Gaze Estimation by Exploring Two-Eye Asymmetry

Yihua Cheng, Xucong Zhang, Feng Lu, *Member, IEEE*, and Yoichi Sato, *Senior Member, IEEE*

***Abstract*— Eye gaze estimation is increasingly demanded by recent intelligent systems to facilitate a range of interactive applications. Unfortunately, learning the highly complicated regression from a single eye image to the gaze direction is not trivial. Thus, the problem is yet to be solved efficiently. Inspired by the two-eye asymmetry as two eyes of the same person may appear uneven, we propose the face-based asymmetric regression-evaluation network (FARE-Net) to optimize the gaze estimation results by considering the difference between left and right eyes. The proposed method includes one face-based asymmetric regression network (FAR-Net) and one evaluation network (E-Net). The FAR-Net predicts 3D gaze directions for both eyes and is trained with the asymmetric mechanism, which asymmetrically weights and sums the loss generated by two-eye gaze directions. With the asymmetric mechanism, the FAR-Net utilizes the eyes that can achieve high performance to optimize network. The E-Net learns the reliabilities of two eyes to balance the learning of the asymmetric mechanism and symmetric mechanism. Our FARE-Net achieves leading performances on MPIIGaze, EyeDiap and RT-Gene datasets. Additionally, we investigate the effectiveness of FARE-Net by analyzing the distribution of errors and ablation study.**

***Index Terms*— Gaze estimation, asymmetric regression, evaluation network, eye appearance.**

I. Introduction

HUMAN eye gaze provides important cues to infer visual attention and cognition. It is wildly demanded by various applications, *e.g.*, people annotates frames with eye gaze in saliency detection [1], [2]. Therefore, the ability to automatically and accurately track human eye gaze is important for many intelligent systems, with direct applications including human-computer interaction [3], [4], saliency detection [5]–[11], virtual reality [12], [13], video surveillance [14], to first-person video analysis [15].

As surveyed in [16], gaze estimation methods can be divided into two categories: model-based and appearance-based. Model-based methods are usually designed to extract small eye features, e.g., infrared reflection points on the corneal surface and pupil center, to compute gaze directions. This group of methods can usually accurate estimate gaze directions. However, they share the common limitations such as 1) requirement on dedicated hardware for active illumination and capturing, 2) high failure rate in the outdoor environment, and 3) limited working distance (typically within 90 cm).

In contrast, appearance-based methods directly learn a mapping function from eye appearances to gaze directions. They can work with a single ordinary RGB camera and estimate gaze directions without explicit eye features detection. Unfortunately, it still remains challenging in handling various eye appearances caused by head pose, illumination conditions, and individual personal differences. These various eye appearances dramatically increase the data space and make the mapping function difficult to learn for conventional methods [17], [18].

Recent works leverage large training data to cover the different eye appearances with the convolutional neural network (CNN) and achieve state-of-the-art performances [19], [20]. Despite the development, most of appearance-based methods still take a single eye image as input [19], [20]. These methods usually flip the eye images of one eye horizontally and then use one mapping function to handle both left and right eyes. Intuitively, it is reasonable to take two eye images as inputs for appearance-based methods. Some two eye images-based methods have been proposed recently [21], [22] and show reasonable accuracy. However, the left and right eye images are seemed as independent features in these methods. In practice, the appearances of two eyes are different due to illuminations and head pose, while the gaze directions of two eyes are approximately the same. This asymmetric problem makes the mapping function complicated to learn.

The goal of this work is to explore how to choose the superior eye, the eye that is more suitable for gaze estimation than another one, to improve the gaze estimation performance in general. Thus, we propose the evaluation-guide asymmetric regression mechanism in this paper. As shown in Fig. 1, the core of the proposed mechanism is the notion of asymmetric mechanism. It is based on our key observation that the estimated gaze directions of left and right eyes can be different, which we define as *two-eye asymmetry*. It suggests that we can train an efficient gaze estimation model if we can find the superior eye.

In order to do so, we consider the following two technical issues: 1) how to design a network that can process both eyes simultaneously as well as asymmetrically, and 2) how to control the asymmetry to optimize the network by finding

Manuscript received September 22, 2019; revised February 23, 2020; accepted March 16, 2020. Date of current version March 27, 2020. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 61972012. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jianbing Shen. *(Corresponding author: Feng Lu.)*

Yihua Cheng is with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: yihua_c@buaa.edu.cn).

Xucong Zhang is with the Department of Computer Science, ETH Zurich, 8006 Zurich, Switzerland (e-mail: xucong.zhang@inf.ethz.ch).

Feng Lu is with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China, also with the Peng Cheng Laboratory, Shenzhen 518000, China, and also with the Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beihang University, Beijing 100191, China (e-mail: lufeng@buaa.edu.cn).

Yoichi Sato is with the Institute of Industrial Science, The University of Tokyo, Tokyo 1538505, Japan (e-mail: ysato@iis.u-tokyo.ac.jp).

Digital Object Identifier 10.1109/TIP.2020.2982828

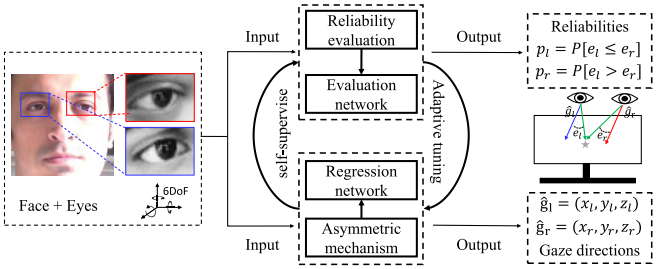


Fig. 1. Overview of the proposed evaluation-guided asymmetric regression mechanism. The processed image is sent into the evaluation network and the regression network simultaneously. The evaluation network performs reliability evaluation and outputs reliabilities of two eye images. The regression network performs asymmetric mechanism to output the gaze directions of two eyes. Moreover, the evaluation network adaptively tunes the asymmetric mechanism of gaze regression, and the regression network provides the ground truth for the self-supervised evaluation network.

the superior eye. Our solution is to ***guide the asymmetric gaze regression by evaluating the performance of the regression mechanism w.r.t. different eyes***, *i.e.* evaluation-guide asymmetric regression mechanism. In particular, we propose a novel model architecture by analyzing the two-eye asymmetry. We propose the asymmetric regression network, which can rely on the superior eye to predict 3D gaze directions for two eyes. We construct the evaluation network, which can adaptively evaluate and adjust the regression strategy of the asymmetric regression network. By integrating the asymmetric regression network and the evaluation network, the proposed asymmetric regression-evaluation network learns to optimize the overall gaze estimation performance. Recent studies [21], [23] demonstrate the effectiveness of facial information for gaze estimation. We further explore the asymmetric regression-evaluation network with facial information and propose the final face-based asymmetric regression-evaluation network (FARE-Net), which contains one face-based asymmetric regression network (FAR-Net) and one evaluation network (E-Net).

Our method makes the following assumptions. First, as also commonly assumed by other methods in this field along this direction [19], [24], the head pose of the user can be obtained by using existing head trackers depending on detected 2D facial landmarks [25]. Second, the user fixates on the same target with both eyes.

The contributions of this work are summarized as follows:

- We propose the FAR-Net to estimate the gaze directions of two eyes. We also propose the E-Net to evaluate and help adjust the regression.
- We observe the two-eye asymmetry, based on which we propose the mechanism of evaluation-guided asymmetric regression. This leads to asymmetric gaze estimation for two eyes.
- Based on the proposed mechanism, we design the FARE-Net which shows promising performance for the gaze estimation task.

An earlier version of this work was published in [26]. We only demonstrate the effectiveness of using two-eye asymmetry in eye image-based gaze estimation methods in the earlier version. Recently, most of proposed gaze estimation methods estimate gaze directions from face images and show reasonable performance. Therefore, it is necessary to explore the two-eye asymmetry with facial information. Compared with the earlier version, we further propose the FARE-Net by integrating face-patch information and provide more comprehensive experiments, *e.g.*, comparison with more state-of-the-arts, analysis on gaze error distribution and selection accuracy of E-Net.

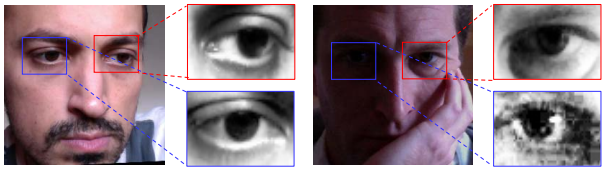


Fig. 2. We select some cases to intuitively illustrate the two-eye asymmetry. It infers that we cannot expect the two eye appearances are the same at any moment despite the user fixates on the same target with both eyes.

II. Related Work

The remote gaze estimation methods can roughly be divided into two major categories: model-based and appearance-based [16], [27].

The model-based methods use geometric eye models [28] to estimate gaze directions. They typically fit the 3D eye models to the detected eye features such as near infrared corneal reflections [28]–[30], pupil center [31], and iris contours [32]–[34]. Although the model-based methods can achieve reasonable accuracy, most of them require dedicated devices such as infrared lights, stereo/high-definition cameras, and RBG-D cameras [32], [33]. Moreover, model-based methods usually only work at a short distance between the user and the camera. These features testify that model-based methods are more suitable to be employed under the controlled environment, *e.g.*, laboratory settings, rather than outdoor settings or with long user-camera distances, *e.g.*, for advertisement analysis [35].

The appearance-based methods attract much attention recently because of the less restrictive requirement compared with model-based methods. They can use a single webcam to capture eye images and learn a mapping function from eye images to the corresponding gaze direction [36]. Neural networks [37], [38], local linear interpolation [36], adaptive linear regression [39], dimension reduction [40] and Gaussian process regression [41] have been proposed to learn the mapping function. Given that the mapping is highly non-linear due to various eye appearances, it still remains challenging to learn a generic mapping to cover all cases.

