Preprint version; final version available at http://ieeexplore.ieee.org
IEEE International Conference on Computer Vision (ICCV) Workshop Proceedings
Published by: IEEE

RT-BENE: A Dataset and Baselines for Real-Time Blink Estimation in Natural Environments

Kévin Cortacero Tobias Fischer Yiannis Demiris
Personal Robotics Laboratory, Department of Electrical and Electronic Engineering
Imperial College London, United Kingdom
cortacero.kevin@gmail.com, {t.fischer, y.demiris}@imperial.ac.uk

Abstract

In recent years gaze estimation methods have made substantial progress, driven by the numerous application areas including human-robot interaction, visual attention estimation and foveated rendering for virtual reality headsets. However, many gaze estimation methods typically assume that the subject's eyes are open; for closed eyes, these methods provide irregular gaze estimates. Here, we address this assumption by first introducing a new open-sourced dataset with annotations of the eye-openness of more than 200,000 eye images, including more than 10,000 images where the eyes are closed. We further present baseline methods that allow for blink detection using convolutional neural networks. In extensive experiments, we show that the proposed baselines perform favourably in terms of precision and recall. We further incorporate our proposed RT-BENE baselines in the recently presented RT-GENE gaze estimation framework where it provides a real-time inference of the openness of the eyes. We argue that our work will benefit both gaze estimation and blink estimation methods, and we take steps towards unifying these methods.

1. Introduction

Accurately estimating a subject's gaze is a challenging task. While in the past specialised devices such as head-mounted eye trackers were used for gaze estimation [25], recently several methods have been proposed that take images captured by a standard webcam as input [5, 13, 28, 33, 48]. Methods that rely on deep learning have been shown to be particularly effective for gaze estimation [13, 48, 50], but their generalisation capabilities largely depend on the underlying datasets that the methods have been trained on. Therefore, a variety of datasets have been proposed, ranging from interactions with phones [28], tablets [23, 28], laptops [48] and computer screens [16] to larger distances that are suitable e.g. in human-robot interactions [13].

Despite the remarkable progress on gaze estimation, the state-of-the-art methods treat images where the subject's eyes are open in the same way as images where the eyes are closed due to blinking [13, 49, 50]. Therefore, even if the eyes are closed, gaze estimates are provided albeit being irregular. Similarly, previous methods for blink detection [4, 14, 15, 29] have not been integrated with gaze estimation methods, and have used separate datasets that significantly differ from those used by gaze estimation methods. In addition, many of these datasets are either not available for the research community, or are captured in "non-wild" scenarios which limit their applicability.

In this work, we present "Real-Time Blink Estimation in Natural Environments" (RT-BENE) to address these shortcomings. We first annotate over 200,000 images of the RT-GENE dataset [13] that was introduced for gaze estimation in natural settings with large camera-subject distances and less constrained subject motion. Our dataset contains more than 10,000 samples of blinking instances, which are significantly more compared to previous datasets. We then use our dataset to train a variety of convolutional neural networks that predict the openness of the eyes.

We evaluate these networks on our proposed dataset and several other datasets (Eyeblink8 [11], Talking Face and Researcher's Night [15]) where we achieve state-of-the-art performance. We also incorporate our networks in the RT-GENE pipeline to allow for joint blink and gaze estimation. As an additional contribution, we extend RT-GENE to estimate the gazes of multiple subjects simultaneously. To foster future research on joint blink and gaze estimation, we make our code and dataset available to the research community[1].

RT-BENE has many possible application areas. Within the computer vision community, gaze has recently been used for visual attention estimation [9], saliency estimation [30] and labelling in the context of video captioning [47]. We argue that incorporation of blinks will further improve performance in these tasks, as it is well known from studies with humans that blinks impact task performance in atten-

[1] www.imperial.ac.uk/Personal-Robotics/

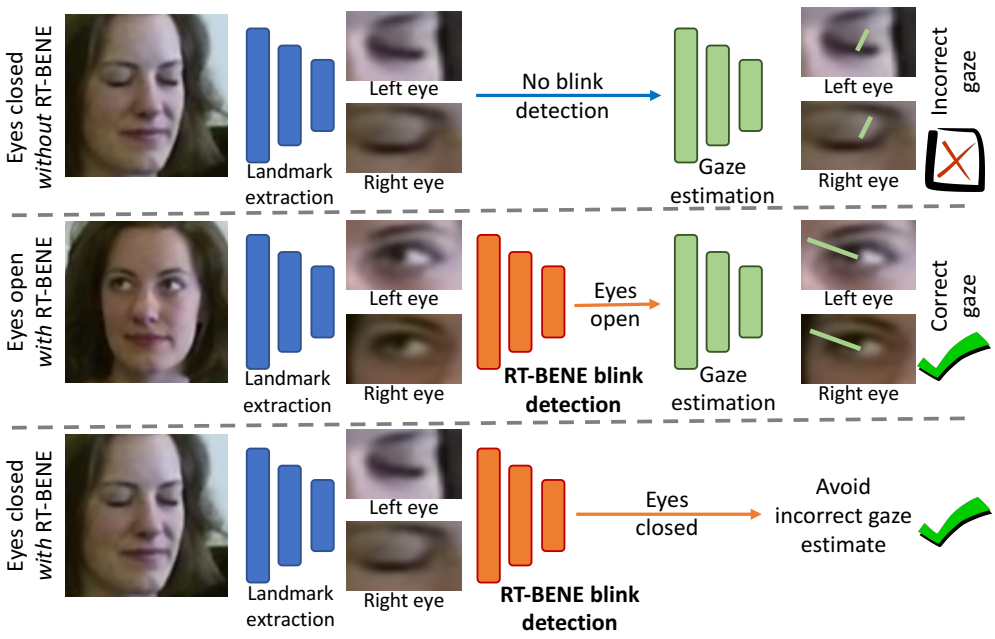

Figure 1: Overview of the proposed RT-BENE method. Top row: in case of a blink, without RT-BENE the gaze was previously incorrectly estimated. Middle & bottom rows: when using RT-BENE, the gaze is correctly estimated when the eyes are open (middle row) and incorrect gaze estimates are avoided when a blink was detected (bottom row).

tion and workload estimation [2] and can be used for user modelling [17].

