Offset Calibration for Appearance-Based Gaze Estimation via Gaze Decomposition

Zhaokang Chen Bertram E. Shi
The Hong Kong University of Science and Technology
zchenbc@connect.ust.hk eebert@ust.hk

Abstract

Appearance-based gaze estimation provides relatively unconstrained gaze tracking. However, subject-independent models achieve limited accuracy partly due to individual variations. To improve estimation, we propose a gaze decomposition method that enables low complexity calibration, i.e., using calibration data collected when subjects view only one or a few gaze targets and the number of images per gaze target is small. Lowering the complexity of calibration makes it more convenient and less time-consuming for the user, and more widely applicable. Motivated by our finding that the inter-subject squared bias exceeds the intra-subject variance for a subject-independent estimator, we decompose the gaze estimate into the sum of a subject-independent term estimated from the input image by a deep convolutional network and a subject-dependent bias term. During training, both the weights of the deep network and the bias terms are estimated. During testing, if no calibration data is available, we can set the bias term to zero. Otherwise, the bias term can be estimated from images of the subject gazing at known gaze targets. Experimental results on three datasets show that without calibration, our method outperforms state-of-the-art by at least 6.3%. *For low complexity calibration sets, our method outperforms other calibration methods. More complex calibration algorithms do not outperform our method until the size of the calibration set is excessively large. Even then, the gains obtained by alternatives are small, e.g., only* 0.1° *lower error for 64 gaze targets. Source code is available at https://github.com/czk32611/Gaze-Decomposition.*

1. Introduction

As an important cue about people's intent, eye gaze has been used in many applications, such as human-computer interfaces [4, 23, 30], human-robot interaction [16], virtual reality [24, 29], social behavioral analysis [15] and health care [12]. These successes have attracted more and more attention to gaze tracking.

To date, most eye trackers have relied upon active illumination, e.g. infrared illumination used in pupil center corneal reflections (PCCR). While these provide high accuracy, accuracy rapidly degrades as the head pose changes. These techniques are commonly used in well-controlled laboratory settings where high accuracy is required. Researchers have proposed many novel methods to alleviate the constraints on head movement, which enable more real-world applications [2, 6, 10, 31, 39]. However, a common disadvantage of active illumination approaches is that they are relatively costly, as they rely upon custom hardware.

Appearance-based gaze estimation estimates gaze directions based on RGB images, providing relatively unconstrained gaze tracking and requiring only commonly available off-the-shelf cameras. However, obtaining high accuracy is very challenging due to large variability caused by factors such as differences in individual appearance, head pose, and illumination [51]. The application of deep convolutional neural networks (CNNs) has reduced estimation error significantly [49]. There are a large number of real and synthetic datasets covering a wide range of these variations [9, 11, 14, 19, 34, 36, 38, 41, 42, 49]. Using these datasets, it has been shown that deep CNNs can learn to compensate for the variability [3, 5, 7, 10, 19, 20, 28, 32, 50].

Unfortunately, the estimation error of subject-independent appearance-based methods is still higher than that achievable using active illumination, e.g. $\sim 5^\circ$ vs $\sim 1^\circ$ achieved by PCCR. Thus, further work must be done to reduce this error. One way to further reduce estimation error is through personal calibration. PCCR-based eye trackers typically require the user to look sequentially at a number of targets for calibration [13]. This enables subject-specific parameters of a geometric 3D eye model to be estimated. Typically, there are nine calibration gaze targets arranged on a three by three grid, but the number of gaze targets is often reduced to five to save time at the expense of poorer accuracy. A similar calibration procedure has been proposed for appearance-based methods, where some parameters of the estimator are fine-tuned [19, 21, 22, 26, 37, 40, 45].

This study focuses on calibrating appearance-based gaze estimators using low complexity calibration sets. We measure the complexity by two factors: the number of gaze targets and the number of images per gaze target. The larger the values, the higher the complexity. For best performance, the multiple images per target should contain a variety of head poses. Low complexity sets are desirable, as they are easier to collect. However, a complex model may easily overfit to low complexity sets, even leading to an increase in estimation error, as reported in [19]. The challenge we address is how to achieve the best accuracy with the lowest complexity and to determine the point at which more complex calibration methods begin to outperform lower complex calibration methods. We target low complexity calibration, as it is more convenient and is more widely applicable. For example, in screen-free applications, it is difficult to divert user's gaze towards a large number of different gaze targets. We are particularly interested in single gaze target calibration. Since the camera should always be visible to the users if it is to get a clear picture of the eyes, it can serve as a convenient gaze target.

We propose a gaze decomposition method based on the assumption that there exists a person-dependent bias that cannot be estimated from images. It is known that there is a deviation between the visual axis and the optic axis of an eye, and that this deviation varies from person to person [1, 13]. Our experimental results confirm this finding. For a subject-independent estimator, the estimation bias varies significantly across subjects but is relatively constant across different gaze angles for the same subject (see Fig. 1). Thus, we decompose the gaze estimate into the sum of a subject-dependent bias term and the output of a subject-independent gaze estimate from images. When no calibration data is available, we set the bias to zero. For calibration, we estimate the bias from images taken as the subject gazes at one or more targets.

Despite its simplicity, our results show that our method outperforms state-of-the-art without calibration on the MPIIGaze [49] and EYEDIAP [11] datasets by more than 6.3%. With only a single gaze target for calibration, our method outperforms all previously proposed calibration methods no matter how many images are used. Using more gaze targets results in further improvements. Although more complex methods eventually outperform ours, this does not occur until the number of gaze targets is prohibitively large (more than 32 gaze targets), which would result in long calibration procedures unlikely to be tolerated by most users.

2. Related work

2.1. Appearance-based gaze estimation

Methods for appearance-based gaze estimation estimates gaze directions from RGB images under visible light. One common approach is the machine learning based method, which directly regress from images to gaze estimates using machine learning techniques. Past approaches to this problem have included k-Nearest Neighbors [33, 38], Support Vector Regression [33] and Random Forests [38]. Recently, deep CNNs have received increasing attention. Zhang *et al*. proposed the first deep CNN for gaze estimation in the wild [49, 51], which improved accuracy significantly. To further improve the accuracy, researchers have proposed enhancements, such as employing the information outside the eye region [19, 50], focusing on the head-eye relationship [7, 25, 32] and extracting better information from the eye images [3, 5, 20, 27, 44].

