Appearance-Based Gaze Estimation via Evaluation-Guided Asymmetric Regression

Yihua Cheng[1], Feng Lu[1,2](✉), and Xucong Zhang[3]

[1] State Key Laboratory of Virtual Reality Technology and Systems,
School of Computer Science and Engineering, Beihang University, Beijing, China
{yihua_c,lufeng}@buaa.edu.cn

[2] Beijing Advanced Innovation Center for Big Data-Based Precision Medicine,
Beihang University, Beijing, China

[3] Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken,
Germany
xczhang@mpi-inf.mpg.de

Abstract. Eye gaze estimation has been increasingly demanded by recent intelligent systems to accomplish a range of interaction-related tasks, by using simple eye images as input. However, learning the highly complex regression between eye images and gaze directions is nontrivial, and thus the problem is yet to be solved efficiently. In this paper, we propose the Asymmetric Regression-Evaluation Network (ARE-Net), and try to improve the gaze estimation performance to its full extent. At the core of our method is the notion of "two eye asymmetry" observed during gaze estimation for the left and right eyes. Inspired by this, we design the multi-stream ARE-Net; one asymmetric regression network (AR-Net) predicts 3D gaze directions for both eyes with a novel asymmetric strategy, and the evaluation network (E-Net) adaptively adjusts the strategy by evaluating the two eyes in terms of their performance during optimization. By training the whole network, our method achieves promising results and surpasses the state-of-the-art methods on multiple public datasets.

Keywords: Gaze estimation · Eye appearance
Asymmetric regression

1 Introduction

The eyes and their movements carry important information that conveys human visual attention, purpose, intention, feeling and so on. Therefore, the ability to automatically track human eye gaze has been increasingly demanded by many recent intelligent systems, with direct applications ranging from human-computer interaction [1,2], saliency detection [3] to video surveillance [4].

This work was supported by NSFC under Grant U1533129, 61602020 and 61732016.

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11218, pp. 105–121, 2018.
https://doi.org/10.1007/978-3-030-01264-9_7

As surveyed in [5], 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, to compute the gaze direction. However, they share common limitations such as (1) requirement on specific hardware for illumination and capture, (2) high failure rate when used in the uncontrolled environment, and (3) limited working distance (typically within 60 cm).

Different with model-based methods, appearance-based methods do not rely on small eye feature extraction under special illumination. Instead, they can work with just a single ordinary camera to capture the eye appearance, then learn a mapping function to predict the gaze direction from the eye appearance directly. Whereas this greatly enlarges the applicability, the challenge part is that human eye appearance can be heavily affected by various factors, such as the head pose, the illumination, and the individual difference, making the mapping function difficult to learn. In recent years, the Convolutional Neural Network (CNN) has shown to be able to learn very complex functions given sufficient training data. Consequently, the CNN-based methods have been reported to outperform the conventional methods [6].

The goal of this work is to further exploit the power of CNNs and improve the performance of appearance-based gaze estimation to a higher level. At the core of our method is the notion of asymmetric regression for the left and the right eyes. It is based on our key observation that (1) the gaze directions of two eyes should be consistent physically, however, (2) even if we apply the same regression method, the gaze estimation performance on two eyes can be very different. Such "two eye asymmetry" implys a new gaze regression strategy that no longer treats both eyes equally but tends to rely on the"high quality eye" to train a more efficient and robust regression model.

In order to do so, we consider the following technical issues, i.e., how to design a network that processes both eyes simultaneously and asymmetrically, and how to control the asymmetry to optimize the network by using the high quality data. Our idea is to ***guide the asymmetric gaze regression by evaluating the performance of the regression strategy w.r.t.different eyes***. In particular, by analyzing the "two eye asymmetry" (Sect. 3), we propose the asymmetric regression network (AR-Net) to predict 3D gaze directions of two eyes (Sect. 4.2), and the evaluation networks (E-Net) to adaptively evaluate and adjust the regression strategy (Sect. 4.3). By integrating the AR-Net and the E-Net (Sect. 4.4), the proposed Asymmetric Regression-Evaluation Network (ARE-Net) learns to maximize the overall performance for the gaze estimator.

Our method makes the following assumptions. First, as commonly assumed by previous methods along this direction [6,7], the user head pose can be obtained by using existing head trackers [8]. Second, the user should roughly fixate on the same targets with both eyes, which is usually the case in practice.

With these assumptions, our method is capable of estimating gaze directions of the two eyes from their images.

In summary, the contributions of this work are threefold:

- We propose the multi-stream AR-Net for asymmetric two-eye regression. 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 which is new.
- Based on the proposed mechanism and networks, we design the final ARE-Net and it shows promising performance in gaze estimation for both eyes.

2 Related Work

There have been an increasing number of recent researches proposed for the task of remote human gaze estimation, which can be roughly divided into two major categories: model-based and appearance-based [5,9].