2. Related work

Gaze estimation In recent years a variety of appearance-based gaze estimation methods have been proposed. Zhang et al. [48, 50] demonstrated that convolutional neural networks (CNNs) can be used to find the mapping between an eye image and the corresponding gaze angle. Cheng et al. [8] have shown that it is beneficial to estimate which eye image (left or right eye) results in better gaze estimation performance, rather than randomly choosing the eye as above. Zhang et al. [49] have demonstrated that the whole face image can be used to estimate the gaze, rather than providing eye images separately. Park et al. [37] propose an elegant approach where the input image is first transformed into a minimal image called gaze map, and the minimal image is then used to estimate the gaze. Fischer et al. [13] demonstrate that deeper architectures lead to better features, and that ensemble models improve gaze accuracy.

Liu et al. [31] have recently addressed the issue that universal models do not capture variabilities in eye shape and eye structure amongst individuals. They propose a linear adaptation method to improve gaze estimation methods using subject-specific models built from a few calibration images. A related issue is that head pose significantly alters the appearance of the eyes. Deng and Zhu [10] propose to combine separate models for the head pose and eyeball movement using a gaze transform layer, which also addresses head pose ambiguity amongst different subjects.

Gaze datasets The generalisation capabilities heavily depend on the datasets that are used to train the methods reviewed above. The EYEDIAP dataset [16] captures subjects in two different scenarios, one where the subject gazes at targets appearing on a screen, and another where the subject gazes at a floating target. The MPIIGaze dataset [48, 50] captures subjects "in the wild" rather than in lab environments. The advantage of MPIIGaze is that lightning conditions and backgrounds vary significantly more compared to previous datasets. While MPIIGaze aims at laptop scenarios, TabletGaze [23] and GazeCapture [28] contain recordings of several hundred participants while using mobile phones or tablets.

Recently, the RT-GENE dataset [13] has been proposed, which takes a different approach compared to all previous datasets that all employ gaze targets appearing on screens. Within RT-GENE, subjects wear eyetracking glasses equipped with motion capture markers to obtain ground truth annotations in free-viewing tasks. The eyetracking glasses are subsequently removed using image inpainting with generative adversarial networks (GANs) to avoid overfitting the CNNs to these images. In our work, we annotate the "natural" sub-dataset of RT-GENE where the subjects do not wear the eyetracking glasses. In the original RT-GENE work, this sub-dataset is used to train the GANs that "remove" the eyetracking glasses from the subject's face.

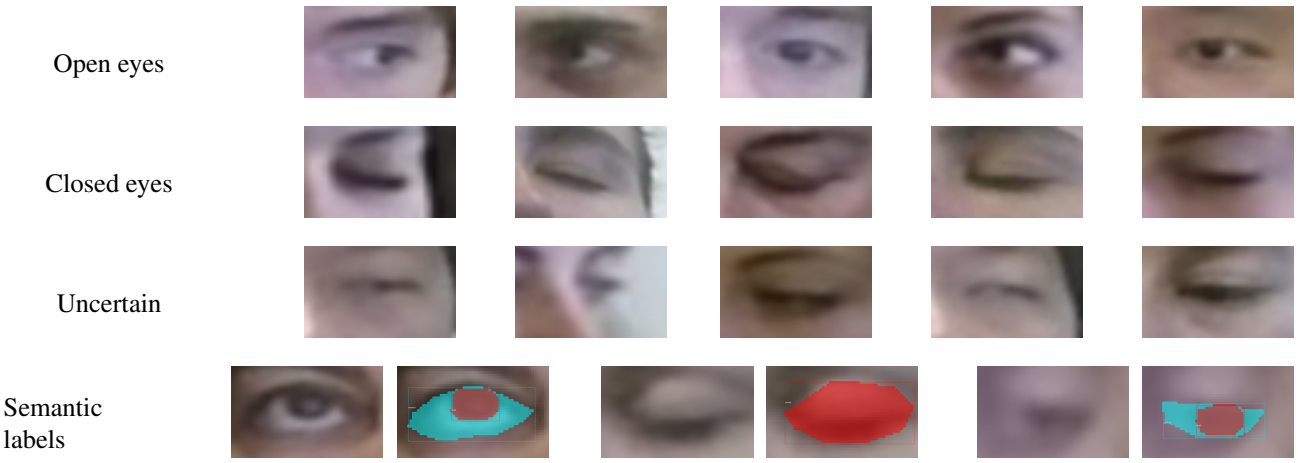

Figure 2: RT-BENE: The first three rows depict five example annotations for each of the three groups (open eyes, closed eyes and uncertain). The bottom row depicts pairs of images: the left image shows the plain eye image and the right image shows the ground truth labels of the semantic labelling of the eye area (sclera in turquoise, iris+pupil in pink and closed eyes in red).

Blink estimation There have been a variety of proposals to estimate whether an eye is closed or open. Traditional methods, such as the one by Lalonde et al. [29] and Schillingmann and Nagai [40], rely on features extracted by SIFT and HOG with a subsequent classification stage. However, the accuracy of these methods is reduced for extreme head angles and under varying skin colour and illumination. To partly overcome these limitations, CNNs have been used to extract more robust features [3, 26] and allow blink estimation in cases where the faces are not full frontal, which is considered challenging [19].

To improve the performance of these methods, it has been proposed to track a sequence of images rather than inputting a single image [4, 29]. Fogelton and Benesova [14] design a state machine for eye blink detection [14] and propose a merging scheme to integrate left and right eye images. In another work, Fogelton and Benesova [15] use dense optical flow to extract features that are inputted into a recurrent neural network.

