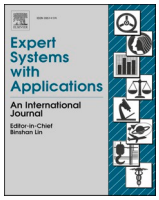
Expert Systems With Applications 238 (2024) 122363

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Review

EG-Net: Appearance-based eye gaze estimation using an efficient gaze network with attention mechanism

Xinmei Wu [a], Lin Li [a,*], Haihong Zhu [a], Gang Zhou [a], Linfeng Li [b], Fei Su [c], Shen He [d], Yanggang Wang [b], Xue Long [e]

[a] School of Resource and Environmental Science, Wuhan University, Wuhan, China
[b] Wuhan Highway Technology Corporation, Building B3, Zone 2, Hangyu, WHU Sci-Park, Wudayuan Road, East Lake Hi-Tech Development Zone, Wuhan, China
[c] School of Surveying and Geo-Informatics, Shandong Jianzhu University, Jinan 250101, China
[d] Wuhan Metro Operation Co., Ltd., No. 99 Jinghan Avenue, Qiaokou District, Wuhan, China
[e] Beijing Institute of Structure and Environment Engineering, No.1, Donggaodi South Dahongmen Road, Fengtai District, Beijing, China

ARTICLE INFO

Keywords:
Gaze estimation
Appearance-based method
EG-Net
Attention mechanism
Compound model scaling

ABSTRACT

Gaze estimation, which has a wide range of applications in many scenarios, is a challenging task due to various unconstrained conditions. As information from both full-face and eye images is instrumental in improving gaze estimation, many multiregion gaze estimation models have been proposed in recent studies. However, most of them simply use the same regression method on both eye and face images, overlooking that the eye region may contribute more fine-grained features than the full-face region, and the variation in the left and right eyes of an individual caused by head pose, illumination, and partially occluded eye may lead to inconsistent estimations. To address these issues, we propose an appearance-based end-to-end learning network architecture with an attention mechanism, named efficient gaze network (EG-Net), which employs a two-branch network for gaze estimation. Specifically, a base CNN is utilized for full-face images, while an efficient eye network (EE-Net), which is scaled up from the base CNN, is used for left- and right-eye images. EE-Net uniformly scales up the depth, width and resolution of the base CNN with a set of constant coefficients for eye feature extraction and adaptively weights the left- and right-eye images via an attention network according to its "image quality". Finally, features from the full-face image, two individual eye images and head pose vectors are fused to regress the eye gaze vectors. We evaluate our approach on 3 public datasets, the proposed EG-Net model achieves much better performance. In particular, our EG-Net-v4 model outperforms state-of-the-art approaches on the MPIIFaceGaze dataset, with prediction errors of 2.41 cm and 2.76 degrees in 2D and 3D gaze estimation, respectively. It also yields a performance improvement to 1.58 cm on GazeCapture and 4.55 degrees on EyeDIAP dataset, with 23.4 % and 14.2 % improvement over prior arts on the two datasets respectively. The code related to this project is open-source and available at https://github.com/wuxinmei/EE_Net.git.

1. Introduction

As a fundamental element of human behaviours, eye gaze estimation is important in different applications (Shic & Scassellati, 2007; Zhao, Lu, Yao, Chen, & Zhang, 2020), such as social interactions (Neilmacrae, Hood, Milne, Rowe, & Mason, 2002; Otsu, Seo, Kitajima, & Chen, 2020; Poulopoulos & Psarakis, 2023), human–computer communications (Fung, Jin, Zhao, & Hoque, 2015; Majaranta & Bulling, 2014; Zhang, Yao, & Cai, 2018), consumer behaviour research (Wedel & Pieters, 2018), visual attention analyses (Asteriadis, Karpouzis, & Kollias, 2014; Liu, Hantao, Heynderickx, & Ingrid, 2011), and virtual environments (Tran, Sen, Haut, Ali, & Hoque, 2020). Gaze directions and their changes are related to the thoughts or mental states of humans (Liu, Yu, Mora, & Odobez, 2021) and can be used to explore the intent and interest of subjects, which have contributed to the advancement of the study and development of eye gaze estimation. While early infrared-based methods typically require specialized devices, visible-light-based techniques applying eye appearance features and/or shapes to regress eye

* Corresponding author.
E-mail addresses: xinmwu@whu.edu.cn (X. Wu), lilin@whu.edu.cn (L. Li), hhzhu@whu.edu.cn (H. Zhu), 2014301130059@whu.edu.cn (G. Zhou), linfengl@hwtc.com.cn (L. Li), sufei21@sdjzu.edu.cn (F. Su), shenhe09@qq.com (S. He), yanggangw@hwtc.com.cn (Y. Wang), dragonme1@126.com (X. Long).

https://doi.org/10.1016/j.eswa.2023.122363
Received 9 May 2023; Received in revised form 26 October 2023; Accepted 26 October 2023
Available online 29 October 2023

gaze reduce the reliance on complex equipment to some extent. These visible-light-based techniques are commonly referred to as feature-based and appearance-based methods (Dan & Qiang, 2010; Sun & Pears, 2023). Feature-based methods can fulfil eye gaze estimation in which head poses and illumination are controlled (Lu, Sugano, Okabe, & Sato, 2014; Williams, Blake, & Cipolla, 2006). The latest appearance-based approaches applying convolutional neural networks (CNNs) have the potential to estimate eye gaze in daily environments with changing lighting conditions, free head poses and large amounts of appearance variation.

Compared to feature-based methods, appearance-based gaze estimation methods regress the mapping function from the eye or face appearance to a gaze and estimate the eye gaze from images captured by consumer cameras, which are usually less expensive and easier to operate. Furthermore, the large number of recently published datasets (Fischer, Chang, & Demiris, 2018; Kellnhofer, Recasens, Stent, Matusik, & Torralba, 2019; Krafka et al., 2016; Zhang, Sugano, Fritz, & Bulling, 2017) allows CNNs to learn appearance invariance from different objects, and the robustness is becoming stronger than that of feature-based methods (Lindén, Sjöstrand, & Proutiere, 2019). Despite the noticeable improvements, previous studies on appearance-based gaze estimation that apply only one or two individual eye images are not very effective in practical applications. To this end, Xucong et al. (2017) took full-face images as input to estimate eye gaze and generated a region importance map of their method; they proved that different facial regions contribute differently to improving gaze estimation performance. Other recent studies further show that information from both full-face images and two individual eye images can benefit gaze estimation performance, and many multiregion gaze estimation works have been proposed (Bao, Cheng, Liu, & Lu, 2020; Guo et al., 2019; He et al., 2019; Krafka et al., 2016). For example, Krafka et al. (2016) developed a multiregion CNN, named iTracker, that takes two individual eye images and full-face images as input to improve the eye gaze estimation performance. However, most multi-region CNN architectures are dedicated to exploring network structures and use the same ConvNet model on both eye and face images, overlooking the fact that different facial regions may contribute differently to gaze estimation performance (Xucong et al., 2017) and that the difference in the left- and right-eye appearance of one individual may also lead to inconsistent estimation results, which should be physically consistent (Cheng, Lu, & Zhang, 2018).

Inspired by attention mechanisms, which dynamically allocate network attention according to the importance of image regions (Guo et al., 2022), we propose a new appearance-based network architecture with attention mechanisms, named EG-Net, to overcome the defects mentioned above. The proposed network architecture applies a two-branch network to gaze estimation, adopting a base CNN for full-face images and an efficient eye network (EE-Net) scaling up from the base CNN for left- and right-eye images. Specifically, we adopt the compound model scaling approach, which has been shown to achieve superior results compared to arbitrary scaling of two or three dimensions, as demonstrated in EfficientNet (Tan & Le, 2019). Accordingly, for the EE-Net, we uniformly scale up the three dimensions (width, depth, and resolution) of the base CNN using a compound coefficient for each eye. Here, depth refers to the number of layers in a network, and width refers to the number of kernels per layer. With such an architecture, more fine-grained features can be captured from eye regions than other facial regions in gaze estimation. Furthermore, considering the inconsistent performance obtained when using two eye images in gaze estimation, EE-Net adaptively weights the left- and right-eye images via an attention network, named Attention-Net, according to the "image quality" differences. This implies that the EE-Net model may no longer treat the left- and right-eye equally instead by paying more attention to the "reliable eye" with higher "image quality". This strategy can address two-eye appearance variation cases that are caused by free head poses, illumination changes, eye occlusion and individual differences (Cheng et al., 2018). Additionally, several datasets with overall high quality and different configurations have been published for this task, i.e., MPIIFaceGaze (Zhang et al., 2017), GazeCapture (Krafka et al., 2016) Gaze360 (Kellnhofer et al., 2019), EyeDIAP(Funes Mora, Monay, & Odobez, 2014), RT-Gene(Fischer et al., 2018), ETH-XGaze(Xucong et al., 2020) and so on. They were recorded with variable lighting conditions and free head poses, and can be used to train the proposed end-to-end network. Based on those datasets, our proposed network takes left- and right-eye and face images as inputs, and explores the mapping relationship between images and the ground-truth eye gaze, so that it can be used to estimate the 2D gaze position and 3D gaze direction from other face images.

The contributions of this study can be summarized as follows. A two-branch network (including the base CNN and EE-Net) architecture, named the efficient gaze network (EG-Net), is proposed for appearance-based eye gaze estimation. This approach can effectively extract more fine-grained features within a fixed resource budget. In addition, an EE-Net model is designed to efficiently allocate network attention between two individual eye images according to their contributions to gaze estimation. It can deal with inconsistent gaze estimations of two eyes. Furthermore, the compound model scaling method, which uniformly scales up the depth, width and resolution of the base CNN for EE-Net, is first applied to the task of gaze estimation and performs better in this field. This method outperforms state-of-the-art approaches on MPIIFaceGaze datasets with errors of 2.41 cm and 2.76 degrees in 2D and 3D gaze estimation, respectively. It also improves the performance with 23.4 % and 9.8 % over prior arts on the GazeCapture and EyeDIAP datasets respectively, and has a better generalization ability when applied to other datasets with comparable accuracy.

2. Related work

We review the previous work considering two aspects that are related to EG-Net, i.e., similar tasks (gaze estimation) and related techniques (model scaling).

Gaze estimation approaches can generally be classified into two main classes according to whether they use infrared light or visible light (Cheng, Wang, Bao, & Lu, 2021; Kar & Corcoran, 2017). Infrared-based techniques are commonly referred to as pupil centre and cornea reflection (PCCR) methods. Traditionally, this has been performed via specific sensors, such as wearable eye trackers or headrests that shine infrared light into a user's eyes to record the reflection (Funes-Mora & Odobez, 2016; Ghiass, Arandjelovic, & Laurendeau, 2018; Ye, Li, Fathi, Han, & Rehg, 2012). Although PCCR methods are often used to obtain ground-truth values due to their high accuracy, these methods are still hardly accessible to users since they require specialized hardware and calibration. Consequently, the gaze estimation ability with fewer special equipment constraints is improved by computer vision and the emergence of various kinds of RGB cameras.

Visible-light-based gaze estimation methods are commonly referred to as feature-based (or shape-based) and appearance-based methods (Cheng et al., 2021; Dan & Qiang, 2010). Although appearance- and feature-based methods have relatively low accuracy compared to PCCR methods, they are not constrained by hardware for measuring gaze relative to a camera. Feature-based methods construct a 3D eye model from the anatomy of the human eye. The gaze direction can be calculated by the geometrical relationship between different facial features and eye features (facial landmarks, cornea reflections, pupil centres, etc.) (Kar & Corcoran, 2017). Feature-based methods have high accuracy and are currently widely used by many professional eye trackers. However, these methods require calibration for the parameters of the human eye, which may be inconvenient for different individuals. They fail to provide an available eye gaze estimation strategy for cases with free head poses and changing illumination. For these defects, recent studies have increasingly focused on appearance-based methods (Deng & Zhu, 2017; Lu, Wang, & Yen-wei, 2008; Yilmaz & Kose, 2016), as they are potentially more robust to low resolution images and have good

generalization performance.

Appearance-based gaze estimation methods directly take a human eye or full-face image as input and learn the mapping function from the eye appearance to the gaze direction or gaze location. Different mapping functions, such as artificial neural networks (ANNs) (Baluja & Pomerleau, 1994), random forest models (Yusuke, Yasuyuki, & Yoichi, 2014), Gaussian process regression (Sugano, Matsushita, & Sato, 2013), local interpolation (Lu, Sugano, Okabe, & Sato, 2012), multimodal models (Lu et al., 2008), support vector regression (SVR) (Wu, Yeh, Hung, & Tang, 2014), incremental learning (Sugano, Matsushita, Sato, & Koike, 2008) and neural networks (Krafka et al., 2016; Zhang, Sugano, Fritz, & Bulling, 2015), have been explored. Among these, deep learning methods for appearance-based gaze estimation have mostly been used in recent years because they can learn features automatically by CNNs from diverse and large volumes of image data (Kellnhofer et al., 2019; Park, Spurr, & Hilliges, 2018; Xucong et al., 2017). For example, Zhang et al. employed one model, like LetNet-5 (Zhang et al., 2015, 2017) and VGG-16 (Zhang et al., 2017) for both eye images and joined the head pose information with the extracted eye features to estimate the gaze. These approaches have shown a relatively lower level of accuracy. On the other hand, models that incorporate dual eyes in separate models (Ali & Kim, 2020; Cheng et al., 2018; Lemley, Kar, Drimbarean, & Corcoran, 2019) have demonstrated enhanced accuracy in predicting gaze estimations. And Lemley et al. (2019) substantiated that a two-channel architecture can enhance accuracy of around 37 % compared to the approach of utilizing a single network for processing both eyes. Furthermore, several scholars have dedicated extensive efforts to explore and have discovered a complementary relationship between the two separate eye images (Cheng et al., 2018; Liu, Yu, Mora, & Odobez, 2018; Liu et al., 2021). For example, Cheng et al. (2018) proposed AR-Net to predict the 3D gaze directions of two eyes and evaluation networks to adaptively evaluate the weight of each eye and adjust the regression strategy. This work achieved promising results and attracted more attention to gaze estimation using both left- and right-eye images. For such cases, other approaches have been proposed for person-independent gaze estimation (Lindén et al., 2019; Park et al., 2019). For instance, Liu et al. (2021) trained a differential network to predict the gaze difference between two eyes of the same object to help eliminate the person-specific bias and alleviate the impact of noise caused by illumination and variabilities in eye shape. In addition to eye gaze estimation from only eye images, other gaze estimation technologies combine both full-face and eye images using a multibranch network (Fischer et al., 2018; Kellnhofer et al., 2019; Krafka et al., 2016; Liu, Liu, Wang, & Lu, 2021; Zhang et al., 2017). Krafka et al. (2016) fed left- and right-eye images, a face image and a face grid related to the location and scale of the detected face into a network (iTracker) to regress the eye gaze. The experimental results showed that this kind of approach with full-face images performs better than the approach in Zhang et al. (2015) without full-face images. To improve the iTracker, system proposed by Kim et al. (2016) included an additional feature, namely, the histogram of gradients (HOG), along with the raw eye and face images. The HOG feature provides a CNN with more semantic information to effectively handle head motion. In summary, these results emphasize the significance of both face and eye images, which may improve the performance of eye gaze estimation considerably and greatly increase the applicability. Nonetheless, appearance-based eye gaze estimation remains a challenging task because various factors can significantly impact the eye and face appearance, such as lighting conditions, head poses, eye decorations, occlusion, and individual differences. To address these challenges, we propose an appearance-based network architecture that utilizes a two-branch network for gaze estimation, with one branch processing full-face images using a base CNN and the other branch using an EE-Net to process left-and right-eye images.