Another common approach is the model based method, which estimates gaze directions using geometric head/eye models whose parameters are estimated from the images (e.g. by facial landmark and pupil center detection) [28, 40].

2.2. Calibration

Existing calibration methods for appearance-based gaze estimation can be divided into differential and adaptation-based approaches.

The differential approach was introduced by [22], where the authors trained a subject-independent Siamese network to estimate the gaze angle difference between two images of the same subject. During testing, they used this network to estimate gaze differences between the input image and nine calibration images taken as the subject gazed at different targets, and averaged the resulting estimates.

In the adaptation-based approach, a subset of the parameters are subject-dependent. Methods following this approach consist of two parts: a subject-independent part, which extracts the common features across different subjects, and an adaptive part, which changes for new subject. These methods are either machine learning based or geometric model based. For machine learning based methods, the subject-independent part is often a CNN-based gaze estimator or feature extractor. These methods differ primarily in the adaptive part. Krafka *et al*. [19] and Strobl *et al*. [37] trained a subject-dependent Support Vector Regression on the features extracted from a subject-independent estimator. Liu *et al*. [22] and Strobl *et al*. [37] learned a homogeneous linear transformation matrix to warp the estimates from a subject-independent estimator. Zhang *et al*. learned a third order polynomial function to warp the estimates [48]. Lindén *et al*. concatenated some subject-dependent parameters to the features extracted from the convolutional layers of a subject-independent network [21]. Zhang *et al*. adapted networks across multiple devices [46]. As an example of a geometric model based method, Wang and Ji adapted subject-dependent parameters of a generic head/eye geometric model [40].

The past work described above used high complexity

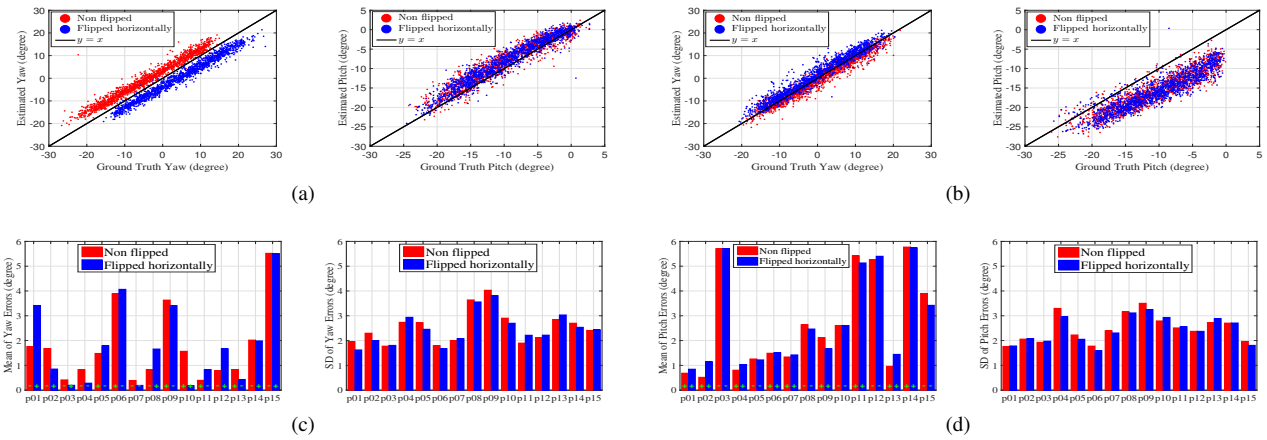

Figure 1. Error analysis of a subject-independent estimator [3] on the MPIIGaze dataset [49]. Upper row shows scatter plots of the estimated angles versus the ground truth of two subjects, (a) p06 and (b) p12. Bottom row shows mean and standard deviation of the (c) yaw and (d) pitch angles for each subject. The green $+$ or $-$ indicate the sign of the mean. Better viewed in color and zoomed in.

calibration sets with multiple gaze targets and a large number of images. Error dropped significantly with high complexity calibration sets, but increased with low complexity sets, most likely due to overfitting [19].

Only recently has attention been paid to reducing the complexity of the calibration set. Yu *et al*. proposed gaze redirection synthesis, which used a generative adversarial network (GAN) to augment the calibration samples used to fine-tune a subject-independent estimator [45]. Park *et al*. proposed FAZE, which trained the adaptive part (the last multi-layer perceptron) using meta-learning to learn good initial weights that can adapt to a few samples without overfitting [26].

3. Methodology

3.1. Analysis of estimation error

Fig. 1 analyzes estimation error on the MPIIGaze dataset [49] made by a state-of-the-art subject-independent estimator [3] by leave-one-subject-out cross-validation. Fig. 1(a) and (b) compare the estimates and the ground truth gaze of two subjects. The plots indicate that the bias is quite constant across gaze angles, but varies between subjects.

Fig. 1(c) and (d) show the mean and standard deviation (SD) of the yaw and pitch error for different subjects. The results show that the errors in both yaw and pitch angles are generally biased. The bias vary across subjects, whereas the SDs are relatively stable. When the images are flipped horizontally, the yaw bias has a similar magnitude, but opposite sign. The pitch bias remains similar. The mean squared bias across subjects (16.2 deg^2) exceeds the mean intra-subject variance (12.9 deg^2), indicating that the bias is a significant contributor to the error. As the data of each subject exhibits considerable variability in illumination and head pose, we hypothesize that the bias is primarily due to difference between subjects.

3.2. Gaze decomposition

Motivated by these findings, we assume that there exists subject-dependent bias that cannot be estimated from images. This assumption is supported by the fact that there exist subject-dependent deviations between the visual and optic axes [1, 13]. The visual axis is the line connecting the nodal point with the fovea, and is what we wish to estimate. The optic axis is the line connecting the nodal point with the pupil center, and is what can be estimated from the images.

