IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 August 2023, accepted 13 September 2023, date of publication 18 September 2023, date of current version 4 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3317013

RESEARCH ARTICLE

Gaze Estimation Based on Attention Mechanism Combined With Temporal Network

GUOJING REN, YANG ZHANG, AND QINGJUAN FENG
School of Automation, Beijing Information Science and Technology University, Beijing 100192, China

Corresponding author: Yang Zhang (zhangyang@bistu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62001035, in part by the Research and Development Program of Beijing Municipal Education Commission under Grant KM202111232018, and in part by the Natural Science Foundation of Beijing under Grant 4232044.

ABSTRACT Due to the continuous and dynamic nature of gaze estimation, the true gaze point at each moment is closely related to the previous moment. Simply detecting individual frames of facial images cannot yield accurate gaze information. In current CNN-based gaze estimation methods, the effective utilization of eye movement temporal information and the ability to capture global relationships in the feature extraction process remain problematic. Addressing these concerns, this paper proposes a novel gaze estimation framework, named FE-net, which incorporates a temporal network. This framework introduces channel attention modules and self-attention modules, enhancing the comprehensive utilization of extracted features and reinforcing the contribution of valuable regions to gaze estimation. We further integrate an RNN structure to learn the temporal dynamics of eye movement processes, significantly improving gaze direction prediction accuracy. This framework predicts the gaze directions of left and right eyes separately using monocular and facial features and computes the overall gaze direction. FE-net achieves state-of-the-art accuracy of 3.19° and 3.16° on the EVE dataset and the MPIIFaceGaze dataset, respectively.

INDEX TERMS Appearance-based gaze estimation, attention mechanism, convolutional neural networks, deep learning.

I. INTRODUCTION

In the current landscape, computer screens have evolved into the primary tools for visualizing external information. To comprehend the content that captivates observers, a more accurate tracking of their points of interest is crucial. The utilization of a camera positioned above the screen to capture changes in the observers' eyes and determine their gaze positions has emerged as a prominent trend. This approach finds extensive applications in domains such as human-computer interaction [1], [2], virtual reality [3], and assisted driving [4], [5], enabling a better understanding of what draws observers' attention.

Over the past few decades, researchers have proposed a plethora of gaze estimation methods, which can be broadly categorized into two main types: model-based methods and appearance-based methods. Model-based methods aim to recover a constructed 3D geometric eye model by identifying specific parameters unique to an individual, and subsequently utilize this model to estimate gaze direction [6], [7], [8], [9]. However, due to the inherent diversity of human eyes, the constructed 3D eye models tend to vary from person to person. These methods often require individual calibration to recover personalized parameters, such as near-infrared corneal reflections, iris contours, iris radius, and kappa angle. The collection of specific data relies on specialized detection equipment, and strict usage conditions, including a constrained working distance between the user and the camera, often limit their applicability to laboratory environments.

Appearance-based methods, on the other hand, do not require specialized equipment. Instead, they utilize regular RGB cameras to directly learn the mapping function from facial appearance to eye gaze direction. These methods have gained popularity and become mainstream due to their simplicity, wide applicability, good generalization, and the maturing of deep learning techniques in recent years.

The associate editor coordinating the review of this manuscript and approving it for publication was Zhe Jin.

Zhang et al. [10] were the first to apply CNN networks for gaze estimation. They extracted eye images from facial images and used them to estimate gaze direction by extracting features specifically from the eye region. Due to the limitations of eye features, a facial-feature-only gaze estimation method was later proposed [18]. Recently, there have been methods that simultaneously use facial images and eye images cropped from the facial region as input, employing three-stream networks to extract features from the face, left eye, and right eye images [11], [12], [13], [14]. However, these methods treat facial and eye images as independent or parallel feature sources and employ simple techniques, such as simple concatenation or fully connected layers, to fuse information from facial and eye images, overlooking their inherent relationships at a granular level of features.

In practical applications, gaze estimation systems typically take video sequences of eye and facial images as input. These videos contain valuable temporal information that can be utilized to improve gaze estimation. In addition to the static features obtained from the images, the temporal information from the videos can contribute to better gaze estimation. Recurrent neural networks (RNNs), such as Long Short-Term Memory (LSTM) [15], [16], have been widely employed in video processing tasks. RNNs automatically capture the temporal information for gaze estimation. However, most existing methods focus on static information and overlook temporal sequence information.

To address the existing challenges in gaze estimation, we propose a novel approach that combines facial features with eye features. In our method, we sequentially predict the gaze direction of each eye and calculate the overall gaze direction by considering both eyes. To overcome the issue of underutilization of features, our method incorporates attention mechanisms to enhance the contribution of relevant regions for gaze estimation. and combined with GRU to incorporate temporal information for gaze estimation. The main contributions of this paper are as follows:

1) We propose a framework called FE-Net that combines facial features with eye features. Compared to state-of-the-art algorithms, FE-Net achieves superior accuracy.
2) We have introduced an AC module to the concatenated features of facial and ocular characteristics, which serves as an attention mechanism integrating channel attention and self-attention. This module facilitates the redistribution of weights to the fused features of the eyes and face, enabling accurate extraction of key features.
3) To capture the temporal dynamics inherent in the observer's eye movement process, we have incorporated GRU (Gated Recurrent Unit) modules into our network architecture. This structure effectively models and captures the temporal information present in eye-tracking data, aiding in our understanding and prediction of eye movement behavior, thereby enhancing gaze accuracy.

II. RELATED WORK

In the past five years, appearance-based gaze estimation methods have been greatly inspired by machine learning and deep learning algorithms. The development of deep learning has provided favorable conditions for the advancement of gaze estimation. Recently, deep learning models, such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), have been employed for gaze estimation.

Zhang et al. [23] proposed the GazeNet framework, which is an appearance-based gaze estimation method utilizing the VGG network. This network combines facial pose and eye region data, injects head pose angles into the first fully connected layer, and trains a regression model on the output layer. The Itracker model introduced by Krafka et al. [11] is based on the AlexNet architecture. This model employs a multi-region network with various inputs, extracting valuable information beyond using eye images alone, and achieves good performance even without calibration. Additionally, methods utilizing dilated convolutions [12], [19] effectively capture feature information at different scales, thereby improving gaze accuracy.