Recently, CNNs achieve great success in the computer vision task [42], [43]. The CNN-based methods have also shown their ability to handle complex gaze estimation tasks with large training data, and thus they have outperformed traditional appearance-based methods [19], [44], [45]. Zhang *et al*. first proposed a CNN-based method to map eye images to gaze directions [19]. Zhang *et al*. took into consideration the full face as input to the CNNs [23]. Deng *et al*. proposed a CNN-based method with geometry constraints [24]. Krafka *et al*. implemented the CNN-based gaze tracker in the mobile devices [21]. Ranjan *et al*. built a branched CNN

architecture model to handle variable head pose [20]. However, these methods always process left and right eyes analogously, while in this paper we try to make further improvement by introducing and utilizing the two-eye asymmetry.

Intuitively, taking face images as input can provide more information for gaze estimation tasks compared with eye images. For example, the information about head pose is implicitly encoded in face images. Previous works extracted such head pose information from face images and took the head pose vector as the additional input for the models [19], [46]. Recently, face image-based methods are increasingly proposed. They directly take face images as input and achieve better performance than eye image-based methods [21], [23]. Cheng *et al.* integrated face and eye images with a coarse-to-fine framework. [47]. Meanwhile, face image-based methods also have various applications. Fan *et al*. extracted features from face images to understand human gaze communication [48]. Chong *et al*. connected gaze with attention estimation, the gaze direction estimated from face images provides evidence for predicting the visual attention of people in images [49]. In this paper, we also explore the effectiveness of the two-eye asymmetry with face images.

III. Two-Eye Asymmetry in Gaze Regression

Before getting into the technical details, we first review the problem of 3D gaze direction estimation, and introduce the two-eye asymmetry that inspires our method.

A. 3D Gaze Estimation via Regression

The eye gaze direction can be denoted as a 3D unit vector $\mathbf{g}$, which represents the eyeball orientation in the camera coordinate system. There is a strong correlation between eyeball orientation and eye appearance, *e.g*., the location of the iris contour and shape of the eyelids. Therefore, the problem of estimating the 3D gaze direction $\mathbf{g} \in \mathbb{R}^3$ from a given eye image $\boldsymbol{I} \in \mathbb{R}^{H \times W}$ with height H and width W can be formulated as a regression problem $\hat{\mathbf{g}} = f(\boldsymbol{I})$, where f is the regression function and $\hat{\mathbf{g}}$ is the estimated value of $\mathbf{g}$.

Besides eyeball orientation, head pose is another major factor in affecting $\boldsymbol{I}$. In order to handle head motion, it is necessary to consider the head pose $\mathbf{h} \in \mathbb{R}^3$ as input for the regression, which results in

$$\hat{\mathbf{g}} = f(\boldsymbol{I}, \mathbf{h}). \tag{1}$$

In the literature, various regression models have been proposed, such as the neural network [37], the Gaussian process regression [50] and the adaptive linear regression [39]. However, the problem is still challenging due to various eye appearances. With the fast development of deep neural networks over the past years, solving such highly complex regression problem becomes possible with newly proposed large datasets. Nevertheless, an efficient network architecture is still crucial for the gaze estimation task.

B. Two-Eye Asymmetry

Most of gaze regression methods generally handle the two eyes analogously, *e.g*. flip right eye images and process it as left eye images [19], [20]. However, two eye appearances are in fact asymmetry in unconstrained gaze estimation. For example, as shown in Fig. 2, due to large head pose variations, the eye far from the camera may be largely deformed; one eye of the two may suffer from poor illumination and thus being degraded. As the result, we observe the two-eye asymmetry regarding the regression accuracy.

Observation: At any moment, we cannot expect the same gaze estimation accuracy for two eyes out of the regression model, and either eye has a chance to be more accurate than the other one.

Intuitively, such two-eye asymmetry problem may have effects on the gaze estimation accuracy when we take the different eyes as input. Therefore, it will bring performance improvement if we can identify and rely on the superior eye of the two eyes for gaze estimation tasks.

IV. Asymmetric Regression-Evaluation Network

Inspired by the two-eye asymmetry, in this section, we deliver the FARE-Net for appearance-based gaze estimation. The FARE-Net is designed to work in an asymmetric way that can rely on the superior eye for gaze estimation. We first provide an overview of the proposed method.

A. Overview

The goal of the proposed FARE-Net is to predict the 3D gaze directions of two eyes. The overall architecture is shown in Fig. 3. The FARE-Net is consisted of two subnetworks: FAR-Net and E-Net. The FAR-Net performs gaze regression of two eyes simultaneously. The outputs of FAR-Net are estimated gaze directions for left eye $\widehat{\mathbf{g}}_l$ and right eye $\widehat{\mathbf{g}}_r$. More importantly, a new asymmetric loss is designed so that the FAR-Net is able to optimize the network in an asymmetric way. The E-Net is an auxiliary network to help optimize the FAR-Net. It outputs two-dimensional probability vector $[p_l, p_r]$, which evaluates the reliability of each eye for gaze estimation task. The E-Net is trained with the output of FAR-Net, where the better performance indicates the larger reliability. Meanwhile, given the reliability of each eye, we can guide the process of optimization for FAR-Net. The asymmetric loss of FAR-Net is guided by the weight ω, which measures the consistency of the outputs between E-Net and FAR-Net. A larger ω indicates a larger consistency and represents a larger learning rate for asymmetric loss. Generally speaking, FAR-Net supervises the training of E-Net and the E-Net provides feedback for guiding the optimization of FAR-Net.

The rests of this section is constructed as follows. The mechanisms in FAR-Net and E-Net are introduced first. Next, we describe the strategy of integrating the FAR-Net and the E-Net. Then, we combine our proposed method with facial information and define the final framework of the proposed method. The whole structure of the proposed method is summarized in the end of this section.

B. Asymmetric Regression

The FAR-Net is a regression network, which estimates the gaze directions of two eyes. The input of FAR-Net contains

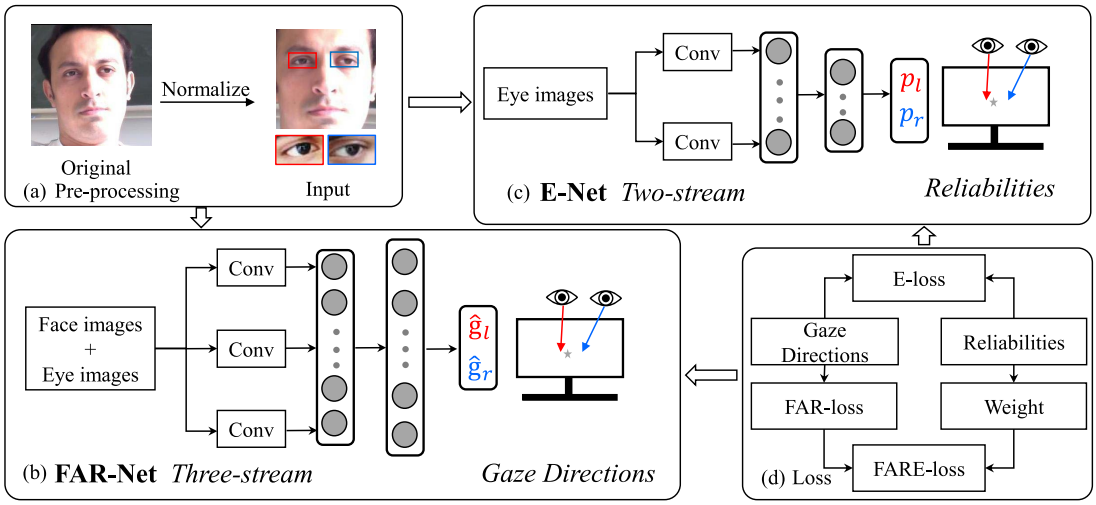


Fig. 3. Overview of the proposed FARE-Net. It consists of two major subnetworks, namely, the FAR-Net and the E-Net. (a) The inputs of FARE-Net are paired eye and face images, where the eye images are croped from the face image. (b) The FAR-Net performs asymmetric regression and predicts the gaze directions of two eyes individually. (c) The E-Net predicts the reliabilities of two eye images in the form of probability. (d) The relations between different loss function. The loss function of E-Net is constructed with the gaze directions that predicted by FAR-Net. The reliabilities estimated by E-Net adjusts the FAR-loss to construct the final loss function of FAR-Net (FARE-loss).

two eye images and head pose, since it predicts the gaze directions of two eyes simultaneously.

In order to handle the asymmetric problem, FAR-Net processes asymmetric regression to estimate the gaze directions of two eyes. The key idea of asymmetric regression is the asymmetric mechanism that we need to rely on the superior eye for gaze estimation tasks. It means the FAR-Net should rely on the eye which can achieve better performance than the other one to optimize the network.

In particular, we first measure the angular errors of the currently predicted 3D gaze directions for left and right eyes by

$$e_l = \arccos\left(\frac{\mathbf{g}_l \cdot \widehat{\mathbf{g}}_l}{\|\mathbf{g}_l\|\|\widehat{\mathbf{g}}_l\|}\right), \tag{2}$$

and

$$e_r = \arccos\left(\frac{\mathbf{g}_r \cdot \widehat{\mathbf{g}}_r}{\|\mathbf{g}_r\|\|\widehat{\mathbf{g}}_r\|}\right), \tag{3}$$

The angular errors e_l and e_r indicate the performance of left and right eyes. The $\mathbf{g}_l$ and $\mathbf{g}_r$ are the ground truth of left and right eyes, and $\widehat{\mathbf{g}}_l$ and $\widehat{\mathbf{g}}_r$ are the predicted gaze directions of left and right eyes.

We then compute the weighted average of the two-eye error

$$e = \lambda_l \cdot e_l + \lambda_r \cdot e_r \tag{4}$$

to represent the loss of gaze prediction accuracy for both eyes. The weights λ_l and λ_r determine which eye should be considered as the superior eye in terms of final gaze estimation accuracy, and the loss becomes asymmetry when $\lambda_l \neq \lambda_r$.