Model scaling. There are many methods for scaling ConvNet with the constraint of computing resources. For example, ResNet (He, Zhang, Ren, & Sun, 2016) can be more effective by scaling up its depth (# layer) from ResNet-18 to ResNet-200. Studies show that deeper networks have prominent advantages compared with shallow networks, while the indexes in deeper convolution networks require exponentially more computing resources than those in shallow networks (Larochelle, Erhan, Courville, Bergstra, & Bengio, 2007). More specifically, scaling up the network depth can derive richer and more complex features and improve performance even after thousands of layers are applied. Unfortunately, deep models can encounter the issue of diminishing feature reuse, potentially leading to slower training times (Zagoruyko & Komodakis, 2016). Additionally, Mobilenets take the depthwise separable convolution as its basic unit to scale up the network width (# channel). Increasing the width of a network can release the training load and capture more fine-grained features, with the drawback that shallow networks with larger widths can capture features only at a relatively low level. The network can capture more fine-grained patterns and has a wider receptive field with higher resolution input images (Huang et al., 2018), while its width and depth are scaled simultaneously. To summarize, the network dimensions, such as the depth, width and resolution, are usually independent; scaling up any single dimension can improve the accuracy to some extent (Lin & Jegelka, 2018; Wu, Shen, & van den Hengel, 2019), while accuracy may start to decline when the network becomes excessively large. Enhancing efficiency and accuracy is achievable by scaling up any of the two or three dimensions individually. Nevertheless, achieving a balance among all three dimensions using a compound coefficient within the allocated resource budget has shown superior performance (Tan & Le, 2019).

Therefore, we uniformly scale up the three dimensions (depth, width and resolution) of the base CNN by a set of fixed scaling coefficients for the EE-Net. Furthermore, we address the issue of inconsistent performance when using only two-eye images in gaze estimation by incorporating an attention network in the EE-Net. This attention network adaptively weights the left and right eye images based on their unobstructedness and lighting conditions, which ensures that the attentions allocated to each eye image are balanced. This approach enables the network to capture more fine-grained features from the eye regions than from other facial regions during gaze estimation.

3. Proposed approach

Although a number of eye gaze estimation approaches have been proposed in recent decades, there still exist some unsolved defects, such as two-eye gaze inconsistencies caused by free head poses, lighting conditions, eye occlusion or individual differences. Additionally, with the emergence of large-scale datasets with free head poses and changing illumination, an efficient network is needed to solve these existing problems. This section describes the appearance-based network architecture and presents the details of EG-Net. The implementation is demonstrated in the final subsection.

3.1. Approach overview

An appearance-based network architecture that applies a two-branch network to full-face images and eye images is proposed for the eye gaze estimation task. Specifically, we employ a base CNN for processing full-face images, emphasizing the extraction of global eye gaze features. In contrast, the EE-Net model is tailored for left- and right-eye images, focusing on the extraction of more detailed features.

A. Task definition

The proposed approach aims to regress the 2D gaze location or 3D gaze direction by learning features from a series of face and two individual eye images. The training set is defined as $D_t = \left\{ (I_i, g_i)|_{i=1}^{N_t} \right\}$, where $I_i \in \mathbb{R}^{H \times W}$ and $g_i \in \mathbb{R}^2$ denote the i-th training images and corresponding ground-truth of gaze vectors, respectively. i is the sample

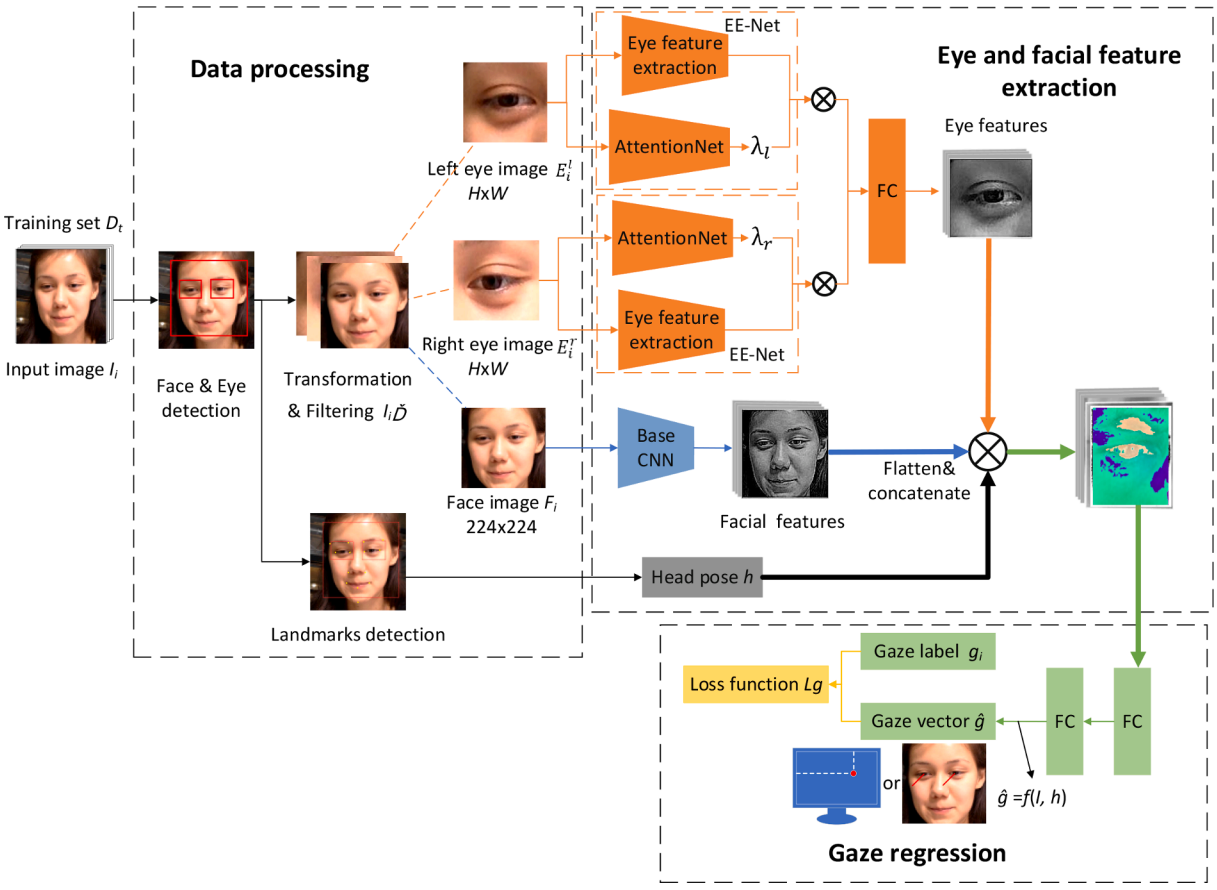


Fig. 1. Overview of the proposed appearance-based eye gaze estimation approach. It consists of three major parts: data processing, eye and facial feature extraction and gaze regression. Data processing for face and eye detection, data transformation and filtering. And two-branch network (base CNN and EE-Net) are applied to capture eye and facial features for eye and facial feature extraction. Finally, the gaze regression module is used to regress the 2D gaze vector.

index. N_t is the number of training images. g_i includes the coordinate values of $\times$ and y (in centimetres) in the screen coordinate system in 2D gaze location estimation, and it includes the pitch and yaw (in degrees) of the gaze direction for 3D gaze direction estimation. The training images I_i are processed into three groups of images, namely, left-eye images $E_i^l \in \mathbb{R}^{H^l \times W^l}$, right-eye images $E_i^r \in \mathbb{R}^{H^r \times W^r}$, and face images $F_i \in \mathbb{R}^{224\times224}$, which can be expressed as $I_i^{'} = \left\{ \left(E_i^l, E_i^r, F_i\right)\big|_{i=1}^{N_t} \right\}$. The validation set is defined as $D_v = \left\{ \left(I_i^v, g_i^v\right)\big|_{i=1}^{N_v} \right\}$, where $I_i^v \in \mathbb{R}^{H\times W}$ and $g_i^v \in \mathbb{R}^2$ denote the i-th validation images and corresponding ground truth of gaze vectors, respectively. N_v is the number of validation images.

The two eye images E_i^l and E_i^r and face images F_i are fed into the proposed network to explore the mapping function between the images and the gaze vectors. The final aim of our approach is to estimate gaze vector $\hat{g}$ from the corresponding unlabelled images I, i.e., $I \rightarrow \hat{g}$. And a mapping function can be formulated as a regression problem. Considering the head pose, the regression function can be expressed as:

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

where $h \in \mathbb{R}^3$ is the head pose vector, and f is the regression function from images to eye gaze. The regression function is usually nonlinear because of the complexity of the eye and facial appearances and the variability in human behaviours. It can be defined as a CNN.

B. Appearance-based eye gaze estimation approach

As shown in Fig. 1, the appearance-based eye gaze estimation approach consists of three major parts: data processing, eye and facial feature extraction and gaze regression. Specifically, the data processing performs face and eye detection, data transformations (crop and resize operations) and invalid image filters. First, for face detection, the multitask cascaded CNN (MTCNN) (Zhang, Zhang, Li, & Qiao, 2016) is optimized to detect and track human faces and left- and right-eye from input images or video frames. Then, the face and eye images are cropped from the original frames and further resized to the size that each ConvNet branch requires. Finally, these images are filtered to remove invalid images (images with closed eyes or some other disqualifying attribute). Meanwhile, since recent head pose estimation and facial landmark detection have already achieved ideal accuracy, a robust and accurate facial landmark detection (Chen, 2021) approach and a generic mean 3D face model are applied to derive the head pose of each of the frames. The face model contains the 3D positions of 14 facial landmarks (eyes, eyebrows, nose and mouth corners). The classical perspective-n-point algorithm is used to derive the 3D rotation matrix and translation vector from the 3D face model coordinate system to the camera coordinate system, and the head pose can be estimated. For eye and facial feature extraction, two networks, namely, the EE-Net and base CNN, are used to extract the eye features from both left- and right-eye images and facial features from face images. EE-Net is designed as a two-branch convolutional network, and it performs eye feature extraction for left- and right-eye images (detailed in 3.2 EG-Net). EE-Net uniformly scales up the depth, width and resolution of the base CNN with a set of constant coefficients for eye feature extraction. It also adaptively weights the left- and right-eye images via an attention network, named Attention-Net, according to the "image quality" affected by head pose, unobstructedness and lighting condition.

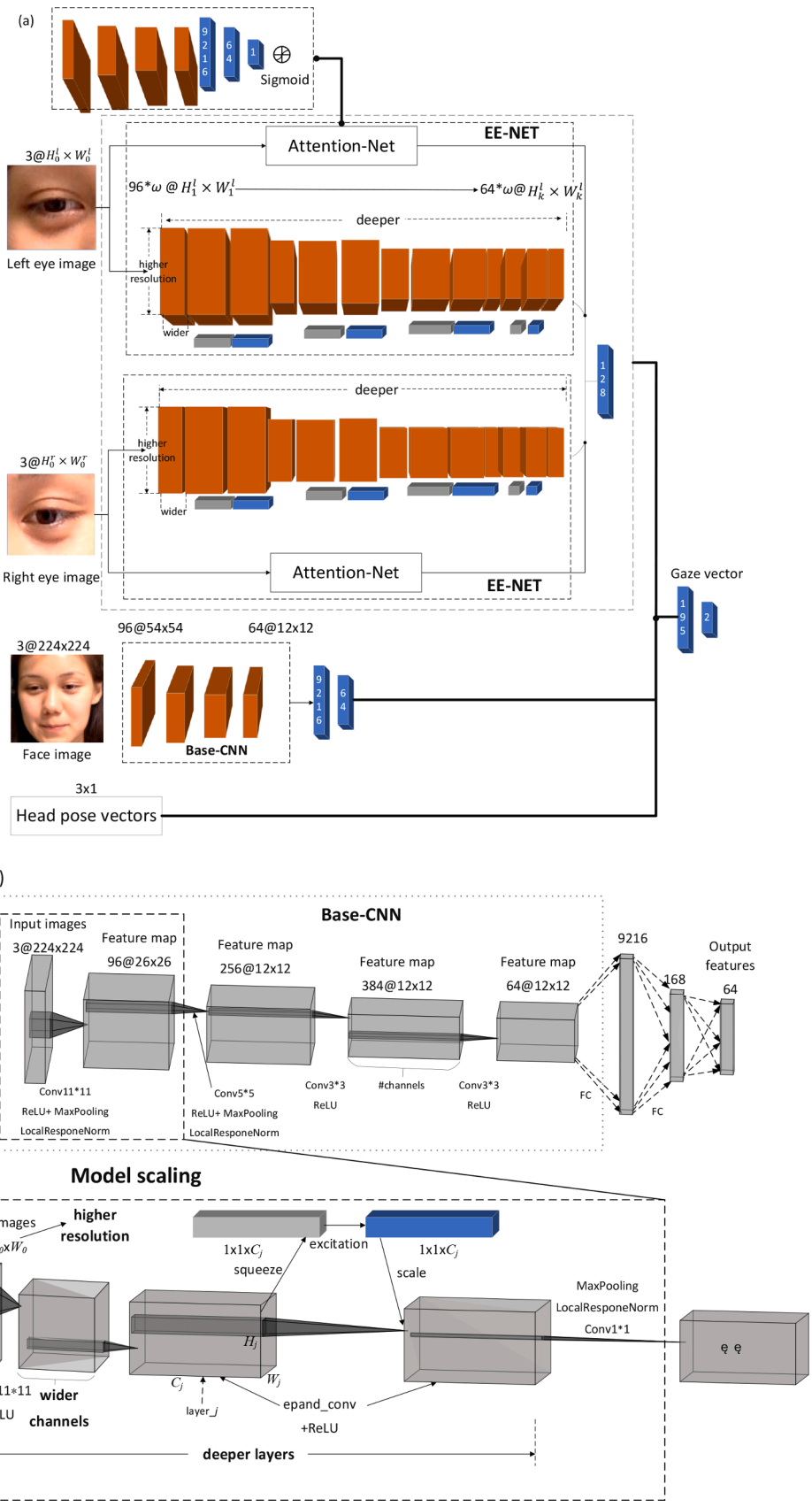


Fig. 2. Architecture of the proposed networks. (a) The whole network architecture. (b) The base CNN is used to extract facial features from face images, and it is the baseline network of EE-Net. (c) The detailed description of the compound model scaling process of the base CNN, which is the element of EE-Net model that takes the left- and right-eye images as inputs to extract eye features.