In addition to CNN networks, the Transformer, initially proposed by Vaswani et al. [20]for natural language processing tasks, has been utilized for gaze estimation tasks. Due to its ability to capture global context, Transformers have demonstrated excellent performance in computer vision tasks. Cheng and Lu introduced the GazeTR model [21], which was the first to apply Transformers to gaze estimation. They utilized a hybrid Vision Transformer (ViT) [32] for appearance-based gaze estimation tasks and achieved state-of-the-art results on multiple datasets. Apart from feature extraction, Cheng et al. [22] proposed an attention module that integrates features from both eyes, ensuring the effective utilization of informative features. Additionally, Cheng et al. [38] proposed a strategy that employs binocular images as input and leverages binocular asymmetry to optimize the entire network.

In RNN architectures, Long Short-Term Memory (LSTM) can be utilized to handle the temporal information in video data. Due to the valuable information contained in video data beyond image data, studies [15], [16], [17] have demonstrated the effectiveness of incorporating temporal features to enhance gaze estimation accuracy. The specific approach involves fusing static features extracted from images with sequential features extracted from videos. Zhou et al. [16] enhanced the Itracker network proposed in [18] by removing the facial grid and utilizing concatenated static features from the eye region images to predict gaze outcomes. They employed a bidirectional LSTM (bi-LSTM) to capture the temporal features between video frames and estimate gaze vectors in the video sequence. Park et al. [17] created a new dataset comprising facial eye-tracking video frames in four directions. They employed GRU, LSTM, and other techniques to extract temporal features between frames. Additionally, they collected and incorporated screen content

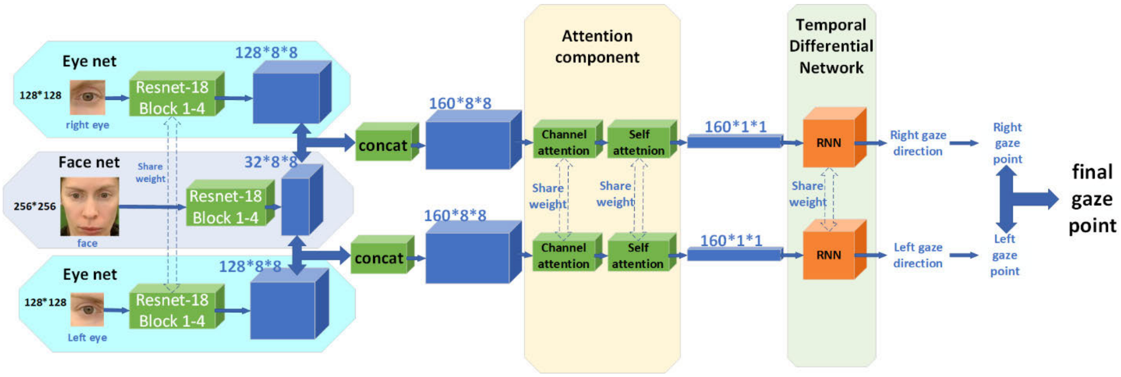


FIGURE 1. The proposed model diagram for gaze prediction combining facial images.

that the observers fixated on for saliency detection, further optimizing gaze prediction.

In recent advancements, addressing the limitation of CNN models in establishing long-range connections for capturing local information to model global images, Wu et al. [24] have introduced a self-supervised approach. This method leverages advanced cues from the eye region to refine facial features in gaze estimation. It uses high-level knowledge to filter the distractive information and bridges the intrinsic relationship between face and eye features. Zhou et al. [25] proposed a weighted network with an adaptive adjustment regression strategy, which learns the varying contributions of different regions to gaze estimation outcomes under free head movement. And there are some methods incorporate attention mechanisms to enhance the consideration of interrelationships between distinct regions. Nagpure and Okuma [26] propose a novel multi-resolution fusion transformer based gaze regression head which is efficient as well as accurate to predict gaze values from multi-resolution features. Song et al. [27] propose an encoder-decoder network with residual blocks and attention blocks. Dai et al. [28] proposed a method based on the convolutional neural network with residual blocks, in which the attention mechanism is integrated into the network to improve the accuracy of gaze tracking. Although the aforementioned methods have incorporated attention mechanisms, their focus has been primarily directed towards entire facial images, neglecting the specific attention to eye-region images and thus failing to comprehensively exploit the reciprocal interplay between facial and ocular information. Regarding the exploration of temporal dynamics, there exists a relative scarcity of approaches that consider the influence of multiple consecutive frames on gaze estimation. Among these, the methods [15], [16], [17] employed time-series networks to extract dynamic temporal information from eye movement processes, yet they overlooked the effective utilization of both ocular and facial features. Our approach, however, simultaneously harnesses ocular and facial features, sequentially enhancing the weight of meaningful features through channel attention and self-attention mechanisms. Moreover, we leverage a time-series network to extract dynamic temporal information from the eye movement process, thereby further improving the accuracy of gaze estimation.

III. PROPOSED MODEL AND ALGORITHM

In this section, we will provide a detailed description of how we designed the end-to-end gaze estimation network. We propose a model called FE-Net, which takes both facial images and eye images as inputs to predict gaze directions. We incorporate channel attention and self-attention modules after the concatenation of facial and eye features. These modules effectively extract and amplify the most informative features from the face and the eyes. The FE-Net network is depicted in Figure 1.

FE-Net consists of three main parts: (1) Eye-net and Face-net, which form the backbone network of the model and utilize the first four layers of the ResNet-18 CNN network. In the Eye-net, the number of channels in the fourth layer is adjusted to facilitate concatenation with the facial features. (2) The AC module, which is an attention mechanism module, consists of the concatenation of the ECA channel attention module and the self-attention module. This module applies an attention mechanism to the feature map obtained by concatenating the facial and eye features, allowing for the selection of relevant features for more effective utilization. (3) The GRU module, which is an RNN network composed of the time-series module GRU. It utilizes the GRU architecture to capture temporal dependencies and sequential information in the input data. Eye-tracking technology is a real-time end-to-end technique, as the process of eye movements unfolds as a coherent sequence. To capture temporal information, we have incorporated an RNN network into our model. In this framework, the Face-net and Eye-net are responsible for extracting static features from a single face image and its corresponding left and right eye images, respectively. These facial features are concatenated with the left and right eye features and then fed into the AC module. The AC module employs its built-in channel attention module and self-attention module to selectively capture important features. Subsequently, the output is passed to the GRU module for learning temporal dynamics. Our proposed model operates in two fundamental states: forward propagation and backward propagation.