Following the key idea of asymmetric regression, we enlarge the weight of the eye in optimizing the network if that eye is more likely to achieve a better performance compared with another one. Hence, the weights are set as:

$$\begin{cases} \lambda_l / \lambda_r = \dfrac{1/e_l}{1/e_r}, \\ \lambda_l + \lambda_r = 1, \end{cases} \tag{5}$$

whose solution is

$$\lambda_l = \frac{1/e_l}{1/e_l + 1/e_r}, \quad \lambda_r = \frac{1/e_r}{1/e_l + 1/e_r}. \tag{6}$$

By substituting the λ_l and λ_r in Eq. (4), the final asymmetric loss becomes

$$\mathcal{L}_{FAR}(e_l, e_r) = 2 \cdot \frac{e_l \cdot e_r}{e_l + e_r}, \tag{7}$$

which encourages the FAR-Net to rely on the superior eye during training.

C. Evaluation Network

The E-Net is designed to predict the reliability of two eyes. It adjusts the asymmetric mechanism of FAR-Net during optimization. The E-Net can be formulated as:

$$[p_l, p_r] = f_E(\boldsymbol{I}_l, \boldsymbol{I}_r; \boldsymbol{\theta}_E), \tag{8}$$

where the θ_E represents the parameters of E-Net.

1) Ground Truth: As for E-Net, the reliabilities of left eye p_l and right eye p_r is needed to supervise the training process. However, it is hard to directly acquire the reliabilities. Therefore, we define the reliabilities as the gaze estimation error e_l and e_r computed from FAR-Net and a lower gaze estimation error indicates a higher reliability. With the definition, the ground truth of E-Net can be derived from the performance which is out of FAR-Net. In particular, during training, the ground truth of the E-Net is given as $p_l = 1$ if the $e_l < e_r$ from the FAR-Net, otherwise $p_l = 0$. Essentially, the E-Net is trained to predict the probability of the efficiency for gaze estimation task from the left and right eye images.

2) Loss Function: In order to train the E-Net to predict the choice of FAR-Net, we set its loss function as below:

$$\begin{aligned} \mathcal{L}_E(e_l, e_r, p_l, p_r, \eta) = -\{&\eta \cdot \|e_l - e_r\|_2 \cdot \log(p_l) \\ &+ (1-\eta) \cdot \|e_l - e_r\|_2 \cdot \log(p_r)\}, \end{aligned} \tag{9}$$

where $\eta = 1$ if $e_l \leq e_r$ and $\eta = 0$ if $e_l > e_r$. $\|e_l - e_r\|_2$ computes the euclidean distance of the angular error for gaze directions of two eyes which are estimated by the FAR-Net.

The loss function of E-Net can be understood intuitively. The E-Net should maximize p_l if the left eye has smaller error out of the FAR-Net, *i.e.*, $e_l < e_r$, and vice versa. The $\|e_l - e_r\|_2$ adds the additional penalty when left and right eyes have large gaze estimation error. In this way, the E-Net is trained to predict the superior eye that can be used to optimize the FAR-Net training.

D. Evaluation-Guided Asymmetric Regression

The ground truth of E-Net is derived from the output of FAR-Net. Meanwhile, an important task of the E-Net is to guide the optimization of FAR-Net by adjusting the asymmetric mechanism.

In particular, the loss function of the FAR-Net in Eq. (7) can be modified as following by integrating the E-Net:

$$\mathcal{L}^*_{FAR}(e_l, e_r, \omega; \beta) = \omega \cdot \mathcal{L}_{FAR}(e_l, e_r) + (1-\omega) \cdot \beta \cdot (\frac{e_l + e_r}{2}), \quad (10)$$

where ω balances the weight between asymmetric learning (the first term) and symmetric learning (the second term). β scales the weight of symmetric learning, and is empirically set to 0.1 in our experiments. In particular, given the output (p_l, p_r) of the E-Net, we compute

$$\omega = \frac{1 + (2\eta - 1) \cdot p_l + (1 - 2\eta) \cdot p_r}{2}. \quad (11)$$

Again, $\eta = 1$ if $e_l \leq e_r$, and $\eta = 0$ if $e_l > e_r$.

Here we omit the derivation of ω. $\omega = 1$ when both the FAR-Net and E-Net have a strong agreement on the superior eye, which means that a highly asymmetric learning strategy can be recommended. $\omega = 0$ when they completely disagree with each other, which means that it is better to merely use a symmetric learning strategy as a compromise. In practice, ω is a decimal number between 0 and 1.

E. Adding Facial Information

Recent works show the ability in tackling appearance-based gaze estimation problem with full-face patch as input [21], [23]. Hence, we further explore the proposed method with facial information.

We use face images as an additional input to the FAR-Net. We remove head pose from the inputs of FAR-Net because the face images contain the information of head pose. The final inputs of FAR-Net are one face image and two eye images. The regression problem of FAR-Net can be formulated as:

$$[\widehat{\mathbf{g}}_l, \widehat{\mathbf{g}}_r] = f_{FAR}(\boldsymbol{I}_l, \boldsymbol{I}_r, \boldsymbol{I}_f; \boldsymbol{\theta}_{FAR}), \quad (12)$$

where the $\boldsymbol{I}_l$ represents the left eye image. The $\boldsymbol{I}_r$ represents the right eye image. The $\boldsymbol{I}_f$ represents the face image. The $\boldsymbol{\theta}_{FAR}$ is the parameter of the FAR-Net.

Meanwhile, there is no change in E-Net since we find adding facial information into E-Net cannot result in obvious performance improvement.

Algorithm 1 Operations During Each Iteration in Training FARE-Net

Input: Eye images $\boldsymbol{I}_l$ and $\boldsymbol{I}_r$, face image $\boldsymbol{I}_f$, gaze directions $\mathbf{g}_l$ and $\mathbf{g}_r$, parameters of FAR-Net and E-Net $\boldsymbol{\theta}_{FAR}$ and $\boldsymbol{\theta}_E$, hyperparameter β

Output: Updated parameters $\boldsymbol{\theta}^*_{FAR}$ and $\boldsymbol{\theta}^*_E$

1: Obtain the predicted gaze direction from FAR-Net, $[\widehat{\mathbf{g}}_l, \widehat{\mathbf{g}}_r] = f_{FAR}(\boldsymbol{I}_l, \boldsymbol{I}_r, \boldsymbol{I}_f; \boldsymbol{\theta}_{FAR})$.
2: Obtain the predicted probability from E-Net, $[p_l, p_r] = f_E(\boldsymbol{I}_l, \boldsymbol{I}_r; \boldsymbol{\theta}_E)$;
3: Calculate e_l by using Eq. (2) and calculate e_r by using Eq. (3). If $e_l \leq e_r$, $\eta = 1$ and if $e_l > e_r$, $\eta = 0$;
4: According to Eq. (9), calculate the loss of E-Net: E-loss $= \mathcal{L}_E(e_l, e_r, p_l, p_r, \eta)$;
5: Calculate the weight ω by using Eq. (11);
6: According to Eq. (10), calculate the loss of FAR-Net: FARE-loss $= \mathcal{L}^*_{FAR}(e_l, e_r, \omega, \beta)$;
7: $\boldsymbol{\theta}^*_{FAR} \leftarrow$ Update the parameter $\boldsymbol{\theta}_{FAR}$ by using E-loss;
8: $\boldsymbol{\theta}^*_E \leftarrow$ Update the parameter $\boldsymbol{\theta}_E$ by using FARE-loss;

F. Guiding Gaze Regression by Evaluation

Following the above description, we summarize again how the FAR-Net and the E-Net are integrated together, and how the E-Net can guide the FAR-Net training.

- **FAR-Net**: takes both eye images and face images as input; loss function is modified by the output of E-Net (p_l, p_r) to adjust the asymmetry adaptively (Eq. (10)).
- **E-Net**: takes both eye images as input; loss function is modified by the errors of FAR-Net (e_l, e_r) to predict the superior eye image for optimization (Eq. (9)).
- **FARE-Net**: the FAR-Net and the E-Net are integrated and trained together. The final gaze estimation results are the output $(\widehat{\mathbf{g}}_l, \widehat{\mathbf{g}}_r)$ from the FAR-Net.

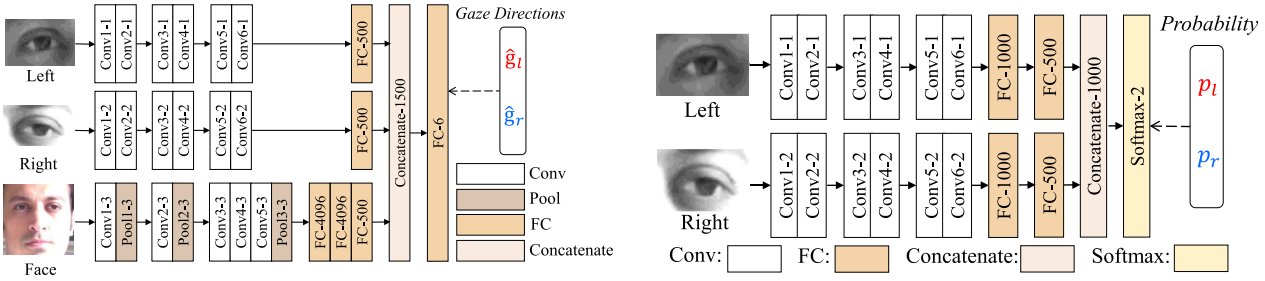
The process of training is also summarized in Algorithm 1.

V. Implementation Details

We describe the implemented details of proposed method in this section. Generally, the architecture of FAR-Net and E-Net can be implemented with any CNN architecture.

A. Data Pre-Processing

The goal of gaze estimation methods is to learn a mapping from the eye image to gaze directions under different head pose. Although head pose has six degrees of freedom, it is possible to cancel the rotation around roll-axis and normalize the distance between cameras and faces. This can reduce the training space and therefore makes the training more efficient. In order to simplify the gaze estimation problem, we use the image normalization method which is proposed in [17]. In a nutshell, the normalization first rotates the camera to cancel the rotation around roll-axis, and then scale the distance between cameras and faces to be the same for all samples. The final eye images can have only two degrees as rotation angles around pitch and yaw axes.



(a) The architecture of FAR-Net (b) The architecture of E-Net

Fig. 4. The architecture of FARE-Net. FARE-Net is consisted of two subnetworks: FAR-Net and E-Net. (a) The FAR-Net estimates 3D gaze directions of two eyes. It is a three-stream network, which produces features from eye images and face images. (b) The E-Net is a two-stream network for evaluating two eyes. The E-Net outputs a 2D probability vector.