Meanwhile, the base CNN is used to perform facial feature extraction tasks. The loss function and other details are given in the corresponding section. Finally, features from the full-face and two individual eye images as well as head pose vectors are fused, and the last fully connected layers are used to regress the eye gaze vectors.

3.2. EG-Net

Eye gaze, as a significant nonverbal fine cue, plays an important role in exploring human behaviour. Eye appearance holds a variety of information related to gaze directions. To exploit this information, we design a network architecture named the efficient gaze network (EG-Net) that can make full use of large-scale datasets and capture richer and more complex eye features to estimate eye gaze.

A. Network architecture

We propose a convolutional neural network with attention mechanism for appearance-based eye gaze estimation. To capture more valuable features from both eyes and face images and address the asymmetric eye appearance issue, EG-Net endeavours to focus on different image region and weigh the left- and right-eye images according to their "image quality". The network architecture is illustrated in Fig. 2(a). It takes left eye image, right eye image, face image as well as head pose vector as inputs. Images are fed into two different branches of CNNs (EE-Net and base CNN), and are represented as some eye and facial feature maps.

For the extraction of eye features, networks scaled up from the base CNN are employed to extract unweighted features from each individual eye image. Subsequently, an Attention-Net, which has a similar architecture with base CNN, embedded in the EE-Net encodes each eye feature map as a weighted vector. The EE-Net determines the weight of each eye using the Attention-Net, accounting for its contribution to gaze estimation. The weighted features of the left and right eyes are then concatenated, followed by fully connected layers that further represent them as a comprehensive eye feature map.

In addition to the left- and right-eye feature representation, the feature maps of the whole face are encoded by the base CNN. Face images cropped from original frames are fed into an independent CNN (the base CNN) followed by two fully connected layers, which produce facial features with 64 dimensions. The so-called base CNN consists of four convolutional layers, all of which are followed by ReLU operations. Max pooling and local response normalization operators are applied after the first two ReLU operations to reduce the image dimensions and enhance the generalization ability of the model. The size and layers of the base CNN are similar to those of AlexNet (Krizhevsky, Sutskever, & Hinton, 2017), and its detailed parameters are shown in Fig. 2(b).

Finally, we train the model using the MSE loss:

$$L_g = \|\widehat{g} - g\|_2^2 \tag{2}$$

Note that although the face image already contains eye regions, the left- and right-eye images are still taken as individual inputs for EE-Net to extract more subtle changes from larger images, and the relatively simple base CNN is used to extract the overall facial features for gaze estimation. Furthermore, the EE-Net processes the two individual eye images independently, a choice justified by the potential inconsistencies in their contributions to gaze estimation. These inconsistencies may arise from factors like eye occlusions, variations in illumination, head poses, and individual differences. Additionally, it's important to note the complementary relationship between the two eyes, and the spatial arrangement of the eyes can have a certain impact on the accuracy of eye gaze estimation. The head pose vector (pitch, yaw, and roll) has some influence on gaze estimation. It has been concatenated with eye and face features by two fully connected layers to regress the eye gaze.

B. Compound model scaling

A convolutional layer j of a CNN can be defined as a function $Y_j = f_j(X_j)$, where f_j denotes the operator of the network, Y_j is the output tensor of that layer, $X_j \in \mathbb{R}^{H_j \times W_j \times C_j}$ is the input tensor, H_j and W_j denote the height and width of the input tensor, respectively, and C_j is the channel dimension. Therefore, the CNN can be expressed as a series of composed layers: $\mathbb{f} = f_k \odot f_{k-1} \cdots \odot f_2 \odot f_1(X_1) = \bigodot_{j=1\cdots k} f_j\left(X_{<H_j,W_j,C_j>}\right)$. In this paper, the base CNN $\mathbb{f}_b$ with four convolutional layers can be defined as:

$$\mathbb{f}_b = f_4 \odot f_3 \odot f_2 \odot f_1(X_1) = \bigodot_{j=1\cdots 4} f_j\left(X_{<H_j,W_j,C_j>}\right) \tag{3}$$

where $< H_j, W_j, C_j >$ denotes the shape of input tensor X of layer j (we defined $H_j = W_j$ in our proposed approach).

Unlike the regular eye gaze estimation approaches that tend to focus on optimizing operators f_j and apply the same ConvNet on both face and two-eye images, the EE-Net model scales up the three dimensions of the network width (C_j), depth (D_j), and resolution (H_j, W_j) by a series of compound coefficients without changing the predefined base CNN (as seen in Fig. 2(c)), which can help capture more fine-grained features from eye regions than other facial regions in gaze estimation. Therefore, the EE-Net can be defined as a nonlinear function relating to the scaling coefficients of the width, depth and resolution for the left-eye network ω_l, d_l, and r_l, as well as the scaling coefficients of the width, depth and resolution for the right-eye network ω_r, d_r, and r_r. Additionally, EE-Net can be expressed as:

$$\mathbb{f}_e(\omega_l, d_l, r_l, \omega_r, d_r, r_r) = \{\mathbb{f}_{el}(\omega_l, d_l, r_l), \mathbb{f}_{er}(\omega_r, d_r, r_r)\} \tag{4}$$

where $\mathbb{f}_e$ denotes the operator of EE-Net, which contains the left-eye operator $\mathbb{f}_{el}$ and the right operator $\mathbb{f}_{er}$. $\mathbb{f}_{el}$ and $\mathbb{f}_{er}$ can be further expressed as:

$$\mathbb{f}_{el}(\omega_l, d_l, r_l) = \bigodot_{j=0\cdots 4} f_j^{d_l}\left(X_{<r_l \bullet H_j, r_l \bullet W_j, \omega_l \bullet C_j>}\right) \tag{5}$$

$$\mathbb{f}_{er}(\omega_r, d_r, r_r) = \bigodot_{j=0\cdots 4} f_j^{d_r}\left(X_{<r_r \bullet H_j, r_r \bullet W_j, \omega_r \bullet C_j>}\right) \tag{6}$$

where $f_j^{d_l}$ and $f_j^{d_r}$ denote that operator f_j is repeated d_l and d_r times in stage j for the left and right eyes, respectively.

The network is designed to adjust the scaling coefficients of the two unweighted eye feature extraction branches of EE-Net and balance the resource distribution between the base CNN and EE-Net to minimize the gaze error for the given resource constraint. It can be formulated as an optimization problem:

$$\min_{\omega_l, d_l, r_l, \omega_r, d_r, r_r} error(\mathbb{f}_e(\omega_l, d_l, r_l, \omega_r, d_r, r_r)) \tag{7}$$

$$\text{Memory}(\mathbb{f}_e(\omega_l, d_l, r_l, \omega_r, d_r, r_r)) \leq \text{target\_memory} - \text{Memory}(\mathbb{f}_b) \tag{8}$$

$$\text{FLOPs}(\mathbb{f}_e(\omega_l, d_l, r_l, \omega_r, d_r, r_r)) \leq \text{target\_flops} - \text{FLOPs}(\mathbb{f}_b) \tag{9}$$

The main challenges of this problem in gaze estimation are that the optimized parameters for the left-eye image ω_l, d_l, and r_l as well as optimized parameters for the right-eye image ω_r, d_r, and r_r depend on each other and change with the given resource constraint. Although some methods tried to scale ConvNet in one of these three dimensions and achieved a relatively better performance, Tan and Le (2019) proved that the accuracy tends to diminish for larger models, and balancing all three dimensions can overcome this defect and achieve better accuracy and efficiency. To this end, we adopt the compound model scaling method (Tan & Le, 2019) and uniformly scale up the three dimensions of the network depth, width, and resolution by a compound coefficient for each eye within a fixed resource budget in the following principled way:

$$\begin{aligned} \text{depth}: d_l &= {\alpha_l}^{\mu_l},\ d_r = {\alpha_r}^{\mu_r} \\ \text{width}: \omega_l &= {\beta_l}^{\mu_l},\ \omega_r = {\beta_r}^{\mu_r} \\ \text{resolution}: r_l &= {\gamma_l}^{\mu_l},\ r_r = {\gamma_r}^{\mu_r} \end{aligned} \tag{10}$$

where μ_l and μ_r denote the compound scaling coefficients for the left-eye model and right-eye model, respectively, and are determined by how many resources are available for model scaling and how to allocate these extra resources to the two individual eye models. α_l, β_l, γ_l, α_r, β_r, and γ_r are constants that identify how to balance the depth, width and resolution components of the left-eye model and right-eye model, respectively.

The FLOPs of a regular convolution operation are proportional to d, ω^2, r^2. Theoretically, if the depth value is scaled by a factor d, FLOPs will increase by the same factor, but if the width is scaled by a factor ω or the resolution by a factor r, FLOPs will accordingly increase by a factor of ω^2 or r^2. Thus, scaling up EE-Net with equation (10) will increase FLOPs by $\left(\alpha_l \bullet {\beta_l}^2 \bullet {\gamma_l}^2\right)^{\mu_l} + \left(\alpha_r \bullet {\beta_r}^2 \bullet {\gamma_r}^2\right)^{\mu_r}$. Since EE-Net scales the same baseline (the base CNN) for both left- and right-eye images, the coefficients of network depth, width and resolution can be defined in the same way for both the left- and right-eye, i.e., $\alpha_l = \alpha_r = \alpha$, $\beta_l = \beta_r = \beta$, and $\gamma_l = \gamma_r = \gamma$. To simplify the calculation process, we refer to the principled method of EfficientNet (Tan & Le, 2019) and further constraints as follow:

$$\alpha_l \bullet {\beta_l}^2 \bullet {\gamma_l}^2 = \alpha_r \bullet {\beta_r}^2 \bullet {\gamma_r}^2 \approx 2$$

$$\alpha_l \geq 1, \beta_l \geq 1, \gamma_l \geq 1, \alpha_r \geq 1, \beta_r \geq 1, \gamma_r \geq 1 \tag{11}$$

Therefore, for any new device, when scaling up EE-Net with equation (10), the FLOPs will approximately increase by $2^{\mu_l} + 2^{\mu_r}$. The computer resources are initially evenly assigned between the left- and right-eye images, where $\mu_l = \mu_r$. In this way, users can specify $2^{\mu_l} + 2^{\mu_r}$ according to how many resources can be used for model scaling. It is defined as 2^{ε}, where ε denotes the overall scaling coefficient of the whole EE-Net model.

C. EE-Net

EE-Net takes left- and right-eye images as inputs. We develop EE-Net by scaling up the base CNN in all three dimensions, i.e., width (# channels), depth (# layers) and resolution (image size), with a set of coefficients, and embed an Attention-Net for weighting the eye feature map. Compared with the base CNN, EE-Net scaling up the network by the compound model scaling method has deeper convolutional layers, a wider network and a larger input image size. This network is designed to extract more subtle changes from left- and right-eye images.

The network attention is initially evenly assigned between the left- and right-eye images by the compound model scaling method. However, the gaze estimation performance of both separate eyes of a person may not always be consistent. Features from individual eye images may not always produce the same gaze values due to the asymmetrical "image quality" of two-eye images caused by factors such as illumination, free head poses, eye occlusion, or other instabilities and uncertainties. In practice, it is possible for either eye to be more "reliable" in gaze estimation or both eyes to be equally "reliable". Thus, inspired by an attention mechanism (Guo et al., 2022), which dynamically allocates computer resources according to the importance of image regions, it is better to assign more resources to the ideal eye image to extract more fine-grained features and fewer resources to the other eye image. To address the problem of this appearance asymmetry between the left- and right-eye images, Attention-Net is embedded in EE-Net to perceive the "image quality" and pay attention to the "reliable" and informative eye images.

The overall architecture of EE-Net is shown in Fig. 2(a). In each eye-specific EE-Net, the left- and right-eye images are processed in two branches. The first branch fed the eye image to some expanded convolutional layers accompanied by some squeeze and excitation blocks (Hu, Shen, Albanie, Sun, & Wu, 2020). Then, a convolutional layer for decreasing the width of the network and batch normalization is applied to the feature maps. The second branch consists of an Attention-Net that estimates a weight according to the gaze estimation performance to denote the importance of each eye image. Finally, the feature maps of the left and right eyes are weighted by the computed weights and combined into a 128-dimensional feature vector at the end of EE-Net.

Mathematically speaking, φ_l and φ_r are defined as the unweighted feature vector extracted from the cropped left- and right-eye images by the compound model scaling method. They can be expressed as follows according to equations (1), (5) and (6):

$$\varphi_l = \bigodot_{j=0\cdots4} f_j^{d_l}\left(E^l{}_{<r_l \bullet H_j, r_l \bullet W_j, \omega_l \bullet C_j>}\right) \tag{12}$$

$$\varphi_r = \bigodot_{j=0\cdots4} f_j^{d_r}\left(E^r{}_{<r_r \bullet H_j, r_r \bullet W_j, \omega_r \bullet C_j>}\right) \tag{13}$$

Attention-Net evaluates the left- and right-eye images and encodes images as weights to help adaptively adjust network attention between two-eye images. θ_l and θ_r are defined as the weights of the left- and right-eye images, respectively. They represent the "reliability" of the corresponding eye images. They can be expressed as:

$$\theta_l = \mathbb{f}_{al}\left(E^l{}_{<H,W,C>}\right) \tag{14}$$

$$\theta_r = \mathbb{f}_{ar}\left(E^r{}_{<H,W,C>}\right) \tag{15}$$

where $\mathbb{f}_{al}(\bullet)$ and $\mathbb{f}_{ar}(\bullet)$ denote the operations in Attention-Net for the left- and right-eye respectively, $E^l{}_{<H,W,C>}$ denotes the left-eye images, and $E^r{}_{<H,W,C>}$ denotes the right-eye images. The size and layers of Attention-Net are similar to those of the base CNN, consisting of four convolutional layers and ReLU operations, two max pooling and local response normalization operators, and sigmoid activation. The sigmoid activation restricts the θ_l and θ_r range to [0, 1], where 1 indicates the most ideal eye images and 0 indicates completely unavailable images.