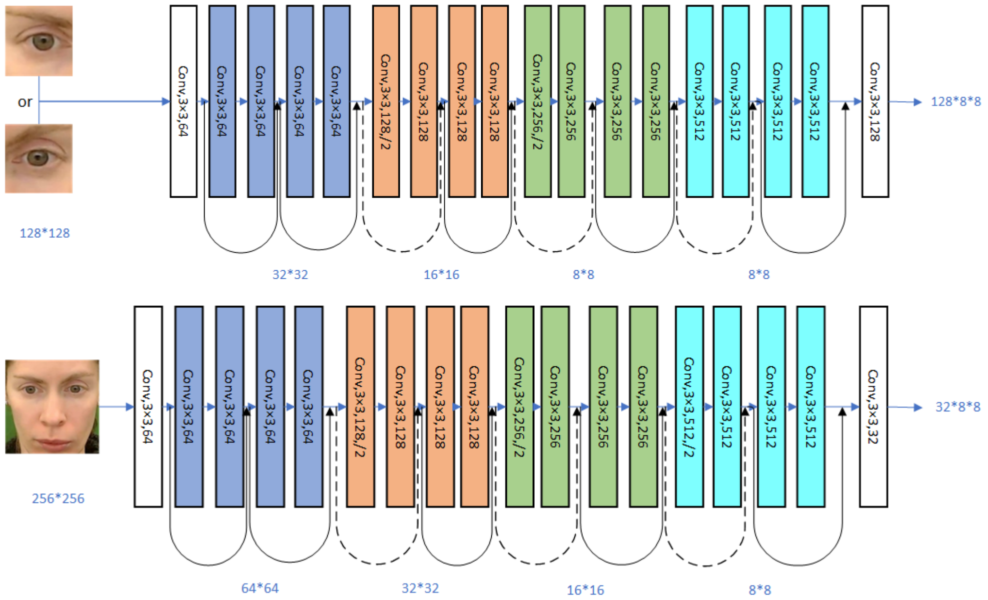


FIGURE 2. Eye-net (top) and Face-net (bottom) network architecture diagrams.

Forward propagation is utilized to compute the intermediate variables at each layer, while backward propagation is employed to calculate the gradients at each layer.

A. FACE-NET AND EYE-NET

The network architecture of Eye-net and Face-net are depicted in Figure 2. Both Face-net and Eye-net utilize the ResNet-18 [29] convolutional neural network. This network introduces residual blocks in traditional convolutional neural networks, which effectively address the issues of gradient vanishing or explosion and degradation. It performs well in feature extraction. The architectures of Face-net and Eye-net are shown in the following figure. The input size for the eye image is 128 × 128 × 8, and for the face image is 256 × 256 × 3. To ensure smooth concatenation of extracted eye features with face features, in the Eye-net network, we adjust the stride to 1 during the process of expanding the channel size to 512 in the fourth layer. We also add a convolutional layer at the end to adjust the feature size to 128. In the Face-net, we add a convolutional layer at the end to adjust the feature size to 32. We believe that eye tracking is more related to eye features and eye features are more important than face features. Therefore, we retain 32 features for face and 128 features for each eye. Finally, we concatenate these features to obtain a feature matrix of size 160 × 8 × 8.

B. AC MODULE

1) EFFICIENT CHANNEL ATTENTION

ECA module is a channel modeling technique based on attention mechanisms. It can be seen as an improved version of SENet [30]. The dimension reduction operation used in SENet has a negative impact on the prediction of channel attention and leads to inefficient and unnecessary dependency modeling. To avoid the influence of dimension reduction on channel attention learning, ECANet proposes a non-reductive local cross-channel interaction strategy. In convolutional neural networks, each channel corresponds to different feature information. The ECA attention mechanism weights each channel to extract the most important features and suppress less important ones. This helps the network to focus more on useful information, thereby improving the expressive power of the features.

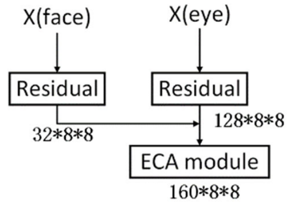


FIGURE 3. The specific location of the ECA channel attention module.

As shown in Figure 3, after the face and eye images are initially processed by the backbone networks to extract features, the face features and eye features are concatenated. To extract the channel features of the fused eye-face representation, we input the concatenated features into the ECA module. Given the input feature map X, $X \in R^{H \times W \times C}$, where H represents height, W represents the width, and C represents the number of channels. First, a non-reductive global average pooling (GAP) operation is applied to the input feature map to obtain aggregated convolutional features. The ECA module adaptively determines the kernel size K using an adaptive function as shown in Equation (1). Then, a one-dimensional convolution is performed to learn the channel attention using the sigmoid function, which calculates the weight G for each channel, as shown in Equation (2). Finally, the generated

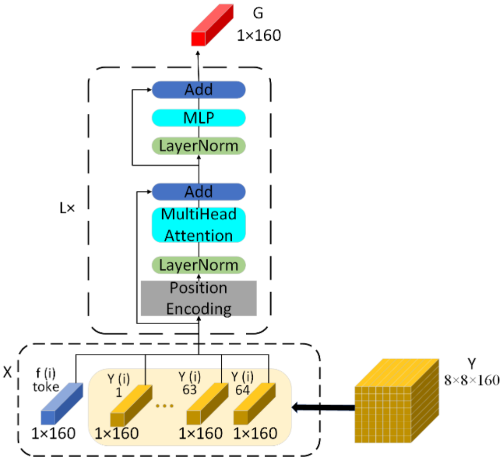


FIGURE 4. **Self-attention module.**

channel attention weights G are multiplied with the original feature map X and scaled to enhance the feature responses of important channels. The weighted feature map Y is obtained as shown in Equation (3).

$$k = \psi\left(C\right) = \left|\frac{\log_2(C)}{\gamma} + \frac{b}{\gamma}\right|_{odd} \tag{1}$$

where $\gamma = 2$, b = 1. C represents the number of channels.

$$G = \sigma\left(W_1 * B + b_1\right) \tag{2}$$

where $W_1 \in R^{C\times C'}$ is the weight matrix, σ represents the Sigmoid activation function, $b_1 \in R^C$ is the bias vector, and B represents the embedded features.

$$Y = X \otimes \mathrm{G} \tag{3}$$

where $\otimes$ denotes element-wise multiplication along the channel dimension.