Blink datasets As in the gaze estimation task, the dataset used to train blink estimation methods has a significant impact on the performance, mainly when applied in in-the-wild scenarios. Furthermore, deep learning models such as CNNs and RNNs benefit from large datasets with many annotated blink samples.

Many previous datasets such as the Talking Face[2] and ZJU [36] datasets are not suitable to train deep networks as they contain fewer than 500 eye blinks. The dataset by Kim et al. [26] contains approximately 5,000 samples of closed eye images, but captures the subjects facing the camera frontally. Furthermore, the dataset by Kim et al. [26] and several other datasets [14, 36] are not readily available to the community or require the submission of licensing agreements. Notable exceptions are the Closed Eyes in the Wild (CEW) dataset [42], which contains 1192 samples of subjects where both eyes are closed, and Eyeblink8 [11] with 804 blinks. We were also able to obtain the challenging Researcher's Night dataset [15].

In summary, the majority of existing datasets is either not easily obtainable by the computer vision community [14, 26, 36], relatively small to train deep networks [36] or not captured in in-the-wild settings [26, 36]. The RT-BENE dataset aims to overcome these limitations and provide a new benchmark for blink estimation in the wild with sufficient samples to train deep networks.

Joint blink and gaze estimation There have been several preliminary works to integrate gaze estimation and blink detection in unified frameworks. Schillingmann and Nagai [40] use HOG features for both pupil location estimation and blink detection. Similarly, Chen et al. [7] employ a heuristic based on the shape of the iris area. Lu et al. [32] propose an elegant method where blinks are equivalent to samples that live outside a low-dimensional manifold consisting of the training data. Another interesting approach was presented by Gou et al. [18], where the eye position and openness are jointly used to iteratively converge towards the true eye position and openness. To the best of our knowledge, there are no deep learning based methods that jointly consider blink estimation and gaze estimation. However, note that Siegfried et al. [41] recently proposed a deep learning based method to classify gaze streams into fixation, saccade, and blink classes.

Semantic labelling of the eye area An interesting approach to combine blink estimation and gaze estimation is in semantically segmenting the eye region, whereby the regions

[2]Dataset available from http://www-prima.inrialpes.fr/FGnet/data/01-TalkingFace/talking_face.html, we use blink annotations kindly provided by [14].

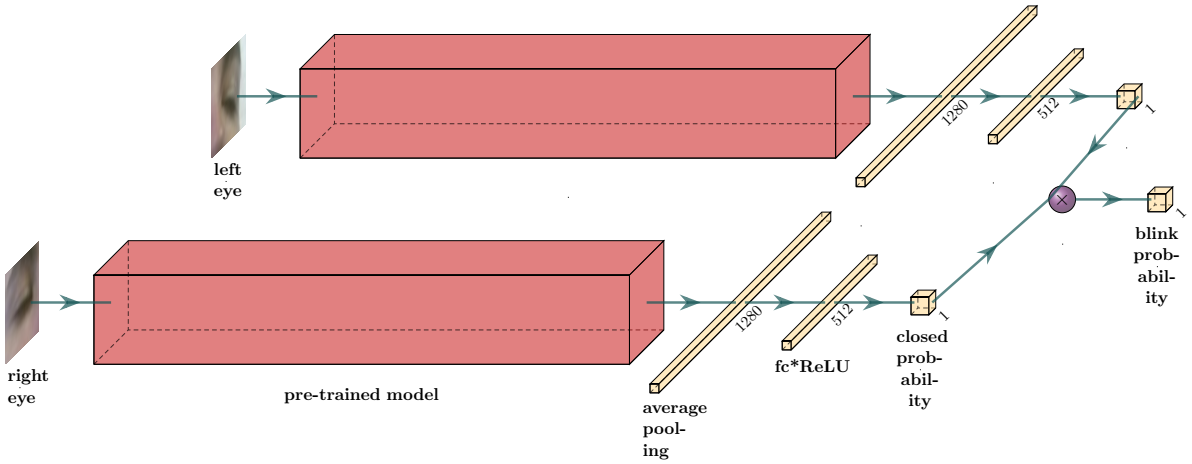

Figure 3: Network architecture. The inputs are the left and right eye images. A pre-trained backbone network is used to extract features for each eye individually. Each backbone network is followed by a fully connected layer of size 512 followed by batch normalization and ReLU6 activation with a dropout of 0.6. The two separate streams share weights and are then concatenated and connected to a single neuron with sigmoid activation, which outputs the probability of the eye openness.

considered can include the skin region surrounding the eye, the sclera (white of the eye), iris (coloured region of the eye) and pupil (black hole in the centre of the iris). Then, a blink can be detected as an absence of the sclera and iris region.

Such a semantic segmentation approach was taken in [46] using traditional image processing methods. Moriyama et al. [35] propose a detailed generative model consisting of 7 regions separated by 8 curves. Mora and Odobez [34] also take a generative approach and use variational Bayes to find person-specific parameters. Wood et al. [45] present a highly realistic 3D morphable model of the eye. The gaze is then estimated by finding a synthesised image that most closely resembles an observed image.

We argue that recent advances in transfer learning applied to instance segmentation methods such as Mask R-CNN [20] can be used to segment the eye area. Mask R-CNN is typically used to identify objects within an image and provide a detailed mask for each recognised image, rather than just a bounding box. In this paper, we adapt pre-trained networks that are trained on large datasets to semantic labelling of the eye area using just a few hundred annotated images.

3. Methodology

In Section 3.1, we describe the annotation strategy for blinks on the RT-GENE dataset; in Section 3.2 we introduce a CNN for blink estimation, and in Section 3.3 we make first steps towards unified blink and gaze estimation by introducing a semantic labelling approach based on Mask R-CNN. Finally, in Section 3.4, we present an extension of the RT-GENE framework to multiple subjects.