B. FAR-Net

As shown in 4(a), The FAR-Net is a three-stream convolutional neural network and the inputs are paired face and eye images. The first two streams are used to extract 500D deep features from each eye independently and the third stream is used to produce 500D feature from face images.

The first and second streams are the same architectures. The first stream takes left eyes as input and the second stream takes right eyes as input. Each stream consists of six convolutional layers and one fully connected layer. We use a gray-scale eye image with the fixed size of 36×60 pixels as the input. All the six convolutional layers use 3×3 filters and the strides are set as (1-2-1-2-1-2). The output channels are 64 for the first and second layers, 128 for the third and fourth layers, and 256 for the fifth and sixth layers. Final, we use a fully connected layer to extract the 500D deep feature of each eye images.

We use one 224*224 pixels RGB face image as the input of the third stream. Meanwhile, in order to extract deep features efficiently, we use AlexNet [51] as the basic component. The AlexNet consists of five convolutional components, three max-pool layers and three fully connected layers. Every convolutional component contains a convolution layer and a local response normalization layer. We replace the local response normalization layer with the batch normalization layer [52] and change the output of the final fully connected layer from 1000D to 500D deep feature.

Finally, we concatenate all the 1500D deep features and utilize the features to regress the gaze directions of two eyes. Each gaze direction is a 3D vector that represents the $[x, y, z]$ of gaze directions. Note that, we do not add the head pose vector because face images already contain the information of head pose.

C. E-Net

The E-Net is a two-stream convolutional neural network as shown in Fig. 4(b). Similarly to FAR-Net, the first stream takes left eyes as input and the second stream takes right eyes as input. Each stream uses one 36×60 gray-scale eye image as input and consists of six convolutional layers and two fully connected layers. The configurations of convolutional layers are set the same as those in the convolutional components in the first stream of FAR-Net. We use an additional 1000D fully connected layer as the hidden layer. Each stream final outputs the 500D feature of each eye. We concatenate the outputs of both streams to be a 1000D feature, and pass the 1000D feature to a softmax layer to generate a 2D vector $[p_l, p_r]$.

D. Training

We implement the proposed method by using TensorFlow [53]. We train the whole network in 200 epochs with a batch size of 100 on the training set. We use the Adam solver [54] with default configuration, which is provided by TensorFlow. The learning rates of FAR-Net and E-Net are respectively set as 0.01 and 0.0005 empirically.

VI. Experimental Evaluation

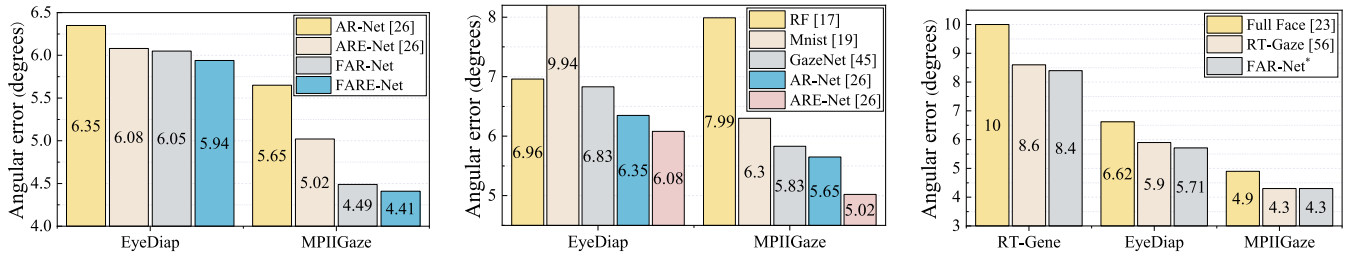
In this section, we first conduct an experiment to prove the advantage of our method in averaged performance. We next analyze the experimental result and show how does our method handle the two-eye asymmetry. Meanwhile, the experiment about the selection accuracy of E-Net is conducted for understanding the role of E-Net. Further, we illustrate the error distribution of gaze regression under different head pose and illumination conditions, the corresponding discussion about the difference in head pose and illumination conditions is given, which can help understand our method. Finally, we perform a case study and show some failure cases of E-Net.

A. Dataset

We conduct experiments on three popular gaze estimation datasets: EyeDiap [55], MPIIGaze [45] and RT-Gene [56].

1) EyeDiap Dataset: EyeDiap dataset [55] contains a set of video clips of 16 participants. The videos are recorded with both free head and fixed head motion under various lighting conditions. Since EyeDiap dataset does not provide a standard subset of evaluation, we sample images from each video to construct the evaluation dataset. The EyeDiap dataset contains two kinds of visual target sessions which are screen target and 3D floating ball. We only use the screen target session for evaluation, since the 3D floating ball sometimes occludes the face image. We sample one image per 15 frames from VGA videos. We obtain the video clip from 14 participants since the other two participants lack of screen target session videos.

2) Modified MPIIGaze Dataset: MPIIGaze dataset [45] is composed of 213, 659 images of 15 participants, which contains a large variety of different illuminations, eye appearances and head pose. It is among the largest datasets for



(a) Ablation study. (b) Comparsion with eye image-based methods. (c) Comparsion with face image-based methods.

Fig. 5. We show the gaze estimation errors of the compared methods. a) We conduct the ablation study about E-Net and facial information. The FARE-Net performs better than the ARE-Net. b) We conduct the comparison with eye image-based methods. The ARE-Net shows the best performance. c) We compare the FAR-Net* with face image-based methods. The FAR-Net* shows the best performance on all datasets.

appearance-based gaze estimation and thus was commonly used in previous works. Note that MPIIGaze dataset provides the image that have already been normalized by the normalization method [17].

The MPIIGaze dataset provides a standard subset for evaluation, which contains 1500 left eye images and 1500 right eye images independently selected from each participant. However, our method requires paired eye images within one face image. Therefore, we modify the evaluation set by finding out the missing eye image of every left-right eye pair from the original dataset. This process doubles the image number in the evaluation set. In our experiments, we use such a modified dataset instead of the original MPIIGaze dataset. Besides, MPIIGaze dataset only provides eye images, while we need full face images as input for our FARE-Net method. Therefore, we obtain the corresponding face image from MPIIFaceGaze [23] which is an expanded dataset of MPIIGaze.

3) RT-Gene Dataset: RT-Gene dataset contains 122,531 images of 15 participants. It provides accurate gaze annotations since they annotate gaze with dedicated eyetracking glasses. In order to remove the eyetracking glass from captured images, they build a generative adversarial network to paint each image. Therefore, RT-Gene dataset provides another inpainted set, which contains 122, 531 painted images.

We do not add the inpainted set into the training set since there are many noises in the inpainted set. We divide the original dataset into 3 subsets for 3-fold cross validation according to the evaluation protocol provided by the dataset. Note that, the image provided by RT-Gene is already processed, in order to provide a fair comparison, we directly use the provided image for evaluation without another data pre-processing.

B. Compared Methods

We mainly conduct experiments between following methods. Results of the following methods are obtained from our implementation or published papers. Note that, AR-Net and ARE-Net are proposed in an earlier version of this work [26]. Compared with FARE-Net, ARE-Net does not use facial information as input while includes the evaluation-guided asymmetric regression mechanism. Thus, We provide experiments for ARE-Net to demonstrate the effectiveness of asymmetric regression and E-Net.

- **RF [17]:** One of the most commonly used regression method. It is shown to be effective for a variety of applications. Similar to [17], multiple RF regressors are trained for each head pose cluster.
- **Mnist [19]:** One of the typical appearance-based gaze estimation methods based on deep neural networks. The input is the image of a single eye. We use the original Caffe codes provided by the authors of [19] to obtain all the result in our experiments.
- **GazeNet [45]:** GazeNet uses a single eye as input and performs better than Mnist. We implement the network as describing in the paper.
- **RT-Gaze [56]:** RT-Gaze uses a two-stream CNN to process two eye images and predicts a single gaze. Up to now, it performs the best performance on the RT-Gene.
- **Full Face [23]:** A deep neural network-based method that takes the full face image as input with a spatial weighting strategy. We also use the original Caffe codes provided by the author to generate result.
- **AR-Net [26]:** An earlier version of FAR-Net, which uses two eye images as input. The AR-Net also performs asymmetric regression to predict the gaze directions of two eyes.
- **ARE-Net [26]:** An earlier version of FARE-Net. It consists of AR-Net and E-Net. Compared with this work, the E-Net in ARE-Net only changes the strides of convolution layers from 2 to 1 and adds max-pool layers.
- **FAR-Net (Ours):** FAR-Net preforms asymmetric regression to estimate gaze directions of two eyes. The loss function of FAR-Net is set as Eq. (7).
- **FARE-Net (Ours):** We final propose FARE-Net in this paper. The FARE-Net uses two eye images and face images as input to estimate the gaze directions of two eyes with evaluation-guide asymmetric regression mechanism.

C. Comparison With Appearance-Based Methods

All the methods are evaluated on the MPIIGaze and EyeDiap datasets. In order to reduce the error caused by overfitting, a leave-one-person-out strategy is applied in the evaluation of the both two datasets. The results are shown in Fig. 5. The accuracy is measured by the averaged angular error of all the test samples including both left and right images.

1) Ablation Study: We first conduct ablation study to demonstrate the effectiveness of the E-Net and facial information. Note that the face and eye images provided by MPIIGaze dataset are normalized with different parameters. In order to preserve the consistency between face images and eye images, we crop two-eye images from normalized face images.

The result is shown in Fig. 5(a). The accuracy of AR-Net is 5.65° on MPIIGaze and 6.35° on EyeDiap. After adding E-Net, ARE-Net significantly outperforms AR-Net with 11% improvement on MPIIGaze and 4% improvement on EyeDiap. This demonstrates that our E-Net is indeed helpful for optimization. After adding facial information, FAR-Net achieves an advanced performance as 4.49° on MPIIGaze, which is more than 20% improvement compared with AR-Net. FAR-Net also shows 6.05° accuracy on EyeDiap, which is 5% improvement compared with AR-Net. By integrating E-Net and facial information, FARE-Net achieves the best performances among all the methods as 4.41° on MPIIGaze and 5.94° on EyeDiap.