The feature vectors φ_l and φ_r are then weighted by θ_l and θ_r, respectively, and the weighted features $\widehat{\varphi_l}$ and $\widehat{\varphi_r}$ can be expressed as follows:

$$\widehat{\varphi_l} = \theta_l \bullet \varphi_l \tag{16}$$

$$\widehat{\varphi_r} = \theta_r \bullet \varphi_r \tag{17}$$

With the attention mechanism embedded in EE-Net, each eye image is weighted according to its image quality. If better performance is achieved by using one of the eyes during the evaluation, the network will pay more attention to this "reliable" and informative eye and less attention to the "unreliable" eye.

3.3. Implementation of EE-Net

EE-Net is designed to uniformly scale up the depth, width, and resolution of the base CNN and embed Attention-Net to evaluate the weights of each input. Assume the training has N steps, the EE-Net is designed to be scaled up by a compound scaling coefficient ε. The computer resource is initially evenly assigned to each of the two eyes, i.e., $\overline{\mu_l} = \overline{\mu_r} = \varepsilon - 1$. The Attention-Net $\mathbb{f}_{al}$ and $\mathbb{f}_{ar}$ are applied to evaluate the overall image reliability of the left- and right-eye, respectively, and this information can further be used to determine the attention allocation strategies according to equations (16) and (17). The training process is divided into M stages. For each stage $0 \leq j \leq M-1$, the networks of the right eye $\mathbb{f}_{er}$ and left eye $\mathbb{f}_{el}$ are trained with image sizes S_i^l and S_i^r, widths C_{ij}^l and C_{ij}^r, and depths D_{ij}^l and D_{ij}^r, respectively. Due to the independence between the convolution layer weights and image sizes, at the beginning of each stage, all weights are inherited from the previous stage. The procedure of EE-Net with attention mechanism for eye feature

Table 1
Architecture of unweighted eye feature extraction branch in EE-Net-v6 compared with the base CNN. Each row describes a stage i with input image resolutions, output channels and stage repeat times. Conv represents convolution layers, and the expand_conv operator of EE-Net-v6 scales up the network channels and contains conv1 × 1, ReLU, squeeze and excitation, and depthwise conv3 × 3 followed by ReLU.

Stage i	Operators in base CNN	Resolution $H_i \times W_i$	# Channels C_i	Operators in EE-Net-v6	Resolution $\widehat{H_i} \times \widehat{W_i}$	# Channels $\widehat{C_i}$	# Repeats $\widehat{d_i}$
0	Conv11 × 11	224 × 224	96	Conv11 × 11	276 × 276	136	1
				expand_conv	67 × 67	816	
				Conv1 × 1, BN	67 × 67	360	
1	MaxPool	54 × 54	96	MaxPool	67 × 67	360	1
2	Conv5 × 5	26 × 26	256	Conv5 × 5	41 × 41	2160	2
				expand_conv	41 × 41	2160	
				Conv1 × 1, BN	41 × 41	540	
3	MaxPool	26 × 26	256	MaxPool	41 × 41	540	1
4	Conv3 × 3	12 × 12	384	Conv3 × 3	33 × 33	3240	2
				expand_conv	33 × 33	3240	
				Conv1 × 1, BN	33 × 33	92	
5	Conv1 × 1	12 × 12	64	Conv1 × 1	16 × 16	552	1
				expand_conv	16 × 16	552	
				Conv1 × 1, BN	16 × 16	92	

extraction is summarized in Algorithm 1.

Algorithm 1 Efficient eye network with attention mechanism for eye feature extraction.

Input: left eye image E^l with image size S_0^l, right eye image E^r with image size S_0^r and ground-truth gaze g
Input: depth coefficient α, width coefficient β, resolution coefficient γ and compound scaling coefficients ε
Input: Number of total training steps N
Initialize: $\mu_l = \mu_r = \varepsilon - 1$
for $i \leftarrow 0$ **to** $N-1$ **do**
  $\theta_l \leftarrow \mathbb{f}_{al}\left(E^l_{<S_0^l>}\right)$, $\theta_r \leftarrow \mathbb{f}_{ar}\left(E^r_{<S_0^r>}\right)$ refer equation (14) and (15)
  Image size: $S_i^l \leftarrow S_0^l \bullet \gamma^{\mu_i^l}$, $S_i^r \leftarrow S_0^r \bullet \gamma^{\mu_i^r}$
  for $j \leftarrow 0$ **to** M **do**
    Width: $C_{ij}^l \leftarrow C_j^l \bullet \beta^{\mu_i^l}$, $C_{ij}^r \leftarrow C_j^r \bullet \beta^{\mu_i^r}$ # C_j^l and C_j^r are the initial channels of base CNN in stage j for left- and right-eye images respectively
    Depth: $D_{ij}^l \leftarrow D_j^l \bullet \alpha^{\mu_i^l}$, $D_{ij}^r \leftarrow D_j^r \bullet \alpha^{\mu_i^r}$ # D_{ij}^l and D_{ij}^r are the repeat time of base CNN in stage j for left- and right-eye images respectively
    Train the left eye model for stage j: $f_j^{D_{ij}^l}(E^l_{<S_i^l,\ C_{ij}^l>})$ refer equation (5)
    Train the right eye model for stage j: $f_j^{D_{ij}^r}(E^r_{<S_i^r,\ C_{ij}^r>})$ refer equation (6)
  end for
  Update $\mathbb{f}_{el}$ and $\mathbb{f}_{er}$ with equation (5) and (6)
  Un-weighed feature vector:
  $\varphi_l = \underset{j=0\cdots4}{\odot} f_j^{D_i^l}\left(E^l_{<S_i^l,C_i^l>}\right)$, $\varphi_r = \underset{j=0\cdots4}{\odot} f_j^{D_i^r}\left(E^r_{<S_i^r,C_i^r>}\right)$
  weighed feature vector: $\widehat{\varphi_l} = \theta_l \bullet \varphi_l$, $\widehat{\varphi_r} = \theta_r \bullet \varphi_r$
  Update $\mathbb{f}_e(\bullet)$ with equation (4)
end for

The model is scaled by the following steps:

(1) Since the compound scaling coefficients ε are user-specified, we first set $\varepsilon = 4$ according to the memory capacity of our experimental platform, which is defined as EE-Net-v6. This means that approximately 2^4 more free resources are allocated to scale up the whole EE-Net model. Table 1 shows the architecture of unweighted eye feature extraction branch in EE-Net-v6.
(2) The depth coefficients α_l and α_r, width coefficients β_l and β_r, and resolution coefficients γ_l and γ_r for the left- and right-eye models can further be determined based on equations (8)–(10). Searching for α_l, β_l, γ_l and α_r, β_r, γ_r around the proposed model using the small grid search method (Tan & Le, 2019), the parameters are set as $\alpha_l = \alpha_r = 1.2$, $\beta_l = \beta_r = 1.1$, and $\gamma_l = \gamma_r = 1.07$ with the constraint in equation (11).
(3) Finally, the two-eye images are initially assumed to be equally reliable, and they can yield consistent estimation results. Therefore, μ_l and μ_r are equal to 3, which means that attention is evenly assigned to each eye image for model scaling.
(4) By fixing the parameters α_l, β_l, γ_l, α_r, β_r, and γ_r as a series of constants, the resource allocation can then be adaptively adjusted by changing the weights of the eye image and compound scaling coefficients ε.

Notably, within the given resource budget, EE-Net can be scaled up with different coefficients ε to develop its other family members from EE-Net-v0 to v6. As large models are prone to overfitting, the dropout ratios linearly increase from 0.2 for EE-Net-v0 to 0.5 for EE-Net-v6.

4. Experiments

In this section, the performance of the proposed EG-Net from v0 to v6 are evaluated on GazeCapture and MPIIFaceGaze for within-dataset evaluation, and GazeCapture, MPIIFaceGaze as well as EyeDIAP datasets for cross-dataset evaluation. Before showing the results, we first describe the experimental settings, including datasets, dataset processing, some training details and the evaluation metric. Then, experiments on real scenes are represented. Furthermore, the importance of some operators embedded in the proposed network is demonstrated by an ablation study. Finally, the proposed appearance-based gaze estimation approach is compared with state-of-the-art methods on MPIIFaceGaze, GazeCapture and EyeDIAP datasets.

4.1. Experimental setup

A. Datasets

MPIIFaceGaze dataset. The MPIIGaze dataset includes a standard subset for evaluation, comprising 1500 left eye images and 1500 right eye images independently selected from each participant. However, our method necessitates paired eye images captured simultaneously. Besides, we also conduct experiments using full face images as input. Consequently, we utilize the MPIIFaceGaze dataset as described in Zhang et al. (2017), which fills the gap by providing the missing face image and image of each left–right eye image pair from the original dataset. The dataset contains 213,659 face images from 15 participants and is widely used for gaze estimation tasks. For this dataset, leave-one-person-out cross validation is performed on the images for all 15 participants. The face images, left-eye images and right-eye images are cropped from the original images according to face bounding boxes and two-eye bounding boxes calculated from facial landmarks from the dataset labels, and resized into the image size required by the proposed model. The dataset provides the ground-truth of the head pose vector and gaze vector, which can be used to estimate the 3D gaze direction (in degrees). It also includes physical screen sizes and 2D gaze positions on

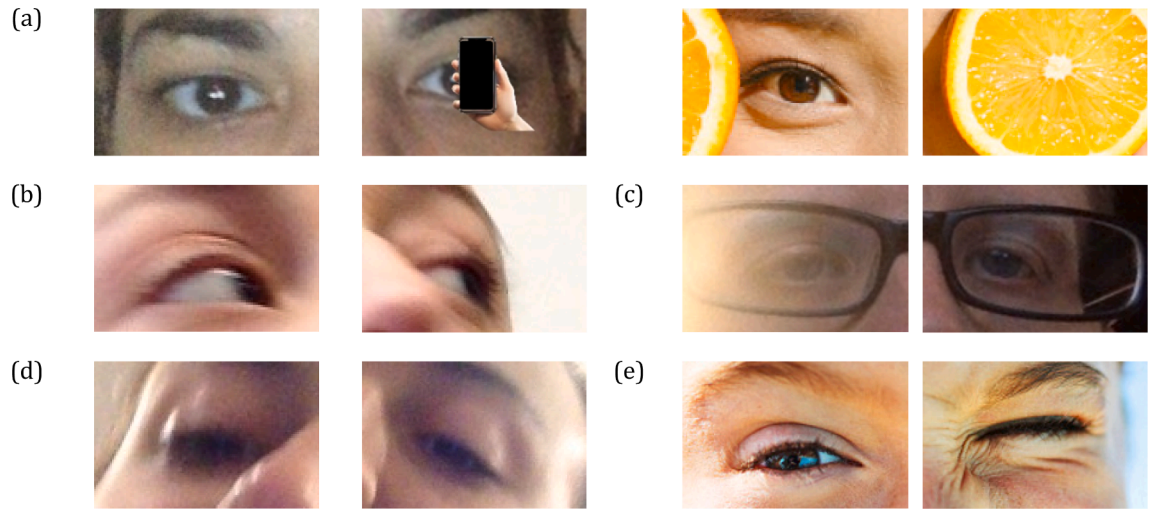


Fig. 3. Examples of the asymmetric eye images. (a) Example of the occluded eye images. (b) Example of eye image with free head pose. (c) Example of eye image with different illumination. (d) Example of blurred eye image. (e) Example of a closed eye image.

the screen coordinates in pixels, which can be applied to evaluate the 2D gaze position in centimetres relative to the screen.

GazeCapture dataset. To evaluate the cross-dataset generalization ability, the model is also trained and tested on the GazeCapture dataset published by Krafka et al. (2016). The dataset contains approximately 2.5 million frames collected from 1,471 participants and has a large variability in head pose, illumination and eye/face appearance, which are very helpful for training a robust and universal model. It contains the following information that we need:

(1) Frames of participants collected from a front-facing camera when the participants were performing a dot tracing task on a mobile phone or tablet.
(2) Bounding boxes around the detected face, left eye and right eye. The full-face images, left-eye images and right-eye images are cropped from the original frames according to the corresponding bounding boxes and then resized to the required image sizes.
(3) The ground-truth values of the gaze vectors (the gaze coordinates in the horizontal and vertical directions in centimetres relative to the camera centre).

We chose approximately 1.5 million valid frames from the complete dataset and ensured that at least one valid frame was selected for all 1,471 subjects. Then, the dataset is divided into a training set, validation set and test set containing images from 1,271, 50 and 150 subjects, respectively.

EyeDIAP dataset. The EyeDIAP (Funes Mora et al., 2014) dataset comprises sixteen participants who participated in three distinct scenarios: discrete screen targets (DS), continuous screen targets (CS), and a 3D floating target (FT). Throughout these scenarios, the participants' gaze was meticulously recorded and monitored using both remote RGB and RGB-D (standard vision and depth) cameras. To enhance the dataset's resilience against diverse head poses, participants were instructed to record two sets of videos for each visual scenario: one with a static head position and another with a mobile head position. For the comparative experiments conducted in this study, we utilize frames captured from the VGA camera component of the RGB-D camera. Note that, we apply the normalization into EyeDIAP following the approach proposed by MPIIFaceGaze (Xucong et al., 2017). And crop eye images from normalized face images with provided landmarks by the dataset.

B. Dataset processing

Appearance-based eye gaze estimation methods take eye or full-face images as inputs and learn the mapping function from eye appearance to gaze. These methods mainly rely on the image, while some real-life scenarios, such as eye occlusions, free head poses and illumination, may lead to differences in the left- and right-eye appearance. Finally, they result in inconsistent gaze estimation between the two eyes of one subject, which should be physically consistent.

To cope with these issues, we synthesized asymmetric eye images by randomly selecting one of two eyes and generating images with occlusion, distortion, blur, and illumination variation. Fig. 3 shows some synthesized asymmetric eye images. Occluded images are generated by manually placing some high-frequency obstructions, such as hair, hat, hand, cup, glasses, medical mask and fruit, in eye images as masks. In addition, distorted eye images are synthesized to deal with the appearance difference in left- and right-eye images with free head poses. There are illumination changes and blurriness when taking pictures in the wild, which also cause asymmetry of the appearance of the two eyes. Thus, we augment the dataset by randomly blurring one of two eyes or changing the brightness of eye images.

C. Training details

The proposed appearance-based eye gaze estimation approach is an end-to-end network that extracts full-face features from face images by base CNN and eye features from two individual eye images by EE-Net. All experiments are performed on a single NVIDIA RTX A6000 GPU with a total memory capacity of 51 GB. The model is implemented in the PyTorch deep learning framework. We initially use the SGD optimizer with a momentum of 0.9 and weight decay of 0.0005 for model training. The model is trained from scratch on the GazeCapture dataset for 18 epochs with 128 batches for each epoch, and the learning rate is set to 10e-3 and decays by 10 every 10 epochs. On the MPIIFaceGaze and EyeDIAP datasets, the model is trained for 18 epochs with 16 batches for each epoch, and the learning rate is set to 0.01 and decays to 0.001 after 10 epochs.

