Kappa Angle Regression with Ocular Counter-Rolling Awareness for Gaze Estimation

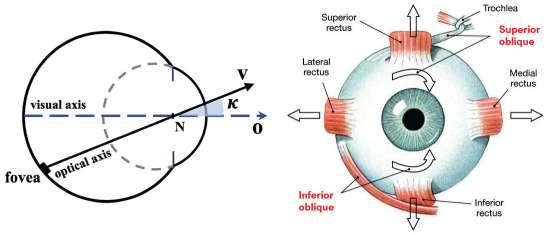
Shiwei Jin, Ji Dai, Truong Nguyen
ECE Dept. UC San Diego.
{sjin, jid046, tqn001}@eng.ucsd.edu

Abstract

Conventional appearance-based 3D gaze estimation methods generally use the roll of the head pose to represent the eyeball's roll status by default. To reduce degrees of freedom of head poses, a normalization step was proposed to apply global transformations to images to make heads upright and eyelids horizontal. However, due to the ocular countering-rolling (OCR) response, the eyeball will rotate in the opposite direction when the head tilts to the side. After normalization, the eyeball will have an extra roll compared to the roll status of the eyeball when the head is not tilted. This roll from the OCR response causes a changed orientation of the eyeball in normalized eye images, which represents the roll status of the anatomical structure inside the eyeball and consequently affects gaze directions. Thus in this work, we propose a pipeline to regress the person-dependent anatomical variation as a calibration process with considering the OCR response, which can work with our proposed eye-image-based person-independent gaze estimator trained with real and synthetic eye images. The proposed method firstly brings the OCR response into the gaze estimation task, achieving better performances on the two benchmark datasets with fewer parameters under the real-time scenarios. With a replacement of a deeper network, compared to state-of-the-art methods, the proposed method is more efficient, achieving a). better average estimate (3.9% and 2.5% improvement), b). much better standard deviation (lower by 59.0% and 44.2%) and c). a much lower number of parameters (reduced by 88.0%).

1. Introduction

Human gaze is an essential indicator for many applications such as human-computer interaction [9, 23], health assessment [14], automotive assistance [29, 32] and virtual reality [26, 45]. Non-invasive appearance-based gaze estimation methods enjoyed significant improvements [20, 27, 39] for in-the-wild settings due to the development of the Convolutional Neural Network (CNN). However, they still struggle with achieving high accuracy due to the challenges caused by variations of head poses [12, 41], noisy and limited annotations [27], eye shapes and anatomical variations of different subjects [15, 21], etc.



(a) Eyeball Structure (b) Eyeball Muscles and Motions

Figure 1. Eyeball structure and muscles. (a) Kappa Angle $\boldsymbol{\kappa}$ is defined as the angle between visual axis $\boldsymbol{V}$ (the line connecting the fovea and nodal point N, which defines gaze and is unobserved) and optic axis $\boldsymbol{O}$ (the line connecting the eyeball center and pupil center, which is related to the observed iris [15]). (b) The arrows show the eyeball motions controlled by the corresponding muscles. The oblique muscles are used for the eyeball roll motion [37].

Several techniques, ranging from normalization for data pre-processing [31, 41] to individual-specific calibration after training [20], were proposed to reduce the variations stated above. Image normalization's fundamental idea is reducing the degrees of freedom of the object pose from six (head poses: pitch, yaw, roll and position: x, y, z) to two (head poses: pitch, yaw) by perspective image warping. This normalization step facilitates mapping from images to gaze directions across different samples or even datasets [44]. Another source of variation causing limited accuracy with a person-independent gaze estimator emerges from the anatomical structures of the eyes. As shown in Fig. 1 (a), the visual axis is not aligned with the optic axis (related to the observed iris) [15], and such alignment differences, called 'Kappa Angle', are subject-specific. Given this unobserved anatomical variation across different subjects, person-dependent calibration methods such as gaze

differences estimation [21], models calibration with meta-learning [27], and personalized parameters regression [20] were proposed, which further improved gaze estimation performance with a few calibration samples.

However, normalization focuses more on global transformations of images according to head poses and ignores independent eyeball motions. It obeys the human eyeball movement response called ocular counter-rolling (OCR). When the head tilts to the side, the OCR response consists of a torsional conjugate eye movement opposite the static head roll direction around the optic axis [8]. As presented in Fig. 2, when the head has a roll motion, the eyeball will have an opposite roll motion to maintain the initial horizontal status instead of rotating together with the head given the indications from iris patterns [24]. After normalization is applied to images, the head and eyelids are transformed to the upright status, but the eyeball's orientation in normalized images still has an extra roll caused by OCR. This extra roll is difficult to acquire from low-resolution eye images and its highly-related variable [25, 28], the roll of the head pose, is abandoned after normalization. Failing to account for the eyeball's counter-rolling movement is undesired because this movement causes different roll status of the eyeball, which implies different fovea locations and consequently changes gaze directions, shown in Fig. 2. Inspired by this observation, we propose a new framework for gaze estimation, which considers OCR during the regression of the person-dependent variable: the Kappa Angle.

Our contributions are:

1) Propose to utilize the OCR response that is obtained by considering the commonly ignored roll of the head pose after normalization, in order to achieve a more precise regression of the Kappa Angle.

2) Integrate the OCR-aware Kappa Angle regression part with a unified eye-image-based gaze estimator to achieve person-dependent calibration during training and evaluation.

3) Present a comparable estimation accuracy and much lower standard deviation with fewer network parameters on benchmark datasets, which indicates the effectiveness of our proposed KAComp-Net.