2) Within Eye Image-Based Methods: The previous experiment demonstrates the effectiveness of the E-Net and facial information. We next consider the scenario where only eye images are used as the input and present experiments within eye image-based methods. The results of all methods are obtained by running the corresponding codes on modified MPIIGaze and EyeDiap.

As shown in Fig. 5(b)., ARE-Net significantly outperforms other methods. GazeNet shows the best performance among compared methods due to the deep network. However, with less computational resource, AR-Net shows 3% improvement on MPIIGaze and 7% improvement on EyeDiap compared with GazeNet. In addition, by introducing the E-Net, the final ARE-Net achieves the best performance with large margins (11%° improvement on MPIIGaze and 4%° improvement on EyeDiap compared with AR-Net). The accuracy of ARE-Net is 5.02° on MPIIGaze and 6.08° on EyeDiap. This demonstrates that our evaluation-guide asymmetric regression mechanism is effective.

3) Within Face Image-Based Methods: The previous experiment shows the comparison among eye image-based methods. Recently, some face image-based methods are proposed and show better performance compared with eye image-based methods [21], [23]. In order to provide a convincing comparison, we further conduct the comparison with face image-based methods. However, most face image-based methods only output a single eye direction [23] and re-define the origin of gaze vector as face center rather than eye center. To make a fair comparison, we have made the following changes in FAR-Net: 1) FAR-Net only outputs a single gaze direction and the origin of gaze vectors is also defined as the face center; 2) we remove the E-Net and use the angular error as the loss function. We use FAR-Net* to represent the modified FAR-Net.

We select Full Face and RT-Gaze as the compared methods. Nota that, although RT-Gaze uses eye images as input, we carefully categorize the method into face image-based methods since it outputs a single eye direction as most face image-based methods and has the best performance on RT-Gene. Meanwhile, since the EyeDiap dataset does not provide the 3D location of face center, the midpoint of the two eye centers is set as the origin of gaze vectors. The experiment on MPIIGaze is also in fact conducted on the the MPIIFaceGaze [23] rather than the modified MPIIGaze dataset for a fair comparison. We also conduct more experiments on RT-Gene, note that we do not conduct the previous two experiments on RT-Gene since RT-Gene only provides a single gaze direction for each face image. We directly report the performance of Full Face on MPIIGaze and RT-Gene from corresponding papers [23], [56] due to the same setting of the experiment. The result of RT-Gaze on RT-Gene is also reported from [56]. The other results are all obtained by running corresponding codes.

TABLE I
COMPARISON WITH APPEARANCE-BASED METHODS

Methods	MPIIGaze	EyeDiap	RT-Gene
RF [17]	7.99°	6.96°	-
Mnist [19]	6.30°	9.94°	-
GazeNet [45]	5.83°	6.83°	-
AR-Net [26]	5.65°	6.35°	-
ARE-Net [26]	5.02°	6.08°	-
FAR-Net	4.49°	6.05°	-
FARE-Net	**4.41°**	**5.94°**	-
Itracker [21]	*6.2°	9.93°	-
Full Face [23]	*4.9°	6.62°	*10.0°
Faze [57]	*5.2°	-	-
MeNet [58]	*4.9°	-	-
Dilated-Net [59]	4.8°	5.9°	-
RT-Gaze [56]	*4.3°	5.9°	*8.6°
FAR-Net*	**4.3°**	**5.71°**	**8.4°**

The results are shown in Fig. 5(c). The RT-Gaze has better performance on all datasets as 4.3° on MPIIGaze, 5.9° on EyeDiap, and 8.6° on RT-Gene than Full Face. Our proposed FAR-Net* achieves the same performance as RT-Gaze on MPIIGaze while performs better on the other two datasets than RT-Gaze. This proves the advantage of FAR-Net* in performance.

4) Additional Comparisons: In order to provide more comprehensive comparisons, we carefully summarize most appearance-based methods and list their performance in Table I for reference. Nota that the methods summarized in the second row estimate gazes for each eye and the methods summarized in the third column estimate one gaze for one face image. We use the star to mark the performance reported from paper and the other performances are obtained by running corresponding codes. It is clear that our method performs well on all datasets.

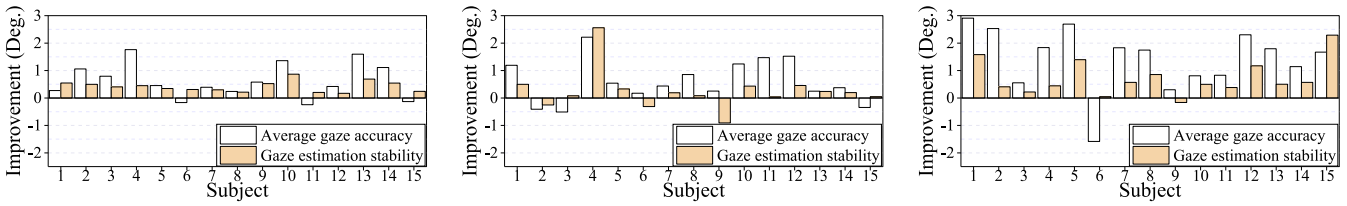
D. Evaluation on Each Individual

The previous experiments show the advantages of our proposed methods in terms of the average performance across subjects, while the robustness of our proposed methods for

TABLE II

COMPARISON WITH THE METHODS WHICH OUTPUT GAZE DIRECTIONS OF TWO EYES REGARDING THEIR ACCURACY ON EACH SUBJECT

	Methods						
	RF	Mnist	GazeNet	AR-Net	ARE-Net	FAR-Net (ours)	FARE-Net (ours)
Subject 1	$7.21 \pm 3.27°$	$4.88 \pm 2.31°$	$5.48 \pm 2.92°$	$4.03 \pm 2.39°$	$3.76 \pm 1.84°$	$3.00 \pm 1.50°$	$\mathbf{2.57 \pm 1.34}°$
Subject 2	$7.77 \pm 3.52°$	$7.08 \pm 3.52°$	$6.29 \pm 2.87°$	$4.41 \pm 2.71°$	$\mathbf{3.35 \pm 2.21}°$	$3.58 \pm 2.52°$	$3.76 \pm 2.47°$
Subject 3	$7.23 \pm 3.25°$	$5.79 \pm 2.53°$	$6.20 \pm 2.13°$	$5.93 \pm 2.40°$	$\mathbf{5.14 \pm 2.00}°$	$5.91 \pm 1.98°$	$5.65 \pm 1.92°$
Subject 4	$8.72 \pm 3.96°$	$6.46 \pm 4.72°$	$4.63 \pm 2.32°$	$6.77 \pm 4.89°$	$5.01 \pm 4.44°$	$\mathbf{2.67 \pm 1.83}°$	$2.79 \pm 1.88°$
Subject 5	$7.47 \pm 3.45°$	$5.90 \pm 2.99°$	$5.39 \pm 2.99°$	$3.69 \pm 2.27°$	$3.24 \pm 1.92°$	$2.96 \pm 1.66°$	$\mathbf{2.70 \pm 1.59}°$
Subject 6	$8.08 \pm 4.34°$	$6.37 \pm 3.38°$	$\mathbf{4.47 \pm 2.69}°$	$6.06 \pm 2.64°$	$6.22 \pm 2.33°$	$5.86 \pm 2.59°$	$6.05 \pm 2.64°$
Subject 7	$8.71 \pm 3.81°$	$5.58 \pm 3.06°$	$5.33 \pm 2.71°$	$4.33 \pm 2.63°$	$3.94 \pm 2.33°$	$3.51 \pm 2.11°$	$\mathbf{3.50 \pm 2.14}°$
Subject 8	$8.64 \pm 4.12°$	$7.55 \pm 4.25°$	$6.49 \pm 4.37°$	$5.84 \pm 3.82°$	$5.60 \pm 3.60°$	$4.80 \pm 3.52°$	$\mathbf{4.75 \pm 3.52}°$
Subject 9	$8.67 \pm 4.09°$	$6.59 \pm 3.94°$	$5.50 \pm 4.04°$	$6.03 \pm 3.82°$	$5.45 \pm 3.29°$	$5.34 \pm 4.18°$	$\mathbf{5.20 \pm 4.20}°$
Subject 10	$8.56 \pm 3.90°$	$7.71 \pm 3.67°$	$5.28 \pm 2.65°$	$7.07 \pm 3.46°$	$5.71 \pm 2.59°$	$4.67 \pm 2.20°$	$\mathbf{4.47 \pm 2.16}°$
Subject 11	$6.70 \pm 3.25°$	$6.00 \pm 3.02°$	$6.09 \pm 2.90°$	$6.49 \pm 2.76°$	$6.73 \pm 2.56°$	$5.82 \pm 2.60°$	$\mathbf{5.26 \pm 2.52}°$
Subject 12	$7.98 \pm 3.54°$	$6.04 \pm 3.00°$	$5.89 \pm 3.01°$	$5.53 \pm 2.46°$	$5.11 \pm 2.29°$	$3.82 \pm 1.91°$	$\mathbf{3.59 \pm 1.84}°$
Subject 13	$7.93 \pm 3.62°$	$6.10 \pm 3.54°$	$5.58 \pm 3.06°$	$5.63 \pm 3.49°$	$4.03 \pm 2.80°$	$3.90 \pm 2.69°$	$\mathbf{3.78 \pm 2.56}°$
Subject 14	$8.01 \pm 3.85°$	$6.93 \pm 3.70°$	$6.45 \pm 3.15°$	$6.80 \pm 3.32°$	$5.68 \pm 2.78°$	$\mathbf{4.99 \pm 2.59}°$	$5.31 \pm 2.59°$
Subject 15	$8.17 \pm 4.22°$	$\mathbf{5.54 \pm 3.09}°$	$8.34 \pm 4.35°$	$6.19 \pm 2.35°$	$6.33 \pm 2.11°$	$6.53 \pm 1.99°$	$6.67 \pm 2.06°$
Average	$7.99°$	$6.30°$	$5.83°$	$5.65°$	$5.02°$	$4.49°$	$\mathbf{4.41}°$



(a) Improvement of ARE-Net based on AR-Net. (b) Improvement of FARE-Net based on ARE-Net. (c) Improvement of FARE-Net based on GazeNet.