We decompose the gaze estimates for the j^{th} image of subject i, $\hat{g}_{i,j} \in \mathbb{R}^2$, into the sum of a subject-independent term $\hat{t}$ estimated from an image $X_{i,j}$ and a subject-dependent bias $\hat{b}_i$, i.e.,

$$\hat{g}_{i,j} = \hat{t}(X_{i,j}; \Phi) + \hat{b}_i, \tag{1}$$

where Φ denotes the parameters of a deep CNN for estimating $\hat{t}$, as described below. All gaze angles are expressed as yaw and pitch. A model that directly estimates the visual axis is equivalent to a model with $\hat{b}_i \equiv 0, \forall i$.

3.3. The proposed network

The architecture of our proposed network is presented in Fig. 2. The general architecture is inspired by iTracker [19] and Dilated-Net [3]. It takes an image of the face and images of both eyes as input. The input images $X_{i,j}$ are first fed to three base CNNs. The architecture of the base CNN is shown in Fig. 2(b). It has five convolutional layers, one max-pooling layer and four dilated-convolutional layers [43] with different dilation rates, r. The strides for all convolutional layers are 1. The two base CNNs that take the eyes as input share the same weights. The feature maps extracted by the base CNNs are then fed to fully-connected

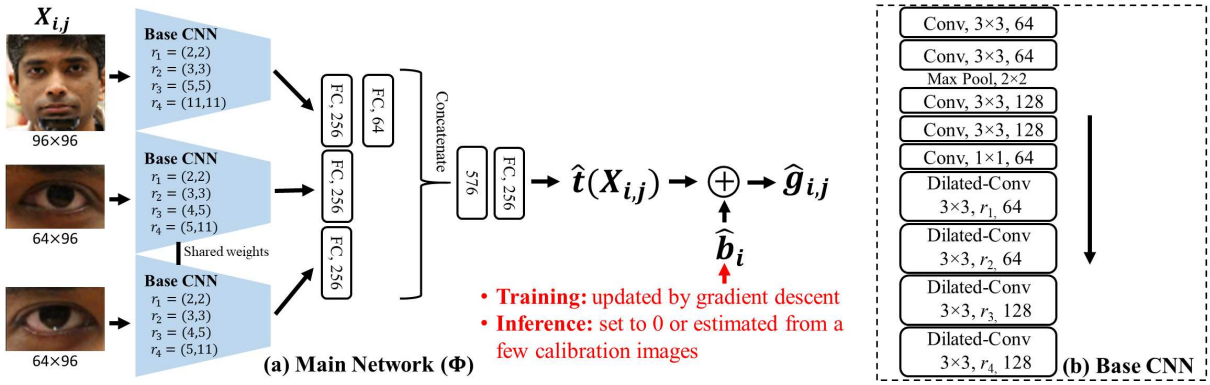

Figure 2. Architecture of the proposed network. (a) The main network that outputs estimate of the gaze of optic axis, $\hat{t}(X_{i,j};\Phi)$, from the input image $X_{i,j}$. $\hat{b}_i$ is the bias of subject i and $\hat{g}_{i,j}$ is estimate of the gaze of visual axis. (b) The base CNN is the basic component of (a). FC/Conv/Dilated-Conv denote fully-connected/convolutional/dilated-convolutional layers. r is the dilation rate.

(FC) layers, concatenated, fed to another FC layer followed by a linear output layer which outputs $\hat{t}(X_{i,j})$. We denote the parameters of this network by Φ.

Rectified Linear Units (ReLUs) are used as the activation functions. Zero-padding is applied to regular convolutional layers. No padding is applied to dilated-convolutional layers. The weights of the first four convolutional layers are initialized from VGG-16 [35] pre-trained on the ImageNet dataset [8]. Batch renormalization layers [17] are applied to all layers trained from scratch. Dropout layers with dropout rates of 0.5 are applied to all FC layers.

Training. Based on Eq. (1), we train the network by solving the following optimization problem:

$$\min_{\Phi,\hat{b}_i}\left(\sum_{i,j}\left\|g_{i,j}-\hat{t}(X_{i,j};\Phi)-\hat{b}_i\right\|_2^2+\lambda\left|\sum_i \hat{b}_i\right|\right). \quad (2)$$

The second term is a regularizer that ensures that the mean subject-dependent bias over the training set is zero. We introduce this constraint because the estimate $\hat{t}$ may have an arbitrary offset which could be canceled by the mean bias. Since this mean bias is arbitrary, the training is insensitive to the value of λ. The $\hat{b}_i$ could also be estimated by calculating the individual mean over the training set, but this would be time-consuming, given the large size of training set, especially since we use online data augmentation.

We use Adam optimizer with default parameters in TensorFlow and a batch size of 64. An initial learning rate of 0.001 is used. It is divided by 10 after every ten epochs. The training proceeds for 35 epochs. We apply online data augmentation including random cropping, scaling, rotation and horizontal flipping. As the bias changes if the images are flipped horizontally, we considered the non-flipped and flipped images as belonging to different subjects.

Testing and calibration. During testing, estimates were computed according to Eq. (1). For a new subject m, if no calibration images were available, we set $\hat{b}_m = 0$.

A calibration set $\mathcal{D}_m$ contains image-gaze pairs for a subject m, $\{(X_{m,j}, g_{m,j}), j = 1, 2, \ldots, |\mathcal{D}_m|\}$, where $|\mathcal{D}_m|$ denotes the cardinality of $\mathcal{D}_m$. We measure the complexity of the calibration set by the number of gaze targets T and the number of images per gaze target S. Thus, $S \cdot T = |\mathcal{D}_m|$. For best performance, the images should capture the expected variability, e.g. in head pose and/or illumination, during testing.

To conduct calibration, we set $\hat{b}_m$ as follows:

$$\hat{b}_m = \frac{1}{|\mathcal{D}_m|}\sum_{(X_{m,j},g_{m,j})\in\mathcal{D}_m}\left(g_{m,j}-\hat{t}(X_{m,j})\right). \quad (3)$$

Unlike [21, 45, 26], which require back propagation for calibration, our proposed calibration only requires forward propagation, reducing computational cost.