The Model-Based Methods estimate gaze directions using certain geometric eye models [10]. They typically extract and use near infrared (IR) corneal reflections [10–12], pupil center [13,14] and iris contours [15,16] from eye images as the input features to fit the corresponding models [17]. Whereas this type of methods can predict gaze directions with a good accuracy, the extraction of eye features may require hardware that may be composed of infrared lights, stereo/high-definition cameras and RBG-D cameras [15,16]. These devices may not be available when using many common devices, and they usually have limited working distances. As a result, the model-based methods are more suitable for being used in the controlled environments, e.g., in the laboratory, rather than in outdoor scenes or with large user-camera distances, e.g., for advertisement analysis [18].

The Appearance-Based Methods have relatively lower demand compared with the model-based methods. They typically need a single camera to capture the user eye images [19]. Certain non-geometric image features are produced from the eye images, and then used to learn a gaze mapping function that maps eye images to gaze directions. Up to now, various mapping functions have been explored, such as neural networks [20,21], local linear interpolation [19], adaptive linear regression [22], Gaussian process regression [23], and dimension reduction [24,25]. Some other methods use additional information such as saliency maps [22,26] to guide the learning process. These methods all aim at reducing the number of required training samples while maintaining the regression accuracy. However, since the gaze mapping is highly non-linear, the problem still remains challenging to date.

The CNNs-Based Methods have already shown their ability to handle complex regression tasks, and thus they have outperformed traditional appearance-based methods. Some recent works introduce large appearance-based gaze datasets [27] and propose effective CNN-based gaze estimators [6,28]. More recently, Krafka *et al.* implement the CNN-based gaze tracker in the mobile devices [29]. Zhang *et al.* take into consideration the full face as input to the CNNs [30]. Deng *et al.* propose a CNN-based method with geometry constraints [7]. In general, these methods can achieve better performance than traditional ones. Note that they all treat the left and the right eyes indifferently, while in this paper we try to make further improvement by introducing and utilizing the two eye asymmetry.

Besides the eye images, recent appearance-based methods may also take the face images as input. The face image can be used to compute the head pose [6,31] or input to the CNN for gaze regression [29,30]. In our method, we only assume available head poses that can be obtained by using any existing head tracker, and we do not require high resolution face images as input for gaze estimation.

3 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.

3.1 3D Gaze Estimation via Regression

Any human gaze direction can be denoted by a 3D unit vector $\mathbf{g}$, which represents the eyeball orientation in the 3D space. Meanwhile, the eyeball orientation also determines the eye appearance in the eye image, e.g., the location of the iris contour and the shape of the eyelids. Therefore, there is a strong relation between the eye gaze direction and the eye appearance in the image. As a result, the problem of estimating the 3D gaze direction $\mathbf{g} \in \mathbb{R}^3$ from a given eye image $\boldsymbol{I} \in \mathbb{R}^{H \times W}$ can be formulated as a regression problem $\mathbf{g} = f(\boldsymbol{I})$.

The regression is usually highly non-linear because the eye appearance is complex. Besides, there are other factors that will affect $\boldsymbol{I}$, and the head motion is a major one. In order to handle head motion, it is necessary to also consider the head pose $\mathbf{h} \in \mathbb{R}^3$ in the regression, which results in

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

where f is the regression function.

In the literature, various regression models have been used, such as the Neural Network [20], the Gaussian Process regression model [32], and the Adaptive Linear Regression model [22]. However, the problem is still challenging. In recent years, with the fast development of the deep neural networks, solving such a highly complex regression problem is becoming possible with the existence of large training dataset, while designing an efficient network architecture is the most important work to do.

3.2 Two Eye Asymmetry

Existing gaze regression methods handles the two eyes indifferently. However, in practice, we observe the two eye asymmetry regarding the regression accuracy.

> ***Observation.*** *At any moment, we cannot expect the same accuracy for two eyes, and either eye has a chance to be more accurate.*

The above "two eye asymmetry" can be due to various factors, e.g., head pose, image quality and individuality. It's a hint that the two eyes' images may have different 'qualities' in gaze estimation. Therefore, when training a gaze regression model, it is better to identify and rely on the high quality eye image from the input to train a more efficient and robust model.

4 Asymmetric Regression-Evaluation Network

Inspired by the "two eye asymmetry", in this section, we deliver the Asymmetric Regression-Evaluation Network (ARE-Net) for appearance-based gaze estimation of two eyes.

4.1 Network Overview

The proposed networks use two eye images $\{\boldsymbol{I}_l^{(i)}\}$, $\{\boldsymbol{I}_r^{(i)}\}$ and the head pose vector $\{\mathbf{h}^{(i)}\}$ as input, to learn a regression that predicts the ground truth $\{\mathbf{g}_l^{(i)}\}$ and $\{\mathbf{g}_r^{(i)}\}$, where $\{\mathbf{g}_l^{(i)}\}$ and $\{\mathbf{g}_r^{(i)}\}$ are 3D gaze directions and i is the sample index. For this purpose, we first introduce the Asymmetric Regression Network (AR-Net), and then propose the Evaluation Network (E-Net) to guide the regression. The overall structure is shown in Fig. 1.