3.1. Dataset generation

As reviewed in Section 2, there is currently little overlap between the datasets being employed for gaze estimation and blink estimation. As we aim for a joint framework for blink and gaze estimation, we annotate the RT-GENE dataset [13], which was recently proposed for gaze estimation, with blink labels. Our proposed RT-BENE dataset has further benefits: it contains sufficient samples to train a deep network, and it is openly available for use by the research community.

In particular, we annotate the 'natural' subset of the dataset, which is the subset where no eyetracking glasses are worn. This subset is subsequently split into three categories: 1) open eyes, 2) closed eyes, and 3) uncertain.

We define open eyes as images where at least some part of the sclera (white part of the eye) or pupil is visible. Closed eyes are those where the eyelids are fully closed. The uncertain category is used when the image cannot be clearly grouped into one of the other categories due to e.g. extreme head poses, or when the two annotators labelled the image differently. Figure 2 (top three rows) shows example images contained in the three categories.

Using this approach, we labelled in total 243,714 images, 218,548 of them where the eyes are open, 10,444 where the eyes are closed and 14,722 uncertain images. These images are used to train the convolutional neural networks in Section 3.2. Note that these numbers are for one eye only, for pairs the numbers have to be halved.

To train the semantic labelling method in Section 3.3, we additionally provide detailed semantic labels for 480 images randomly sampled from the RT-GENE dataset (400 training images and 80 images for evaluation)[3]. We annotate the following segments using the VGG Image Annotator [12]: eyelid, sclera, and iris+pupil (combined). For closed eyes, the sclera and iris+pupil areas are empty. For open eyes, the

[3]Note that labelling these images is a very time-consuming task, hence not the whole dataset has been labelled.

eyelid is not labelled. Figure 2 (bottom row) shows example annotations.

3.2. Convolutional neural networks for blink estimation

We provide a set of baseline CNNs for blink estimation. As input, the eye images of the left and right eyes are provided. Rather than using complex architectures, we use standard backbones, namely MobileNetV2 [39], ResNet50 [21] and DenseNet121 [22], to extract features.

The feature extraction step is followed by fully connected layers that output the probabilities of the openness of each eye. The last step consists in fusing the individual probabilities to an overall probability. Note that the last step can be skipped if required. We use the binary cross-entropy loss function. To further improve the performance, we make use of ensemble models where we average the predictions of multiple models, similarly to RT-GENE [13]. Also note that the dataset is highly imbalanced, i.e., there are many more images with open eyes compared to those with blinks ($N_{\text{open}} \gg N_{\text{closed}}$). Therefore, we make use of oversampling [24, 27], where the weights of open eye samples is $N_{\text{total}}/(2N_{\text{open}})$ and accordingly for blink samples $N_{\text{total}}/(2N_{\text{closed}})$.

In the experimental results (Section 4), we evaluate the speed-accuracy trade-off of the different backbones. The detailed network architecture is shown in Figure 3. Please refer to our code release for further details on the training of our method.

3.3. Semantic labelling of the eye region

In this section, we use Mask R-CNN to make a step towards joint gaze and blink estimation. Typically, Mask R-CNN is used to identify objects within an image and provide a pixel-by-pixel mask for each of the objects. In our work, we propose to treat the various eye areas as different "objects".

We use Mask R-CNN as implemented by Matterport [1] to train ResNet101 backbone networks. We adjust the network architecture to cope with the small image sizes of just 36x60. The anchors ("bounding box" equivalents) are of sizes 8, 16, 32, and 64. We use the semantically labelled images described in Section 3.1 to fine-tune the networks.

At inference time, for each label, we consider the mask that has the highest corresponding confidence value. If the "eyelid" segment is dominant, the eye is considered to be closed. To find the centre of the pupil, we compute the centre of mass of the iris+pupil segment. Note that there are some cases where the sclera segment is detected, but no iris+pupil segment. In this case, we currently do not compute the pupil centre but consider the eye to be open.

In future work, we plan to provide the pupil position as an additional input to appearance-based gaze estimators to improve their gaze estimation performance.

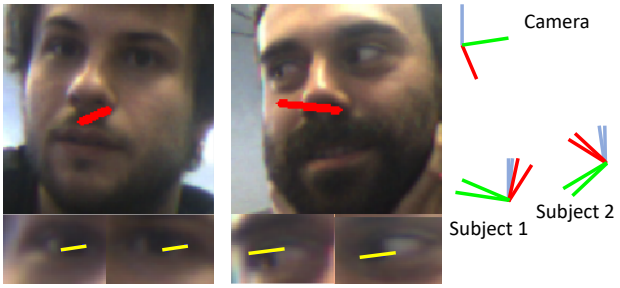

Figure 4: Multi-person gaze estimation. Two subjects are detected, and their head poses are estimated (top images). The bottom images depict the estimated gaze for both subjects. On the right, we show the relative transforms between the camera and each of the subjects. For each subject, two transforms are shown: one for the head pose and another for the eye gaze.

3.4. Multi-person gaze estimation

As an additional contribution, we allow for multi-person gaze estimation within the RT-GENE framework [13]. Multi-person gaze estimation is beneficial for a range of scenarios, including the detection of attention in group scenarios [6].

The RT-GENE framework has been restructured so that the messages exchanged between the face detection, landmark extraction and gaze estimation nodes contain a list of subjects (with corresponding images) rather than a single subject. We use the Hungarian method to keep track of subjects over time. An example of two subjects who interact with each other is shown in Figure 4.

4. Experimental results

We first evaluate our proposed RT-BENE framework for blink estimation within Section 4.1. The evaluation is performed on the following datasets: our proposed RT-BENE, Eyeblink8 [11], and Researcher's Night [15] (see Suppl. Material for example images). We compare to Google's ML-Kit[4] and the method by Anas et al. [3].