2) SELF-ATTENTION MODULE

The self-attention mechanism is the core idea of the Transformer architecture [20], and it is used to extract global information from the input by computing weighted sums of the feature maps. As shown in Figure 4, the Transformer consists of three components: Multi-Head Self-Attention (MSA), Multi-Layer Perceptron (MLP), and Layer Normalization (LN). Here, we refer to the positional encoding and patch embedding process from [32] and [33]. We create a learnable f_{token}^i and encode it as 0. We obtain the feature matrix $Y^{(i)} \in R^{H\times W\times C}$ after the channel attention mechanism, where H and W represent the size of the feature matrix, C represents the number of channels, and i represents the feature matrix of the i-th frame in the input video, where i≤3. Here, H=W=8 and C=160. The feature matrix $Y^{(i)}$ is divided into 64 one-dimensional vectors $Y_1^{(i)} \sim Y_{64}^{(i)}$, and these vectors are encoded with positions 1 to 64. Finally, we concatenate f_{token}^i with these vectors to obtain the feature matrix $X = \left[f_{token}^i; Y_1^{(i)}; Y_2^{(i)}; \ldots Y_{64}^{(i)}\right]$.

The feature matrix X is then fused using the multi-head self-attention mechanism (MSA). It is further mapped to (Queries) $Q \in R^{n\times d_k}$, (Keys) $K \in R^{n\times d_k}$, and (Values) $V \in R^{n\times d_v}$ using the multi-layer perceptron (MLP), where n is the batch size, d_k and d_v are the dimensions of each feature. In this model, we have $d_k = d_v = 8$. The output of the self-attention module is computed as follows Equation (4):

$$Attention\left(Q, K, V\right) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

where *soft max* denotes the softmax function, QK^T represents the dot product between Queries and Keys, and the scaling factor $\sqrt{d_k}$ is used to normalize the dot product. The result is then multiplied element-wise with V.

Transformers also employ the skip connection idea [29]. The input feature matrix is first processed with Layer Normalization (LN) to stabilize training and facilitate faster convergence. Then, it goes through the multi-head self-attention (MSA) mechanism and the multi-layer perceptron (MLP) to obtain the output Y. LN is used in each layer of the Transformer to ensure stable training. The Transformer layer can be represented as follows:

$$\mathrm{x}' = \mathrm{MSA}\left(\mathrm{LN}\left(\mathrm{X}\right)\right) + \mathrm{X} \tag{5}$$

$$\mathrm{G} = \mathrm{MLP}\left(\mathrm{LN}\left(\mathrm{x}'\right)\right) + \mathrm{x}' \tag{6}$$

where X represents the input feature matrix, which has the same dimensions as the output feature matrix G of each layer in the Transformer. The dimension of G is given by $G \in R^{1\times d}$, where d = 160.

C. TEMPORAL NETWORK

After the AC module, we utilized a Gate Recurrent Unit (GRU) cell, which is a type of recurrent neural network [34]. Similar to Long-Short Term Memory (LSTM) [35], GRU was introduced to address issues related to long-term memory and gradient vanishing or exploding during backpropagation. Compared to LSTM, GRU has a simplified internal structure by incorporating update gates and reset gates to control the flow and retention of information, leading to improved accuracy.

When dealing with sequential data, the computation process of Gated Recurrent Unit (GRU) can be simplified into two steps: update gate and reset gate computation. The update gate z_t and reset gate r_t are calculated based on the input x_t and the previous hidden state h_{t-1}. Here, σ represents the sigmoid function, and W_z and W_r are the corresponding weight matrices. The formulas for calculating the update gate and reset gate are shown in equations (7) and (8) respectively.

$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right) \tag{7}$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right) \tag{8}$$

By using the update gate and reset gate, we can compute the new hidden state h_t by combining the candidate hidden state $(h_{t'})$ and the previous hidden state h_{t-1}. The formulas for calculating the candidate hidden state and the new hidden

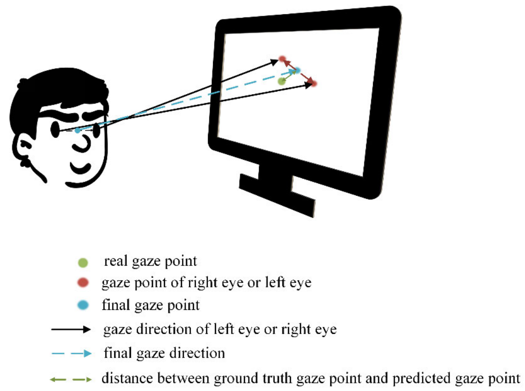

● real gaze point
● gaze point of right eye or left eye
● final gaze point
→ gaze direction of left eye or right eye
– –→ final gaze direction
←– –→ distance between ground truth gaze point and predicted gaze point

FIGURE 5. The determination of the final direction.

state are given by equations (9) and (10) respectively.

$$h_{t'} = \tanh\left(W_h \cdot [r_t \odot h_{t-1}, x_t]\right) \tag{9}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot h_{t'} \tag{10}$$

where tanh represents the hyperbolic tangent function, W_h is the weight matrix, and $\odot$ denotes element-wise multiplication (Hadamard product). In the FE-Net model, three GRU units are set up to predict the gaze direction of the third video frame given three consecutive video frames as input.