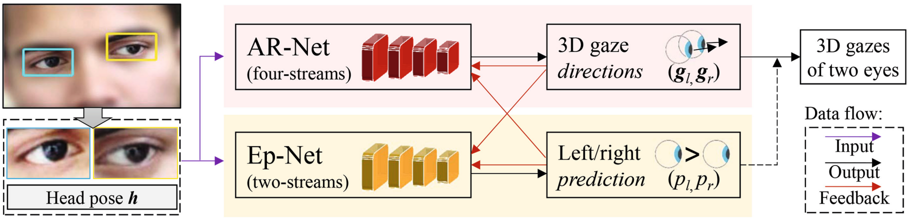


Fig. 1. Overview of the proposed Asymmetric Regression-Evaluation Network (ARE-Net). It consists of two major sub-networks, namely, the AR-Net and the E-Net. The AR-Net performs asymmetric regression for the two eyes, while the E-Net predicts and adjust the asymmetry to improve the gaze estimation accuracy.

Asymmetric Regression Network (AR-Net). It is a four-stream convolutional network and it performs 3D gaze direction regression for both the left and the right eyes (detailed in Sect. 4.2). Most importantly, it is designed to be able to optimize the two eyes in an asymmetric way.

Evaluation Network (E-Net). It is a two stream convolutional network that learns to predict the current asymmetry state, i.e., which eye the AR-Net tends to optimize at that time, and accordingly it adjusts the degree of asymmetry (detailed in Sect. 4.3).

Network training. During training, parameters of both the AR-Net and the E-Net are updated simultaneously. The loss functions and other details will be given in the corresponding sections.

Testing stage. During test, the output of the AR-Net are the 3D gaze directions of both eyes.

4.2 Asymmetric Regression Network (AR-Net)

The AR-Net processes two eye images in a joint and asymmetric way, and estimates their 3D gaze directions.

Architecture. The AR-Net is a four-stream convolutional neural network, using the "base-CNN" as the basic component followed by some fully connected layers, as shown in Fig. 2(a). Follow the idea that both the separate features and joint feature of the two eyes should be extracted and utilized, we design the first two streams to extract a 500D deep features from each eye independently, and the last two streams to produce a joint 500D feature in the end.

Note that the head pose is also an important factor to affect gaze directions, and thus we input the head pose vector (3D for each eye) before the final regression. The final 1506D feature vector is produced by concatenating all the outputs from the previous networks, as shown in Fig. 2(a).

The Base-CNN. The so called "base-CNN" is the basic component of the proposed AR-Net and also the following E-Net. It consists of six convolutional layers, three max-pooling layers, and a fully connected layer in the end. The structure of the base-CNN is shown in Fig. 2(c). The size of each layer in the base-CNN is set to be similar to that of AlexNet [33].

The input to the base-CNN can be any gray-scale eye image with a fixed resolution of 36 × 60. For the convolutional layers, the learnable filters size is 3 × 3. The output channel number is 64 for the first and second layer, 128 for the third and fourth layer, and 256 for the fifth and sixth layer.

Loss Function. We measure the angular error of the currently predicted 3D gaze directions for the two eyes by

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

and

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

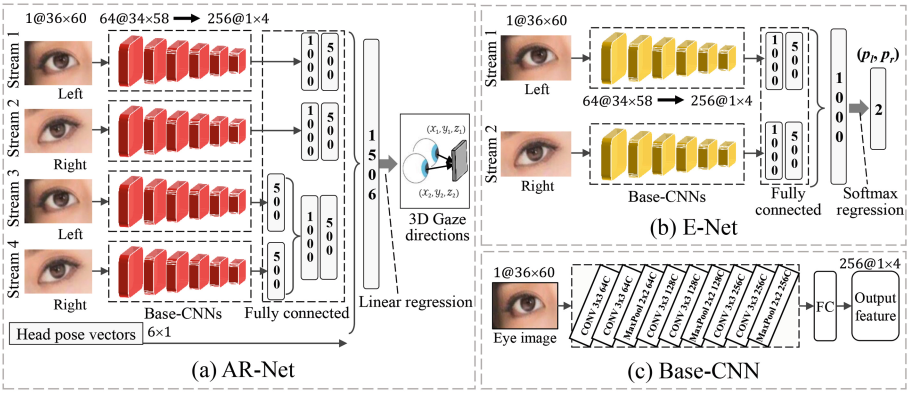


Fig. 2. Architecture of the proposed networks. (a) The AR-Net is a four-stream network to produce features from both the eye images. A linear regression is used to estimate the 3D gaze directions of the two eyes. (b) The E-Net is a two-stream network for two eye evaluation. The output is a two-dimensional probability vector. (c) The base-CNN is the basic component to build up the AR-Net and the E-Net. It uses an eye image as input. The output is a 1000D feature after six convolutional layers.

where $f(\cdot)$ indicates the gaze regression. Then, we compute the weighted average of the two eye errors

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

to represent the loss in terms of gaze prediction accuracy of both eyes.