2. Related Work

2.1. Appearance-Based Gaze Estimation

Appearance-based gaze estimation methods aim at mapping eye-containing images the gaze directions (2D screen locations or 3D gaze direction vectors), which achieved significant improvements [12, 42] compared with geometric approaches [11, 33, 36] given supports from several large-scale datasets [10, 18, 19, 44] and constantly evolving deep learning techniques. GazeNet [42] was the first learning-based 3D gaze estimation method that took one eye image

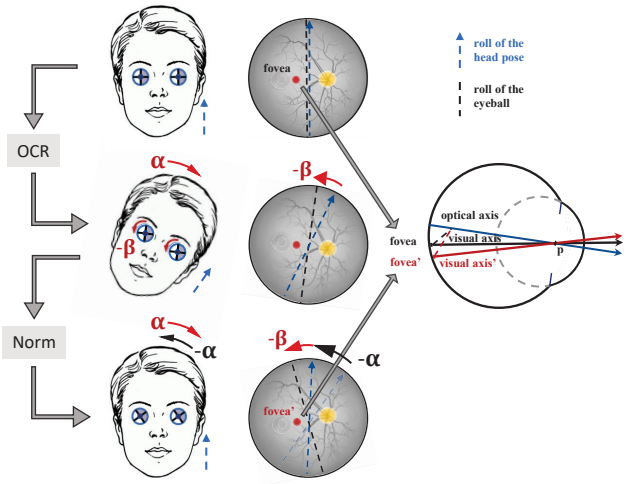


Figure 2. Changed gaze directions caused by the processes of OCR and normalization. **OCR stage**: When a subject rotates his head with a roll of α, the eyeball will rotate in the opposite direction with a roll of $-\beta$. **Norm stage**: After normalization, the roll of the head pose is normalized to zero by an rotation of $-\alpha$. But the current eyeball still has a roll with $-\beta$, leading to different fovea locations and altering the direction of the visual axis.

as the input. Except for the single input, multi-branch's inputs included both eyes inputs [6, 7, 21]; full-face inputs [38, 43]; and multi-model inputs [5, 7, 19] were proposed and achieved improvements given extra informative data, which were at the cost of calculation complexity and memory requirements. More recently, person-dependent calibration [4, 20, 21, 27, 40] (or domain adaptation [2, 22, 34]) approaches were proposed, which attempted to remove personal variations (or domain gaps) with a few annotated (or unannotated) samples. Strobl *et al.* [30] utilized the features from a person-independent model over the test subject's data to further train a person-specific Support Vector Regression for personalized gaze estimation. Liu *et al.* [21] proposed learning the gaze difference between two images of the same eye to remove the unobserved person-dependent variables. Chen *et al.* [4] decomposed gaze into a person-independent component estimated from images and a person-dependent bias regressed as network parameters. Liu *et al.* [22] used an ensemble of networks for collaborative learning, guided by outliers. Bao *et al.* [2] introduced the constraint of rotation consistency for unsupervised domain adaptation. Our method follows this gaze decomposition idea. A unified gaze estimator was utilized for estimating the person-independent component of gaze and the person-dependent part was regressed by including OCR.

2.2. Gaze Redirection

Given the need for large amounts of labeled data for training a robust gaze estimator, several conditional image

synthesis methods were proposed to generate images with desired gaze directions. Ganin *et al.* [13] proposed learning warping fields to rearrange the pixels' locations for gaze redirection given the input images. Yu *et al.* [39] introduced a cycle pipeline with semantic segmentation consistency to supervise warping-field-based gaze redirection and they [40] further extended the warping-field-based methods with an unsupervised learning strategy by representation learning. Besides warping-field-based methods, He *et al.* [17] first applied the Generative Adversarial Network to the gaze redirection task, generating photo-realistic eye images with desired gaze directions. Park *et al.* [27] proposed FAZE: an encoder-decoder structure to change gaze directions and head orientations in feature space with desired labels. Zheng *et al.* [46] presented ST-ED, which first advanced the encoder-decoder method from eye images to full-face images. However, these gaze redirection methods did not consider modeling person-independent components of gaze. Much more accurate gaze estimation results trained and tested only with synthetic data also proved it. Thus in our work, we utilized ST-ED to generate synthetic face images with desired gaze directions for learning the person-independent component of the gaze.

2.3. Ocular Counter-Rolling

Ocular Counter-Rolling (OCR) is a partially compensatory torsional eye movement only when the head is tilted toward the shoulder [8]. In particular, the OCR response of human eyeballs is controlled separately by the surrounding muscles called superior and inferior obliques [37], shown in Fig. 1 (b). When the head is tilted with α toward the shoulder in a natural pose, these muscles make the eyeball rotate in the opposite direction with β, as shown in Fig. 2. After normalization, eye images look horizontally orientated, but the eyeball and the fovea location are tilted with an extra roll (β) owing to the OCR response. This extra roll (β) caused by OCR ***doesn't change*** the absolute value of the Kappa Angle, which is always invariant for the same subject. It only ***redistributes*** the pitch and yaw components of the Kappa Angle. Thus we can compensate this redistribution (counteract OCR) on pitch and yaw components of the Kappa Angle by applying a rotation matrix built by β. Given this, we proposed a Kappa Angle compensation method with OCR awareness, elaborated in Section 3.

3. Method

In this section, we will firstly discuss the cases without or with considering ocular counter-rolling (OCR). Secondly, we will show the difference between real and synthetic data based on some simulation results. Thirdly, we will introduce the training and evaluation pipeline with considering OCR and the person-dependent part of gaze. Lastly, we will introduce loss functions for supervising the whole process.

3.1. Processes W/O or W/ OCR

Fig. 1 (a) illustrates that the Kappa Angle ($\boldsymbol{\kappa}$) represents the angle between the optic axis ($\boldsymbol{O}$) and the visual axis ($\boldsymbol{V}$), and is dependent on the individual.

$$\boldsymbol{O} + \boldsymbol{\kappa} = \boldsymbol{V}, \tag{1}$$

where $\boldsymbol{O}$, $\boldsymbol{V}$ and $\boldsymbol{\kappa} \in \mathbb{R}^{2\times 1}$ (2D vectors representing **pitch and yaw**), and hence we can use the addition to depict the 3D relationship of these variables. According to Atchison's study [1], the absolute angle value (norm of pitch and yaw) of the Kappa Angle remains constant for the same subject. However, if the eyeball's roll status changes with respect to the head coordinate system, the pitch and yaw of the Kappa Angle will adjust accordingly, as depicted in Fig. 2.