D. DETERMINATION OF GAZE DIRECTION

After obtaining the gaze directions for the left eye and right eye separately using FE-Net, we further calculate the overall gaze direction by determining the final point of gaze (PoG). Figure 5 illustrates the process of determining the final gaze direction. Firstly, the model sequentially obtains the gaze directions for the left eye and right eye. Next, we intersect the gaze rays with the screen plane. By applying camera transformations relative to the screen plane, we compute the intersection point of the gaze rays with the screen plane, which represents the coordinates of the point of gaze (PoG). The function$f$ represents the calculation process. The gaze direction for the left eye is denoted as $\hat{g}_l$, and the gaze direction for the right eye is denoted as $\hat{g}_r$. The calculation of the final PoG viewpoint $\hat{p}$ is given by Equation (11).

$$\hat{p} = \frac{1}{2}\left(f\left(\hat{g}_l\right) + f\left(\hat{g}_r\right)\right) \tag{11}$$

Finally, we convert $\hat{p}$ back to the final gaze direction $\hat{g}$ using the inverse function f^{-1}. The calculation process of $\hat{g}$ is described by Equation (12).

$$\hat{g} = f^{-1}\left(\hat{p}\right) \tag{12}$$

IV. EXPERIMENTAL RESULTS

A. DATASET

The EVE dataset [17] provides a total of 12,308,334 frames of video data, including natural eye movements of 54 participants while fixating on a screen, along with information about the screen content being observed. In this study, we only utilized the eye-tracking data from the participants in this dataset. The gaze angles in the dataset range from −60 to +60 degrees vertically and −70 to +70 degrees horizontally. The dataset also includes a significant amount of head motion. We followed the standard segmentation and reported the final results based on the test sequence as outlined in [17]. The test set labels for this dataset were not publicly available, and our testing process was conducted on the official website of the dataset.

The MPIIFaceGaze dataset [23] is one of the most widely used datasets for appearance-based gaze estimation methods. It consists of 213,659 images captured over several months in the daily lives of 15 subjects, with no restriction on head pose. The images are collected from real-world environments, providing a diverse range of lighting conditions and head poses.

B. EVALUATION METRICS

In the MPIIFaceGaze dataset, this study employed the commonly used leave-one-out criterion for gaze estimation evaluation. We used 14 subjects as the training dataset, 1 subject as the testing dataset, and calculated the average error accuracy across 15 experiments as the performance metric. For the validation of the EVE dataset, we utilized the pre-defined splits of the dataset, consisting of 39 subjects for training, 5 subjects for validation, and 10 subjects for testing. The evaluation metric used in this study is the angular error. A higher angular error indicates lower accuracy of the model. The angular error can be defined using Equation (13):

$$L_{angular} = \arccos\left(\frac{g \cdot \hat{g}}{\|g\| \, \|\hat{g}\|}\right) \tag{13}$$

where g represents the angle of the true gaze direction, while $\hat{g}$ represents the angle of the predicted gaze direction.

The gaze point distance error is represented as shown in Equation (14), which denotes the Euclidean distance between the estimated gaze point and the target gaze point.

$$L_{dist} = \|G_E - G_T\| \tag{14}$$

where G_T represents the coordinates of the actual gaze position, and G_E signifies the coordinates of the predicted gaze position.

C. MODEL CONVERGENCE ANALYSIS

We performed a convergence analysis of our FE-Net on the EVE dataset, and the results are shown in Figure 6. The error on the validation set is consistently higher than the training set, not due to parameter settings or overfitting issues, but rather due to the intrinsic nature of the problem. Each individual has a specific offset in their eye's optical axis and gaze direction, which the model can only learn relatively general features from the training set. The angular error for both the left and right eyes rapidly decreases throughout the entire iteration process and reaches its minimum value at approximately 3500 steps. Overall, the proposed network has demonstrated fast and robust convergence.

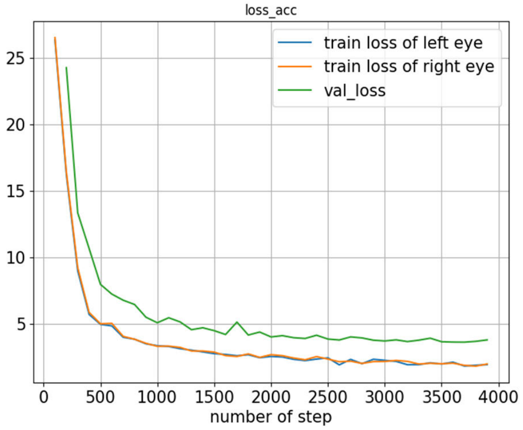


FIGURE 6. Loss curve during training.

TABLE 1. Comparison results with Other methods on the EVE dataset.

Dataset / Method	EVE	
	2D gaze location error (cm)	3D gaze direction error (degree)
Hybrid-VIT [21]	3.92	3.54
EyeNetGRU[17]	3.85	3.64
Gaze360 [15]	3.83	3.45
FE-Net (ours)	**3.55**	**3.19**

D. COMPARATIVE EXPERIMENTS ON THE EVE DATASET

The use of the EVE dataset for testing is relatively limited among existing methods. Thus, we attempted to replicate the techniques proposed in papers [21] and [15] to conduct comparative analysis. The corresponding test results are presented in Table 1.

In the method presented in paper [21], facial images were utilized as inputs, and a self-attention mechanism was employed to process the feature maps. Both paper [17] and [15] integrated time series RNN networks into their methods. The EyeNetGRU method exclusively focused on extracting temporal information from eye video frames, while the Gaze360 method solely extracted time information from facial video frames. Neither of these approaches incorporated attention mechanisms. In contrast, our proposed method integrates fused features from both facial and eye sources and includes a GRU module for dynamic feature extraction. The achieved angular error result is 3.19°, which is 0.26° lower than the error from the Gaze360 method. The 2D gaze position error is 3.55cm, showing a reduction of 0.28cm in error compared to the Gaze360 method. These final results highlight the superiority of our proposed approach.