Preprocessing. We apply the modified data normalization method introduced in [47]. This method rotates and scales an image so that the resulting image is taken by a virtual camera facing a reference point on the face from a fixed distance and canceling the roll angle of the head. The images are normalized by perspective warping, converted to gray scale and histogram-equalized. To automatically detect landmarks we use dlib [18].

4. Experiments

We evaluated our proposed network through cross-subject evaluation both within- and cross-datasets as well as both with and without calibration. We evaluated two calibration scenarios: single gaze target calibration (**SGTC**) with multiple images per target ($T = 1$, S variable) and multiple gaze target calibration (**MGTC**) with multiple targets and with a single image ($S = 1$, T variable).

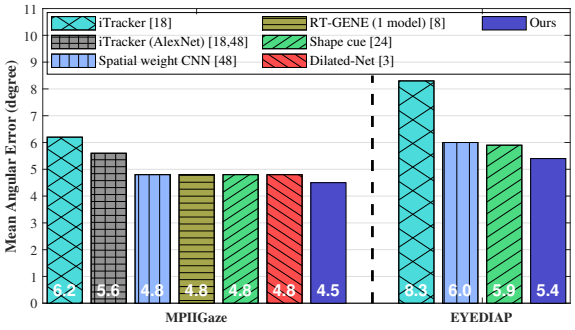

Figure 3. Mean angular error of estimation without calibration on MPIIGaze and EYEDIAP.

4.1. Datasets

We used the MPIIGaze [49] and the EYEDIAP [11] datasets for within-dataset evaluation. We trained on the MPIIGaze and tested on the ColumbiaGaze [36] for cross-dataset evaluation.

The MPIIGaze dataset contains full face images of 15 subjects (six females, five with glasses). We train and test on the "Evaluation Subset", which contains $3,000$ images for each subject. Within this subset, half of the images are flipped horizontally. The reference point for image normalization is set to the center of the face.

The EYEDIAP dataset contains full face videos with three gaze targets (continuous screen target, discrete screen target and floating target) and two types of head pose (static and dynamic). We use the data from continuous and discrete screen targets, which include 14 subjects (three females, none with glasses). The reference point for image normalization is set to the midpoint of both eyes.

The ColumbiaGaze dataset contains $5,800$ full face images of 56 subjects (24 females, 21 with glasses). For each subject, images are collected for each combination of five horizontal head poses $(0°, \pm15°, \pm30°)$, seven horizontal gaze directions $(0°, \pm5°, \pm10°, \pm15°)$ and three vertical gaze directions $(0°, \pm10°)$. We exclude the images with $10°$ vertical gaze directions from ColumbiaGaze for these evaluations, since the MPIIGaze dataset mainly covers pitch angles from $-20°$ to $0°$.

4.2. Within-dataset evaluation

Estimation without calibration. For MPIIGaze, we conducted 15-fold leave-one-subject-out cross-validation as in [3, 9, 50]. For EYEDIAP, we followed the protocol described in [50], i.e., five-fold cross-validation on four VGA videos (both screen targets with both types of head pose) sampled at 2 fps (about 1,200 images per subject).

We compared with several baselines: iTracker [19, 50], spatial weights CNN [50], RT-GENE [9], CNN with shape cue [25] and Dilated-Net [3]. All of these methods used face images (or face plus eye images) as input.

The results are shown in Fig. 3. On MPIIGaze, our proposed network achieved $4.5°$ mean angular error, which outperformed the state-of-the-art $4.8°$ [3, 9, 25, 50] by 6.3%. On the EYEDIAP dataset, it achieved $5.4°$, which outperformed the state-of-the-art $5.9°$ [25] by 8.5%. The re-implemented result of the CNN with shape cue on EYEDIAP was worse than that reported in [25] because we used different subset of the data and we did not filter the data as in [25]. For consistency with the other models, we compare only with RT-GENE without ensembling. Using an ensemble of four models, RT-GENE achieved an accuracy of $4.3°$ on MPIIGaze. Applying ensembling to our model did not improve performance.

Calibration procedures. To evaluate calibration, for each subject, we first sampled a calibration set $\mathcal{D}_m$ from the test set and calibrated using Eq. (3). We then tested on the images not belonging to $\mathcal{D}_m$.

For multiple gaze target calibration (MGTC), we randomly selected T images from the test set as $\mathcal{D}_m$. For single gaze target calibration (SGTC), we first randomly selected a calibration target in gaze space. We then randomly selected S images whose gaze angles were within $2°$ of the target as $\mathcal{D}_m$. We discarded a calibration gaze target if less than S images met the $2°$ requirement.

We also calculated the lower bound performance of our proposed method, $\underline{E}$, by estimating the bias from all images in the test set. We denote the mean error of the subject-independent estimator without calibration by $\overline{E}$.

Calibration results on MPIIGaze. To evaluate the overall performance of MGTC and SGTC across different calibration sets, for each value of S or T, we report average errors computed over $5,000$ trials for each subject.

We compared our approach with several existing calibration methods: fine-tuning the last FC layer (**FC**), linear adaptation (**LA**) [22, 37], third order polynomial adaptation (**PA**) [48], the differential method (**DF**) [22], fine-tuning the latent parameters (**LP**) [21], gaze redirection synthesis (**GRS**) [45] and **FAZE** [26]. FC, LA and PA were directly applied to our network trained with gaze decomposition, where the Moore–Penrose inverse was used for LA and PA. DF and LP were re-implemented using the same architecture as in Fig. 2. For GRS and FAZE, we used their reported results for within-dataset leave-one-subject-out cross-validation. As both GRS and FAZE only used eye images as input, we trained our network without the face component in Fig. 2 for comparison.