We treat the blink detection as a binary problem: either the eyes are closed or open. As the main performance metric, we use the F_1-score (closed eyes are treated as positive class for all evaluations). We also report precision, recall, average precision (AP, a score to summarise the precision-recall curve[5]) as well as the computational speed (frames per second).

In Section 4.2, we show that our method generalises well to different scenarios. We train our method on RT-BENE and apply the trained model to the Talking Face dataset (cross-dataset evaluation). In Section 4.3, we investigate

[4] https://developers.google.com/ml-kit/

[5] Details in https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html and [38].

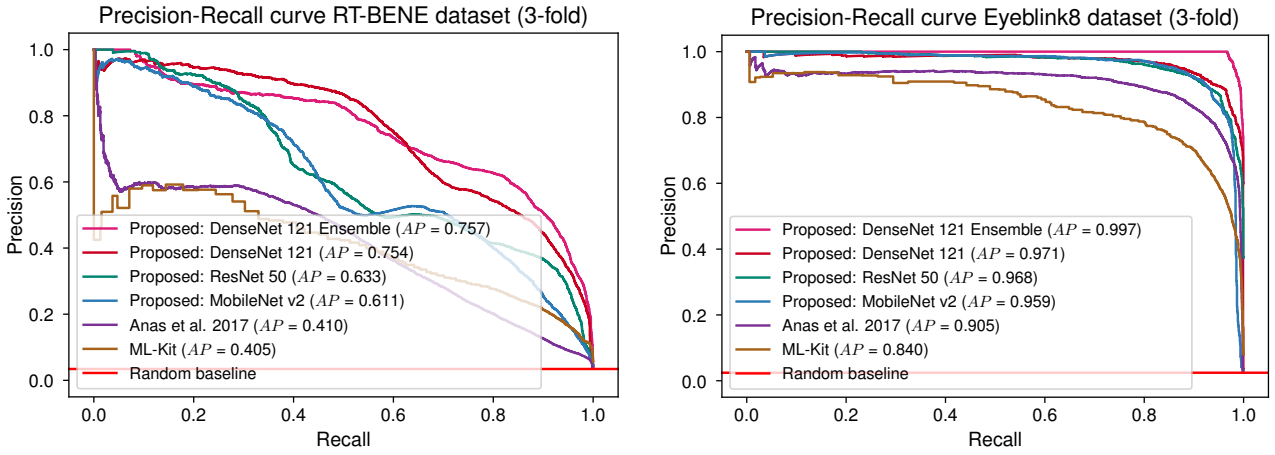

Figure 5: Precision-recall curves for the RT-BENE dataset (left; 3-fold evaluation) and Eyeblink8 dataset (right; 3-fold evaluation).

the performance depending on the number of training images. This investigation shows whether the acquisition of additional images will further improve our method's performance. In Section 4.4, we then show preliminary results for the use of Mask R-CNN to semantically label the eye region. This currently allows for blink detection, and we show promising results for gaze estimation using the same method.

The network training and evaluation is performed on a machine with 3 Nvidia GeForce 1080 GPUs, an Intel i9-7960X CPU with 16 cores and 96GB RAM.

4.1. Blink detection using CNNs

Evaluations on the RT-BENE dataset follow the same training, test and validation split as in the RT-GENE dataset [13], i.e. a 3-fold evaluation is used. We also use a 3-fold evaluation on the Eyeblink8 dataset. Our networks are trained for 15 epochs with the batch size being 64.

Figure 5 shows the precision-recall curves of various methods, and detailed quantitative results for all metrics are contained in Tables 2 and 3 (RT-BENE dataset and Eyeblink8 dataset respectively).

In experiments on RT-BENE, we observe the following: Our Ensemble performs best ($F_1 = 0.721 \pm 0.044$). DenseNet ($F_1 = 0.658 \pm 0.028$) outperforms ResNet ($F_1 = 0.598 \pm 0.107$), which in turn outperforms MobileNet v2 ($F_1 = 0.588 \pm 0.064$). This ranking remains relatively stable regardless of which performance metric (precision, recall, AP or F_1) is used. However, increased performance comes with an increase in computational cost (27.5 ± 4.8FPS for DenseNet compared to 42.2 ± 3.6FPS for MobileNet v2).

Note that our best performing model outperforms ML-Kit ($F_1 = 0.290 \pm 0.036$) and Anas et al. [3] ($F_1 = 0.529 \pm 0.075$; 36.3% performance increase). Note that ML-Kit is closed-source and was hence not fine-tuned to RT-BENE.

The performance on Eyeblink8 is similar. Specifically, our Ensemble ($F_1 = 0.976 \pm 0.018$) outperforms the method by Anas et al. [3] ($F_1 = 0.834 \pm 0.077$; 17% better performance) and ML-Kit (37.8% better performance). Detailed evaluations are contained in Table 3.

Finally, we evaluate on Researcher's Night (given their train, test and validation split). There, our Ensemble also performs very accurately ($F_1 = 0.913$) and outperforms all other models (see Supplementary Material).

Table 1 shows the confusion matrices of our best performing model on the RT-BENE, Eyeblink8 and Researcher's Night datasets.

4.2. Cross-dataset evaluation

To show that our trained models perform well in other scenarios, we conduct the following cross-dataset evaluations. We train a model on all subjects of the RT-BENE dataset (13 subjects of the three folds combined, however, note that there are four subjects that are used for model validation). The trained model is then applied to the Talking Face dataset.

IWe achieve $F_1 = 0.966$, compared to $F_1 = 0.695$ using ML-Kit (39.0% performance increase).

This shows that our proposed model performs still very well in scenarios that were not seen at training time. In the next section, we evaluate whether more training samples to train our model would lead to a further increase in performance.

4.3. Performance with increasing dataset size

Increases in the number of training samples have the potential to significantly improve machine learning algorithms, in particular so with deep neural networks [43].