E. ABLATION EXPERIMENTS ON THE EVE DATASET

The previous results indicated that our method achieved favorable results on the EVE dataset. However, they did not provide evidence for the effectiveness of the AC module and GRU module. Therefore, we conducted ablation experiments specifically targeting the AC module and GRU module on the EVE dataset to demonstrate their effectiveness. During the ablation process, the ECA module was replaced with an FC layer. Since the self-attention module reduces the dimensionality of the feature maps, we introduced an average pooling layer as a substitute for the self-attention module. Finally, the GRU module was replaced with an FC module.

TABLE 2. Ablation experiments based on the EVE dataset.

ECA attention	Self-attention	GRU	3D gaze direction error (degree)	2D gaze location error (cm)
√	√	√	3.19	3.55
√	√		3.27	3.64
	√	√	3.30	3.67
	√		3.38	3.76
		√	3.34	3.72
			3.46	3.85

The results of the ablation experiments in Table 2 demonstrate that the addition of the channel attention mechanism, self-attention mechanism, and GRU temporal sequence reduces the testing error from the initial 3.46° to 3.19°, resulting in a total error reduction of 7.8%. When only the GRU module was added, the error decreased by 0.12° compared to using the backbone network alone. In the presence of both attention mechanisms, the difference in error between adding and not adding the GRU module was 0.08° (3.27°-3.19°), indicating that the GRU module captures more inter-frame temporal features, leading to improved model performance. Furthermore, the addition of the AC module to the original backbone network reduced the error from 3.46° to 3.27°, resulting in a total error reduction of 0.19°. When the AC module was added in combination with the GRU module, the difference in error between adding and not adding the AC module was 0.15° (3.34°-3.19°).

In order to further demonstrate the impact of the ECA channel attention mechanism and self-attention mechanism in the AC module on the model performance, Table 3 is provided, which illustrates the effects of adding the channel attention mechanism and self-attention mechanism in the presence of GRU. Based on the testing data from the EVE dataset, it can be observed that the percentage of errors below 2°, after adding the self-attention mechanism, increased from 28.2% (FE-Net (only GRU)) to 31.5% (FE-Net (self-attention + GRU)). Furthermore, after adding the channel attention mechanism, the percentage of errors below 2° further increased to 33.5% (FE-Net (all attention)), resulting in a cumulative improvement of 5.3% compared to the case with only GRU. Additionally, when observing the percentage of

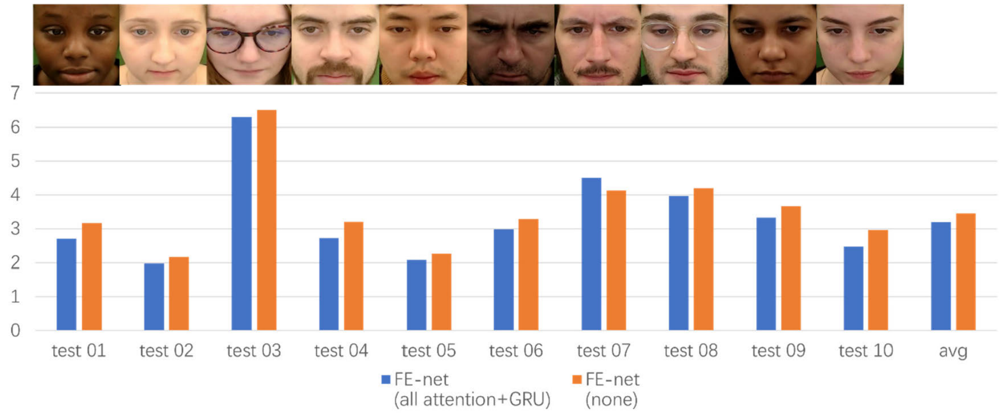


FIGURE 7. The effects of incorporating the AC module and GRU module on each individual.

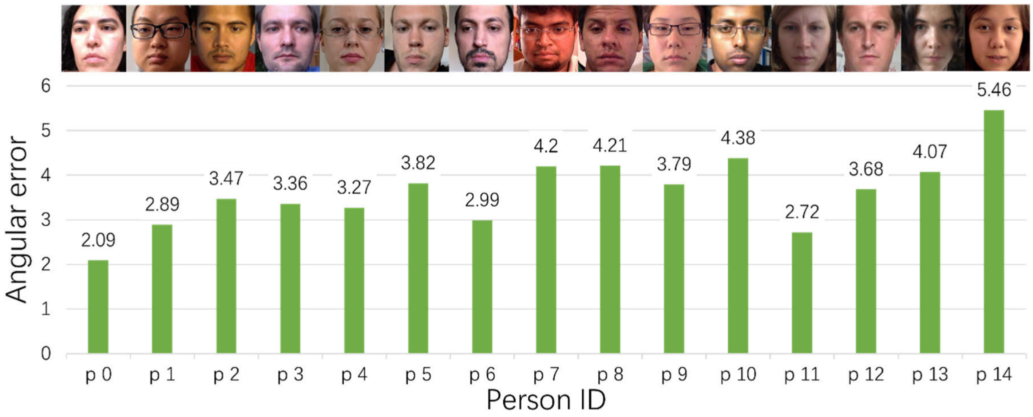


FIGURE 8. For each individual in the MPIIFaceGaze dataset.

TABLE 3. Impact of Channel Attention and Self-Attention in the AC Module with GRU on Model Performance (based on EVE dataset).

Method	Testing Error < 2° (%)	Testing Error < 4° (%)
FE-Net (only GRU)	28.2	70.5
FE-Net (self-attention + GRU)	31.5	73.0
FE-Net (all attention + GRU)	33.5	76.0

errors below 4°, it was found that adding the self-attention mechanism increased the percentage from 70.5% to 73%. Moreover, after adding the channel attention mechanism, the percentage of errors below 4° further increased from 73% to 76%, resulting in a cumulative improvement of 5.5% compared to the case with only GRU. These results provide evidence that the addition of the channel attention mechanism and self-attention mechanism contributes to the improvement of model performance.

We also conducted individual testing and comparisons on each person in the test set. Figure 7 presents the comparative results of FE-Net with the addition of the AC module and GRU module, compared to FE-Net with only the backbone network, for each test subject. The evaluation results of gaze estimation for the ten test subjects indicate a significant improvement in accuracy with the addition of the AC module and GRU module for nine of them, while only test 07 showed a slight decrease in performance. Due to individual differences, we cannot guarantee that our proposed method is suitable for every person. However, based on the data results, it can be observed that our proposed approach, which incorporates attention mechanisms and the GRU module, is effective for the majority of individuals, resulting in a cumulative average error reduction of 7.8% compared to FE-Net with only the backbone network. These findings provide strong evidence for the meaningfulness of adding the AC module and GRU module.