Table 1 and Table 2 show that for all methods, as the complexity of calibration set increased (T or S increased), the error decreased. Our proposed method performed the best for low complexity calibration sets. It outperformed or

	Backbone	Number of images per gaze target S			
		1	5	9	16
FC	Dilated+GD	4.8 ± 0.8	7.1 ± 1.2	8.0 ± 1.2	8.4 ± 1.1
LA [22, 37]	Dilated+GD	*NA*	15.1 ± 2.7	14.9 ± 2.5	14.3 ± 2.1
PA [48]	Dilated+GD	*NA*	*NA*	14.7 ± 2.4	14.3 ± 2.1
LP [21]	Dilated	4.2 ± 1.3	3.9 ± 0.8	3.8 ± 0.7	3.7 ± 0.6
DF [22]	Dilated	4.2 ± 1.5	3.5 ± 0.7	3.4 ± 0.5	3.3 ± 0.4
Ours (w/o GD)	Dilated	4.2 ± 1.4	3.6 ± 0.7	3.5 ± 0.5	3.4 ± 0.4
Ours	Dilated+GD	$\mathbf{3.7} \pm 1.4$	$\mathbf{3.1} \pm 0.6$	$\mathbf{3.0} \pm 0.4$	$\mathbf{2.9} \pm 0.3$

Table 1. Estimation Error (mean $\pm$ SD in degree) of SGTC ($T = 1$) on MPIIGaze.

Face+Eye	Backbone	Number of gaze targets T						
		1	5	9	16	32	64	128
FC	Dilated+GD	4.8 ± 0.8	5.5 ± 1.5	4.5 ± 1.0	3.5 ± 0.5	2.9 ± 0.2	2.7 ± 0.1	2.5 ± 0.1
LA [22, 37]	Dilated+GD	*NA*	4.7 ± 3.0	3.1 ± 0.7	2.8 ± 0.3	$\mathbf{2.6} \pm 0.1$	$\mathbf{2.5} \pm 0.1$	$\mathbf{2.4} \pm 0.0$
PA [48]	Dilated+GD	*NA*	*NA*	9.2 ± 3.5	3.8 ± 1.3	2.8 ± 0.3	2.6 ± 0.1	$\mathbf{2.4} \pm 0.0$
LP [21]	Dilated	4.2 ± 1.3	3.3 ± 0.4	3.2 ± 0.3	3.0 ± 0.1	3.0 ± 0.1	2.9 ± 0.1	2.9 ± 0.0
DF [22]	Dilated	4.2 ± 1.5	3.2 ± 0.4	3.1 ± 0.3	3.0 ± 0.1	$\mathbf{2.6} \pm 0.1$	2.6 ± 0.0	2.6 ± 0.0
Ours (w/o GD)	Dilated	4.2 ± 1.4	3.3 ± 0.4	3.1 ± 0.2	3.0 ± 0.1	3.0 ± 0.1	2.9 ± 0.0	2.9 ± 0.0
Ours	Dilated+GD	$\mathbf{3.7} \pm 1.4$	$\mathbf{2.9} \pm 0.4$	$\mathbf{2.7} \pm 0.2$	$\mathbf{2.7} \pm 0.1$	$\mathbf{2.6} \pm 0.1$	2.6 ± 0.0	2.6 ± 0.0
Eye only								
GRS [45]*	VGG16	5.0	4.2	4.0				
FAZE [26]*	DenseNet	4.7	4.0	3.9	3.8	3.8	3.7	3.7
Ours*	Dilated+GD	$\mathbf{4.6} \pm 1.2$	$\mathbf{3.6} \pm 0.5$	$\mathbf{3.4} \pm 0.3$	$\mathbf{3.4} \pm 0.3$	$\mathbf{3.3} \pm 0.1$	$\mathbf{3.3} \pm 0.0$	$\mathbf{3.3} \pm 0.0$

$*$: The three methods at the bottom only used eye images as input, while the others used face+eye images as input.

Table 2. Estimation Error (mean $\pm$ SD in degree) of MGTC ($S = 1$) on MPIIGaze.

matched the performance of other methods for all cases in SGTC and for MGTC when the number of images was less than or equal to 32. For example, for SGTC with 16 samples, our method reduced the error by 12.1% compared to the second best method (DF). We attribute this better performance to the small number of adaptive parameters, which avoids overfitting. Unlike LP, our method acts directly on the gaze estimates, making it more effective since the primary cause of error is a subject-dependent bias. As the number of gaze targets increased beyond 32, other methods (e.g. LA and PA) eventually outperformed ours. However, we believe that collecting images for more than 32 gaze targets for calibration would be too time-consuming for most real-world applications.

For our proposed method, SGTC reduced the estimation error significantly. For example, when calibrated on 16 samples, it reduced the error by 1.6° (35.6%) in comparison to the estimator without calibration. MGTC led to further reductions. On average, given the same number of images, the gap between SGTC and MGTC was about 0.2° (4.4% of the error of the estimator without calibration).

Robustness of SGTC to the location of gaze target. We further evaluated the robustness of SGTC to the location of gaze target. Fig. 4 presents the average errors for different gaze target locations. For each $5^\circ \times 5^\circ$ region in Fig. 4, we created a 10×10 square grid of calibration targets spaced by 0.5° along each axis and computed the average error of SGTC across all targets. We set $S = 9$, since the error reduction begins to saturate at this point. Since half of the images were flipped and we considered them separately when computing the bias and the accuracy, each subject generates two accuracies (30 in total). Note that none of the flipped/non-flipped images of the test subject were included in the training set.

The results in Fig. 4 show that the error achieved by calibrating at a target in the center of the gaze range is lower than the error achieved by calibrating at a target in the boundary. However, the standard deviation over locations is only 0.15°, indicating that SGTC is quite robust to the location of calibration target. The errors of at least 22 out of 30 subjects (73.3%) were reduced by calibration, dependent on the location of gaze target. For those calibrations where the error increased, the average increase was only 0.2°.