For the blink estimation tasks, the datasets are relatively small, with many of them containing fewer than 500 blinks (see Section 2). While RT-BENE contains considerably more

RT-BENE dataset

	Ground truth: Open	Ground truth: Closed	Total
Predicted: Open	2236	1293	3529
Predicted: Closed	625	71153	71778
Total	2861	72446	75307

Eyeblink8 dataset

	Ground truth: Closed	Ground truth: Open	Total
Predicted: Closed	1673	12	1685
Predicted: Open	54	69172	69226
Total	1727	69184	70911

Researcher's night dataset

	Ground truth: Closed	Ground truth: Open	Total
Predicted: Closed	986	4	990
Predicted: Open	184	104547	104731
Total	1170	104551	105721

Table 1: Confusion matrices of our proposed method (DenseNet 121) on the RT-BENE (top) and Eyeblink8 (middle) and Researcher's Night (bottom) datasets; 3-fold evaluations combined.

images (10,444 closed eye images and 218,548 open eye images), the dataset size is still several orders of magnitude smaller compared to other tasks such as object recognition.

In this section, we evaluate the performance depending on the number of training images. We follow a similar 3-fold evaluation as in Section 4.1, but rather than using all images from the subjects in the training set, we randomly sample $X = \{25, 50, 75, 100\}\%$ of each subject's image set. Also, we use only one of the three folds for this study.

Our hypothesis is that increased training sample size leads to an increase in performance. Indeed, the performance increases from $F_1 = 0.581$ when using 25% of the training set to $F_1 = 0.614$ (50% of the dataset), $F_1 = 0.693$ (75% of the dataset) and finally $F_1 = 0.706$ (full training set). These results, further detailed in Table 4, indicate that even larger datasets could further improve the blink detection performance of CNNs. A similar trend can be observed in Table 5 that investigates the impact of the dataset size on the Eyeblink8 dataset. In future work, we think it would be beneficial to jointly train on multiple datasets to mitigate this issue.

4.4. Semantic labelling of the eye region

Here, we show preliminary results on the use of Mask R-CNN to semantically label the eye region. Note that only 400 images are used to fine-tune Mask R-CNN as annotating each image requires a significant amount of time (see Section 3.1).

In quantitative evaluations, our proposed method does not currently perform as well as the CNN baselines that were evaluated in Section 4.1: In the 3-fold evaluation on RT-BENE, we achieve $F_1 = 0.222$.

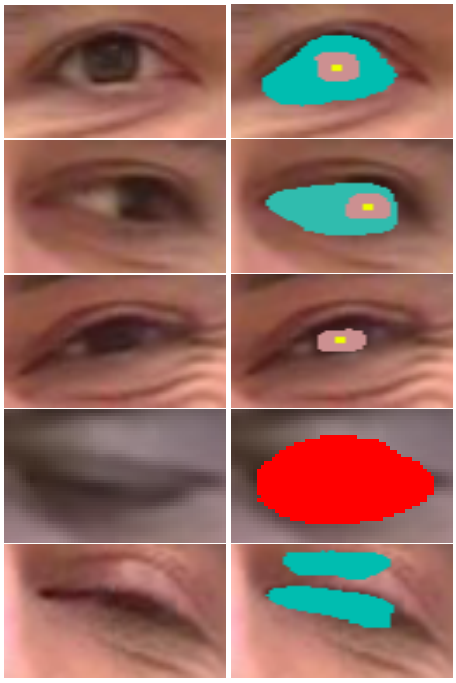

Figure 6: Mask R-CNN results exemplified on several eye images. Turquoise masks indicate the sclera and pink masks indicate the iris+pupil. The yellow dots indicate the centres of the iris+pupil masks and could be used for gaze estimation. The red mask in the fourth row indicates a closed eye. The last row indicates a failure case.

Despite the low performance in blink detection, we believe that semantic labelling is a fruitful endeavour for future research. In Figure 6, we show sample segmentation masks predicted by our model.

Interestingly, Mask R-CNN also predicts the pupil position. The pupil is visible in 54 out of the 80 evaluation images. We manually annotated the ground truth pupil position in these images. We then compare the ground truth position with that estimated by Mask R-CNN. The Euclidean distance is 1.7 ± 1.6px in the 51 images where the pupil was detected by Mask R-CNN, which is a promising result.

The results demonstrate that our proposed method can be used to simultaneously detect whether an eye is open or closed, and if it is open, the pupil location can be reliably estimated.

5. Conclusions, Limitations and Future Works

In this work, we introduced a highly accurate framework for blink estimation, which we called "RT-BENE". The framework consists of two main parts: a new dataset and a set of convolutional neural networks trained on this dataset. Different backbone networks were evaluated, and it was shown that "deeper" networks perform better but come

Method	Precision	Recall	AP	F_1	FPS
Google ML-Kit	0.172 ± 0.025	0.946 ± 0.033	0.439 ± 0.091	0.290 ± 0.036	–[a]
Anas et al. [3]	0.533 ± 0.127	0.537 ± 0.033	0.486 ± 0.114	0.529 ± 0.075	408.3 ± 10.7
Proposed MobileNetV2	0.579 ± 0.099	0.604 ± 0.026	0.642 ± 0.072	0.588 ± 0.064	42.2 ± 3.6
Proposed ResNet	0.595 ± 0.148	0.610 ± 0.060	0.649 ± 0.096	0.598 ± 0.107	41.8 ± 3.9
Proposed DenseNet	0.613 ± 0.055	0.717 ± 0.048	0.728 ± 0.066	0.658 ± 0.028	27.5 ± 4.8
Proposed Ensemble	0.664 ± 0.064	0.791 ± 0.017	0.774 ± 0.016	0.721 ± 0.044	20.5 ± 3.7

Table 2: Performance on the RT-BENE dataset

[a]There is no fair speed comparison for ML-Kit, as ML-Kit cannot run on desktop machines. On modern tablets, ML-Kit runs between 15-20 FPS.