Asymmetric Loss. The weights λ_l and λ_r determine whether the accuracy of the left or the right eye should be considered more important. In the case that $\lambda_l \neq \lambda_r$, the loss function becomes asymmetric. According to the "two eye asymmetry" discussed in Sect. 3.2, if one of the two eyes is more likely to achieve a smaller error, we should enlarge its weight in optimizing the network. Following this idea, we propose to set the weights according to the following:

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

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

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

which encourages to rely on the high quality eye in training.

4.3 Evaluation Network (E-Net)

As introduced above, the AR-Net can rely on the high quality eye image for asymmetric learning. In order to provide more evidence on which eye it should be, we design the E-Net to learn to predict the choice of the AR-Net, and also guide its asymmetric strategy during optimization.

Architecture. The E-Net is a two-stream network with the left and the right eye images as input. Each of the two stream is a base-CNN followed by two fully connected layers. The output 500D features are then concatenated to be a 1000D feature, as shown in Fig. 2(b).

Finally, the 1000D feature is sent to the Softmax regressor to output a 2D vector $[p_l, p_r]^{\mathrm{T}}$, where p_l is the probability that the AR-Net chooses to rely on the left eye, and p_r for the right eye.

During training, the ground truth for p is set to be 1 if $e_l < e_r$ from the AR-Net, otherwise p is set to be 0. In other words, the evaluation network is trained to predict the probability of the left/right eye image being more efficient in gaze estimation.

Loss Function: In order to train the E-Net to predict the AR-Net's choice, we set its loss function as below:

$$\begin{aligned}\mathcal{L}_E = -\{&\eta \cdot \arccos(f(\boldsymbol{I}_l) \cdot f(\boldsymbol{I}_r)) \cdot \log(p_l) + \\ &(1-\eta) \cdot \arccos(f(\boldsymbol{I}_l) \cdot f(\boldsymbol{I}_r)) \cdot \log(p_r)\},\end{aligned} \tag{8}$$

where $\eta = 1$ if $e_l \leq e_r$, and $\eta = 0$ if $e_l > e_r$. Besides, $\arccos(f(\boldsymbol{I}_l) \cdot f(\boldsymbol{I}_r))$ computes the angular difference of the two eye gaze directions estimated by the AR-Net, which measures the inconsistency of $\mathbf{g}_l$ and $\mathbf{g}_r$.

This loss function can be intuitively understood as follows: if the left eye has smaller error in the AR-Net, i.e., $e_l < e_r$, the E-Net should choose to maximize p_l to learn this fact in order to adjust the regression strategy of the AR-Net, especially in the case when $\mathbf{g}_l$ and $\mathbf{g}_r$ are inconsistent. In this way, the E-Net is trained to predict the high quality eye that can help optimize the AR-Net.

Modifying the Loss Function of AR-Net. An important task of the E-Net is to adjust the asymmetry of the AR-Net, with the aim to improve the gaze estimation accuracy, as explained before. In order to do so, by integrating the E-Net, the loss function of the AR-Net in Eq. (7) can be modified as

$$\mathcal{L}^*_{AR} = \omega \cdot \mathcal{L}_{AR} + (1-\omega) \cdot \beta \cdot (\frac{e_l + e_r}{2}), \tag{9}$$

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 was set to 0.1 in our experiments. In particular, given the output (p_l, p_r) of the E-Net, we compute

$$\omega = \frac{1 + (2\eta - 1) \cdot p_l + (1 - 2\eta) \cdot p_r}{2}. \tag{10}$$

Again, $\eta = 1$ if $e_l \leq e_r$, and $\eta = 0$ if $e_l > e_r$. Here we omit the derivation of ω, while it is easy to see that $\omega = 1$ when both the AR-Net and E-Net have a strong agreement on the high quality eye, meaning that a heavily asymmetric learning strategy can be recommanded; $\omega = 0$ when they completely disagree, meaning that it is better to just use a symmetric learning strategy as a compromise. In practice, ω is a decimal number between 0 and 1.

4.4 Guiding Gaze Regression by Evaluation

Following the explanations above, we summarize again how the AR-Net and the E-Net are integrated together (Fig. 1), and how the E-Net can guide the AR-Net.

- **AR-Net**: takes both eye images as input; loss function modified by the E-Net's output (p_l, p_r) to adjust the asymmetry adaptively (Eq. (9)).
- **E-Net**: takes both eye images as input; loss function modified by the AR-Net's output $(f(\boldsymbol{I}_l), f(\boldsymbol{I}_r))$ and the errors (e_l, e_r) to predict the high quality eye image for optimization (Eq. (8)).
- **ARE-Net**: as shown in Fig. 1, the AR-Net and the E-Net are integrated and trained together. The final gaze estimation results are the output $(f(\boldsymbol{I}_l), f(\boldsymbol{I}_r))$ from the AR-Net.

5 Experimental Evaluation

In this section, we evaluate the proposed Asymmetric Regression-Evaluation Network by conducting multiple experiments.

5.1 Dataset

The proposed is a typical appearance-based gaze estimation method. Therefore, we use the following datasets in our experiments as previous methods do. Necessary modification have been done as described.

