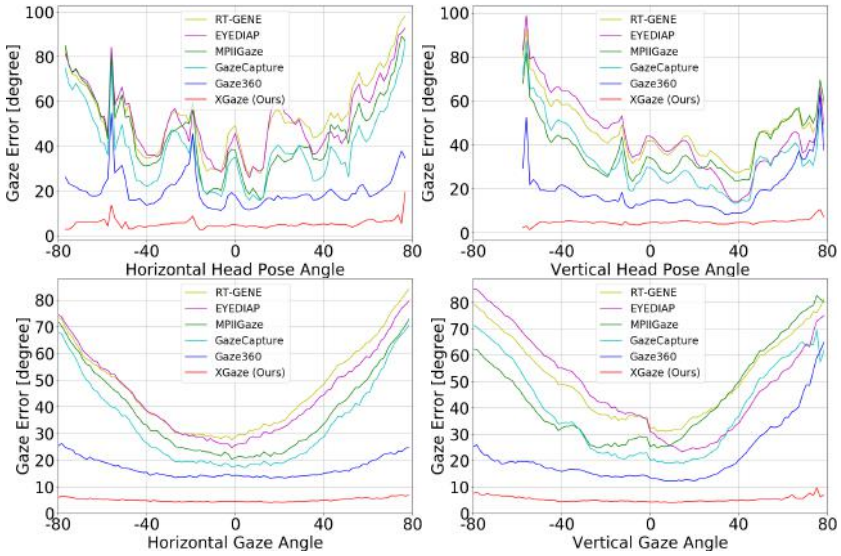

Fig. 8: Gaze estimation error distribution across head poses (first row) and gaze directions (second row) in horizontal and vertical directions respectively. The colored curves represent results with different training sets tested on $\mathbb{TE}$.

directions is 82.1%, and constraining both head poses and gaze directions is 206.4%.

Illumination condition. In the center of Fig. 9, we show results by training the baseline with all lighting conditions or only with the full-lighting condition. The performance drop (9% from 7.8 degrees to 8.5 degrees) indicates the impact of lighting conditions on gaze estimation performance.

Personal appearance. In [19], the authors show gaze estimation performance with different numbers of participants. Our repeated experiment with our baseline on ETH-XGaze shows the same trend as increasing number of participants improves the performance (see Fig. 9, right).

Input resolution. The image resolution analysis in [42] was only for eye images and the highest resolution was 60×36. The default input face patch image size to ResNet is 224×224 which we used in our baseline. We resized the input image to be 112×112 and 448×448 and then fed them into the baseline. Since there is an average pooling layer at the end of the ResNet convolutional layers, we do not need to modify the architecture with respect to different resolutions.

The results of resolution variation are shown in Tab. 4. The performance is improved when training and testing on higher resolutions, which indicates the potential of high-resolution gaze estimation. However, different with results in [42], the model trained on one size achieves much worse results on other sizes. This can be caused by the much higher image resolution in ETH-XGaze with large appearance differences compared to the MPIIGaze in [42]. We did not specifically develop the method to handle cross-resolution input images and expect future works can properly deal with cross-resolution training.

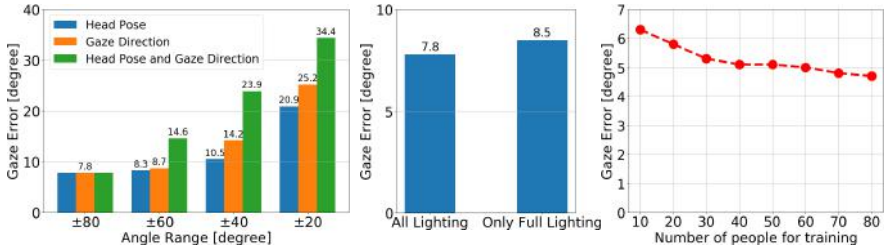

Fig. 9: Gaze estimation error distribution by constraining head poses and gaze directions (left), lighting conditions (middle), and number of people (right) during training. The number above each bar is the gaze estimation error in degrees.

Train \ Test	112 × 112	224 × 224	448 × 448
112 × 112	**5.4**	25.3	37.2
224 × 224	20.2	**4.5**	42.1
448 × 448	65.1	54.7	**4.2**

Table 4: Gaze estimation errors in degrees generated by models trained with different input image sizes in pixels.

6 Conclusion

We present a new large-scale gaze estimation dataset ETH-XGaze, featuring large variation in head pose and gaze direction, high-resolution imagery, varied subject appearance, systematic illumination conditions, as well as accurate ground-truth gaze vectors. Evaluation using a baseline method shows that training on ETH-XGaze significantly improves robustness towards variation in head pose and gaze direction compared to existing datasets, adding a very valuable resource for future work on gaze estimation. In addition, we propose a standardized experimental protocol and evaluation framework that will be made available via the benchmark website alongside the dataset, allowing for fair comparison of gaze estimation algorithms on neutral ground.

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.

Fig. 10: ERC logo.