The EVE dataset consists of eye images and facial images captured from various angles, with generally good lighting conditions. This enables obtaining decent accuracy even in scenarios where there are significant variations in the test subjects' poses. However, the dataset exhibits limited robustness to changes in lighting conditions. For example, in Figure 7. test03 demonstrates a case where the experimental participant

TABLE 4. Performance of FE-Net on the MPIIFaceGaze dataset.

Method	Input image	MPIIFaceGaze 3D gaze direction error (degree)
GEDD-Net [31]	Face and eyes	4.5°
CA-Net [22]	Face and eyes	4.1°
AGE-Net [36]	Face and eyes	4.09°
Gaze360 [15]	Face	4.06°
Hybrid-VIT [21]	Face	4.00°
Bot2L-Net [37]	Face	3.97°
GazeNAS-ETH[26]	Face	3.96°
AFF-Net [16]	Face and eyes	3.73°
STTDN [33]	Face and eyes	3.73°
FE-Net-left eye (ours)	Face and left eye	3.74°
FE-Net-right eye (ours)	Face and right eye	3.76°
FE-Net(ours)	Face and eyes	**3.66°**

wore tinted glasses. The occlusion caused by the glasses resulted in lower brightness in the eye image. Furthermore, the refraction and reflection of light caused by the glasses' lenses further deteriorated the test performance, leading to poor results. Apart from this particular scenario, the differences in error rates among other test participants were relatively small.

F. CROSS-DATASET VALIDATION

To further demonstrate the effectiveness of FE-Net, we conducted retraining and testing on the MPIIFaceGaze dataset and compared it with the latest existing methods based on the same dataset. In Table 4, we summarized the research content of eight other methods for gaze estimation to conduct a comparison. We particularly focused on the inputs used by these eight methods and their corresponding error results. Notably, the trend in recent gaze estimation is leaning towards utilizing both facial and ocular images as inputs. Hence, we evaluated the accuracy of single-eye gaze prediction using both facial and individual ocular images, followed by the computation of the final gaze point. Our FE-Net achieved lower angular error compared to the other eight methods. Specifically, it achieved a reduction of 0.07° in error compared to the highly accurate STTDN. Clearly, our method outperforms the others.

In addition to testing the final gaze direction of both eyes, we also examined the gaze directions of the left and right eyes separately for comparison. By converting the monocular gaze lines to monocular screen gaze points, we obtained the midpoint of the predicted gaze points for both eyes and calculated the final gaze direction accordingly. As shown in the table, the error in monocular gaze direction is relatively large. However, after applying our gaze calculation strategy, the average error of the final gaze direction decreased by 0.09° compared to the monocular error. This geometric-based gaze determination strategy is computationally simple, efficient, and highly effective.

In Figure 8, we present the results of our predictions for each individual in the MPIIFaceGaze dataset. The data highlights the performance of our FE-Net model on each person's gaze estimation. Notably, FE-Net performed best on person ID p0 with an angular error of 2.09° and worst on person ID p14 with an error of 5.46°. Compared to the EVE dataset, the predictions on the MPIIFaceGaze dataset show smaller variations in errors among different individuals. We attribute this to the richer variety and lighting conditions present in the MPIIFaceGaze dataset, which contributes to its enhanced robustness and diversity.

V. CONCLUSION

Addressing the limitations of CNN's capacity to capture global relationships during feature extraction and the underutilization of temporal dynamics that lead to lower gaze estimation accuracy, this study introduces a novel gaze estimation network called FE-net, integrating channel attention and self-attention mechanisms. Through experimentation, we empirically demonstrated the effectiveness of incorporating attention mechanisms in enhancing gaze accuracy. Additionally, we incorporated a GRU module to learn the impact of temporal dynamics on gaze estimation. Comparative evaluations against state-of-the-art methods indicated that FE-net achieved cutting-edge accuracy on the publicly available EVE and MPIIFaceGaze datasets, with errors of 3.19° and 3.66° respectively. However, facial-based gaze estimation remains challenged by factors like lighting conditions, facial differences among observers, and head postures, leading to still unsatisfactory accuracy. Future endeavors will focus on robustly addressing individual calibration issues to enhance gaze estimation's robustness, and we intend to explore the integration of observed screen saliency information to further improve estimation precision.

REFERENCES

[1] P. Chakraborty, S. Ahmed, M. A. Yousuf, A. Azad, S. A. Alyami, and M. A. Moni, "A human–robot interaction system calculating visual focus of human's attention level," *IEEE Access*, vol. 9, pp. 93409–93421, 2021, doi: 10.1109/access.2021.3091642.

[2] D. Strazdas, J. Hintz, A. Khalifa, A. A. Abdelrahman, T. Hempel, and A. Al-Hamadi, "Robot system assistant (RoSA): Towards intuitive multi-modal and multi-device human–robot interaction," *Sensors*, vol. 22, no. 3, p. 923, 2022.

[3] S. N. Moral-Sánchez, M. T. Sánchez-Compaña, and I. Romero, "Geometry with a STEM and gamification approach: A didactic experience in secondary education," *Mathematics*, vol. 10, no. 18, p. 3252, Sep. 2022.

[4] G. Yuan, Y. Wang, J. Peng, and X. Fu, "A novel driving behavior learning and visualization method with natural gaze prediction," *IEEE Access*, vol. 9, pp. 18560–18568, 2021, doi: 10.1109/access.2021.3054951.

[5] L. Alam, M. M. Hoque, M. A. A. Dewan, N. Siddique, I. Rano, and I. H. Sarker, "Active vision-based attention monitoring system for non-distracted driving," *IEEE Access*, vol. 9, pp. 28540–28557, 2021, doi: 10.1109/access.2021.3058205.