Calibration results on EYEDIAP. We conducted leave-one-subject-out cross-validation on two VGA videos

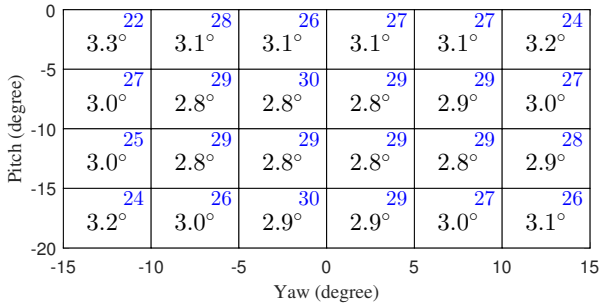

Figure 4. Mean angular error of SGTC ($S = 9$) when the gaze target is located within different $5° \times 5°$ regions on the MPIIGaze dataset. $\overline{E} = 4.5°$, $\underline{E} = 2.6°$. The upper-right value in each box indicates the number of subjects whose mean errors were reduced by calibration (30 subjects in total).

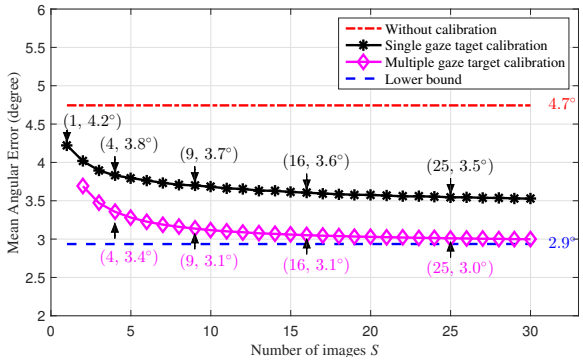

Figure 5. Mean angular error after calibration as a function of number of calibration samples on the EYEDIAP dataset.

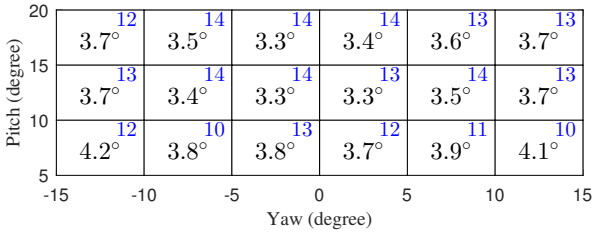

Figure 6. Mean angular error of SGTC ($S = 9$) when the gaze target is located within different $5° \times 5°$ regions on the EYEDIAP dataset. $\overline{E} = 4.7°$, $\underline{E} = 2.9°$, $S = 9$. 14 subjects in total.

(continuous screen target with static and dynamic head pose) sampled at 15 fps (about $3,500$ images per subject). We chose this data because it had sufficient annotation to remove outliers, e.g. images during blinks.

The results are shown in Fig. 5 and Fig. 6. The findings are generally consistent with those of MPIIGaze. The overall performance is worse than that on MPIIGaze, most likely due to lower image resolution and larger head pose variability. As shown in Fig. 5, SGTC reduced the error significantly. A $1.2°$ (25.5%) improvement was achieved when calibrated on 25 samples. As shown in Fig. 6, $0.5°$ (10.6%) to $1.4°$ (29.8%) improvement versus without calibration ($4.7°$) was achieved across different locations. The standard deviation of the error across locations was small ($0.26°$). The errors of at least 10 out of 14 subjects (71.4%) were reduced by calibration. For subjects whose error increased, the average increase was $0.4°$.

Without calibration, the average error achieved in this experiment ($4.7°$) was better than that in the previous experiment ($5.4°$, Fig. 3), most likely due to the larger size of the training set. Videos were sampled at 15 fps for this experiment, but only 2fps for the experiment of Fig. 3. In addition, outliers were removed in this experiment using the extra annotation available for this subset.

4.3. Cross-dataset evaluation

We trained on MPIIGaze and tested on ColumbiaGaze. Fig. 7 shows the performance of SGTC ($S = 5$) for each calibration gaze target, which is higher than the within dataset results of MPIIGaze (Fig. 4) and EYEDIAP (Fig. 6), indicating the difficulty of cross-dataset evaluation. Consistent with our previous results, the performance was poor when calibrated at the four gaze targets at the boundary, i.e., samples with horizontal directions $\pm 15°$. However, for the middle ten calibration targets, SGTC reduced the error by $0.8°$ (14.5%) to $1.1°$ (20.0%) in comparison to that without calibration. For 46 to 51 out of the 56 subjects (82.1% to 91.1%), errors were reduced by calibration. For subjects whose error increased, the average increase was $0.4°$.

Our method also outperformed LP by $0.8°$ (14.5%) and DF by $0.6°$ (11.3%) in this cross-dataset evaluation (see Table A1 in the supplementary materials).

4.4. Ablation studies

We evaluated the importance of applying the gaze decomposition (**GD**) during training by training a network without gaze decomposition, i.e., by minimizing the squared error between the estimated and ground truth gaze angles without the bias term $\hat{b}$ in Eq. (2), which we refer to as **ND**. We calibrated this network in the same way as the network trained with gaze decomposition by adding the bias estimated by Eq. (3). As shown in Fig. 8, applying gaze decomposition reduced the error on MPIIGaze both without calibration (from $4.7°$ to $4.5°$, 4.3%) and with SGTC (by about $0.35°$ across all values of S). We obtained similar results on EYEDIAP. For example, the error without calibration was reduced from $5.8°$ to $5.4°$ (6.9%).

We also repeated the cross-dataset evaluation. The results of ND are shown in Fig. 9, where $\overline{E} = 5.5°$ and $\underline{E} = 4.5°$. Both GD and ND achieved the same $5.5°$ in estimation without calibration. However, on average, a gain of $0.4°$ was achieved by gaze decomposition after calibration ($T = 1$, $S = 5$).

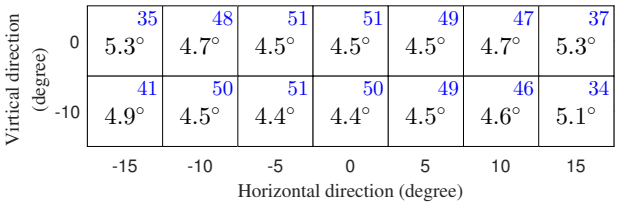

Figure 7. Mean angular error when calibrated at different gaze targets ($T = 1, S = 5$). Trained on MPIIGaze and tested on Columbia. $\overline{E} = 5.5°, \underline{E} = 4.1°$. 56 subjects in total.