W/O OCR: Diff-NN [21] is a typical method without considering OCR in the gaze estimation task. While the optical axis $\boldsymbol{O}$ can be estimated from images using a unified model, there is no ground truth available. On the other hand, the gaze direction $\boldsymbol{V}$ does have ground truth. However, because the Kappa Angle is person-specific and not directly observable from images, it is not possible to estimate $\boldsymbol{V}$ using images alone. To address this, Diff-NN estimates the difference in gaze by subtracting the unobservable Kappa Angle and leveraging the available ground truth. Given two images ($\boldsymbol{I}_1, \boldsymbol{I}_2$) from the same eye, the gaze difference is

$$\boldsymbol{V}_1 - \boldsymbol{V}_2 = (\boldsymbol{O}_1 + \boldsymbol{\kappa}_1) - (\boldsymbol{O}_2 + \boldsymbol{\kappa}_2), \tag{2}$$

where the subscripts denote variables related to the respective images. If we don't consider OCR during gaze estimation, the pitch and yaw of the Kappa Angle maintain constant regardless of different head poses between images. In this case, Eq. 2 simplifies to

$$\boldsymbol{V}_1 - \boldsymbol{V}_2 = \boldsymbol{O}_1 - \boldsymbol{O}_2 \text{ if } \boldsymbol{\kappa}_1 = \boldsymbol{\kappa}_2, \tag{3}$$

indicating the scenario without considering OCR.

W/ OCR: Due to the presence of OCR response, the eyeball undergoes an additional roll ($-\beta$) that counteracts the roll motion of the head (α), as illustrated in Fig. 2. Even after normalization where the head roll is removed, the extra roll ($-\beta$) of the eyeball persists in the normalized eye images. As a result, the pitch and yaw of the Kappa Angle vary for the same subject's data and consequently Eq. 3 is no longer valid. We can update Eq. 1 to

$$\boldsymbol{O} + \mathcal{T}^{-1}\left[\boldsymbol{R}_{OCR} \cdot \mathcal{T}(\underline{\boldsymbol{\kappa}})\right] = \boldsymbol{V}, \tag{4}$$

where $\boldsymbol{R}_{OCR} \in \mathbb{R}^{3\times 3}$ is a roll rotation matrix built given the OCR response; $\underline{\boldsymbol{\kappa}}$ represents the Kappa Angle with invariant pitch and yaw components; $\mathcal{T}$ is a function to transform pitch and yaw to a 3D directional unit vector and $\mathcal{T}^{-1}$ represents the inverse process. As reported in the statistics [25, 28], the roll (β) from OCR is around $1/7.5$ of the roll motion (α) of the head.

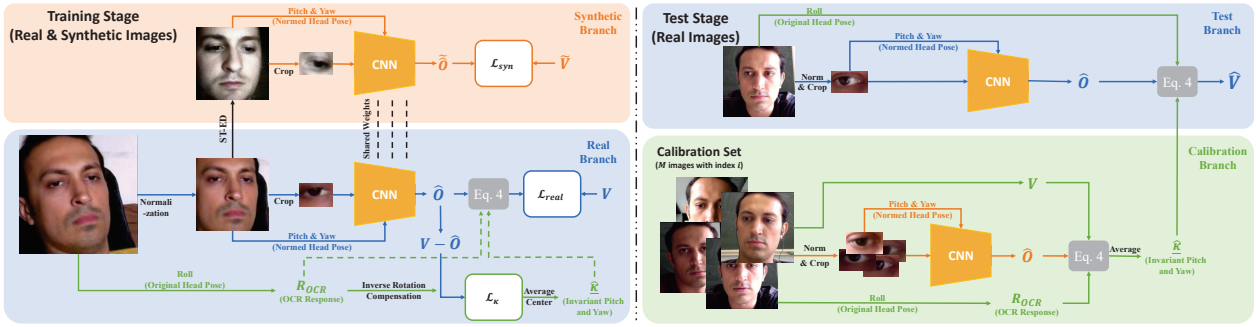


Figure 3. Training (left) and test (right) pipelines of KAComp-Net. Training stage has two branches: synthetic branch aids the CNN in learning the optic axis directions by generated data from ST-ED with manually assigned labels, while real branch focuses on estimating the OCR-compensated Kappa Angle $\underline{\widehat{\kappa}}$ with invariant pitch and yaw components. Test stage consists of calibration and test branches. Calibration branch estimates the Kappa Angle of the test subject using M labeled data. Then, test branch estimates the final gaze directions using the output from CNN and the estimated Kappa Angle. $\boldsymbol{V}$ ($\boldsymbol{O}$) denotes the visual (optic) axis directions. $\widehat{(\cdot)}$ denotes the estimated variables and $\widetilde{(\cdot)}$ denotes the synthetic variables used by ST-ED. $\mathcal{L}$ denotes the loss functions which are elaborated in Section 3.4.

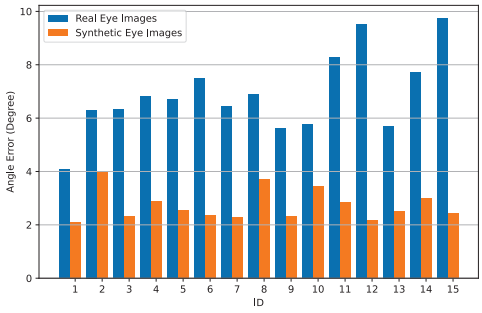


Figure 4. Comparison of angular errors given real and synthetic data from MPIIFaceGaze with the leave-one-subject-out protocol. The network has a single branch with three convolutional layers.

3.2. Real and Synthetic Data

The optic axis is the line connecting the nodal point with the pupil center, which is related to the observed iris [15] and can be estimated from images [4]. However, the optic axis is not provided in gaze datasets. The visual axis defines gaze, which is what we want to estimate. However, due to the subject-dependent unobservable deviations between visual and optic axes [1, 15], a unified gaze estimator does not work well on new subject data. Considering this, we proposed to utilize synthetic redirected eye images with manually set gaze directions to learn the approximation of the optic axis given the simulation results. We quantitatively evaluated the subject-dependent variations across different subjects' eye images between real and synthetic data. Synthetic data was generated based on real data with assigned conditions by using ST-ED [46]. We trained a three-layer CNN with either real or synthetic eye images and tested its performance with the 'leave-one-subject-out' protocol. The mean angular errors are $6.89°$ and $2.72°$ on the real and synthetic data from MPIIFaceGaze [43], shown in Fig. 4. The mean angular errors of EYEDIAP [10] are $7.31°$ (real data) and $2.30°$ (synthetic data). We noted that there existed a **large gap** in angle errors between real and synthetic data. In other words, the subject-dependent unobservable deviations from the Kappa Angle across different subjects were largely removed after gaze redirection. Based on this, we utilized synthetic eye images to learn the person-independent part of gaze, viewed as the approximation of the optic axis.