Modified MPIIGaze Dataset: the MPIIGaze dataset [6] is composed of 213659 images of 15 participants, which contains a large variety of different illuminations, eye appearances and head poses. It is among the largest datasets for appearance-based gaze estimation and thus is commonly used. All the images and data in the MPIIGaze dataset have already been normalized to eliminate the effect due to face misalignment.

The MPIIGaze dataset provides a standard subset for evaluation, which contains 1500 left eye images and 1500 right eye images independently selected from each participants. However, our method requires paired eye images captured at the same time. Therefore, we modify the evaluation set by finding out the missing image of every left-right eye image pair from the original dataset. This doubles the image number in the evaluation set. In our experiments, we use such a modified dataset instead of the original MPIIGaze dataset.

Besides, we also conduct experiments to compare with methods using full face images as input. As a result, we use the same full face subset from the MPIIGaze dataset as described in [30].

UT Multiview Dataset [34]**:** it contains dense gaze data of 50 participants. Both the left and right eye images are provided directly for use. The data normalization is done as for the MPIIGaze dataset.

EyeDiap Dataset [27]**:** it contains a set of video clips of 16 participants with free head motion under various lighting conditions. We randomly select 100 frames from each video clip, resulting in 18200 frames in total. Both eyes can be obtained from each video frame. Note that we need to apply normalization for all the eye images and data in the same way as the MPIIGaze dataset.

5.2 Baseline Methods

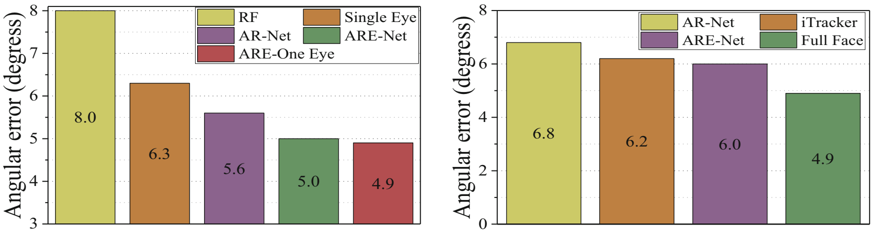
For comparison, we use the following methods as baselines. Results of the baseline methods are obtained from our implementation or the published paper.

- **Single Eye** [6]**:** One of the typical appearance-based gaze estimation method 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 [6] to obtain all the results in our experiments. Note that another method [28] also uses the same network for gaze estimation and thus we regard [6,28] to be the same baseline.
- **RF:** One of the most commonly used regression method. It is shown to be effective for a variety of applications. Similar to [34], multiple RF regressors are trained for each head pose cluster.
- **iTracker** [29]**:** A multi-streams method that takes the full face image, two individual eye images, and a face grid as input. The performance of iTracker has already been reported in [30] on the MPIIGaze dataset and thus we use the reported numbers.
- **Full Face** [30]**:** A deep neuroal network-based method that takes the full face image as input with a spatial weighting strategy. Its performance has also been tested and reported on the same MPIIGaze dataset.

5.3 Within Dataset Evaluation

We first conduct experiments with training data and test data from the same dataset. In particular, we use the modified MPIIGaze dataset as described in Sect. 5.1 since it contains both eye images and the full face images of a large amount. Note that because the training data and test data are from the same dataset, we use the leave-one-person-out strategy to ensure that the experiments are done in a fully person-independent manner.

Eye image-Based Methods. We first consider the scenario where only eye images are used as the input. The accuracy is measured by the average gaze error of all the test samples including both the left and right images. The results



(a) v.s. eye image-based methods.

(b) v.s. full face image-based methods.

Fig. 3. Experimental results of the within-dataset evaluation and comparison.

of all the methods are obtained by running the corresponding codes on our modified MPIIGaze dataset with the same protocol. The comparison is shown in Fig. 3(a). The proposed method clearly achieves the best accuracy. As for the AR-Net, the average error is 5.6°, which is more than 11% improved compared to the Single Eye method, and also 30% improved compared to the RF method. This is benefited from both our new network architecture and loss fuction design. In addition, by introducing the E-Net, the final ARE-Net further improves the accuracy by a large margin. This demonstrates the effectiveness of the proposed E-Net as well as the idea of evaluation-guided regression. The final accuracy of 5.0° achieves the state-of-the-art for eye image-based gaze estimation.

Full Face Image-Based Methods. Recent methods such as [30] propose to use the full face image as input. Although our method only requires eye images as input, we still make a comparison with them. As for the dataset, we use the face image dataset introduced previously, and extract the two eye images as our input. Note that following [30], the gaze origin is defined at the face center for both the iTracker and Full Face methods. Therefore, in order to make a fair comparison, we also convert our estimated two eye gaze vectors to have the same origin geometrically, and then take their average as the final output.

As shown in Fig. 3(b), the Full Face method achieves the lowest error, while the proposed AR-Net and ARE-Net also show good performance which is comparable with the iTracker. Note the fact that our method is the only one that does not need full face image as input, its performance is quite satisfactory considering the save of computational cost (face image resolution 448 × 448 v.s. eye image resolution 36 × 60).