Method	Precision	Recall	AP	F_1
Google ML-Kit	0.584 ± 0.184	0.965 ± 0.020	0.838 ± 0.100	0.708 ± 0.155
Anas et al. [3]	0.754 ± 0.108	0.941 ± 0.022	0.883 ± 0.060	0.834 ± 0.077
Proposed MobileNet v2	0.884 ± 0.008	0.919 ± 0.036	0.955 ± 0.024	0.901 ± 0.021
Proposed ResNet	0.842 ± 0.065	0.939 ± 0.037	0.950 ± 0.042	0.887 ± 0.051
Proposed DenseNet	0.883 ± 0.038	0.954 ± 0.026	0.969 ± 0.021	0.916 ± 0.019
Proposed Ensemble	0.995 ± 0.004	0.958 ± 0.035	0.997 ± 0.002	0.976 ± 0.018

Table 3: Performance on the Eyeblink8 dataset

% of train set	Prec.	Rec.	AP	F_1
25	0.662	0.517	0.723	0.581
50	0.703	0.546	0.782	0.614
75	0.747	0.646	0.774	0.693
100	0.700	0.712	0.796	0.706

Table 4: Performance depending on dataset size (DenseNet model; RT-BENE dataset; one fold only)

% of train set	Prec.	Rec.	AP	F_1
25	0.908	0.804	0.893	0.853
50	0.968	0.797	0.940	0.874
75	0.889	0.904	0.970	0.897
100	0.928	0.895	0.965	0.911

Table 5: Performance depending on dataset size (DenseNet model; Eyeblink8 dataset; one fold only)

with an additional computational cost. Furthermore, a significant performance improvement over previous state-of-the-art has been achieved.

We also introduced a preliminary method for semantic eye labelling based on Mask R-CNN and demonstrated promising results in this direction. We hope that by using crowdsourcing, we can obtain more semantic annotations to train the Mask R-CNN and improve its performance.

While the methods based on CNNs currently outperform the semantic labelling approach, we believe that the semantic labelling based on Mask R-CNN will lead to a unified blink and gaze estimation method in the future. Rather than a purely geometric approach, we will integrate the ellipse denoting the sclera as well as the detected pupil location with appearance-based gaze methods by providing additional inputs to these methods.

Other future works include the following. Firstly, while our current blink dataset contains images of the RT-GENE dataset where no eyetracking glasses are worn, we will incorporate the inpainted images of the RT-GENE dataset. In these images, the subject originally had worn eyetracking glasses to obtain ground-truth gaze locations, which were then subsequently removed using GANs. This will further increase the size of our RT-BENE dataset.

Secondly, we aim for even finer-grained semantic segmentation. We hypothesise that this will allow the estimation of the pupil dilation, which is known to be a predictor of task difficulty [44] amongst others.

While we provided an extension for the RT-GENE framework to provide gaze estimates of multiple subjects simultaneously, this currently does not scale well as each additional subject linearly increases the required computational time. While the single subject and dual subject estimates are provided in real-time, the computational speed drops considerably for more than three subjects.

We make the dataset and code available to the community (`www.imperial.ac.uk/Personal-Robotics/`) to foster future research on joint gaze and blink estimation.

Acknowledgements We thank the Personal Robotics Lab members at Imperial College for their support during this research. This work was supported by the European Union H2020 Framework Programme (Project PAL, H2020-PHC-643783), and a Royal Academy of Engineering Chair in Emerging Technologies.

References

[1] W. Abdulla. Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow. https://github.com/matterport/Mask_RCNN, 2017. 5

[2] H. Admoni and B. Scassellati. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1):25–63, 2017. 2

[3] E. R. Anas, P. Henriquez, and B. J. Matuszewski. Online Eye Status Detection in the Wild with Convolutional Neural Networks. In *Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 88–95, 2017. 3, 5, 6, 8

[4] T. Appel, T. Santini, and E. Kasneci. Brightness- and motion-based blink detection for head-mounted eye trackers. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 1726–1735, 2016. 1, 3

[5] T. Baltrusaitis, P. Robinson, and L. P. Morency. OpenFace: An open source facial behavior analysis toolkit. In *IEEE Winter Conference on Applications of Computer Vision*, 2016. 1

[6] F. Bartoli, G. Lisanti, L. Seidenari, S. Karaman, and A. Del Bimbo. Museumvisitors: a dataset for pedestrian and group detection, gaze estimation and behavior understanding. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 19–27, 2015. 5

[7] B. C. Chen, P. C. Wu, and S. Y. Chien. Real-time eye localization, blink detection, and gaze estimation system without infrared illumination. In *International Conference on Image Processing*, pages 715–719, 2015. 3

[8] Y. Cheng, F. Lu, and X. Zhang. Appearance-Based Gaze Estimation via Evaluation-Guided Asymmetric Regression. In *European Conference on Computer Vision*, pages 105–121, 2018. 2

[9] E. Chong, N. Ruiz, Y. Wang, Y. Zhang, A. Rozga, and J. M. Rehg. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *European Conference on Computer Vision*, pages 383–398, 2018. 1

[10] H. Deng and W. Zhu. Monocular Free-Head 3D Gaze Tracking with Deep Learning and Geometry Constraints. In *IEEE International Conference on Computer Vision*, pages 3162–3171, 2017. 2

[11] T. Drutarovsky and A. Fogelton. Eye Blink Detection Using Variance of Motion Vectors. In *European Conference on Computer Vision Workshops*, Computer Science, pages 436–448, 2016. 1, 3, 5

[12] A. Dutta and A. Zisserman. The VIA Annotation Software for Images, Audio and Video. *arXiv preprint arXiv:1904.10699*, 2019. 4