3.3. Pipeline

Training Stage. To begin with, given a real face image, we apply the preprocessing step utilized in ST-ED to obtain a normalized face image with the corresponding normalized ground truth gaze $\boldsymbol{V}$, the normalized head pose $\boldsymbol{H}$ and the roll motion $\boldsymbol{H}^{roll}$ of the head before normalization. Next, we input the normalized face image into ST-ED to generate redirected face images using the provided condition as pseudo labels: gaze $\widetilde{\boldsymbol{V}}$ and head pose $\widetilde{\boldsymbol{H}}$. We then crop the same-side eye images from the normalized real and synthetic face images and feed them into a single-stream convolutional neural network (CNN) that takes one eye image at a time as input. The normalized head pose is attached to the intermediate features of the eye images, similar to GazeNet [44]. The output of the CNN is the estimated direction (pitch and yaw) of the subject-independent component of the gaze, which approximates the optic axis.

We use the roll motion $\boldsymbol{H}^{roll}$ to calculate the roll status of the eyeball and build the rotation matrix $\boldsymbol{R}_{OCR}$, as described in Section 3.1. We can then estimate the Kappa Angle for each instance using Eq. 4, based on the estimated optic axis, the ground-truth gaze and the roll status of the eyeball. To ensure that the pitch and yaw of the estimated Kappa Angle are identical for the same subject, we

employ the Kappa Angle loss, which is built on the Center Loss [35]. This loss aims at reducing standard deviations of the estimated instance-wise Kappa Angle within each subject's data and iteratively updates the average center as the subject-wise Kappa Angle. Finally, the estimated subject-wise Kappa Angle is used to update the unobservable subject-dependent part and combined with the estimated optic axis for final gaze estimation.

Evaluation Stage. During evaluation, we randomly pick a certain number (M) of calibrated samples with ground-truth labels from the same test subject. We input these samples into the trained CNN to estimate their optic axis directions. Using the provided gazes and the calculated OCR response, we derive several instance-wise Kappa Angles $\widehat{\underline{\boldsymbol{\kappa}}}_i$ from calibrated samples based on Eq. 4. We then calculate their average as the estimated subject-wise Kappa Angle $\widehat{\underline{\boldsymbol{\kappa}}}$. Finally, we use the estimated optic axis directions of the target samples and $\widehat{\underline{\boldsymbol{\kappa}}}$ to determine the visual axis direction as the final gaze. Since we apply the Kappa Angle to the compensation of the subjects' variation, we call it '*Kappa Angle Compensation Neural Network*' and *KAComp-Net* in short.

3.4. Loss Functions

We trained our proposed KAComp-Net using multi-objective loss functions defined as

$$\mathcal{L} = \lambda_{syn} \cdot \mathcal{L}_{syn} + \lambda_{real}\mathcal{L}_{real} + \lambda_{\kappa} \cdot \mathcal{L}_{\kappa}, \tag{5}$$

where we empirically set $\lambda_{syn}, \lambda_{real} = 1.0$ and $\lambda_{\kappa} = 0.5$. In order to balance real and synthetic eye image groups, we use the same number of real and synthetic eye images during the training. The details of each loss component are elaborated on in the following paragraphs.

Kappa Angle loss. There is no ground truth to supervise the learning of the Kappa Angle and the only clue to restrict it is that the pitch and yaw of $\underline{\boldsymbol{\kappa}}$ keep identical across samples within the same subject's data. Thus we propose the Kappa Angle loss, which aims at making the standard deviation (SD) of the calculated Kappa Angle with considering OCR within every subject's data as small as possible. Inspired by the Center Loss [35] designed for classification, which narrowed the intra-class distances from data points to the class center, we applied it to the Kappa Angle loss.

There are K subjects' data included in the training set. Each subject has N_k real eye images $\boldsymbol{I}^e$ and N_k synthetic eye images $\widetilde{\boldsymbol{I}}^e$. The Kappa Angle loss is defined as

$$\mathcal{L}_{\kappa} = \frac{1}{2N_k}\sum_{i=1}^{N_k}\left(||\widehat{\underline{\boldsymbol{\kappa}}}_i - \boldsymbol{c}_k||_2^2 + ||\widetilde{\widehat{\underline{\boldsymbol{\kappa}}}}_i - \boldsymbol{c}_{K+1}||_2^2\right), \tag{6}$$

where $k = 1, \cdots, K$. The former part, $||\widehat{\underline{\boldsymbol{\kappa}}}_i - \boldsymbol{c}_k||_2^2$, in Eq. (6) is designed for **real** eye images, where $\boldsymbol{c}_k$ represents the center point (mean values) of the calculated $\widehat{\underline{\boldsymbol{\kappa}}}_i$ over all samples from the subject with identity number k and $\widehat{\underline{\boldsymbol{\kappa}}}_i$ is calculated given Eq. 4 as

$$\widehat{\underline{\boldsymbol{\kappa}}}_i = \mathcal{T}^{-1}\left\{\boldsymbol{R}_{OCR,i}^{-1} \cdot \mathcal{T}\left[\boldsymbol{g}^{gt}(\boldsymbol{I}_i^e) - \psi(\boldsymbol{I}_i^e)\right]\right\}, \tag{7}$$