5.4 Cross-Dataset Evaluation

We then present our evaluation results in a cross-dataset setting. For the training dataset, we choose the UT Multiview dataset since it covers the largest variation of gaze directions and head poses. Consequently, we use data from the other two datasets, namely the MPIIGaze and EyeDiap datasets, as test data. As for the test data from the Eyediap dataset, we extract 100 images from each video clip, resulting in 18200 face images for test.

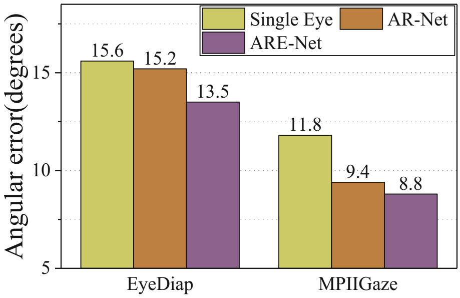


Fig. 4. Experimental results of the cross-dataset evaluation. The proposed methods outperform the Single Eye method on the EyeDiap and MPIIGaze datasets.

We first compare our method with the Single Eye method, which is a typical CNN-based method. As shown in Fig. 4, the proposed ARE-Net outperforms the Single Eye method on both the MPIIGaze and the EyeDiap datasets. In particular, compared with the Single Eye method, the performance improvement is 13.5% on the EyeDiap dataset, and 25.4% on the MPIIGaze dataset. This demonstrates the superior of the proposed ARE-Net. Note that our basic AR-Net also achieves a better accuracy than the Single Eye method. This shows the effectiveness of the proposed four-stream network with both eyes as input.

5.5 Evaluation on Each Individual

Previous experiments show the advantage of the proposed method in terms of the average performance. In this section, we further analyse its performance for each subject. As shown in Table 1, results for all the 15 subjects in the MPIIGaze dataset are illustrated, with a comparison to the Single Eye method. The proposed ARE-Net and AR-Net outperform the Single Eye method for almost every subject (with only one exception), and the ARE-Net is also consistently better than the AR-Net. This validates our key idea and confirms the robustness of the proposed methods.

Table 1. Comparison of the Single Eye, AR and ARE methods regarding their accuracy on each subject.

Method	Subject															Avg.
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
Single Eye	4.9	7.1	5.8	6.5	5.9	6.4	5.6	7.6	6.6	7.7	6.0	6.0	6.1	6.9	5.5	6.3
AR-Net	4.0	4.4	5.9	6.8	3.7	6.1	4.3	5.8	6.0	7.1	6.5	5.5	5.6	6.8	6.2	5.7
ARE-Net	3.8	3.4	5.1	5.0	3.2	6.2	3.9	5.6	5.5	5.7	6.7	5.1	4.0	5.7	6.3	**5.0**

5.6 Analysis on E-Net

The proposed E-Net is the key component of our method and thus it is important to know how it benefits the method. To this end, we make further analysis based on the initial results obtained in Sect. 5.3. According to the comparisons shown in Table 2, we have the following conclusions:

- Regarding the overall gaze error, the existence of the E-Net improves the accuracy greatly in all cases compared to other methods.
- The E-Net can still select the relatively better eye to some extent from the already very ballanced output of the ARE-Net, while those other strategies cannot make more efficient selection.
- With the E-net, the difference between the better/worse eyes reduces greatly (to only 0.4°). Therefore, the major advantage of the E-Net is that it can optimize both the left and the right eyes simultaneously and effectively.
- Even if compared with other methods with correctly selected better eyes, the ARE-Net still achieves the best result without selection.

Table 2. Analysis on average gaze errors of: (left to right) average error of two eyes/E-Net's selection/the better eye/the worse eye/difference between the better and worse eyes/the eye near the camera/the more frontal eye.

Methods	Two eyes	E-Net select	Better eye	Worse eye	Δ	Near	Frontal
RF	8.0	–	6.7	9.4	2.7	8.1	8.1
Single Eye	6.3	–	5.0	7.6	2.6	6.2	6.4
AR-Net	5.7	–	5.3	6.0	0.7	5.6	5.7
ARE-Net	**5.0**	**4.9**	**4.8**	**5.2**	**0.4**	5.0	5.0

5.7 Additional Anaysis

Additional analyses and discussions on the proposed method are presented in this section.

Convergency. Figure 5 shows the convergency analysis of the proposed ARE-Net tested on the MPIIGaze dataset. During iteration, the estimation error tends to decrease guadually, and achieves the minimum after around 100 epochs. In general, during our experiments, the proposed network is shown to be able to converge quickly and robustly.