D. Evaluation metric

We report the gaze estimation performance on both the MPIIFaceGaze, GazeCapture and EyeDIAP datasets. For MPIIFaceGaze, the accuracy of the 3D gaze direction (yaw and pitch) in degrees and the 2D gaze position in centimetres relative to the screen are adopted as performance metrics. For GazeCapture dataset collected from mobile phone and tablet devices, we adopt the 2D gaze position in centimetres relative to the screen as performance metrics. And we also calculate the 3D gaze

Table 2
EG-Net-v0 to EG-Net-v6 performance results on the GazeCapture dataset, the MPIIFaceGaze dataset and its augmented dataset. The values in the first row were reproduced by using the method of iTracker (Krafka et al., 2016). And EG-Net denoted by* represents model with Attention-Net. Notably, we excluded the identical operations for face feature extraction and head pose regression from the number of parameters and GFLOPs. However, the values attributed to Attention-Net have been factored into their respective number of parameters and GFLOPs.

Model	Image resolution	Scaling coefficient ε	# params (million)	GFLOPs	Euclidean error (cm)	Angular error (degrees)	
						Original datasets	Augment datasets
iTracker	224 × 224	\	2.52	0.88	2.55	6.2	
EG-Net-v0	224 × 224	0	4.64	2.7	2.53	2.98	3.17
EG-Net-v0*	224 × 224	0	7.16	3.58	1.83	2.86	3.02
EG-Net-v1	231 × 231	1.5	11.28	4.82	1.75	2.93	3.10
EG-Net-v1*	231 × 231	1.5	13.8	5.7	1.72	2.82	2.97
EG-Net-v2	240 × 240	2	14.14	6.38	1.72	2.83	3.08
EG-Net-v2*	240 × 240	2	16.66	7.26	1.67	2.80	2.95
EG-Net-v3	248 × 248	2.5	15.52	7.6	1.68	2.82	2.93
EG-Net-v3*	248 × 248	2.5	18.04	8.48	1.66	2.79	2.80
EG-Net-v4	257 × 257	3	16.74	8.7	1.60	2.80	2.91
EG-Net-v4*	257 × 257	3	19.26	9.58	1.58	2.76	2.78
EG-Net-v5	267 × 267	3.5	19.88	11.48	1.82	2.88	2.94
EG-Net-v5*	267 × 267	3.5	22.4	12.36	1.78	2.81	2.81
EG-Net-v6	276 × 276	4	22.74	14.68	1.88	2.96	3.00
EG-Net-v6*	276 × 276	4	25.98	15.55	1.80	2.84	2.85

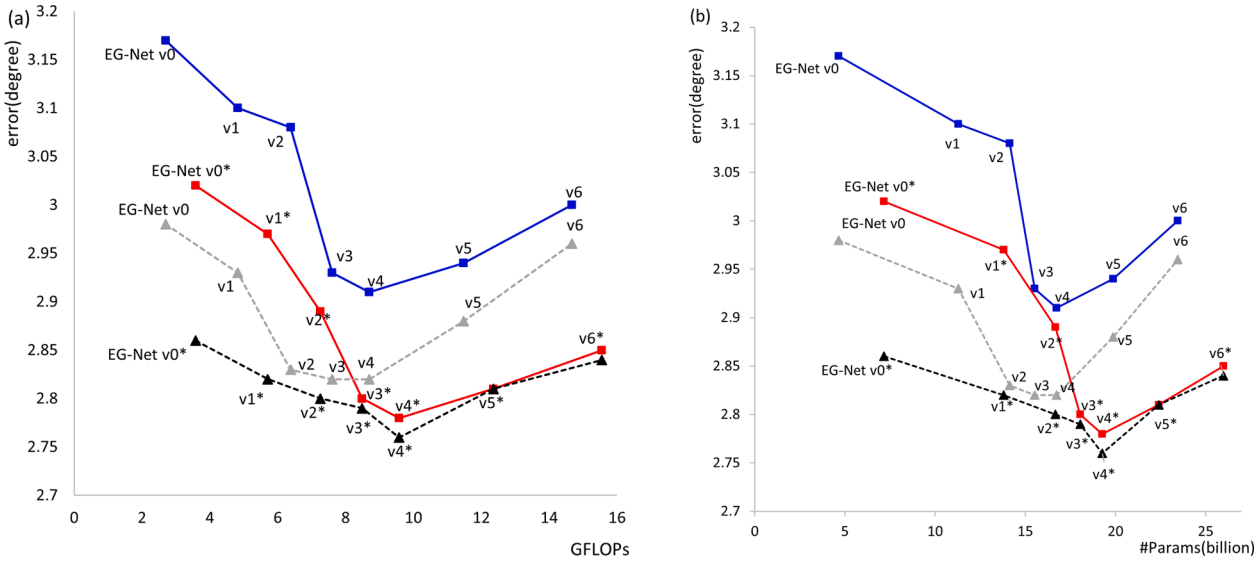


Fig. 4. Gaze estimation errors of EG-Net-v0 to v6 on the original MPIIFaceGaze dataset and its augmented dataset in centimetres (Euclidean error). EG-Net with * represents Attention-Net applied according to the image qualities. The solid line represents the eye gaze estimation results for the augmented MPIIFaceGaze dataset. The dotted line represents the eye gaze estimation results for the original MPIIFaceGaze dataset. (a) GFLOPs vs. eye gaze estimation errors. (b) Parameter numbers vs. eye gaze estimation errors.

direction (yaw and pitch) according to the physical screen sizes provided by this dataset for the cross-dataset evaluation. For EyeDIAP dataset, the accuracy of the 3D gaze direction (yaw and pitch) in degrees are adopted as performance metrics. Additionally, since some methods do not make source code available, we compare the gaze estimation accuracy with some state-of-the-art methods and expose the number of parameters and FLOPs involved in EG-Net for later comparison.

4.2. EG-Net evaluation

A. Within-dataset evaluation

To demonstrate the efficiency and effectiveness of the proposed method, we conduct experiments to evaluate the number of parameters, gigaFLOPs (GFLOPs) and gaze estimation errors of 2D gaze location and 3D gaze direction for EG-Net from v0 to v6. All of the EG-Net family models take a face image, two individual eye images and head pose as inputs. We use the base CNN to extract features from face images and the EE-Net family models to extract features from two individual eye images. The EE-Net family models are scaled up from the base CNN using different compound efficiencies ε. We train the models on the GazeCapture dataset, MPIIFaceGaze dataset and its augmented dataset using the settings illustrated in section 3.3 and evaluate the model on the validation set. The experimental results are summarized in Table 2, and models with similar scaling coefficients ε are grouped together in the table for an intuitive accuracy comparison. Furthermore, the GFLOPs error curves and parameter error curves of our EG-Net-v0 to v6 models are plotted in Fig. 4(a) and (b) for comparison, respectively.

As seen in Table 2, the EG-Net scaled-up iTracker model for the eye feature extraction task with different coefficients from v0 to v6 perform better in all three datasets, with errors of 1.58 ~ 2.53 cm on the GazeCapture dataset, 2.76 ~ 2.98 degrees on the MPIIFaceGaze dataset, and 2.78 ~ 3.17 degrees on its augmented dataset. EG-Net-v4 with a scaling coefficient of 3, that is, $\mu_l = \mu_r = 2$, achieved the best performance

Table 3
Gaze estimation errors in degrees on cross-data evaluation of the EG-Net-v4. All numbers are gaze estimation error in degrees.

Test Train	GazeCapture	MPIIFaceGaze	EyeDIAP
GazeCapture	–	4.37	9.53
MPIIFaceGaze	5.53	–	11.02
EyeDIAP	12.01	9.98	–

compared with other EG-Net family models, with an error of 1.58 cm on GazeCapture and 2.76 degrees on MPIIFaceGaze. In addition, considering the difference between the left- and right-eye appearances, we attempt to reallocate resources for the feature extraction task between the two eyes by Attention-Net. Experimental results show that the EG-Net models with Attention-Net have consistently higher accuracy than models without Attention-Net. EG-Net-v4 without Attention-Net has a 2D Euclidean error of 1.6 cm and a 3D angular error of 2.8 degrees on the two datasets, which are approximately 1.27 % and 1.43 % larger than its counterparts of EG-Net-v4, respectively.

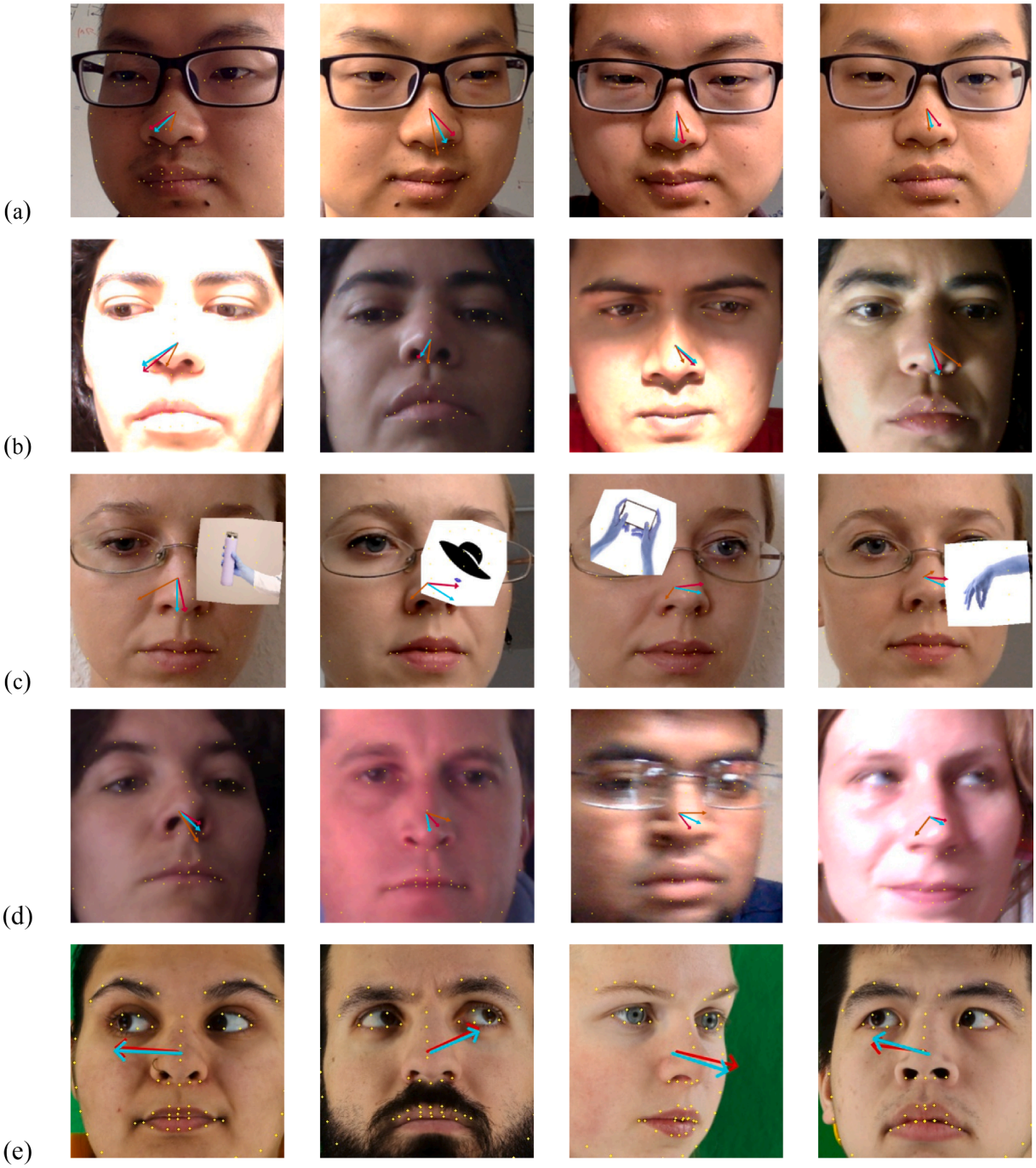


Fig. 5. Gaze estimation results of the EG-Net-v4 and iTracker models on real scenes. The red arrow is the ground-truth value of the eye gaze. The cyan arrow is the eye gaze estimated by EG-Net-v4. The orange arrow is the eye gaze estimated by iTracker. (a) Gaze estimation performance with different head poses. (b) Gaze estimation performance with changing illumination. (c) Gaze estimation performance with eye occlusion. (d) Gaze estimation performance for blurred images. (e) Gaze estimation performance of different individuals. Noted that, the last column is trained on the MPIIFaceGaze dataset and verified on the ETH_XGaze dataset, whereas the other columns involve training and testing on the MPIIFaceGaze dataset with a leave-one-person-out settings.

Fig. 4 demonstrates the gaze estimation errors of EG-Net-v0 to v6 on the original MPIIFaceGaze dataset and its augmented dataset. As seen in Fig. 4(a) and (b), the gaze estimation errors initially decrease as the compound scaling coefficients increase. The gaze estimation error of the model achieves the minimum value and starts to increase from EG-Net-v4, where the scaling coefficient ε equals 3. These results imply that the improved performance is not caused by an increase in the number of parameters or GFLOPs. When the model is scaled up to a certain size, the gaze estimation accuracy will no longer be improved. The EG-Net-v4 model achieves a 3D gaze direction estimation error of 2.76 degrees on the MPIIFaceGaze dataset. Although MPIIFaceGaze is augmented by randomly blurring one of two eyes, changing the brightness of eye images or occluding one of two eyes, EG-Net-v4 achieves an error of 2.78 degrees, which is lower than its other counterparts. Furthermore, the Attention-Net designed to make the model pay more attention to the eye appearance with high "image quality" outperforms the EG-Net without Attention-Net on two datasets. Although the EG-Net with Attention-Net did not perform well at first on the original dataset, the error gap of EG-Net on MPIIFaceGaze and its augmented dataset become narrow from EG-Net v3 to v6, while it always maintains a larger value on EG-Net without Attention-Net. In addition to that, the whole architecture of the EG-Net-v4 with parameters of around 21.78 million outperform the AGE-Net (L R D & Biswas, 2021) by 32.5 % (4.09 degrees) with ~105 million parameters on MPIIFaceGaze. This result demonstrates that we can improve gaze error with less number of parameters. Such smaller models with less memory or GFLOPs can be useful for achieving less latency and less stringent requirements of high end GPUs for real-time gaze estimation.