where $\boldsymbol{R}_{OCR,i}^{-1}$ is the inverse rotation matrix given the OCR response with regard to the i-th real eye image; $\psi(\cdot)$ denotes the output from *KAComp-Net*, which is the estimated direction of the optic axis; and $\boldsymbol{g}^{gt}(\cdot)$ denotes the ground truth gaze direction given the image. The latter part, $||\widetilde{\widehat{\underline{\boldsymbol{\kappa}}}}_i - \boldsymbol{c}_{K+1}||_2^2$, in Eq. (6) is designed for **synthetic** eye images. Since the Kappa Angles are no longer varied across different subjects' synthetic data, we assign only one center point $\boldsymbol{c}_{K+1}$ to all synthetic data. The subscript $K+1$ means a new center point different from the previous K center points of real data. $\widetilde{\widehat{\underline{\boldsymbol{\kappa}}}}_i$ is defined as

$$\widetilde{\widehat{\underline{\boldsymbol{\kappa}}}}_i = \boldsymbol{g}^{gt}(\widetilde{\boldsymbol{I}}_i^e) - \psi(\widetilde{\boldsymbol{I}}_i^e), \tag{8}$$

where we don't consider OCR in synthetic data.

Gaze loss for synthetic images. This loss aims at supervising the network learning the manually designed gaze from synthetic eye images, which have smaller and less varied Kappa Angles across different subjects' data. In other words, this loss guides the network to learn synthetic cases with nearly overlapped optic axis and visual axis.

$$\mathcal{L}_{syn} = \frac{1}{N_k}\sum_{i=1}^{N_k}\left|\left|\boldsymbol{g}^{gt}(\widetilde{\boldsymbol{I}}_i^e) - \psi(\widetilde{\boldsymbol{I}}_i^e)\right|\right|_1. \tag{9}$$

Gaze loss for real images. The aim of importing this gaze loss is to balance real and synthetic data influences. Since we have center points for every subject, which represent the estimated Kappa Angle, we can remove this unobservable subject-dependent part from ground truth gaze to acquire the optic axis directions for real eye images as the ground truth. To be specific,

$$\begin{aligned}\mathcal{L}_{real} &= \frac{1}{N_k}\sum_{i=1}^{N_k}\left|\left|\psi(\boldsymbol{I}_i^e) - \hat{\boldsymbol{O}}(\boldsymbol{I}_i^e)\right|\right|_1, \\ \hat{\boldsymbol{O}}(\boldsymbol{I}_i^e) &= \boldsymbol{g}^{gt}(\boldsymbol{I}_i^e) - \mathcal{T}^{-1}\left[\boldsymbol{R}_{OCR,i} \cdot \mathcal{T}(\boldsymbol{c}_k)\right].\end{aligned} \tag{10}$$

4. Experiments

In this section, we thoroughly evaluated the performance of the proposed algorithm with other state-of-the-art methods on published datasets. We also elaborated several impacts on the proposed algorithm, such as the numbers of references in calibration, the proportion of synthetic images and the estimated Kappa Angles distribution.

4.1. Datasets

MPIIGaze [44] is a widely used benchmark dataset for the appearance-based in-the-wild gaze estimation task. In

CAL	Methods	Backbone	Num of Channels	Input Image(s)	Estimated Gaze Type	Mean Angle Error (Degree)	
						MPIIFaceGaze	EYEDIAP
No	GazeNet [44]	VGG-16	1	Single-Eye	Eye	5.83	6.83
	ARE-Net [7]	6 CL	4	Both-Eyes	Face	5.02	6.08
	CA(Eye)-Net [6]	10 CL	2	Both-Eyes	Face	5.01	5.30
	AGE-Net [3]	9 CL + 4 Dilated CL	2	Both-Eyes	Face	4.64	-
Yes	Diff-NN [21]	3 CL	2	Same-Eyes	Eye	4.72±0.40	4.51±0.52
	KAComp-Net[1]	3 CL	1	Single-Eye	Eye	**4.21±0.28**	**3.89±0.25**
	RedFTAdap [40]	VGG-16	1	Single-Eye	Eye	4.01	-
	Faze [27]	DenseNet	1	Both-Eyes	Face	3.90	-
	DAGEN[2] [16]	ResNet-18	1	Both-Eyes	Face	3.74	4.30
	Diff-NN-VGG [21]	VGG-16	2	Same-Eyes	Eye	3.80±0.61	3.53±0.52
	KAComp-Net-VGG[1]	VGG-16	1	Single-Eye	Eye	**3.65±0.25**	**3.44±0.29**

Table 1. Quantitative comparison (MPIIFaceGaze and EYEDIAP) with state-of-the-art eye-image-based gaze estimation methods, which are classified given the need for the calibration (**CAL**) step. The network size can be described by the backbone category of a single channel and the number of channels. The types of input image(s) are: one single (**Single-**) eye, both-side (**Both-**) eyes from one face image and same-side (**Same-**) eyes from different face images of one subject. The estimated gaze type '**Eye**' represents that the origin of gaze directions is the center of the eye. The estimated gaze type '**Face**' represents that the origin of gaze directions is the center between two eyes' centers. [1]Results were acquired with additional synthetic data for training. Details are discussed in Section 4.4. [2]Calibration needed methods' results are based on nine calibrated samples except DAGEN since it achieves best performance based on four references samples.

our experiments, due to the need to generate synthetic face images, we utilized its subset MPIIFaceGaze [43], which contains 37667 full-face images captured from 15 participants' images (nine males and six females). EYEDIAP [10] contains 94 full-face videos from 16 subjects with labeled outliers (blinking or distraction) of each frame. We utilized the data from discrete and continuous screen targets with both static (SP) and dynamic (DP) head poses, covering 14 participants (11 males and 3 females).

Since raw images in both datasets contain the upper torso and the provided data collection information indicates a horizontal camera position, we estimate the roll of the head pose in raw images as the actual roll of the head to eliminate any ambiguity arising from the camera pose.