Case Study. We show some representative cases that explain why the proposed method is superior to the previous one, as shown in Fig. 6. In these cases, using only a single eye image, e.g., as the Single Eye method, may perform well for one eye but badly for the other eye, and the bad one will affect the final accuracy

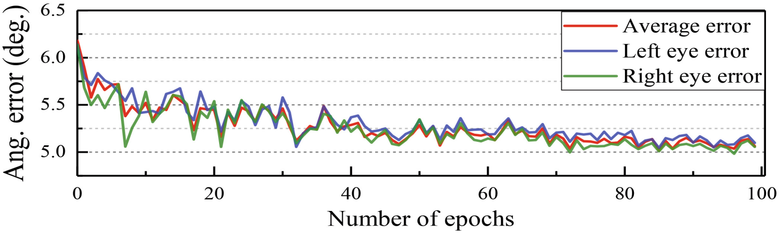


Fig. 5. Validation on the convergency of the ARE-Net.

greatly. On the other hand, the ARE-Net performs asymmetric optimization and helps improve both the better eye and the worse eye via the designed evaluation and feedback strategy. Therefore, the output gaze errors tend to be small for both eyes and this results in a much better overall accuracy. This is also demonstrated in Table 2.

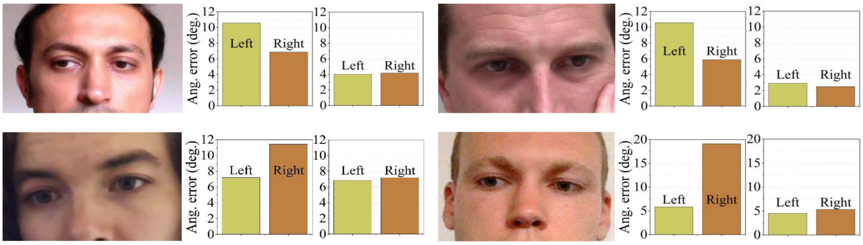


Fig. 6. Comparison of two eyes' gaze errors. The Single Eye method (left plot of each case) usually produces large errors in one eye while the proposed ARE-Net (right plot of each case) reduces gaze errors for both eyes.

Only One Eye Image as Input. Our method requires both the left and the right eye images as input. In the case that only one of the eye images is available, we can still test our network as follows.

Without loss of generality, assume we only have a left eye image. In order to run our method, we need to feed the network with something as the substitute for the right eye. In our experiment, we use (1) **0** matrix, i.e., a black image, (2) a copy of the left eye, (3) a randomly selected right eye image from a different person in the dataset, and (4) a fixed right eye image (typical shape, frontal gaze) from a different person in the dataset.

We test the trained models in Sect. 5.3 in the same leave-one-person-out manner. The average results of all the 15 subjects on the modified MPIIGaze dataset are shown in Table 3. It is interesting that if we use a black image or a copy of the input image to serve as the other eye image, the estimation errors are quite good (∼6°). This confirms that our network is quite robust even if there is a very low quality eye image.

Table 3. Gaze estimation errors using only one eye image as input to the ARE-Net.

Input image	Substitute for the missing eye image			
	0 matrix	Copy input	Random eye	Fixed eye
Left eye	6.3° (left)	6.1°(left)	8.5°(left)	10.7°(left)
Right eye	6.2° (right)	6.1°(right)	7.9°(right)	9.3°(right)

6 Conclusion and Discussion

We present a deep learning-based method for remote gaze estimation. This problem is challenging because learning the highly complex regression between eye images and gaze directions is nontrivial. In this paper, we propose the Asymmetric Regression-Evaluation Network (ARE-Net), and try to improve the gaze estimation performance to its full extent. At 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. Accordingly, we design the multi-stream ARE-Net. It contains one asymmetric regression network (AR-Net) to predict 3D gaze directions for both eyes with an asymmetric strategy, and one evaluation networks (E-Net) to adaptively adjust the strategy by evaluating the two eyes in terms of their quality in optimization. By training the whole network, our method achieves good performances on public datasets.

There are still future works to do along this line. First, we consider extending our current framework to also exploit the full face information. Second, since our current base-CNN is simple, it is possible to further enhance its performance if we use more advanced network structures.

References

1. Zhang, X., Sugano, Y., Bulling, A.: Everyday eye contact detection using unsupervised gaze target discovery. In: Proceedings of the ACM Symposium on User Interface Software and Technology (UIST), pp. 193–203 (2017)
2. Sugano, Y., Zhang, X., Bulling, A.: Aggregaze: collective estimation of audience attention on public displays. In: Proceedings of the ACM Symposium on User Interface Software and Technology (UIST), pp. 821–831 (2016)
3. Sun, X., Yao, H., Ji, R., Liu, X.M.: Toward statistical modeling of saccadic eye-movement and visual saliency. IEEE Trans. Image Process. **23**(11), 4649 (2014)
4. Cheng, Q., Agrafiotis, D., Achim, A., Bull, D.: Gaze location prediction for broadcast football video. IEEE Trans. Image Process. **22**(12), 4918–4929 (2013)
5. Hansen, D., Ji, Q.: In the eye of the beholder: A survey of models for eyes and gaze. IEEE Trans. PAMI **32**(3), 478–500 (2010)
6. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4511–4520 (2015)
7. Zhu, W., Deng, H.: Monocular free-head 3D gaze tracking with deep learning and geometry constraints. In: The IEEE International Conference on Computer Vision (ICCV) (2017)