B. Cross-dataset evaluation

Cross-dataset evaluation is essential as it can indicate the generalization ability of the model. This task is quite challenging, given the notable disparities in face and eye appearance, image resolution, illumination, and head pose across various datasets. To further validate the generalization ability of the features learned by EG-Net-v4, we define the cross-dataset evaluation as training the model on GazeCapture, MPIIFaceGaze and EyeDIAP datasets and testing on these three datasets. The cross-dataset evaluation results are shown in Table 3.

As can be seen from Table 3, training on GazeCapture dataset and testing on two datasets have better performance than training on other two datasets. Additionally, training on this dataset and testing on MPIIFaceGaze even outperform the current best performing method in a similar cross-dataset evaluation (Xucong et al., 2020) (with error of 4.5 degrees, 2.89 % improvement). Training our EG-Net-v4 on MPIIFaceGaze and testing on GazeCapture also achieve a higher accuracy than on EyeDIAP. Additionally, we also train the model on both three datasets and test on RT-Gene (Fischer et al., 2018), with angular errors of 13.54 degrees, 14.75 degrees and 13.97 degrees. Despite lower overall ranking when comparing with other cross-dataset ranking of ours, they achieve around 6.5 % improvement over instead the EE-Net by ResNet-50 (Xucong et al., 2020). In summary, these cross-data evaluation results demonstrate the generalization ability of our proposed EG-Net.

4.3. Experiments on real scenes

There are many factors, such as head poses, illumination, glasses, blurring, eye occlusion, and even individual differences, that may affect the accuracy of eye gaze estimation. In order to verify the effectiveness of our model in these scenarios, we tested gaze estimation results of EG-Net-v4 and compare it with iTracker model. The gaze estimation results are shown in Fig. 5. Compared with iTracker, EG-Net-v4 achieves better eye gaze estimation performance. It is because the compound model scaling strategy in EE-Net enable the model to capture more subtle features from two eye images, and the Attention-Net in EG-Net-v4 enable the model to focus on the eye with higher "image quality".

Table 4
The test errors of gaze estimation on the MPIIFaceGaze and EyeDIAP datasets for the ablation study of EG-Net-v6 components.

Model	MPIIFaceGaze		EyeDIAP
	2D gaze location (cm)	3D gaze direction (degrees)	3D gaze direction (degrees)
Baseline model (EG-Net-v0)	2.77	3.17	5.76
Depth Scaling ($d_l = d_r = 1.4$)	2.73	3.12	5.58
Width Scaling ($\omega_l = \omega_r = 1.4$)	2.69	3.08	5.66
Resolution Scaling ($r_l = r_r = 1.4$)	2.69	3.09	5.32
EG-Net-v6 (no resolution scaling)	2.68	3.06	5.29
EG-Net-v6 (no width scaling)	2.66	3.03	5.18
EG-Net-v6 (no depth scaling)	2.66	3.04	5.24
EG-Net-v6 (no Attention-Net)	2.61	2.96	5.13
EG-Net-v6	**2.48**	**2.84**	**4.84**

Note: Bold values indicate the best results.

Fig. 5(a) reports the effect of head pose on gaze estimation. EG-Net-v4 outperforms iTracker when the object has a large head pose (first and second columns in Fig. 5(a)). The performance gap between EG-Net-v4 and iTracker becomes narrow when the subject faces the camera directly (third and fourth column in Fig. 5(a)). As shown in Fig. 5(b), EG-Net-v4 performs well under strong light (first column), low light (second column) and partial uneven light environments (third column). The objects in the fourth column of Fig. 5(b) have different appearances between the left- and right-eye because of the light coming from the right side of the object. The gaze estimated by EG-Net-v4 is still close to the ground-truth value, while the error of iTracker becomes relatively large. Furthermore, compared with iTracker, EG-Net-v4 performs relatively better on images with occluded eyes (Fig. 5(c)). This is because the model with the attention mechanism can better capture subtle features from the non-occluded eye appearance. Fig. 5(d) illustrates some eye gaze estimation results of blurring examples. Although the EG-Net-v4 eye gaze estimation results for blurred images are not as ideal as those for clear images, it still has an approximately 43 % accuracy improvement compared with iTracker. As shown in Fig. 5(e), we train our EG-Net-v4 model on MPIIFaceGaze dataset and test it on the ETH-XGaze (Xucong et al., 2020). The experimental results demonstrate that our model exhibits robustness when it comes to estimating the eye gaze of different individuals.

4.4. Ablation study

We use the MPIIFaceGaze and EyeDIAP datasets to demonstrate the significance of different components of the proposed gaze estimation method in ablation experiments. The ablation study mainly focuses on the scaling up operations of depth/width/resolution, the Attention-Net, as well as the inputs, namely, head pose vectors, face image and two-eye images.

A. Ablation study of compound model scaling and Attention-Net

Tan and Le (2019) proved that scaling up any dimension of the network can improve accuracy to some extent, but the accuracy diminished when the scaling coefficient increased to a certain value; and that balancing all dimensions of network depth, width and resolution in a principled way can achieve better accuracy and efficiency. Based on these observations, Table 4 further summarizes the experimental results obtained by comparing the EG-Net-v6 model with EG-Net-v0, the base CNN scaled by a single dimension with a factor of 1.4 for both left- and

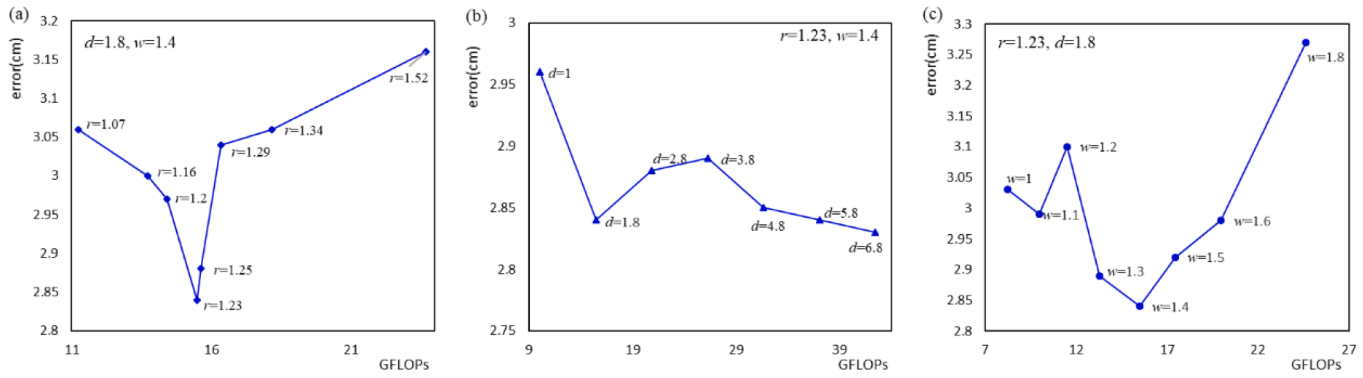


Fig. 6. Scaling up EE-Net-v6 with different network resolution(r), depth (d) and width (w) coefficients. Larger networks with larger width, depth, or resolution tend to achieve higher accuracy at first, but the accuracy gain quickly saturates. The architecture of unweighted eye feature extraction branch in EE-Net-v6 is described in Table 1.

right-eye images, EG-Net-v6 without depth scaling, EG-Net-v6 without width scaling, and EG-Net-v6 without resolution scaling, EG-Net-v6 without Attention-Net. For example, resolution scaling with $r_l = r_r = 1.4$ means that the image sizes of two individual eyes are scaled up to 310 × 310. EG-Net-v6 without resolution scaling means that the resolution scaling operation is ablated by using eye images with a size of 224 × 224 as input and keeping the other components the same. Additionally, errors in each of the rows are reported when the resources between the left- and right-eye images are equally assigned, except for in the last row, which depicts complete EG-Net-v6 embedded with Attention-Net.

As shown in Table 4, scaling up any single dimension of the network depth, width or resolution by 1.4 contributes to the error reduction, with 1.6 %, 2.8 % and 2.5 % accuracy improvements on MPIIFaceGaze dataset, respectively, and 3.1 %, 1.7 % and 7.6 % accuracy improvements on EyeDIAP dataset, respectively, while EG-Net-v0 without any scaling up operations achieves errors of 2.77 cm and 3.17 degrees for 2D and 3D gaze estimation on MPIIFaceGaze dataset, respectively, and 5.76 degrees on EyeDIAP dataset. EG-Net-v6 with Attention-Net achieves the highest accuracy with a 2D Euclidean error of 2.48 cm and a 3D angular error of 2.84 degrees on MPIIFaceGaze and 4.84 degrees on EyeDIAP. In addition, when compared with EG-Net-v6 without Attention-Net with errors of 2.96 degrees and 4.84 degrees on MPIIFaceGaze and EyeDIAP datasets, respectively, the other three methods without one of the scaling operations have significantly lower accuracy. This is because the three dimensions are not independent; for example, if the input image size is larger, more network layers can increase the receptive field, and more channels are needed to capture fine-grained patterns on the larger input image. Ablating Attention-Net also increases the model error by approximately 4.5 % for 2D gaze estimation and 3D gaze estimation on MPIIFaceGaze dataset and 7.46 % on EyeDIAP dataset.

We also compare resolution scaling, depth scaling and width scaling under the EG-Net-v6 architecture on MPIIFaceGaze dataset, as shown in Fig. 6. With higher resolution input images, models may capture more fine-grained patterns. Fig. 6(a) illustrates the results of scaling network resolutions ($r = 1$ denotes resolution 224 × 224, $r = 1.07$ denotes resolution 240 × 240, and $r = 1.52$ denotes resolution 340 × 340). The gaze estimation gains more accurate result at first, but diminishes for very high resolutions. Same as resolution scaling, scaling the width can also reduce the error, but when the number of channels continues to increase, the error starts to increase again (Fig. 6(c)). As is shown in Fig. 6 (b), if we only scale network depth d without changing resolution ($r = 1.23$) and width ($w = 1.4$), the error decreases to 2.84 degrees quickly and then becomes saturated. With deeper layers ($d = 6.8$), depth scaling achieves almost the same accuracy as EG-Net-v6; unfortunately, it also consumes more FLOPs.

B. Ablation study of inputs

Table 5
The test errors of gaze estimation on MPIIFaceGaze and EyeDIAP datasets for the ablation study of inputs.

2Model	MPIIFaceGaze	EyeDIAP	
	2D gaze location (cm)	3D gaze direction (degrees)	3D gaze direction (degrees)
EG-Net-v6 (without head pose)	3.06	3.50	4.97
EG-Net-v6 (without face image)	4.67	5.34	9.41
EG-Net-v6 (without two eye images)	2.94	3.38	5.02
EG-Net-v6 (without left-eye image)	2.92	3.34	4.95
EG-Net-v6 (without right-eye image)	2.94	3.36	4.97
EG-Net-v6	**2.48**	**2.84**	**4.84**

Note: Bold values indicate the best results.

For a comparison, the ablating images such as face image, two-eye images, left-eye image and right-eye image are replaced by blank images with a pixel value of 255, and EG-Net-v6 without head pose takes [0, 0, 0] as its head pose vector. Table 5 summarizes the gaze estimation results of the ablation study of inputs on MPIIFaceGaze and EyeDIAP dataset. EG-Net-v6 gains lower error than other networks that ablate the head pose, face image, two-eye images, and left- and right-eye images. The full-face image plays the most important role in gaze estimation with approximately 45 % error reduction on MPIIFaceGaze and 48.6 % on EyeDIAP dataset. Head pose also makes great sense in gaze estimation with 0.5 cm 2D gaze location error reduction, and 0.57 degrees and 0.13 degrees in its 3D gaze direction on MPIIFaceGaze and EyeDIAP datasets, respectively. Meanwhile, ablating one of the two-eye images or both eyes will increase the gaze estimation error, with around 15 % and 3.6 % error reduction for two-eye image ablations on MPIIFaceGaze and EyeDIAP datasets, respectively. While ablating one of the two-eye images can result in higher accuracy compared to ablating both two-eye images simultaneously.

4.5. Comparison with State-of-the-Art methods

We evaluate EG-Net on the three datasets, MPIIFaceGaze, GazeCapture and EyeDIAP, and compare it with the state-of-the-art methods on these three datasets.

A. Performance comparison on the MPIIFaceGaze dataset

First, the model is applied to the MPIIFaceGaze dataset and

Table 6
Error for 2D gaze estimation in centimetres (Euclidean error) and 3D gaze estimation in degrees (angular error) compared with state-of-the-art methods on the MPIIFaceGaze dataset.

Model	2D gaze location (cm)	3D gaze direction (degrees)
iTracker (Krafka et al., 2016)	4.57	6.2
ARE-Net (Cheng et al., 2018)	4.29	4.9
Spatial weight CNNs (Xucong et al., 2017)	4.2	4.8
RT-GENE (Fischer et al., 2018)	4.2	4.8
AFF-Net (Bao et al., 2020)	3.9	4.4
L2CS-Net (Abdelrahman et al., 2022)	–	3.92
Ali & Kim (2020)	–	2.80
Wang et al. (2023)	–	2.76
EG-Net-v4 (ours)	**2.41**	**2.76**

Note: The method proposed by Wang et al. (2023) and Ali & Kim (2020) are performed on MPIIGaze dataset.

compared with other methods that have already shown their good performance in gaze estimation tasks. The approach proposed by L2CS-Net (Abdelrahman, Hempel, Khalifa, & Al-Hamadi, 2022) which achieved the state-of-the-art performance on the MPIIFaceGaze dataset is included as a comparison method. The results are shown in Table 6 and Fig. 7. EG-Net-v4 achieves a Euclidean error of 2.41 cm for 2D gaze location estimation and an angular error of 2.76 degrees for 3D gaze direction estimation. The error gaps between iTracker and EG-Net-v4 are 2.16 cm and 3.44 degrees, respectively. Comparing with the 3.92 degrees error attained by the most recent L2CS-Net (Abdelrahman et al., 2022), EG-Net-v4 decrease the 3D angular error by 29.6 %.

Additionally, we incorporate the methodologies presented by Wang et al. (2023) and Ali & Kim (2020) as comparative methods due to their demonstrated state-of-the-art performance in gaze estimation using the MPIIGaze dataset (as is shown in Table 6). In comparison to the method proposed by Ali and Kim (2020) accuracy with a 2.8-degree error in 3D gaze estimation, our proposed EG-Net-v4, which exhibits an approximate 1.43 % reduction in 3D angular error, yields better performance. Wang et al. (2023) achieves comparable accuracy to EG-Net-v4. Nevertheless, the training samples used by Wang et al. (2023) for their model align with the testing samples, whereas our EG-Net-v4 exhibits an error of 2.76 degrees in a leave-one-out strategy to ensure that the experiments are done in a fully person-independent manner. This different setting suggests that our model have more robustness across various individuals.