We evaluated the extent to which the gains relative to the state-of-the-art reported here were due to the use of gaze decomposition or due to the use of dilated-convolution by training a network similar to Fig. 2 but without dilated-convolution. Its base CNNs had eight convolutional layers and four max-pooling layers. In leave-one-subject-out cross-validation on MPIIGaze, the standard CNN without calibration had errors of $5.3°$ when trained without the gaze decomposition and $4.7°$ with gaze decomposition. In comparison, the dilated CNN had errors of $4.7°$ without gaze decomposition and $4.5°$ with gaze decomposition. For SGTC ($T = 1, S = 9$), the errors were $4.1°$ (standard CNN/ND), $3.6°$ (standard CNN/GD), $3.5°$ (dilated CNN/ND) and $3.0°$ (dilated CNN/GD). For MGTC ($T = 9, S = 1$), the errors were $3.6°$ (standard CNN/ND), $3.2°$ (standard CNN/GD), $3.1°$ (dilated CNN/ND) and $2.7°$ (dilated CNN/GD). We conclude that the use of gaze decomposition and the use of dilated CNNs are equally effective in reducing the error (by about $0.5°$), and that their effect is cumulative. Applying one after the other reduces error further by $0.2°$ to $0.5°$.

4.5. Consistency of the learned bias

We evaluated whether the learned biases were consistent for the same subject using the data from leave-one-subject-out (15 fold) cross-validation on MPIIGaze. We compared the intra-subject variance computed from the 14 folds where the subject was in the training set with the inter-subject variance computed from the means of the estimated biases. For yaw, the average intra-subject variance was 0.03 deg^2 in comparison to the inter-subject variance of 5.40 deg^2. For pitch, the variances were 0.05 deg^2 and 3.66 deg^2. The intra-subject variance was a small percentage (0.56%-1.4%) of the inter-subject variance, indicating that the bias is learned consistently and reliably during training. The mean and SD for each subject is provided in Table A2 of the supplementary materials.

5. Conclusions

We proposed a novel gaze decomposition method for appearance-based gaze estimation. We conducted experiments on the MPIIGaze, the EYEDIAP and the ColumbiaGaze datasets. Without calibration, the proposed method outperformed state-of-the-art methods without ensembling on the MPIIGaze and EYEDIAP datasets. Further work must be done to determine why ensembling did not help our method as much as it did in [9]. The proposed method reduced the estimation error significantly with low complexity calibration sets. For example, it reduced the estimation error by 33.3% while only requiring 9 images looking at one target for calibration.

For best performance, the multiple images per gaze target should contain variations in head pose. In most real-world applications, it is easy to ask the user to move his/her head to provide variations in head pose. Including variations in illumination may also help if we expect to encounter those during use. MPIIGaze includes significant variability in illumination per subject, as data collection was over multiple days. However, EYEDIAP and ColumbiaGaze were collected under fairly constant illumination. Our proposed method provided significant reduction in error in all cases.

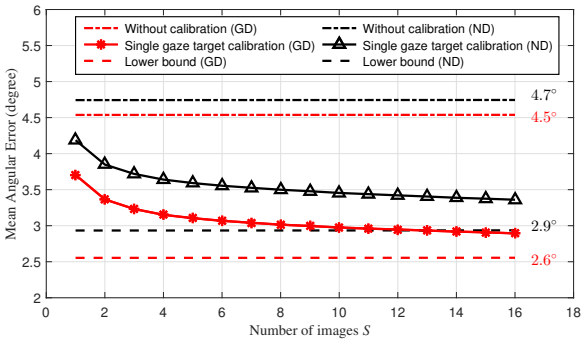

Figure 8. Comparison of the networks with/without gaze decomposition (GD and ND) on MPIIGaze.

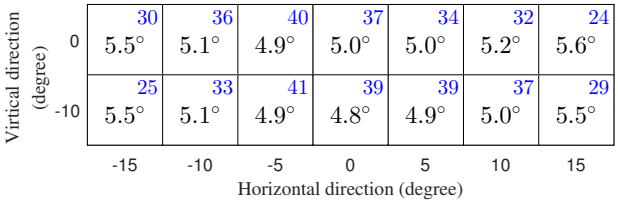

Figure 9. Mean angular error achieved by the network without gaze decomposition when calibrated at different gaze targets ($T = 1, S = 5$). Trained on MPIIGaze and tested on Columbia. $\overline{E} = 5.5°, \underline{E} = 4.5°$. 56 subjects in total.

Acknowledgements

This work was supported in part by the Hong Kong Innovation and Technology Fund under Grant ITS/406/16FP.

References

[1] D. A. Atchison and G. Smith. *Optics of the human eye*. Butterworth-Heinemann, Oxford, 2000.

[2] E. Brau, J. Guan, T. Jeffries, and K. Barnard. Multiple-gaze geometry: Inferring novel 3d locations from gazes observed in monocular video. In *Proceedings of the European Conference on Computer Vision*, pages 641–659. Springer, 2018.

[3] Z. Chen and B. E. Shi. Appearance-based gaze estimation using dilated-convolutions. In *Asian Conference on Computer Vision*, pages 309–324. Springer, 2018.

[4] Z. Chen and B. E. Shi. Using variable dwell time to accelerate gaze-based web browsing with two-step selection. *International Journal of Human–Computer Interaction*, 35(3):240–255, 2019.

[5] Y. Cheng, F. Lu, and X. Zhang. Appearance-based gaze estimation via evaluation-guided asymmetric regression. In *Proceedings of the European Conference on Computer Vision*, pages 105–121. Springer, 2018.

[6] 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 *Proceedings of the European Conference on Computer Vision*, pages 397–412. Springer, 2018.

[7] H. Deng and W. Zhu. Monocular free-head 3d gaze tracking with deep learning and geometry constraints. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3162–3171. IEEE, 2017.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[9] T. Fischer, H. J. Chang, and Y. Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of European Conference on Computer Vision*, pages 339–357. Springer, 2018.

[10] W. Fuhl, D. Geisler, T. Santini, T. Appel, W. Rosenstiel, and E. Kasneci. Cbf: Circular binary features for robust and real-time pupil center detection. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, page 8. ACM, 2018.