Fig. 6. We show the improvement on gaze accuracy and gaze estimation stability. The improvement of average gaze accuracy is equal to the decrease in the average angular error, and the improvement of gaze estimation stability is equivalent to the reduction in the standard deviation of angular errors. Fig (a) show the improvement after adding E-Net. Fig (b) show the improvement after adding facial information. Fig (c) show the comparison between our proposed FARE-Net and GazeNet.

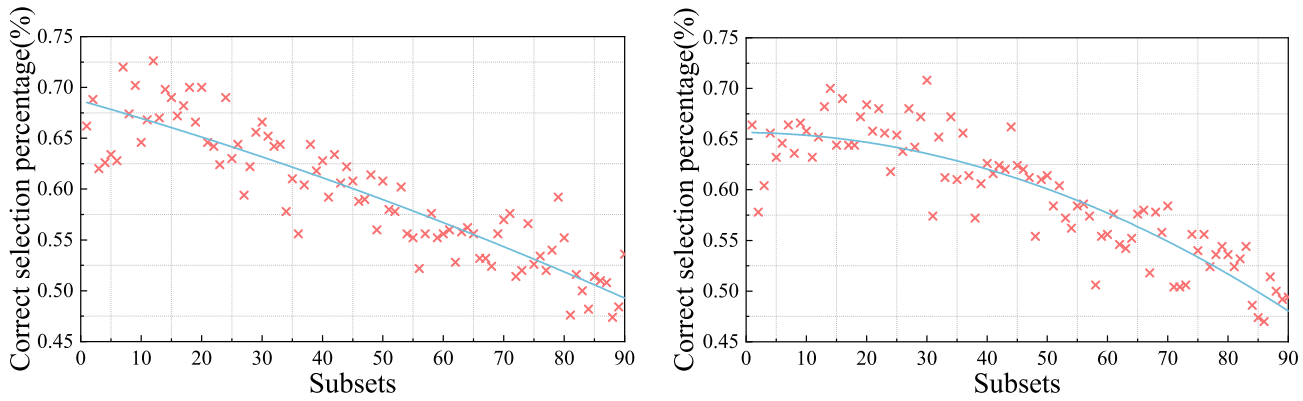
individuals is still unknown. Hence, we further analyze the performance of each subject in this section.

We select the method which outputs the gaze directions of two eyes and conduct experiments with leave-one-person-out strategy on MPIIGaze as described in VI-C. We compute the mean value of angular error for each subject respectively, and summarize the results of all the 15 subjects in Table II. In order to demonstrate the stability of proposed methods, we compute the standard deviation of angular error for each subject and add the result into Table II. Finally, The result in Table II is conducted as (Mean Value±Standard deviation). Note that both mean value and standard deviation are the lower the better.

In the Table II, the bold font indicates the method which shows the best accuracy for corresponding subjects. It is obvious that most of bold fonts appeared in the columns of our proposed methods especially FARE-Net. We first analyze the results of ARE-Net and AR-Net to investigate the advantage of E-Net. With E-Net, the accuracy of ARE-Net is improved on 12 subjects and the stability is improved on all subjects compared with AR-Net. In addition, compared with FAR-Net, FARE-Net also improves the accuracy on 10 subjects and the stability on 8 subjects (the standard deviation of FARE-Net and FAR-Net on subject 8 and subject 14 are equal). Those facts show that the existence of E-Net not only improves the accuracy but also the stability. Next, we analyze the results among our proposed methods and other methods. Compared with GazeNet, ARE-Net improves the accuracy of 11 subjects and stability of 14 subjects. Meanwhile, FARE-Net further improves the accuracy of 14 subjects and the stability of 14 subjects compared with GazeNet. These demonstrate that the proposed FARE-Net is robust.

The performance improvement in mean value and standard deviation are also illustrated in Fig. 6. The two different colors indicate mean value and standard deviation respectively. The average gaze accuracy and gaze estimation stability in legend represent mean value and standard deviation respectively. The horizontal axis represents different subjects, and vertical axis represents the performance improvement compared with baseline. Note that the performance improvement is equal to the decrease in mean value and standard deviation.

Fig. 6(a) shows the improvement with E-Net. We take AR-Net as the baseline and analyze the result of the ARE-Net. As show in Fig. 6(a), the accuracy and the stability of ARE-Net both show clear improvement after adding E-Net. Fig. 6(b) shows the improvement after adding



(a) Relationship between gaze error and selection accuracy

(b) Relationship between gaze error's difference and selection accuracy

Fig. 7. We show the correct selection percentage of the E-Net. We sort the gaze estimation result of FARE-Net according to average performance or performance difference of left and right eyes from large to small, and divide the sorted result into 90 subsets where each subset contains 500 images. The cross represents the average correct selection percentage of subsets and the line is with quadratic polynomial curve fitting.

TABLE III
THE ASYMMETRIC STATE OF DIFFERENT METHODS

	Mnist	GazeNet	AR	ARE	FAR	FARE
Worse eye	7.6°	7.33°	6.0°	5.2°	4.72°	4.57°
Superior eye	5.0°	4.32°	5.3°	4.8°	4.26°	4.24°
Difference (Δ)	2.6°	3.01°	0.7°	0.4°	0.46°	0.33°

facial information. We use ARE-Net as the baseline and compare the result of FARE-Net with ARE-Net. After adding facial information, the accuracy is clearly improved for most subjects, while the improvements of stability are small except subject 4. This shows that facial information can not effectively improve the stability. It is reasonable because the face image is more complex than eye image. We use GazeNet as the baseline in Fig. 6(c) and compare GazeNet with the proposed FARE-Net. The proposed FARE-Net significantly outperforms the GazeNet almost on all subjects.

E. Asymmetric State

The two-eye asymmetry is a key observation in this paper. Thus, we investigate the asymmetric state of different methods in this section. We make further analysis based on the initial results obtained in Sec. VI-C. The result is shown in Table III, where we use the averaged gaze error to present the performance.

In particular, we calculate the angular errors in advance. The *Superior eye* selects the eye image showing less angular error than the other one from paired eye images. Correspondingly, the *Worse eye* selects the eye image showing larger angular error than another one. The difference Δ shows the difference between the results of the *Superior eye* and the *Worse eye*.

According to the comparison shown in Table III, we have the following conclusions:

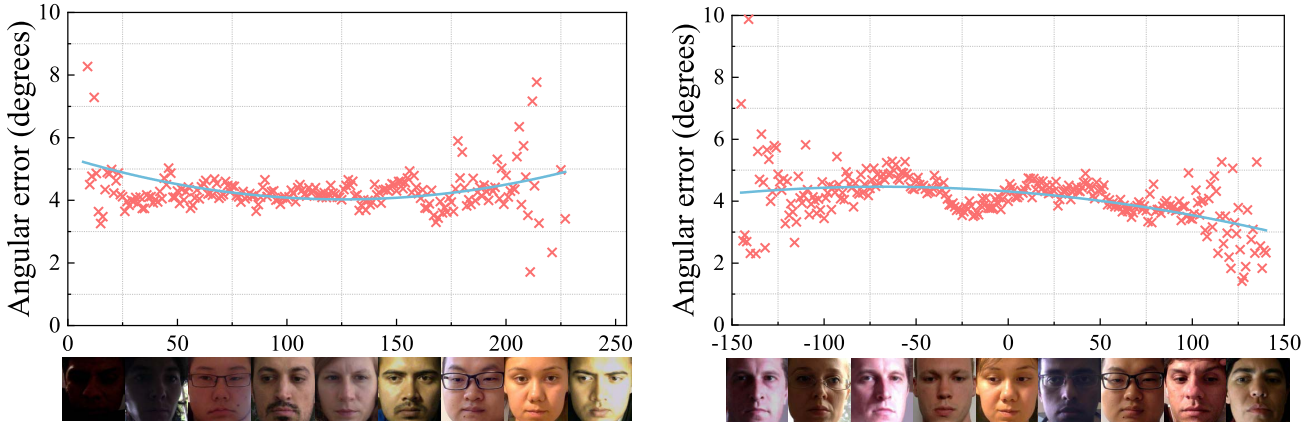
- The asymmetric regression can effective improve the worse eye performance.
- With the E-Net, ARE-Net and FARE-Net not only improve the superior eye performance but also improve the worse eye performance.
- With the asymmetric regression and the E-Net, the two-eye asymmetry can be effective handled. The difference between the superior/worse eyes reduces greatly which is 0.4° in ARE-Net and 0.33 in FARE-Net. Therefore, the major advantage of the E-Net is that it can optimize both left and right eyes simultaneously and effectively.

F. Additional Analysis on E-Net

E-Net is a key component of the proposed method. Thus, it is essential to conduct experiments for understanding the E-Net. In this section, we conduct two experiments including selection accuracy and comparison with other selection mechanisms. The experiments can provide more information for understanding the usage and advantage of the E-Net.

1) Selection Accuracy: We first conduct the experiment about the selection accuracy of E-Net. We compute the selection accuracy based on the result of FARE-Net on MPIIGaze. The correct selection percentage of E-Net is 59.6%, which is not very high. The reason is that, the role of E-Net is to handle the cases where the two eyes cannot be consistent during training, rather than picking out every good estimate during test. In other words, it helps train the FAR-Net with imperfect training data and thus the estimates can be eventually good for both eyes. To validate the statement, we further analyze the result of selection accuracy regarding the gaze prediction error.