4.2. Evaluation Protocol

We cross-validated the methods' performance within the published datasets. In detail, we utilized the 'leave-one-subject-out' protocol when we evaluated the models within MPIIFaceGaze or EYEDIAP. Each time we select one subject's data as the test set, and the rest was viewed as the training set. Note that only real data was utilized as the test set, and the synthetic data generated from the test subject's data was not included in the training set in case of data leakage. At test time, we needed to choose several eye images for calibration. In order to alleviate the bias from some calibrated samples, we repeated testing the same trained model 200 times with random combinations of samples and calculated the mean angular errors of the predicted gazes and the standard deviations as the corresponding trained models' performance on the test subject's data. We looped all subjects' data as the test set one by one and reported the average of mean angular errors and the standard deviations. Since the proposed KAComp-Net aims at predicting the single-eye gaze, we trained and evaluated the models on left and right eye images separately.

4.3. Comparison with Eye-image-based Methods

We listed several state-of-the-art eye-image-based gaze estimation methods in Table 1, which were categorized into two groups according to the needs for calibration. There were two kinds of outputs: *eye gazes* and *face gazes*. The *eye gaze* represents the direction from the eye center to the gazing target, usually used for single-eye gaze estimation. The *face gaze* represents the direction from the center of two eye centers to the gazing target, which requires more inputs (e.g. left and right eye images).

Effectiveness of Calibration. It was straightforward to notice that with the assistance of a few ($M = 9$) calibration samples from the test set, the methods achieved significant improvements even if the networks were much simpler. However, better performance and lower calculation complexity were at the cost of the need for several calibrated samples and labeled gazes, which required extra ef-

forts (e.g., calibration before use) under practical scenarios.

Methods with Calibration. We first compared two shallow networks' (Diff-Net [21], KAComp-Net) performance, which were designed for real-time purposes. Then we replaced the proposed KAComp-Net backbone from three-layer CNN to the VGG-16 backbone for fair comparisons with other calibration-needed state-of-the-art methods.

Compared with Diff-NN, the proposed KAComp-Net worked better with only around half of the multiply-accumulate (MAC) operations. The KAComp-Net performed a $0.51°(10.81\%)$ boost on the MPIIFaceGaze and a $0.62°(13.75\%)$ boost on the EYEDIAP. Apart from decreasing the mean angular error, KAComp-Net maintained a more stable performance given different calibration sets. The standard deviation of KAComp-Net was reduced by $0.12°(30.00\%)$ and $0.27°(51.92\%)$ compared to those of Diff-NN in MPIIFaceGaze and EYEDIAP, respectively. These results fully demonstrated that the estimated Kappa Angle with considering OCR is a more unified and robust characteristic than the gaze difference from the same eye when the network depth (learning ability) is limited. The differential-based method assumed that the pitch and yaw of the Kappa Angle were invariant concerning different head poses if the input eye images were normalized, which violated OCR and introduced the undesired person-dependent bias when calculating gaze differences. KAComp-Net considered OCR and used it to derive the invariant pitch and yaw of the Kappa Angle for the optic axis direction estimation, which essentially removed the bias caused by OCR and guided the network to learn a more unified feature from the eye images. The lower standard deviation meant a lower dependence on the calibrated samples, whose impact was further discussed in Section 4.5. In order to make the proposed KAComp-Net competitive compared to other state-of-the-art methods, we replaced the three-layer CNN with the pre-trained VGG-16 backbone for better feature extraction ability. The training parameters remained identical after we changed the backbone. KAComp-Net-VGG achieved $8.98\%, 6.41\%$ and 3.95% improvements on MPIIFaceGaze compared with RedFTAdap [40], FAZE [27] and Diff-NN-VGG [21], respectively.

4.4. Impacts of Synthetic Images

We discussed the effects from synthetic images based on experiments of Diff-NN and KAComp-Net in this section. Synthetic data were generated from real training data only.

Evaluation with Diff-NN. We utilized Diff-NN to investigate synthetic data impacts on the differential-based network. Since Diff-NN needed pairs of eye images from the same subject for training, we implemented three experiments according to the source of paired images: 1) real samples only; 2) separated real or synthetic samples within pairs; 3) Mixture of real and synthetic samples within pairs.

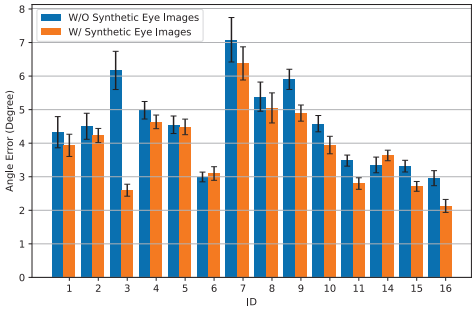


Figure 5. Comparison of mean angular errors and standard deviations (200 repeated experiments) of the gaze by KAComp-Net with and without synthetic data according to the leave-one-subject-out protocol in EYEDIAP.

The inference process was taken only on real data with 200 repeated evaluations and the number of calibration samples $M = 9$. The performance on real data only (RO), real and synthetic data independently (RS_I) and the mixture of real and synthetic data (RS_M) were $4.51 \pm 0.52°$, $4.27 \pm 0.50°$ and $5.99 \pm 0.68°$ on the EYEDIAP, respectively. RO and RS_M performance were similar, which meant that the synthetic data maintained the same gaze difference property as the real data. The mixture of them achieved worse performance than the other two, which further demonstrated that the Kappa Angle variation of the synthetic data was no longer kept as the real data did.

Evaluation with KAComp-Net. We did the experiments on KAComp-Net with or without synthetic data. Fig. 5 elaborates on the impacts of synthetic samples on KAComp-Net. The mean angle error was $4.54 \pm 0.32°$ without synthetic data and $3.89 \pm 0.25°$ with synthetic data in EYEDIAP. Synthetic data played an important role during the training of the KAComp-Net because it helped supervise the network learning a unified characteristic and further improved the accuracy for the Kappa Angle regression.