B. Performance comparison on the GazeCapture dataset

To further evaluate the generalization ability and performance of the proposed approach, we compare EG-Net-v4 with five methods: iTracker (Krafka et al., 2016), SD (Yang, Xie, Su, & Yuille, 2019), SAGE (He et al., 2019), TAT (Guo et al., 2019) and AFF-Net (Bao et al., 2020), on GazeCapture. Table 7 and Fig. 8 summarize the comparison results for 2D gaze estimation on mobile phones and tablets. As seen from Table 7, EG-Net-v4 achieves lower gaze estimation error than these methods on GazeCapture with errors of 1.28 cm on images from mobile phones and 1.71 cm on images from tablets. For mobile phones, the errors decrease to a similar level by SD, SAGE and TAT in 2019, with an approximately 4.8 % improvement from iTracker (error of 1.86 cm), while AFF-Net significantly improves the accuracy by 12.9 %. Moreover, for the gaze estimation task on tablet-captured images, these methods achieve consistently higher errors than their mobile phone-captured counterparts, as shown in Fig. 8. AFF-Net made great progress in reducing the error by 18.1 % compared with the iTracker error. To the best of our knowledge, AFF-Net achieves state-of-the-art approaches on GazeCapture with errors of 1.62 cm and 2.30 cm for mobile phone- and tablet-captured images, respectively. Our EG-Net-v4 model scales up iTracker and achieves a statistically significant performance improvement of 21 % on mobile phone-captured images and 25.7 % on tablet-captured images over AFF-Net. These results demonstrate that EG-Net-v4 has better gaze estimation performance on both datasets compared to previous state-of-the-art methods.

C. Performance comparison on the EyeDIAP dataset

For the evaluation of the EyeDIAP dataset, the screen target sessions, as elaborated in Section 4.1-A, are employed. And one image of per 15 frames from four VGA videos of each participant is selected for training and testing. As two subjects lacked videos in the screen target session, we ultimately obtain images from 14 subjects. From this dataset, we crop the left and right eye images according to head pose and eye centres annotations provided by the dataset and normalized them by the same way as MPIIFaceGaze dataset. Noted that, for this dataset, we also apply the leave-one-person-out strategy to obtain robust results. Consequently, the dataset is divided into a total of 1181*13 frames for training and 1181 frames for testing.

EG-Net-v4 is trained and evaluated on the EyeDIAP dataset and subjected to comparison with other methods that have already demonstrated strong performance in gaze estimation tasks: iTracker (Krafka

Table 7
Gaze estimation error in centimetres (Euclidean error) compared with state-of-the-art methods on the GazeCapture dataset.

Model	Mobile phone	Tablet
iTracker (Krafka et al., 2016)	1.86	2.81
SD (Yang et al., 2019)	1.81	2.61
SAGE (He et al., 2019)	1.78	2.72
TAT (Guo et al., 2019)	1.77	2.66
AFF-Net (Bao et al., 2020)	1.62	2.30
EG-Net-v4 (ours)	**1.28**	**1.71**

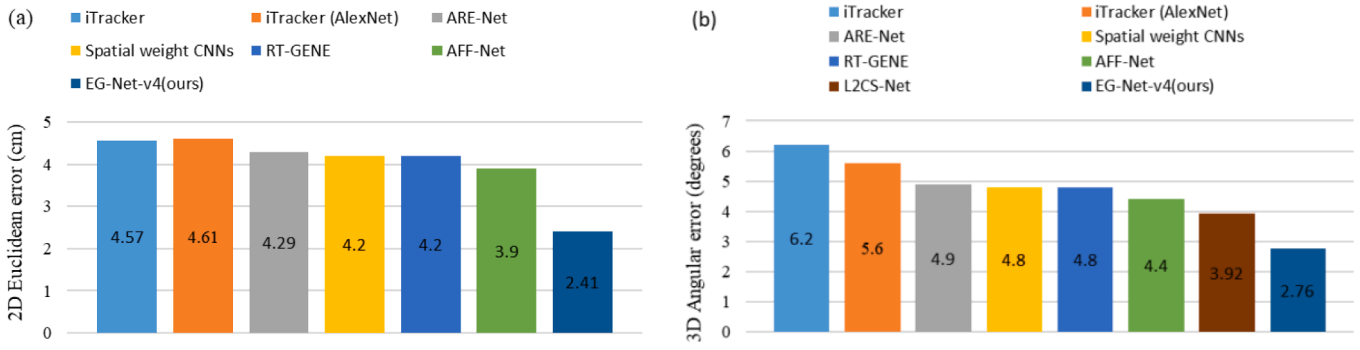


Fig. 7. Error of gaze estimation compared with state-of-the-art methods on the MPIIFaceGaze dataset. (a) 2D gaze estimation error in centimetres (Euclidean error). (b) 3D gaze estimation error in degrees (angular error).

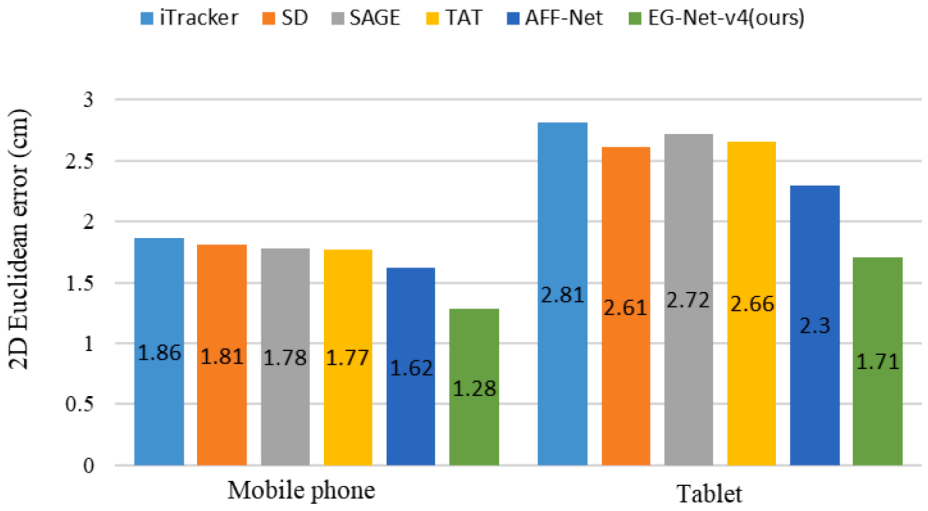


Fig. 8. Gaze estimation error in degrees (angular error) compared with state-of-the-art methods on the GazeCapture dataset.

Table 8
Comparison of the results with the state-of-the-art methods on the EyeDIAP dataset.

Model	3D gaze direction (degrees)
iTracker (Krafka et al., 2016)	8.3
MPIIGaze (Zhang et al., 2017)	6.3
Spatial weight CNNs (Xucong et al., 2017)	6.0
MSGazeNet (Mahmud et al., 2022)	5.86
CA-Net (Cheng et al., 2020)	5.3
EG-Net-v4 (ours)	**4.55**

et al., 2016), MPIIGaze (Zhang et al., 2017), Spatial weight CNNs (Xucong et al., 2017), MSGazeNet (Mahmud, Hungler, & Etemad, 2022) and CA-Net (Cheng, Shiyao, Fei, Qian, & Lu, 2020). Table 8 and Fig. 9 summarize the comparison results for 3D gaze direction errors in degree. As seen from Table 8, given that the state-of-the-art method of CA-Net (Cheng et al., 2020) with error of 5.3 degrees, our EG-Net-v4 improved it by 14.15 % with error of 4.55 degrees on EyeDIAP dataset.

5. Conclusion and discussion

In this study, we demonstrate that inconsistent performance in gaze estimation tasks exists between left- and right-eye images, which is caused by free head poses, illumination changes, eye occlusion and individual differences. Therefore, the recent methods that use the same ConvNet on both eye and face images have inherent deficiencies. To overcome these deficiencies, an appearance-based network architecture is proposed that applies two-branch network, namely, a base CNN and EE-Net, for the face images and the two individual eye images, respectively. In particular, EE-Net, which is designed for eye image feature extraction, uniformly scales up the depth, width and resolution of the base CNN by a set of constant coefficients and adaptively adjusts the network attention to the left- and right-eye images by Attention-Net.

This study shows that, compared to the recent methods that treat all regions with the same network, the proposed method is more robust for capturing face and eye appearance variations caused by free head poses, illumination changes, eye occlusion and individual differences. This method is evaluated and compared with state-of-the-art methods. The conducted experiments showed that EG-Net achieves relatively low errors in gaze estimation tasks, with a 2D Euclidean error of 1.58 ~ 2.53 cm on the GazeCapture dataset, a 3D angular error of 2.76 ~ 2.98 degrees on the MPIIFaceGaze dataset, and 2.78 ~ 3.17 degrees on its augmented dataset. EG-Net-v4 even outperforms state-of-the-art approaches on the MPIIFaceGaze dataset, with prediction errors of 2.41 cm and 2.76 degrees in 2D and 3D gaze estimation, respectively. It achieves results with an error of 1.28 cm on mobile phone-captured images and 1.71 cm on tablet captured images on the GazeCapture dataset, indicating its good performance in the gaze estimation task. It also yields a

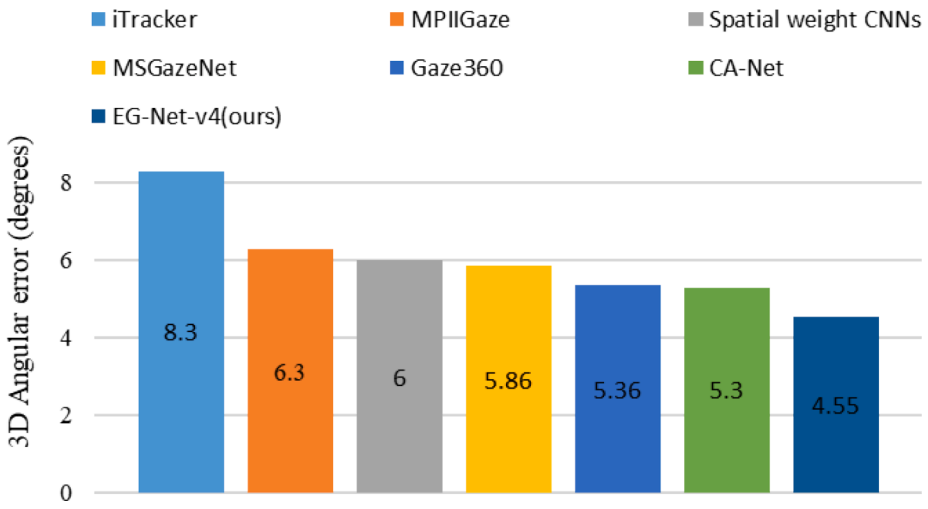


Fig. 9. Gaze estimation error compared with state-of-the-art methods on the EyeDIAP dataset.

performance improvement of 4.55 degrees on EyeDIAP dataset, with 14.2 % improvement over prior arts.

In our future work, we intend to explore the incorporation of more advanced networks to replace the current simple base CNN, with a focus on scaling them for eye image feature extraction. Additionally, given the challenges posed by relatively long distances, unpredictable scenes, and the potential absence of human faces in wild scenes, we believe it's essential to investigate the simultaneous integration of face appearance, head poses, human body posture, and depth information into the model. This approach holds the potential to effectively address these challenges and potentially broaden the application scenarios in wild. Furthermore, human eye gaze tracking is a time-related continuous task; thus, as part of our future research, we plan to incorporate time series factors into our approach to achieve more accurate gaze estimation across consecutive video frames. Finally, in view of practical applications, we will test the generalization ability of the proposed model in an intelligent cockpit to ensure that it is robust in the driver eye gaze estimation task.

CRediT authorship contribution statement

Xinmei Wu: Conceptualization, Methodology, Investigation, Software, Visualization, Validation, Writing – review & editing. **Lin Li:** Funding acquisition, Investigation, Methodology, Supervision, Writing – review & editing. **Haihong Zhu:** Project administration, Resources. **Gang Zhou:** Data curation, Methodology, Resources, Writing – review & editing. **Linfeng Li:** Project administration, Resources. **Fei Su:** Validation. **Shen He:** Data curation. **Yanggang Wang:** Visualization. **Xue Long:** Visualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Key R&D Program of China (No. 2021YFB2501101), Technology Application Project (2021-4201-21-000661), and Natural Science Foundation of Shandong Province (ZR2021QD105).

References

Abdelrahman, A. A., Hempel, T., Khalifa, A., & Al-Hamadi, A. (2022). *L2CS-Net: Fine-Grained Gaze Estimation in Unconstrained Environments*. Paper presented at the IEEE International Conference on Image Processing (ICIP) 2022from https://doi.org/10.48550/arXiv.2203.03339.

Ali, A., & Kim, Y. (2020). Deep fusion for 3D gaze estimation from natural face images using multi-stream CNNs. *IEEE Access, 8*, 69212–69221. https://doi.org/10.1109/ACCESS.2020.2986815

Asteriadis, S., Karpouzis, K., & Kollias, S. (2014). Visual focus of attention in non-calibrated environments using gaze estimation. *International Journal of Computer Vision, 107*(3), 293–316. https://doi.org/10.1007/s11263-013-0691-3

Baluja, S., & Pomerleau, D. (1994). *Non-Intrusive Gaze Tracking Using Artificial Neural Networks*. Paper presented at the Proceedings of the 6th International Conference on Neural Information Processing Systems, San Francisco, CA, USAfrom https://dl.acm.org/doi/abs/10.5555/2987189.2987284.

Bao, Y., Cheng, Y., Liu, Y., & Lu, F. (2020). *Adaptive Feature Fusion Network for Gaze Tracking in Mobile Tablets*. Paper presented at the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italyfrom https://ieeexplore.ieee.org/document/9412205.

Chen, C. (2021). {PyTorch Face Landmark}: A Fast and Accurate Facial Landmark Detector. (Reprinted.

Cheng, Y., Lu, F., & Zhang, X. (2018). *Appearance-Based Gaze Estimation via Evaluation-Guided Asymmetric Regression*. Paper presented at the ECCV 2018, Chamfrom https://go.exlibris.link/tGShWl1K.

Cheng, Y., Shiyao, H., Fei, W., Qian, C., & Lu, F. (2020). *A Coarse-to-Fine Adaptive Network for Appearance-Based Gaze Estimation*. Paper presented at the Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) from https://doi.org/10.48550/arXiv.2001.00187.