We compute the average gaze estimation errors of both eyes and sort the average errors from large to small. We divide the sorted result into 90 subsets, where each subset contains 500 cases. The selection accuracies are computed inside each of the 90 subsets. We show the relationship between selection accuracy and gaze error in Fig. 7(a). It is clearly that the selection accuracy is higher for larger gaze estimation errors and is lower for smaller gaze estimation errors. The selection accuracy is around 70% for the top-500 samples and is around 47% for the last-500 samples. In addition, we alternatively



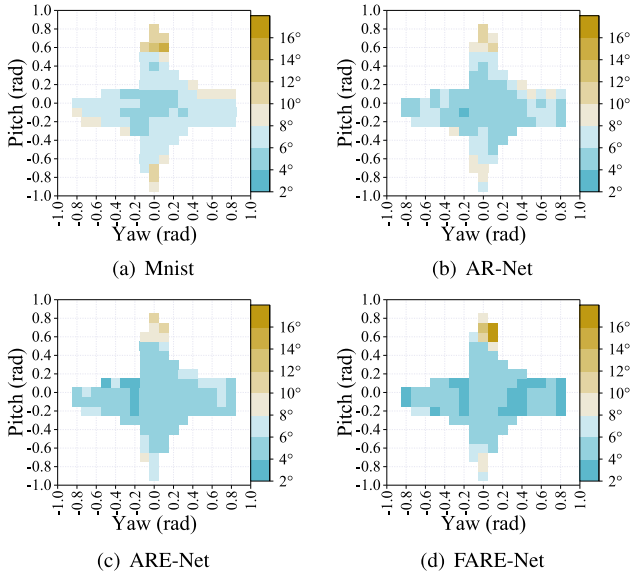
(a) The error distribution across different intensities

(b) The error distribution across different horizontal angles

Fig. 8. We illustrate the gaze error distribution under different illumination conditions. The horozental axis is the mean intensity in Fig (a) and is the mean intensity difference of the left and right face region in Fig (b). The line is with quadratic polynomial curve fitting.

TABLE IV
ANALYSIS ON AVERAGE GAZE ERRORS OF DIFFERENT SELECTION MECHANISMS

Mechanisms	Mnist	GazeNet	AR-Net	FAR-Net
Two eyes	6.3°	5.8°	5.7°	4.5°
Near	6.2°	5.8°	5.6°	4.5°
Frontal	6.4°	5.8°	5.7°	4.5°
E-Net	–	–	5.0°	4.4°
E-Net selection	–	–	4.9°	4.4°



(a) Mnist (b) AR-Net

(c) ARE-Net (d) FARE-Net

Fig. 9. Gaze estimation angular error distributions of four methods across head poses distributions. Our methods decrease the error caused by the variance of yaw effectively.

sort all the results according to the difference between two eye's gaze estimation errors rather than their individual errors. The result is shown in Fig. 7(b), which shows that larger differences between two eyes' errors are related to to higher selection accuracies.

Consequently, the following conclusions can be reached. 1) It is not the high selection accuracy of E-Net that produces low gaze estimation errors. In fact, small gaze errors can be related to low selection accuracy. 2) The selection accuracy is high when the two eyes have very different gaze estimation errors. These conclusions are also consistent with our previous statement, that the E-Net is not expected to reduce the gaze error by selecting very good eye in the output, but aims at handling large difference in two eyes' gaze estimation to help train the FAR-Net.

2) Comparison With Other Selection Mechanisms: It is important to know how is the E-Net different from other selection mechanisms. To this end, we make further analysis based on the initial results obtained in Sec. VI-C. As shown in Table IV, we count the averaged gaze error under different selection mechanisms. All the selection mechanisms select one or two eye images from paired eye images, the introduction of different selection mechanisms is presented as follows.

The *Two eyes* selects both two eye images. The *Near* selects the eye nearer to the camera than another one according to the location in camera coordinate system. The *Frontal* selects the more frontal eye than another one according to head pose. In addition, we show the result with using E-Net during training, and present it in the row of *E-Net*. The result is the average angular error of two eyes. As we can see from Table IV, the AR-Net with *E-Net* mechanism is equal to ARE-Net, and FAR-Net with *E-Net* mechanism is equal to FARE-Net. With using E-Net during training, the prediction of E-Net indicates which eye image may show more accurate result. According to the prediction of E-Net, the *E-Net selection* further selects the eye image showing more accurate estimation result. Note that, the accuracy shown in *E-Net selection* is based on the result in *E-Net*.

It is obvious that *Near* and *Frontal* mechanisms can not effective improve the performance, while the performance can be improved with using E-Net during training. It is because

Fig. 10. Comparison between two eyes' gaze errors. The ARE-Net and FARE-Net show lower gaze errors of both eyes than other methods, and also smaller differences between each pair of two eyes.

common selection mechanisms do not involve the training stage and just simply sample a subset from results. However, the E-Net can adaptively tune the objective function to help optimize the regression network during training. Meanwhile, although the E-Net is not designed to improve the performance with selecting the superior eye, *E-Net selection* can also bring little performance improvement compared with *E-Net*. Those are the difference between E-Net and those selection mechanisms. The proposed E-Net can not be simply replaced with other selection mechanisms.

G. Further Analysis

1) Gaze Error w.r.t. Head Pose: Head pose is an important factor that results in asymmetry in eye appearance. In order to show the effectiveness of proposed methods in detail, we illustrate the error distribution across head pose. We first obtain the predicted gaze direction of MPIIGaze and cluster all the gaze directions with head pose. Next, we select the cluster larger than 10 and count the mean angular error. The result are summarized in the Fig. 9. The horizontal axis and vertical axis respectively represent yaw and pitch of head pose.

As shown in Fig. 9, the Mnist method achieves larger error for the extreme head pose compared with the frontal head pose, which is intuitive. However, as for the other methods, the gaze errors do not increase with extreme head pose in yaw, and even decrease a little than the error with frontal head pose. The result shows the advantage of our methods in handing two-eye asymmetry caused by head pose. Meanwhile, from the distribution of FARE-Net, we can find that the FARE-Net not only improve the performance with extreme head pose in yaw which can cause the two-eye asymmetry but also the overall performance. This indicates that our methods are generalized for different head pose.

2) Gaze Error w.r.t. Illumination: Illumination is another important factor which can cause the two-eye asymmetry. Therefore, we next investigate the effectiveness of the proposed method in different illumination conditions. We respectively conduct experiments to show the performance distribution of FARE-Net on MPIIGaze across different mean intensities and horizontal angles. Note that, we use the mean intensity difference between left and right face regions to represent the horizontal angle since most datasets do not provide the angle of illumination. The result is shown in Fig. 8.

The performance distribution shows difference between the two illumination conditions. The performance of FARE-Net becomes worse with extreme light intensities, *i.e.*, very dark or very bright, which is quite intuitive. However, the average error of FARE-Net does not increase with extreme horizontal angles of illumination, and even decrease a little. It is reasonable since the extreme horizontal angles of illumination causes the two-eye asymmetry, which can be handled by FARE-Net. This interesting fact infers the effectiveness of our method in handling two-eye asymmetry.

H. Case Study

We show some representative cases that explain why our proposed methods are superior than the previous works, as shown in Fig. 10. In these cases, using only a single eye image as the Mnist method and the GazeNet method may perform well for one eye but badly for the other eye, and the bad one will affect the final accuracy greatly. On the other hand, ARE-Net and FARE-Net perform asymmetric optimization and help improve both the better eye and the worse eye via the designed evaluation and feedback strategy. Therefore, the output gaze errors are small for both eyes which result in a much better overall accuracy. This is also demonstrated in Table III.

Some failure cases can also help understand the proposed method and highlight the potential future work. Therefore, we select the most representative cases with largest gaze errors and show them in Fig. 11. The cases show that large errors are associated with the factors like very low light scene, large head rotation and the existence of glasses. These factors deform the eye appearance and make the gaze regression difficult to learn. They are also the common issues for most appearance-based gaze estimation methods. Meanwhile, we observe that in such cases both eyes are with low quality, even though the E-Net can select the superior eye, the gaze accuracy cannot be

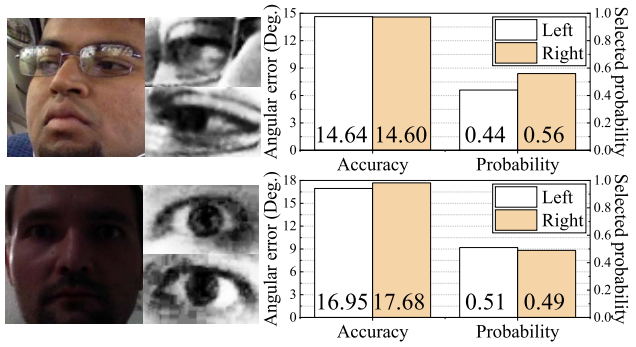


Fig. 11. We select failure cases with largest gaze errors. They are selected form the result of FARE-Net on MPIIGaze. In each case, the top is the left eye image and the bottom is the right eye image. The E-Net selects the eye had larger probability than the other.

improved. These observations suggest that we need to further improve the robustness of gaze estimation with low quality input images in future works.

VII. Conclusion

In this paper, we proposed a novel appearance-based gaze estimation FARE-Net and try to improve the gaze estimation performance to its full extent. The core of our method is the notion of "two-eye asymmetry" which can be observed on the performance of the left and the right eyes during gaze estimation. The FARE-Net contains one face-based asymmetric regression network, which use face images and eye images as input to predict 3D gaze directions for both eyes with an asymmetric strategy, and one evaluation networks to adaptively adjust the strategy by evaluating the two eyes in terms of their reliability during optimization. Our experiments show the proposed FARE-Net achieves the leading performance in three public datasets.

References

[1] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8554–8564.

[2] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, and A. Borji, "Revisiting video saliency prediction in the deep learning era," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jun. 24, 2019, doi: 10.1109/TPAMI.2019.2924417.

[3] X. Zhang, Y. Sugano, and A. Bulling, "Everyday eye contact detection using unsupervised gaze target discovery," in *Proc. 30th Annu. ACM Symp. User Interface Softw. Technol. (UIST)*, 2017, pp. 193–203.

[4] Y. Sugano, X. Zhang, and A. Bulling, "AggreGaze: Collective estimation of audience attention on public displays," in *Proc. ACM Symp. User Interface Softw. Technol. (UIST)*, 2016, pp. 821–831.

[5] X. Sun, H. Yao, R. Ji, and X.-M. Liu, "Toward statistical modeling of saccadic eye-movement and visual saliency," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4649–4662, Nov. 2014.