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

[7] A. Nakazawa and C. Nitschke, "Point of gaze estimation through corneal surface reflection in an active illumination environment," in *Computer Vision—ECCV*. Berlin, Germany: Springer, 2012, pp. 159–172.

[8] K. A. Funes Mora and J.-M. Odobez, "Geometric generative gaze estimation (G^{3}E) for remote RGB-D cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1773–1780.

[9] Y. Xiong, H. J. Kim, and V. Singh, "Mixed effects neural networks (MeNets) with applications to gaze estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7743–7752.

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

[11] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2176–2184.

[12] Z. Chen and B. E. Shi, "Appearance-based gaze estimation using dilated-convolutions," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2018, pp. 309–324.

[13] Z. Yu, X. Huang, X. Zhang, H. Shen, Q. Li, W. Deng, J. Tang, Y. Yang, and J. Ye, "A multi-modal approach for driver gaze prediction to remove identity bias," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2020, pp. 768–776.

[14] Z. Wang, J. Zhao, C. Lu, H. Huang, F. Yang, L. Li, and Y. Guo, "Learning to detect head movement in unconstrained remote gaze estimation in the wild," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 3443–3452.

[15] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, "Gaze360: Physically unconstrained gaze estimation in the wild," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6912–6921.

[16] X. Zhou, J. Lin, J. Jiang, and S. Chen, "Learning a 3D gaze estimator with improved itracker combined with bidirectional LSTM," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 850–855.

[17] S. Park, E. Aksan, X. Zhang, and O. Hilliges, "Towards end-to-end video-based eye-tracking," in *Computer Vision—ECCV*. Glasgow, U.K.: Springer, 2020, pp. 747–763.

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

[19] Y. Luo, J. Chen, and J. Chen, "CI-Net: Appearance-based gaze estimation via cooperative network," *IEEE Access*, vol. 10, pp. 78739–78746, 2022.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Vol. 30, 2017, pp. 1–11.

[21] Y. Cheng and F. Lu, "Gaze estimation using transformer," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 3341–3347.

[22] Y. Cheng, S. Huang, F. Wang, C. Qian, and F. Lu, "A coarse-to-fine adaptive network for appearance-based gaze estimation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 10623–10630.

[23] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "MPIIGaze: Real-world dataset and deep appearance-based gaze estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 162–175, Jan. 2019.

[24] Y. Wu, G. Li, Z. Liu, M. Huang, and Y. Wang, "Gaze estimation via modulation-based adaptive network with auxiliary self-learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5510–5520, Aug. 2022, doi: 10.1109/TCSVT.2022.3152800.

[25] X. Zhou, J. Jiang, Q. Liu, J. Fang, S. Chen, and H. Cai, "Learning a 3D gaze estimator with adaptive weighted strategy," *IEEE Access*, vol. 8, pp. 82142–82152, 2020, doi: 10.1109/access.2020.2990685.

[26] V. Nagpure and K. Okuma, "Searching efficient neural architecture with multi-resolution fusion transformer for appearance-based gaze estimation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, doi: 10.1109/wacv56688.2023.00095.

[27] X. Song, S. Guo, Z. Yu, and J. Dong, "An encoder–decoder network with residual and attention blocks for full-face 3D gaze estimation," in *Proc. 7th Int. Conf. Image, Vis. Comput. (ICIVC)*, Jul. 2022, doi: 10.1109/icivc55077.2022.9886734.

[28] L. Dai, J. Liu, Z. Ju, and Y. Gao, "Attention-mechanism-based real-time gaze tracking in natural scenes with residual blocks," *IEEE Trans. Cognit. Develop. Syst.*, vol. 14, no. 2, pp. 696–707, Jun. 2022, doi: 10.1109/tcds.2021.3064280.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[30] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.

[31] Z. Chen and B. E. Shi, "Towards high performance low complexity calibration in appearance based gaze estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 1174–1188, Jan. 2023.

[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and S. Gelly, "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, Singapore, Mar. 2021, pp. 29–30.

[33] Y. Li, L. Huang, J. Chen, X. Wang, and B. Tan, "Appearance-based gaze estimation method using static transformer temporal differential network," *Mathematics*, vol. 11, no. 3, p. 686, Jan. 2023.

[34] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. NIPS*, 2014, pp. 1–9.

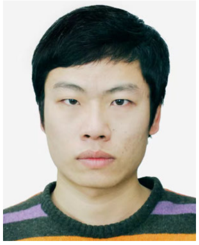
[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[36] P. Biswas, "Appearance-based gaze estimation using attention and difference mechanism," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3143–3152.

[37] X. Wang, J. Zhou, L. Wang, Y. Yin, Y. Wang, and Z. Ding, "BoT2L-Net: Appearance-based gaze estimation using bottleneck transformer block and two identical losses in unconstrained environments," *Electronics*, vol. 12, no. 7, p. 1704, Apr. 2023, doi: 10.3390/electronics12071704.

[38] Y. Cheng, F. Lu, and X. Zhang, "Appearance-based gaze estimation via evaluation-guided asymmetric regression," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2018, pp. 105–121, doi: 10.1007/978-3-030-01264-9_7.

GUOJING REN received the M.S. degree from Qingdao Agricultural University, in 2020. She is currently pursuing the master's degree in electronic and information science with Beijing Information Science and Technology University. Her current research interests include gaze estimation and machine learning.



YANG ZHANG received the B.S. and Ph.D. degrees from the Ocean University of China, Qingdao, China, in 2010 and 2016, respectively. From 2014 to 2016, he was a Visiting Scholar with the Department of Electrical and Computer Engineering, University of Miami, FL, USA. From 2017 to 2019, he was a Postdoctoral Researcher with Tsinghua University, Beijing, China. He is currently an Associate Professor with the College of Automation, Beijing Information Science and Technology University, Beijing. His current research interests include underwater vision, video coding, and the segmentation of satellite components.

QINGJUAN FENG received the M.S. and Ph.D. degrees from the Beijing Institute of Technology, in 2005 and 2009, respectively. She joined Beijing Information Science and Technology University, in 2014. She is currently an Associate Professor with the School of Automation, Beijing Information Science and Technology University. Her current research interests include human–computer interaction and image processing.

• • •