Cheng, Y., Wang, H., Bao, Y., & Lu, F. (2021). Appearance-based gaze estimation with deep learning. Paper presented at the from *A Review and Benchmark*. https://arxiv.org/abs/2104.12668.

Dan, W. H., & Qiang, J. (2010). In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*(3), 478–500. https://doi.org/10.1109/TPAMI.2009.30

Deng, H., & Zhu, W. (2017). Monocular free-head 3D gaze tracking with deep learning and geometry constraints. *Paper presented at the 2017 IEEE International Conference on Computer Vision (ICCV)from*.

Fischer, T., Chang, H. J., & Demiris, Y. (2018). *RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments*. Paper presented at the, Chamfrom https://link.springer.com/chapter/10.1007/978-3-030-01249-6_21.

Funes Mora, K. A., Monay, F., & Odobez, J. (2014). *EYEDIAP: a database for the development and evaluation of gaze estimation algorithms from RGB and RGB-D cameras*. Paper presented at the from https://doi.org/10.1145/2578153.2578190.

Funes-Mora, K. A., & Odobez, J. (2016). Gaze estimation in the 3D space using RGB-D sensors. *International Journal of Computer Vision, 118*(2), 194–216. https://doi.org/10.1007/s11263-015-0863-4

Fung, M., Jin, Y., Zhao, R. J., & Hoque, M. E. (2015). ROC speak: Semi-automated personalized feedback on nonverbal behavior from recorded videos. In *Paper presented at the 2015 ACM International Joint Conference from*. https://doi.org/10.1145/2750858.2804265

Ghiass, R. S., Arandjelovic, O., & Laurendeau, D. (2018). Highly Accurate and fully automatic 3D head pose estimation and eye gaze estimation using RGB-D sensors and 3D morphable models. *Sensors, 18*(12), 4280. https://doi.org/10.3390/s18124280

Guo, M., Xu, T., Liu, J., Liu, Z., Jiang, P., Mu, T.,... Hu, S. (2022). Attention Mechanisms in Computer Vision: A Survey. *Computational Visual Media*(8), 331-368. doi: 10.1007/s41095-022-0271-y.

Guo, T., Liu, Y., Zhang, H., Liu, X., Kwak, Y., Yoo, B. I.,... Choi, C. (2019). *A Generalized and Robust Method Towards Practical Gaze Estimation on Smart Phone*. Paper presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCV 2019 Workshop), Seoulfrom https://go.exlibris.link/l043xkP.

He, J., Pham, K., Valliappan, N., Xu, P., Roberts, C., Lagun, D.,... Navalpakkam, V. (2019). *On-Device Few-Shot Personalization for Real-Time Gaze Estimation*. Paper presented at the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea (South) from https://ieeexplore.ieee.org/document/9021975.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep Residual Learning for Image Recognition*. Paper presented at the from https://doi.org/10.48550/arXiv.1512.03385.

Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2020). Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(8), 2011–2023. https://doi.org/10.1109/TPAMI.2019.2913372

Huang, Y., Cheng, Y., Bapna, A., Firat, O., Chen, M. X., Chen, D.,... Le, Q. V. (2018). *GPipe: Efficient Training of Giant Neural Networks using Pipeline Parallelism*. Paper presented at the Proceedings of the 33rd International Conference on Neural Information Processing Systems from https://dl.acm.org/doi/10.5555/3454287.3454297.

Kar, A., & Corcoran, P. (2017). A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms. *IEEE Access, 5*, 16495–16519. https://doi.org/10.1109/ACCESS.2017.2735633

Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., & Torralba, A. (2019). *Gaze360: Physically Unconstrained Gaze Estimation in the Wild*. Paper presented at the ICCV, Seoul, Korea (South)from https://ieeexplore.ieee.org/document/9010825.

Kim, M., Wang, O., & Ng, N. (2016). *Convolutional neural network architectures for gaze estimation on mobile devices*. Standford Univ.

Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W.,... Torralba, A. (2016). *Eye Tracking for Everyone*. Paper presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) from https://ieeexplore.ieee.org/document/7780608.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM, 60*(6), 84–90. https://doi.org/10.1145/3065386

L R D, M., & Biswas, P. (2021). *Appearance-based Gaze Estimation using Attention and Difference Mechanism*. Paper presented at the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) from https://go.exlibris.link/HkypJLb2.

Larochelle, H., Erhan, D., Courville, A., Bergstra, J., & Bengio, Y. (2007). An empirical evaluation of deep architectures on problems with many factors of variation. In *Paper presented at the Proceedings of the 24th international conference on Machine learning*. https://doi.org/10.1145/1273496.1273556

Lemley, J., Kar, A., Drimbarean, A., & Corcoran, P. (2019). Convolutional neural network implementation for eye-gaze estimation on low-quality consumer imaging systems. *IEEE Transactions on Consumer Electronics, 65*(2), 179–187. https://doi.org/10.1109/TCE.2019.2899869

Lin, H., & Jegelka, S. (2018). *ResNet with One-Neuron Hidden Layers is a Universal Approximator*. Paper presented at the Proceedings of the 32nd International Conference on Neural Information Processing Systems, Red Hook, NY, USAfrom https://dlnext.acm.org/doi/10.5555/3327345.3327515.

Lindén, E., Sjöstrand, J., & Proutiere, A. (2019). *Learning to Personalize in Appearance-Based Gaze Tracking*. Paper presented at the 2019 IEEE/CVF International

Conference on Computer Vision Workshop (ICCVW), Seoul, Korea (South) from https://ieeexplore.ieee.org/document/9022231.
Liu, H., Heynderickx, & Ingrid. (2011). Visual attention in objective image quality assessment: based on eye-tracking data. *IEEE Transactions on Circuits & Systems for Video Technology, 21*(7), 971–982. https://doi.org/10.1109/TCSVT.2011.2133770
Liu, G., Yu, Y., Mora, K. A. F., & Odobez, J. (2018). *A Differential Approach for Gaze Estimation with Calibration*. Paper presented at the BMVC 2018from http://bmvc2018.org/contents/papers/0792.pdf.
Liu, G., Yu, Y., Mora, K., & Odobez, J. M. (2021). A Differential Approach for Gaze Estimation. [Journal Article; Research Support, Non-U.S. Gov't]. *IEEE Trans Pattern Anal Mach Intell*, 43(3), 1092-1099. doi: 10.1109/TPAMI.2019.2957373.
Liu, Y., Liu, R., Wang, H., & Lu, F. (2021). *Generalizing Gaze Estimation with Outlier-guided Collaborative Adaptation*. Paper presented at the ICCV2021from https://go.exlibris.link/Nvl4jpsx.
Lu, F., Sugano, Y., Okabe, T., & Sato, Y. (2012). *Head pose-free appearance-based gaze sensing via eye image synthesis*. Paper presented at the, Tsukuba, Japan from https://ieeexplore.ieee.org/document/6460306.
Lu, F., Sugano, Y., Okabe, T., & Sato, Y. (2014). Adaptive linear regression for appearance-based gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 10*(36), 2033–2046. https://doi.org/10.1109/TPAMI.2014.2313123
Lu, H., Wang, C., & Yen-wei, C. (2008). *Gaze tracking by Binocular Vision and LBP features*. Paper presented at the International Conference on Pattern Recognition (ICPR 2008) from https://ieeexplore.ieee.org/document/4761019.
Mahmud, Z., Hungler, P., & Etemad, A. (2022). Multistream Gaze Estimation with Anatomical Eye Region Isolation by Synthetic to Real Transfer Learning. *ArXiv*, abs/2206.09256. doi: 10.48550/arXiv.2206.09256.
Majaranta, P., & Bulling, A. (2014). Eye Tracking and Eye-Based Human–Computer Interaction *Human–Computer Interaction Series book series (HCIS)* (39-65): Springer London. (Reprinted.
Neilmacrae, C., Hood, B. M., Milne, A. B., Rowe, A. C., & Mason, M. F. (2002). Are you looking at me? Eye gaze and person perception. *Psychological Science, 13*(5), 460–464. https://doi.org/10.1111/1467-9280.00481
Otsu, K., Seo, M., Kitajima, T., & Chen, Y. (2020). *Automatic Generation of Eye Gaze Corrected Video Using Recursive Conditional Generative Adversarial Networks*. Paper presented at the, Kobe, Japanfrom https://go.exlibris.link/RFwhbhzz.
Park, S., Mello, S. D., Molchanov, P., Iqbal, U., Hilliges, O.,... Kautz, J. (2019). *Few-Shot Adaptive Gaze Estimation*. Paper presented at the, Seoul, Korea (South) from https://ieeexplore.ieee.org/document/9008783.
Park, S., Spurr, A., & Hilliges, O. (2018). *Deep Pictorial Gaze Estimation*. Paper presented at the Computer Vision – ECCV 2018, Chamfrom https://go.exlibris.link/Bf30VZMc.
Poulopoulos, N., & Psarakis, E. Z. (2023). *Few-shot Gaze Estimation via Gaze Transfer*. Paper presented at the 18th International Conference on Computer Vision Theory and Applications, Lisbon, Portugalfrom https://www.scitepress.org/Link.aspx?doi=10.5220/0011789800003417.
Shic, F., & Scassellati, B. (2007). A behavioral analysis of computational models of visual attention. *International Journal of Computer Vision, 73*(2), 159–177. https://doi.org/10.1007/s11263-006-9784-6
Sugano, Y., Matsushita, Y., Sato, Y., & Koike, H. (2008). *An Incremental Learning Method for Unconstrained Gaze Estimation*. Paper presented at the ECCV 2008: Computer Vision – ECCV 2008 from https://doi.org/10.1007/978-3-540-88690-7_49.
Sugano, Y., Matsushita, Y., & Sato, Y. (2013). Appearance-based gaze estimation using visual saliency. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(2), 329–341. https://doi.org/10.1109/TPAMI.2012.101
Sun, H., & Pears, N. (2023). Accurate Gaze Estimation using an Active-gaze Morphable Model. *CoRR*, abs/2301.13186. doi: 10.48550/arXiv.2301.13186.
Tan, M., & Le, Q. V. (2019). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. Paper presented at the PMLR 97from https://go.exlibris.link/xsjx0lyd.
Tran, M., Sen, T., Haut, K., Ali, M. R., & Hoque, M. E. (2020). Are you really looking at me? A feature-extraction framework for estimating interpersonal eye gaze from conventional video. *IEEE Transactions on Affective Computing, 13*(2), 912–925. https://doi.org/10.1109/TAFFC.2020.2979440
Wang, Q., Wang, H., Dang, R., Zhu, G., Pi, H., … Shic, F.,... Hu, B.. (2023). Style transformed synthetic images for real world gaze estimation by using residual neural network with embedded personal identities. *Applied Intelligence, 53*(2), 2026–2041. https://doi.org/10.1007/s10489-022-03481-9
Wedel, M., & Pieters, R. (2018). A Review of Eye-Tracking Research in Marketing. In N. K. Malhotra (Ed.), (4, pp. 123-147): Emerald Group Publishing Limited. (Reprinted).
Williams, O., Blake, A., & Cipolla, R. (2006). *Sparse and Semi-supervised Visual Mapping with the S3GP*. Paper presented at the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06) from https://doi.org/10.1109/CVPR.2006.285.
Wu, Y., Yeh, C., Hung, W., & Tang, C. (2014). Gaze direction estimation using support vector machine with active appearance model. *Multimedia Tools and Applications, 70*(3), 2037–2062. https://doi.org/10.1007/s11042-012-1220-z
Wu, Z., Shen, C., & van den Hengel, A. (2019). Wider or Deeper: Revisiting the ResNet model for visual recognition. *Pattern Recognition, 90*, 119–133. https://doi.org/10.1016/j.patcog.2019.01.006
Xucong, Z., Park, S., Beeler, T., Bradley, D., Tang, S.,... Hilliges, O. (2020). *ETH-XGaze: A Large Scale Dataset for Gaze Estimation Under Extreme Head Pose and Gaze Variation*. Paper presented at the ECCV 2020, Berlin, Heidelbergfrom https://doi.org/10.1007/978-3-030-58558-7_22.
Xucong, Z., Yusuke, S., Mario, F., & Andreas, B. (2017). *It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation*. Paper presented at the Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conferencefrom https://www.computer.org/csdl/proceedings-article/cvprw/2017/0733c299/12OmNzaQoPr.
Yang, C., Xie, L., Su, C., & Yuille, A. L. (2019). *Snapshot Distillation: Teacher-Student Optimization in One Generation*. Paper presented at the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USAfrom https://doi.ieeecomputersociety.org/10.1109/CVPR.2019.00297.
Ye, Z., Li, Y., Fathi, A., Han, Y., & Rehg, J. M. (2012). *Detecting eye contact using wearable eye-tracking glasses*. Paper presented at the Proceedings of the 2012 ACM Conference on Ubiquitous Computing.
Yilmaz, C. M., & Kose, C. (2016). *Local Binary Pattern Histogram features for on-screen eye-gaze direction estimation and a comparison of appearance based methods*. Paper presented at the 2016 39th International Conference on Telecommunications & Signal Processing(TSP)from https://ieeexplore.ieee.org/document/7760973.
Yusuke, S., Yasuyuki, M., & Yoichi, S. (2014). *Learning-by-Synthesis for Appearance-Based 3D Gaze Estimation*. Paper presented at the, Columbus, OH, USAfrom https://ieeexplore.ieee.org/document/6909631?arnumber=6909631.
Zagoruyko, S., & Komodakis, N. (2016). *Wide Residual Networks*. Paper presented at the Proceedings of the British Machine Vision Conference (BMVC), York, Francefrom https://dx.doi.org/10.5244/C.30.87.
Zhang, C., Yao, R., & Cai, J. (2018). Efficient eye typing with 9-direction gaze estimation. *Multimedia Tools and Applications, 77*(15), 19679–19696. https://doi.org/10.1007/s11042-017-5426-y
Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters, 23*(10), 1499–1503. https://doi.org/10.1109/LSP.2016.2603342
Zhang, X., Sugano, Y., Fritz, M., & Bulling, A. (2015). *Appearance-Based Gaze Estimation in the Wild*. Paper presented at the from https://go.exlibris.link/wNN02t1g.
Zhang, X., Sugano, Y., Fritz, M., & Bulling, A. (2017). MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation. [Journal Article; Research Support, Non-U.S. Gov't]. *IEEE Trans Pattern Anal Mach Intell*, 41(1), 162-175. doi: 10.1109/TPAMI.2017.2778103.
Zhao, H., Lu, M., Yao, A., Chen, Y., & Zhang, L. (2020). Learning to draw sight lines. *International Journal of Computer Vision, 128*(5), 1076–1100. https://doi.org/10.1007/s11263-019-01263-4