4.5. Impacts of Calibrated Samples

Fig. 6 illustrates the impact of the number of calibrated samples in EYEDIAP. The evaluation protocol is illustrated in Section 4.2. When the number of calibrated samples was less than three, Diff-NN had similar performance compared with no-calibration-needed methods. Especially when the number of calibrated samples $M = 1$, Diff-NN achieved $1.03°$ worse than GazeNet [44], mainly due to large gaze differences between limited calibrated samples and target ones, and the number of network layers. As the number of calibrated samples increased, the prediction errors and the standard deviations of Diff-NN dropped significantly because more calibrated samples with similar gaze directions to target ones were acquired. Given the same number of

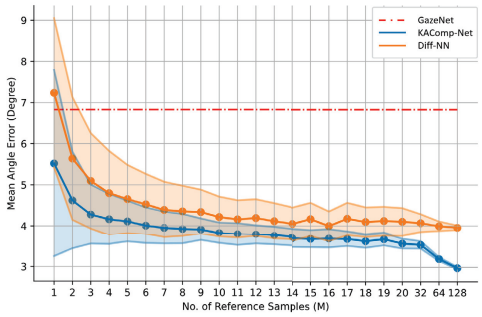
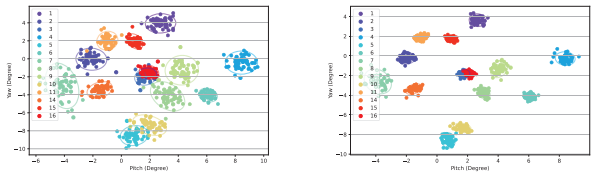


Figure 6. Comparison of mean angular errors and standard deviations (200 repeated experiments) among the state-of-the-art methods given different numbers of calibrated samples in EYEDIAP.



(a) $M = 9$ (b) $M = 64$

Figure 7. Distribution maps of the estimated Kappa Angles given different subjects with KAComp-Net given M calibrated samples. The legend number means subject ID in EYEDIAP.

calibrated samples, KAComp-Net achieved more accurate and more stable results than Diff-NN, proving the higher tolerance to the calibrated samples. Even with only one calibrated sample, KAComp-Net can achieve $1.32°(19.33\%)$ improvement compared with GazeNet. When the number of calibrated samples was larger than 64, KAComp-Net can further improve the estimation accuracy, unlike the plateauing performance of Diff-NN, shown in Fig. 6. The main reason for this phenomenon was given the estimation accuracy of the Kappa Angle from calibrated samples, which was discussed in detail in Section 4.6.

4.6. Estimated Kappa Angle Distribution

During the inference, we first calculated the Kappa Angles from the calibrated samples of the test subject. The estimated Kappa Angle distribution maps with different numbers (9 and 64) of calibrated samples were shown in Fig. 7 based on 50 repeated experiments. Note that with more calibrated samples for calibration, the estimated Kappa Angles had smaller standard deviations, which yielded smaller angular errors, shown in Fig. 5. An obvious comparison was found by the distribution maps between the subjects with ID 7 and 16. The estimated Kappa Angle range of the ID 7 subject was over $6° \times 2°$, and the corresponding predicted angle error was $6.38°$, which was 64% higher than the mean angular error over all subjects. However, the distribution map of the ID 16 subject had less than a $2° \times 2°$ area, which achieved $2.13°$ angular error (45% lower than the mean angular error).

Methods	Mean Angle Error (Degree)	
	SP	DP
Diff-NN	3.46±0.40	4.76±0.41
KAComp-Net	**3.16±0.26**	**4.37±0.26**

Table 2. Estimated mean angle errors given static (SP) and dynamic (DP) head pose data in EYEDIAP.

	Diff-NN	KAComp-Net
Params (M)	42.015	**5.044**
MACs (M)	89.148	**28.581**

Table 3. Complexity Comparison between the Differential Method and Kappa Angle Compensation Method

4.7. Impacts of Head Pose Variations

KAComp-Net is designed to remove the variance caused by OCR, but it doesn't depend on various head poses (or rolls) to trigger OCR for estimating the Kappa Angle. This is because OCR only affects whether it is needed to compensate for the redistribution of the pitch and yaw of the Kappa Angle before regressing this anatomical variable within each subject's data. Table 2 shows consistent improvements compared with Diff-NN under different levels of head pose variations in EYEDIAP, which also proves the importance of considering OCR.

4.8. Algorithm Complexity

Diff-Net and KAComp-Net share the same three-convolutional-layer backbone, which aims at achieving real-time gaze estimation. Table 3 compares the size of the network and the number of multiply–accumulate (MAC) operations. We observe that KAComp-Net reduced 87.99% (67.94%) parameters (MACs) compared with Diff-NN.

5. Conclusion

In this work, we derived and proposed a pipeline to regress the pitch and yaw of the Kappa Angle under the head coordinate system given the ocular counter-rolling response. This person-dependent Kappa Angle regression works with an eye-image-based person-independent gaze estimator trained with real and synthetic eye images for person-dependent calibration with a few samples. Several experiments on the benchmark datasets showed the effectiveness and robustness of the proposed methods with limited calibration samples.

References

[1] David A Atchison, George Smith, and George Smith. *Optics of the human eye*, volume 35. Butterworth-Heinemann Oxford, 2000. 3, 4

[2] Yiwei Bao, Yunfei Liu, Haofei Wang, and Feng Lu. Generalizing gaze estimation with rotation consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4207–4216, 2022. 2

[3] Pradipta Biswas et al. Appearance-based gaze estimation using attention and difference mechanism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3143–3152, 2021. 6

[4] Zhaokang Chen and Bertram Shi. Offset calibration for appearance-based gaze estimation via gaze decomposition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 270–279, 2020. 2, 4

[5] Zhaokang Chen and Bertram E Shi. Appearance-based gaze estimation using dilated-convolutions. In *Asian Conference on Computer Vision*, pages 309–324. Springer, 2018. 2

[6] Yihua Cheng, Shiyao Huang, Fei Wang, Chen Qian, and Feng Lu. A coarse-to-fine adaptive network for appearance-based gaze estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10623–10630, 2020. 2, 6

[7] Yihua Cheng, Feng Lu, and Xucong Zhang. Appearance-based gaze estimation via evaluation-guided asymmetric regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 100–115, 2018. 2, 6

[8] Shirley G Diamond and Charles H Markham. Ocular counterrolling as an indicator of vestibular otolith function. *Neurology*, 33(11):1460–1460, 1983. 2, 3

[9] Lex Fridman, Bryan Reimer, Bruce Mehler, and William T Freeman. Cognitive load estimation in the wild. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–9, 2018. 1