[11] 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 *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 255–258. ACM, 2014.

[12] A. Grillini, D. Ombelet, R. S. Soans, and F. W. Cornelissen. Towards using the spatio-temporal properties of eye movements to classify visual field defects. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, page 38. ACM, 2018.

[13] E. D. Guestrin and M. Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on Biomedical Engineering*, 53(6):1124–1133, 2006.

[14] Z. He, A. Spurr, X. Zhang, and O. Hilliges. Photo-realistic monocular gaze redirection using generative adversarial networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[15] S. Hoppe, T. Loetscher, S. A. Morey, and A. Bulling. Eye movements during everyday behavior predict personality traits. *Frontiers in Human Neuroscience*, 12:105, 2018.

[16] C.-M. Huang and B. Mutlu. Anticipatory robot control for efficient human-robot collaboration. In *ACM/IEEE International Conference on Human Robot Interaction*, pages 83–90. IEEE, 2016.

[17] S. Ioffe. Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. In *Advances in Neural Information Processing Systems*, pages 1942–1950, 2017.

[18] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758, 2009.

[19] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba. Eye tracking for everyone. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2176–2184, 2016.

[20] D. Lian, L. Hu, W. Luo, Y. Xu, L. Duan, J. Yu, and S. Gao. Multiview multitask gaze estimation with deep convolutional neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2018.

[21] E. Lindén, J. Sjöstrand, and A. Proutiere. Learning to personalize in appearance-based gaze tracking. *IEEE International Conference on Computer Vision Workshops*, 2019.

[22] G. Liu, Y. Yu, K. A. Funes-Mora, J.-M. Odobez, and E. T. SA. A differential approach for gaze estimation with calibration. In *British Machine Vision Conference*, 2018.

[23] R. Menges, C. Kumar, D. Müller, and K. Sengupta. Gazetheweb: A gaze-controlled web browser. In *Proceedings of the Web for All Conference on The Future of Accessible Work*, page 25. ACM, 2017.

[24] B. I. Outram, Y. S. Pai, T. Person, K. Minamizawa, and K. Kunze. Anyorbit: Orbital navigation in virtual environments with eye-tracking. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, page 45. ACM, 2018.

[25] C. Palmero, J. Selva, M. A. Bagheri, and S. Escalera. Recurrent cnn for 3d gaze estimation using appearance and shape cues. In *British Machine Vision Conference*, 2018.

[26] S. Park, S. De Mello, P. Molchanov, U. Iqbal, O. Hilliges, and J. Kautz. Few-shot adaptive gaze estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[27] S. Park, A. Spurr, and O. Hilliges. Deep pictorial gaze estimation. In *Proceedings of the European Conference on Computer Vision*, pages 741–757. Springer, 2018.

[28] S. Park, X. Zhang, A. Bulling, and O. Hilliges. Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, page 21. ACM, 2018.

[29] A. Patney, M. Salvi, J. Kim, A. Kaplanyan, C. Wyman, N. Benty, D. Luebke, and A. Lefohn. Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions on Graphics*, 35(6):179, 2016.

[30] J. Pi and B. E. Shi. Probabilistic adjustment of dwell time for eye typing. In *International Conference on Human System Interactions*, pages 251–257. IEEE, 2017.

[31] J. Pi and B. E. Shi. Task-embedded online eye-tracker calibration for improving robustness to head motion. In *Pro-*

ceedings of the ACM Symposium on Eye Tracking Research & Applications, page 8. ACM, 2019.

[32] R. Ranjan, S. De Mello, and J. Kautz. Light-weight head pose invariant gaze tracking. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2156–2164. IEEE, 2018.

[33] T. Schneider, B. Schauerte, and R. Stiefelhagen. Manifold alignment for person independent appearance-based gaze estimation. In *International Conference on Pattern Recognition*, pages 1167–1172. IEEE, 2014.

[34] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2242–2251. IEEE, 2017.

[35] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[36] B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar. Gaze locking: Passive eye contact detection for human-object interaction. In *Proceedings of the ACM Symposium on User Interface Software and Technology*, pages 271–280. ACM, 2013.

[37] M. A. Strobl, F. Lipsmeier, L. R. Demenescu, C. Gossens, M. Lindemann, and M. De Vos. Look me in the eye: Evaluating the accuracy of smartphone-based eye tracking for potential application in autism spectrum disorder research. *Biomedical engineering online*, 18(1):51, 2019.

[38] Y. Sugano, Y. Matsushita, and Y. Sato. Learning-by-synthesis for appearance-based 3D gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1821–1828. IEEE, 2014.

[39] H. Wang, J. Pi, T. Qin, S. Shen, and B. E. Shi. SLAM-based localization of 3D gaze using a mobile eye tracker. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, page 65. ACM, 2018.

[40] K. Wang and Q. Ji. Real time eye gaze tracking with 3d deformable eye-face model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1003–1011, 2017.

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

[42] E. Wood, T. Baltrušaitis, L.-P. Morency, P. Robinson, and A. Bulling. Learning an appearance-based gaze estimator from one million synthesised images. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, pages 131–138. ACM, 2016.

[43] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[44] Y. Yu, G. Liu, and J.-M. Odobez. Deep multitask gaze estimation with a constrained landmark-gaze model. In *European Conference on Computer Vision Workshops*, pages 456–474. Springer, 2018.

[45] Y. Yu, G. Liu, and J.-M. Odobez. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11937–11946, 2019.

[46] X. Zhang, M. X. Huang, Y. Sugano, and A. Bulling. Training person-specific gaze estimators from user interactions with multiple devices. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, page 624. ACM, 2018.

[47] X. Zhang, Y. Sugano, and A. Bulling. Revisiting data normalization for appearance-based gaze estimation. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, page 12. ACM, 2018.

[48] X. Zhang, Y. Sugano, and A. Bulling. Evaluation of appearance-based methods and implications for gaze-based applications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, page 416. ACM, 2019.

[49] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4511–4520, 2015.

[50] 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. IEEE, 2017.

[51] 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.