[13] T. Fischer, H. J. Chang, and Y. Demiris. RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments. In *European Conference on Computer Vision*, pages 339–357, 2018. 1, 2, 4, 5, 6

[14] A. Fogelton and W. Benesova. Eye blink detection based on motion vectors analysis. *Computer Vision and Image Understanding*, 148:23–33, 2016. 1, 3

[15] A. Fogelton and W. Benesova. Eye blink completeness detection. *Computer Vision and Image Understanding*, 176-177:78–85, 2018. 1, 3, 5

[16] K. A. Funes Mora, F. Monay, and J.-M. Odobez. EYEDIAP: A Database for the Development and Evaluation of Gaze Estimation Algorithms from RGB and RGB-D Cameras. In *ACM Symposium on Eye Tracking Research and Applications*, pages 255–258, 2014. 1, 2

[17] T. Georgiou and Y. Demiris. Adaptive user modelling in car racing games using behavioural and physiological data. *User Modeling and User-Adapted Interaction*, 27(2):267–311, 2017. 2

[18] C. Gou, Y. Wu, K. Wang, K. Wang, F. Y. Wang, and Q. Ji. A joint cascaded framework for simultaneous eye detection and eye state estimation. *Pattern Recognition*, 67:23–31, 2017. 3

[19] D. W. Hansen and Q. Ji. In the Eye of the Beholder: A Survey of Models for Eyes and Gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):478–500, 2010. 3

[20] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision*, pages 2961–2969, 2017. 4

[21] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5

[22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely Connected Convolutional Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017. 5

[23] Q. Huang, A. Veeraraghavan, and A. Sabharwal. TabletGaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. *Machine Vision and Applications*, 28(5-6):445–461, 2017. 1, 2

[24] J. M. Johnson and T. M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):27, 2019. 5

[25] M. Kassner, W. Patera, and A. Bulling. Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 1151–1160, 2014. 1

[26] K. W. Kim, H. G. Hong, G. P. Nam, and K. R. Park. A study of deep CNN-based classification of open and closed eyes using a visible light camera sensor. *Sensors*, 17(7):1–21, 2017. 3

[27] G. King and L. Zeng. Logistic regression in rare events data. *Political analysis*, 9(2):137–163, 2001. 5

[28] K. Krafka, A. Khosla, P. Kellnhofer, and H. Kannan. Eye Tracking for Everyone. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2176–2184, 2016. 1, 2

[29] M. Lalonde, D. Byrns, L. Gagnon, N. Teasdale, and D. Laurendeau. Real-time eye blink detection with GPU-based SIFT tracking. In *Canadian Conference on Computer and Robot Vision*, pages 481–487, 2007. 1, 3

[30] G. Leifman, D. Rudoy, T. Swedish, E. Bayro-Corrochano, and R. Raskar. Learning gaze transitions from depth to improve video saliency estimation. In *IEEE International Conference on Computer Vision*, pages 1698–1707, 2017. 1

[31] G. Liu, Y. Yu, K. A. F. Mora, and J.-M. Odobez. A Differential Approach for Gaze Estimation with Calibration. In *British Machine Vision Conference*, 2018. 2

[32] F. Lu, Y. Sugano, T. Okabe, and Y. Sato. Adaptive Linear Regression for Appearance-Based Gaze Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(10):2033–2046, 2014. 3

[33] B. Masse, S. Ba, and R. Horaud. Tracking Gaze and Visual Focus of Attention of People Involved in Social Interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2711–2724, 2018. 1

[34] K. A. F. Mora and J. M. Odobez. Geometric generative gaze estimation (G3E) for remote RGB-D cameras. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1773–1780, 2014. 4

[35] T. Moriyama, T. Kanade, J. Xiao, and J. F. Cohn. Meticulously detailed eye region model and its application to analysis of facial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(5):738–752, 2006. 4

[36] G. Pan, L. Sun, Z. Wu, and S. Lao. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In *IEEE International Conference on Computer Vision*, 2007. 3

[37] S. Park, A. Spurr, and O. Hilliges. Deep Pictorial Gaze Estimation. In *European Conference on Computer Vision*, pages 741–757, 2018. 2

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 5

[39] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018. 5

[40] L. Schillingmann and Y. Nagai. Yet Another Gaze Detector: An Embodied Calibration Free System for the iCub Robot. In *IEEE-RAS International Conference on Humanoid Robots*, pages 8–13, 2015. 3

[41] R. Siegfried, Y. Yu, and J.-M. Odobez. A Deep Learning Approach for Robust Head Pose Independent Eye Movements Recognition from Videos. In *ACM Symposium on Eye Tracking Research & Applications*, pages 31:1–31:5, 2019. 3

[42] F. Song, X. Tan, X. Liu, and S. Chen. Eyes closeness detection from still images with multi-scale histograms of principal oriented gradients. *Pattern Recognition*, 47:2825–2838, 2014. 3

[43] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *IEEE International Conference on Computer Vision*, pages 843–852, 2017. 6

[44] P. van der Wel and H. van Steenbergen. Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin & Review*, 25(6):2005–2015, 2018. 8

[45] E. Wood, T. Baltrušaitis, L.-P. Morency, P. Robinson, and A. Bulling. A 3D Morphable Eye Region Model for Gaze Estimation. In *European Conference on Computer Vision*, pages 297–313, 2016. 4

[46] H. Yamazoe, A. Utsumi, T. Yonezawa, and S. Abe. Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions. In *Symposium on Eye tracking Research & Applications*, pages 245–250, 2008. 4

[47] Y. Yu, J. Choi, Y. Kim, K. Yoo, S.-H. Lee, and G. Kim. Supervising neural attention models for video captioning by human gaze data. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 490–498, 2017. 1

[48] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Appearance-Based Gaze Estimation in the Wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4511–4520, 2015. 1, 2

[49] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2299–2308, 2017. 1, 2

[50] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):162–175, 2019. 1, 2