References

1. Baluja, S., Pomerleau, D.: Non-intrusive gaze tracking using artificial neural networks. In: Advances in Neural Information Processing Systems. pp. 753–760 (1994)
2. Beeler, T., Bickel, B., Beardsley, P., Sumner, B., Gross, M.: High-quality single-shot capture of facial geometry. In: ACM Transactions on Graphics (TOG), pp. 1–9 (2010)
3. Bérard, P., Bradley, D., Gross, M., Beeler, T.: Lightweight eye capture using a parametric model. ACM Transactions on Graphics (TOG) **35**(4), 1–12 (2016)
4. Bérard, P., Bradley, D., Gross, M., Beeler, T.: Practical person-specific eye rigging. In: Computer Graphics Forum. vol. 38, pp. 441–454. Wiley Online Library (2019)
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: Proceedings of the IEEE International Conference on Computer Vision. pp. 1021–1030 (2017)
6. Cheng, Y., Huang, S., Wang, F., Qian, C., Lu, F.: A coarse-to-fine adaptive network for appearance-based gaze estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 10623–10630 (2020)
7. Demiris, Y.: Prediction of intent in robotics and multi-agent systems. Cognitive processing **8**(3), 151–158 (2007)
8. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. IEEE transactions on pattern analysis and machine intelligence **34**(4), 743–761 (2011)
9. Fischer, T., Jin Chang, H., Demiris, Y.: Rt-gene: Real-time eye gaze estimation in natural environments. In: Proceedings of the European Conference on Computer Vision. pp. 334–352 (2018)
10. 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: Proceedings of the ACM Symposium on Eye Tracking Research & Applications. pp. 255–258. ACM (2014)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
12. He, Z., Spurr, A., Zhang, X., Hilliges, O.: Photo-realistic monocular gaze redirection using generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6932–6941 (2019)
13. Huang, Q., Veeraraghavan, A., Sabharwal, A.: Tabletgaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. Machine Vision and Applications **28**(5-6), 445–461 (2017)
14. 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: Proceedings of the 11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (2016)
15. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(7), 1325–1339 (jul 2014)
16. Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., Torralba, A.: Gaze360: physically unconstrained gaze estimation in the wild. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6912–6921 (2019)

17. Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E.: The megaface benchmark: 1 million faces for recognition at scale. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4873–4882 (2016)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
19. Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., Torralba, A.: Eye tracking for everyone. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2176–2184 (2016)
20. Liu, G., Yu, Y., Mora, K.A.F., Odobez, J.M.: A differential approach for gaze estimation with calibration. In: British Machine Vision Conference. vol. 2, p. 6 (2018)
21. Lu, F., Sugano, Y., Okabe, T., Sato, Y.: Inferring human gaze from appearance via adaptive linear regression. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 153–160. IEEE (2011)
22. Lu, F., Sugano, Y., Okabe, T., Sato, Y.: Adaptive linear regression for appearance-based gaze estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(10), 2033–2046 (2014)
23. Majaranta, P., Bulling, A.: Eye tracking and eye-based human–computer interaction. In: Advances in physiological computing, pp. 39–65. Springer (2014)
24. Nech, A., Kemelmacher-Shlizerman, I.: Level playing field for million scale face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7044–7053 (2017)
25. Park, S., Mello, S.D., Molchanov, P., Iqbal, U., Hilliges, O., Kautz, J.: Few-shot adaptive gaze estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9368–9377 (2019)
26. Park, S., Spurr, A., Hilliges, O.: Deep pictorial gaze estimation. In: Proceedings of the European Conference on Computer Vision. pp. 721–738 (2018)
27. Ruch, T.C., Fulton, J.F.: Medical physiology and biophysics. Academic Medicine **35**(11), 1067 (1960)
28. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
29. Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. vol. 1, pp. 519–528. IEEE (2006)
30. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2107–2116 (2017)
31. Smith, B.A., Yin, Q., Feiner, S.K., Nayar, S.K.: Gaze locking: passive eye contact detection for human-object interaction. In: Proceedings of the 26th annual ACM symposium on User interface software and technology. pp. 271–280 (2013)
32. Soo Park, H., Jain, E., Sheikh, Y.: Predicting primary gaze behavior using social saliency fields. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3503–3510 (2013)
33. Sugano, Y., Matsushita, Y., Sato, Y.: Learning-by-synthesis for appearance-based 3d gaze estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1821–1828 (2014)

34. Wang, K., Zhao, R., Ji, Q.: A hierarchical generative model for eye image synthesis and eye gaze estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 440–448 (2018)
35. Wang, K., Zhao, R., Su, H., Ji, Q.: Generalizing eye tracking with bayesian adversarial learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11907–11916 (2019)
36. Wood, E., Baltrušaitis, T., Morency, L.P., Robinson, P., Bulling, A.: A 3d morphable eye region model for gaze estimation. In: Proceedings of the European Conference on Computer Vision. pp. 297–313. Springer (2016)
37. Wood, E., Baltrušaitis, T., Morency, L.P., Robinson, P., Bulling, A.: Learning an appearance-based gaze estimator from one million synthesised images. In: Proceedings of the ACM Symposium on Eye Tracking Research & Applications. pp. 131–138 (2016)
38. Yu, Y., Liu, G., Odobez, J.M.: Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11937–11946 (2019)
39. Yu, Y., Odobez, J.M.: Unsupervised representation learning for gaze estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7314–7324 (2020)
40. Yu, Z., Yoon, J.S., Venkatesh, P., Park, J., Yu, J., Park, H.S.: Humbi 1.0: Human multiview behavioral imaging dataset (June 2020)
41. Zhang, X., Sugano, Y., Bulling, A.: Revisiting data normalization for appearance-based gaze estimation. In: Proceedings of the ACM Symposium on Eye Tracking Research & Applications. p. 12. ACM (2018)
42. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(1), 162–175 (2019)
43. Zimmermann, C., Ceylan, D., Yang, J., Russell, B., Argus, M., Brox, T.: Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 813–822 (2019)