[6] D. Fan, M. Cheng, J. Liu, S. Gao, Q. Hou, and A. Borji, "Salient objects in clutter: Bringing salient object detection to the foreground," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 186–202.

[7] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 20–33, Jan. 2018.

[8] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 38–49, Jan. 2018.

[9] W. Wang, J. Shen, L. Shao, and F. Porikli, "Correspondence driven saliency transfer," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5025–5034, Nov. 2016.

[10] W. Wang, J. Shen, X. Dong, A. Borji, and R. Yang, "Inferring salient objects from human fixations," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Mar. 18, 2019, doi: 10.1109/TPAMI.2019.2905607.

[11] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2368–2378, May 2018.

[12] Y. Zhang, L. Zhang, W. Hamidouche, and O. Deforges, "A fixation-based 360° benchmark dataset for salient object detection," 2020, *arXiv:2001.07960*. [Online]. Available: http://arxiv.org/abs/2001.07960

[13] Y. Xu *et al.*, "Gaze prediction in dynamic 360° immersive videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 5333–5342.

[14] Q. Cheng, D. Agrafiotis, A. M. Achim, and D. R. Bull, "Gaze location prediction for broadcast football video," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4918–4929, Dec. 2013.

[15] H. Yu, M. Cai, Y. Liu, and F. Lu, "What I see is what you see: Joint attention learning for first and third person video co-analysis," in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, 2019, pp. 1358–1366.

[16] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 478–500, Mar. 2010.

[17] Y. Sugano, Y. Matsushita, and Y. Sato, "Learning-by-synthesis for appearance-based 3D gaze estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1821–1828.

[18] T. Schneider, B. Schauerte, and R. Stiefelhagen, "Manifold alignment for person independent appearance-based gaze estimation," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 1167–1172.

[19] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4511–4520.

[20] R. Ranjan, S. De Mello, and J. Kautz, "Light-weight head pose invariant gaze tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 2156–2164.

[21] K. Krafka *et al.*, "Eye tracking for everyone," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2176–2184.

[22] Q. Huang, A. Veeraraghavan, and A. Sabharwal, "TabletGaze: Dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets," *Mach. Vis. Appl.*, vol. 28, nos. 5–6, pp. 445–461, Aug. 2017.

[23] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "It's written all over your face: Full-face appearance-based gaze estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops)*, 2017, pp. 51–60.

[24] H. Deng and W. Zhu, "Monocular free-head 3D gaze tracking with deep learning and geometry constraints," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3143–3152.

[25] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EPnP: An accurate O(n) solution to the PnP problem," *Int. J. Comput. Vis.*, vol. 81, no. 2, pp. 155–166, Feb. 2008.

[26] Y. Cheng, F. Lu, and X. Zhang, "Appearance-based gaze estimation via evaluation-guided asymmetric regression," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 100–115.

[27] C. H. Morimoto and M. R. M. Mimica, "Eye gaze tracking techniques for interactive applications," *Comput. Vis. Image Understand.*, vol. 98, no. 1, pp. 4–24, Apr. 2005.

[28] E. D. Guestrin and M. Eizenman, "General theory of remote gaze estimation using the pupil center and corneal reflections," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 6, pp. 1124–1133, Jun. 2006.

[29] Z. Zhu and Q. Ji, "Novel eye gaze tracking techniques under natural head movement," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 12, pp. 2246–2260, Dec. 2007.

[30] A. Nakazawa and C. Nitschke, "Point of gaze estimation through corneal surface reflection in an active illumination environment," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 159–172.

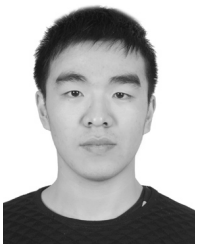
[31] R. Valenti, N. Sebe, and T. Gevers, "Combining head pose and eye location information for gaze estimation," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 802–815, Feb. 2012.

[32] K. A. Funes Mora and J.-M. Odobez, "Geometric generative gaze estimation (G3E) for remote RGB-D cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1773–1780.

[33] X. Xiong, Z. Liu, Q. Cai, and Z. Zhang, "Eye gaze tracking using an RGBD camera: A comparison with a RGB solution," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput., Adjunct Publication*, 2014, pp. 1113–1121.

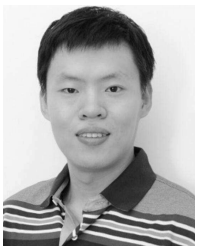
[34] F. Lu, Y. Gao, and X. Chen, "Estimating 3D gaze directions using unlabeled eye images via synthetic iris appearance fitting," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1772–1782, Sep. 2016.

[35] A. T. Duchowski, "A breadth-first survey of eye-tracking applications," *Behav. Res. Methods, Instrum., Comput.*, vol. 34, no. 4, pp. 455–470, Nov. 2002.
[36] K.-H. Tan, D. J. Kriegman, and N. Ahuja, "Appearance-based eye gaze estimation," in *Proc. 6th IEEE Workshop Appl. Comput. Vis. (WACV)*, Dec. 2002, pp. 191–195.
[37] S. Baluja and D. Pomerleau, "Non-intrusive gaze tracking using artificial neural networks," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 1994, pp. 753–760.
[38] L.-Q. Xu, D. Machin, and P. Sheppard, "A novel approach to real-time non-intrusive gaze finding," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 1998, pp. 428–437.
[39] F. Lu, Y. Sugano, T. Okabe, and Y. Sato, "Adaptive linear regression for appearance-based gaze estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 10, pp. 2033–2046, Oct. 2014.
[40] F. Lu, X. Chen, and Y. Sato, "Appearance-based gaze estimation via uncalibrated gaze pattern recovery," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1543–1553, Apr. 2017.
[41] O. Williams, A. Blake, and R. Cipolla, "Sparse and semi-supervised visual mapping with the S^ 3GP," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2006, pp. 230–237.
[42] X. Dong, J. Shen, D. Wu, K. Guo, X. Jin, and F. Porikli, "Quadruplet network with one-shot learning for fast visual object tracking," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3516–3527, Jul. 2019.
[43] W. Wang, J. Shen, and H. Ling, "A deep network solution for attention and aesthetics aware photo cropping," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1531–1544, Jul. 2018.
[44] E. Wood, T. Baltrušaitis, L.-P. Morency, P. Robinson, and A. Bulling, "Learning an appearance-based gaze estimator from one million synthesised images," in *Proc. 9th Biennial ACM Symp. Eye Tracking Res. Appl. (ETRA)*, 2016, pp. 131–138.
[45] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "MPIIGaze: Real-world dataset and deep appearance-based gaze estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 162–175, Jan. 2017.
[46] F. Lu, Y. Sugano, T. Okabe, and Y. Sato, "Gaze estimation from eye appearance: A head pose-free method via eye image synthesis," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3680–3693, Nov. 2015.
[47] Y. Cheng, S. Huang, F. Wang, C. Qian, and F. Lu, "A coarse-to-fine adaptive network for appearance-based gaze estimation," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Feb. 2020, pp. 1–8.
[48] L. Fan, W. Wang, S.-C. Zhu, X. Tang, and S. Huang, "Understanding human gaze communication by spatio-temporal graph reasoning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5724–5733.
[49] E. Chong, N. Ruiz, Y. Wang, Z. Yun, and J. Rehg, "Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 383–398.
[50] Y. Sugano, Y. Matsushita, and Y. Sato, "Appearance-based gaze estimation using visual saliency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 329–341, Feb. 2013.
[51] K. Alex, S. Ilya, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2012, pp. 1097–1105.
[52] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, p. 448–456.
[53] M. Abadi *et al.* (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. [Online]. Available: https://www.tensorflow.org/
[54] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–13.
[55] K. A. F. 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 *Proc. Eye Tracking Res. Appl. Symp. (ETRA)*, 2014, pp. 255–258.
[56] T. Fischer, H. Chang, and Y. Demiris, "RT-GENE: Real-time eye gaze estimation in natural environments," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 334–352.
[57] S. Park, S. D. Mello, P. Molchanov, U. Iqbal, O. Hilliges, and J. Kautz, "Few-shot adaptive gaze estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9368–9377.
[58] Y. Xiong, H. J. Kim, and V. Singh, "Mixed effects neural networks (MeNets) with applications to gaze estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7743–7752.
[59] Z. Chen and B. E. Shi, "Appearance-based gaze estimation using dilated-convolutions," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2018, pp. 309–324.



Yihua Cheng received the B.S. degree from the College of Computer Science and Technology, Beijing University of Posts and Telecommunications, in 2017. He is currently pursuing the Ph.D. degree with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University. His research interests include computer vision and human gaze analysis.

Xucong Zhang received the B.Sc. degree from China Agriculture University in 2007, the M.Sc. degree from Beihang University in 2010, and the Ph.D. degree from the Max-Planck-Institute for Informatics in 2018. He is currently a Postdoctoral Researcher with Advanced Interactive Technologies Laboratory, Swiss Federal Institute of Technology (ETH) Zurich, Zurich, Switzerland. His research interests include computer vision, human–computer interaction, and learning-based gaze estimation.



Feng Lu (Member, IEEE) received the B.S. and M.S. degrees in automation from Tsinghua University, in 2007 and 2010, respectively, and the Ph.D. degree in information science and technology from The University of Tokyo in 2013. He is currently a Professor with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University. His research interests include computer vision, human–computer interaction, and augmented intelligence.

Yoichi Sato (Senior Member, IEEE) received the B.S. degree from The University of Tokyo in 1990 and the M.S. and Ph.D. degrees in robotics from the School of Computer Science, Carnegie Mellon University, in 1993 and 1997, respectively. He is currently a Professor with the Institute of Industrial Science, The University of Tokyo. His research interests include physics-based vision, reflectance analysis, first-person vision, and gaze sensing and analysis. He served/is serving in several conference organization and journal editorial roles, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the *International Journal of Computer Vision*, *Computer Vision and Image Understanding*, the CVPR 2023 General Co-Chair, the ICCV 2021 Program Co-Chair, the ACCV 2018 General Co-Chair, the ACCV 2016 Program Co-Chair, and the ECCV 2012 Program Co-Chair.