[10] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc 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, 2014. 2, 4, 6

[11] Kenneth Alberto Funes Mora and Jean-Marc Odobez. 3d gaze tracking and automatic gaze coding from rgb-d cameras. In *IEEE Conference in Computer Vision and Pattern Recognition, Vision Meets Cognition Workshop*, number CONF, 2014. 2

[12] Kenneth A Funes-Mora and Jean-Marc Odobez. Gaze estimation in the 3d space using rgb-d sensors. *International Journal of Computer Vision*, 118(2):194–216, 2016. 1, 2

[13] Yaroslav Ganin, Daniil Kononenko, Diana Sungatullina, and Victor Lempitsky. Deepwarp: Photorealistic image resynthesis for gaze manipulation. In *European conference on computer vision*, pages 311–326. Springer, 2016. 3

[14] Alessandro Grillini, Daniel Ombelet, Rijul S Soans, and Frans W Cornelissen. Towards using the spatio-temporal properties of eye movements to classify visual field defects. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pages 1–5, 2018. 1

[15] Elias Daniel Guestrin and Moshe Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on biomedical engineering*, 53(6):1124–1133, 2006. 1, 4

[16] Zidong Guo, Zejian Yuan, Chong Zhang, Wanchao Chi, Yonggen Ling, and Shenghao Zhang. Domain adaptation gaze estimation by embedding with prediction consistency. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 6

[17] Zhe He, Adrian Spurr, Xucong Zhang, and Otmar Hilliges. Photo-realistic monocular gaze redirection using generative adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6932–6941, 2019. 3

[18] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6912–6921, 2019. 2

[19] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2176–2184, 2016. 2

[20] Erik Lindén, Jonas Sjostrand, and Alexandre Proutiere. Learning to personalize in appearance-based gaze tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1, 2

[21] Gang Liu, Yu Yu, Kenneth Alberto Funes Mora, and Jean-Marc Odobez. A differential approach for gaze estimation with calibration. In *BMVC*, volume 2, page 6, 2018. 1, 2, 3, 6, 7

[22] Yunfei Liu, Ruicong Liu, Haofei Wang, and Feng Lu. Generalizing gaze estimation with outlier-guided collaborative adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3835–3844, 2021. 2

[23] Raphael Menges, Chandan Kumar, Daniel Müller, and Korok Sengupta. Gazetheweb: A gaze-controlled web browser. In *Proceedings of the 14th International Web for All Conference*, pages 1–2, 2017. 1

[24] Kwang-Keun Oh, Byeong-Yeon Moon, Hyun Gug Cho, Sang-Yeob Kim, and Dong-Sik Yu. Measurement of ocular counter-roll using iris images during binocular fixation and head tilt. *Journal of International Medical Research*, 49(3):0300060521997329, 2021. 2

[25] Jorge Otero-Millan, Carolina Treviño, Ariel Winnick, David S Zee, John P Carey, and Amir Kheradmand. The video ocular counter-roll (vocr): a clinical test to detect loss of otolith-ocular function. *Acta oto-laryngologica*, 137(6):593–597, 2017. 2, 3

[26] Benjamin I Outram, Yun Suen Pai, Tanner Person, Kouta Minamizawa, and Kai Kunze. Anyorbit: Orbital navigation in virtual environments with eye-tracking. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pages 1–5, 2018. 1

[27] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-shot adaptive

gaze estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9368–9377, 2019. 1, 2, 3, 6, 7

[28] Millard F Reschke, Scott J Wood, and Gilles Clément. Ocular counter rolling in astronauts after short-and long-duration spaceflight. *Scientific reports*, 8(1):1–9, 2018. 2, 3

[29] Julian Schwehr and Volker Willert. Driver's gaze prediction in dynamic automotive scenes. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–8, 2017. 1

[30] Maximilian AR Strobl, Florian Lipsmeier, Liliana R Demenescu, Christian Gossens, Michael Lindemann, and Maarten 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):1–12, 2019. 2

[31] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi 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, 2014. 1

[32] Ashish Tawari, Kuo Hao Chen, and Mohan M Trivedi. Where is the driver looking: Analysis of head, eye and iris for robust gaze zone estimation. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 988–994. IEEE, 2014. 1

[33] Kang Wang and Qiang Ji. Real time eye gaze tracking with kinect. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2752–2757. IEEE, 2016. 2

[34] Yaoming Wang, Yangzhou Jiang, Jin Li, Bingbing Ni, Wenrui Dai, Chenglin Li, Hongkai Xiong, and Teng Li. Contrastive regression for domain adaptation on gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19376–19385, 2022. 2

[35] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016. 5

[36] Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. A 3d morphable eye region model for gaze estimation. In *European Conference on Computer Vision*, pages 297–313. Springer, 2016. 2

[37] Kenneth W Wright. Anatomy and physiology of eye movements. In *Pediatric Ophthalmology and Strabismus*, pages 125–143. Springer, 2003. 1, 3

[38] Yunyang Xiong, Hyunwoo J Kim, and Vikas Singh. Mixed effects neural networks (menets) with applications to gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7743–7752, 2019. 2

[39] Yu Yu, Gang Liu, and Jean-Marc Odobez. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11937–11946, 2019. 1, 3

[40] Yu Yu and Jean-Marc Odobez. Unsupervised representation learning for gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7314–7324, 2020. 2, 3, 6, 7

[41] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Revisiting data normalization for appearance-based gaze estimation. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pages 1–9, 2018. 1

[42] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4511–4520, 2015. 2

[43] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 51–60, 2017. 2, 4, 6

[44] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):162–175, 2017. 1, 2, 4, 5, 6, 7

[45] Yi Zhang, Lu Zhang, Wassim Hamidouche, and Olivier Deforges. A fixation-based 360° benchmark dataset for salient object detection. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 3458–3462. IEEE, 2020. 1

[46] Yufeng Zheng, Seonwook Park, Xucong Zhang, Shalini De Mello, and Otmar Hilliges. Self-learning transformations for improving gaze and head redirection. *Advances in Neural Information Processing Systems*, 33:13127–13138, 2020. 3, 4