8. Lepetit, V., Moreno-Noguer, F., Fua, P.: EPNP: an accurate o(n) solution to the pnp problem. Int. J. Comput. Vis. **81**(2), 155 (2008)
9. Morimoto, C., Mimica, M.: Eye gaze tracking techniques for interactive applications. CVIU **98**(1), 4–24 (2005)
10. Guestrin, E., Eizenman, M.: General theory of remote gaze estimation using the pupil center and corneal reflections. IEEE Trans. Biomed. Eng. **53**(6), 1124–1133 (2006)
11. Zhu, Z., Ji, Q.: Novel eye gaze tracking techniques under natural head movement. IEEE Trans. Biomed. Eng. J. **54**(12), 2246–2260 (2007)
12. Nakazawa, A., Nitschke, C.: Point of gaze estimation through corneal surface reflection in an active illumination environment. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, pp. 159–172. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33709-3_12
13. Valenti, R., Sebe, N., Gevers, T.: Combining head pose and eye location information for gaze estimation. IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc. **21**(2), 802–815 (2012)
14. Jeni, L.A., Cohn, J.F.: Person-independent 3d gaze estimation using face frontalization. In: Computer Vision and Pattern Recognition Workshops, pp. 792–800 (2016)
15. Funes Mora, K.A., Odobez, J.M.: Geometric generative gaze estimation (g3e) for remote RGB-D cameras. In: IEEE Computer Vision and Pattern Recognition Conference, pp. 1773–1780 (2014)
16. Xiong, X., Liu, Z., Cai, Q., Zhang, Z.: Eye gaze tracking using an RGBD camera: a comparison with a RGB solution. The 4th International Workshop on Pervasive Eye Tracking and Mobile Eye-Based Interaction (PETMEI 2014), pp. 1113–1121 (2014)
17. Wang, K., Ji, Q.: Real time eye gaze tracking with 3d deformable eye-face model. In: The IEEE International Conference on Computer Vision (ICCV) (2017)
18. Duchowski, A.T.: A breadth-first survey of eye-tracking applications. Behav. Res. Methods Instrum. Comput. **34**(4), 455–470 (2002)
19. Tan, K., Kriegman, D., Ahuja, N.: Appearance-based eye gaze estimation. In: WACV, pp. 191–195 (2002)
20. Baluja, S., Pomerleau, D.: Non-Intrusive Gaze Tracking Using Artificial Neural Networks. Carnegie Mellon University (1994)
21. Xu, L.Q., Machin, D., Sheppard, P.: A novel approach to real-time non-intrusive gaze finding. In: BMVC, pp. 428–437 (1998)
22. Lu, F., Sugano, Y., Okabe, T., Sato, Y.: Adaptive linear regression for appearance-based gaze estimation. IEEE Trans. Pattern Anal. Mach. Intell. **36**(10), 2033–2046 (2014)
23. Williams, O., Blake, A., Cipolla, R.: Sparse and semi-supervised visual mapping with the S^3GP. In: CVPR, pp. 230–237(2006)
24. Schneider, T., Schauerte, B., Stiefelhagen, R.: Manifold alignment for person independent appearance-based gaze estimation. In: International Conference on Pattern Recognition (ICPR), pp. 1167–1172 (2014)
25. Lu, F., Chen, X., Sato, Y.: Appearance-based gaze estimation via uncalibrated gaze pattern recovery. IEEE Trans. Image Process. **26**(4), 1543–1553 (2017)
26. Sugano, Y., Matsushita, Y., Sato, Y., Koike, H.: Appearance-based gaze estimation with online calibration from mouse operations. IEEE Trans. Hum. Mach. Syst. **45**(6), 750–760 (2015)

27. Mora, K.A.F., Monay, F., Odobez, J.M.: Eyediap:a database for the development and evaluation of gaze estimation algorithms from RGB and RGB-D cameras. In: Symposium on Eye Tracking Research and Applications, pp. 255–258 (2014)
28. Wood, E., Morency, L.P., Robinson, P., Bulling, A.: Learning an appearance-based gaze estimator from one million synthesised images. In: Biennial ACM Symposium on Eye Tracking Research & Applications, pp. 131–138 (2016)
29. Krafka, K., et al.: Eye tracking for everyone. In: Computer Vision and Pattern Recognition, pp. 2176–2184 (2016)
30. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: 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 (CVPRW) (2017)
31. Lu, F., Sugano, Y., Okabe, T., Sato, Y.: Head pose-free appearance-based gaze sensing via eye image synthesis. In: International Conference on Pattern Recognition, pp. 1008–1011 (2012)
32. Sugano, Y., Matsushita, Y., Sato, Y.: Appearance-based gaze estimation using visual saliency. IEEE Trans. Pattern Anal. Mach. Intell. **35**(2), 329 (2013)
33. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (2012)
34. Sugano, Y., Matsushita, Y., Sato, Y.: Learning-by-synthesis for appearance-based 3D gaze estimation. In: Computer Vision and Pattern Recognition, pp. 1821–1828 (2014)