Binocular Feature Fusion and Spatial Attention Mechanism based Gaze Tracking

Lihong Dai, Jinguo Liu, *Senior Member*, *IEEE*, and Zhaojie Ju, *Senior Member*, *IEEE*

***Abstract*—Gaze tracking is widely used in driver safety driving, visual impairment detection, virtual reality, human robot interaction, and reading process tracking. However, varying illumination, various head poses, different distances between human and cameras, occlusion of hair or glasses, and low quality images, pose huge challenges to accurate gaze tracking. In this paper, based on binocular feature fusion and convolution neural network (CNN), a novel method of gaze tracking is proposed, in which local binocular spatial attention mechanism (LBSAM) and global binocular spatial attention mechanism (GBSAM) are integrated into the network model to improve the accuracy. Furthermore, the proposed method is validated on the GazeCapture database. In addition, four groups of comparative experiments have been conducted: between binocular feature fusion model and binocular data fusion model; among the local binocular spatial attention model, the local binocular channel attention model and the model without local binocular attention mechanism; between the model with GBSAM and that without GBSAM; and between the proposed method and other state-of-the-art approaches. The experimental results verify the advantages of binocular feature fusion, LBSAM and GBSAM, and the effectiveness of the proposed method.**

***Index Terms*—Gaze tracking, attention mechanism, feature fusion, CNN**

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1304600, in part by the Natural Science Foundation of China (Grant 51775541,51575412,52075530), in part by the CAS Interdisciplinary Innovation Team under Grant JCTD-2018-11, and in part by the AiBle project co-financed by the European Regional Development Fund. *(Corresponding author: Jinguo Liu, Zhaojie Ju.)*

L. Dai is with the State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, and also with institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110169, China. Moreover, L. Dai is with University of the Chinese Academy of Sciences, Beijing 100049, and also with School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan 114051, China (e-mail: dailihong2004@163.com).

J. Liu is with the State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, and also with institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110169, China (e-mail: liujinguo@sia.cn).

Z. Ju is with the State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, and with institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110169,

I. Introduction

Gaze tracking is widely used in driver safety driving[1], visual impairment detection[2], virtual reality[3], human robot interaction[4] and reading process tracking[5]. However, accurate gaze tracking is a challenging task. On the one hand, when the distance between the person and the camera becomes longer, the image resolution will decrease, and the important information contained in the image will also decrease, which shows that it is undoubtedly difficult to track the gaze in low-quality images [6]. On the other hand, due to the influence of light changes, head pose changes, hair or glasses occlusion or other factors, it brings challenges to accurate gaze tracking [7]. Therefore, how to accurately carry out gaze tracking in low quality images with variable illumination and pose is a issue to be addressed. In view of this, the methods of gaze tracking are studied. Besides, because the attention mechanism is conducive to the extraction of useful important information, it is conductive to improving the performance of gaze tracking. Therefore, the attention mechanism is also studied.

A. Gaze Tracking Related Work

Gaze tracking approach can be classified into feature regression-based one, model-based one and appearance-based one.

In the method based on feature regression, some features need to be explicitly extracted, and polynomial is normally used to fit the relationship between the features and gaze points. Then by multiple groups of known features and their corresponding gaze points, the parameters in the polynomial are determined by regression method. Finally, the gaze points can be predicted by the features and the polynomial. Zheng et al. adopt the regression based approach to determine the correspondence between head pose and gaze direction by second-order polynomial fitting, and obtain the gaze by head pose[8].

In the model-based method, the 3D geometric model of the human eye needs to be established, and the relevant parameters in the model need to be determined by calibration, and then the gaze point is obtained by the model. Mora et al. adopt the model-based method to deal with the head pose and gaze direction under the universal framework, and conduct semantic segmentation for the eye image, which has certain adaptability to different conditions [9].

The above two methods require higher image quality, so it

is difficult to accurately locate the gaze when the distance between the person and the camera is far or the resolution of the camera is low. Thus, the appearance-based method attracts more and more researchers.

In the appearance-based method, the image is taken as the input and the gaze as the output. The relationship between the image and the gaze is determined by learning. Compared with the former two methods, in the appearance-based approach, the features will be automatically extracted during the training process. Moreover, there is no need to explicitly establish the model, and no calibration is required. Furthermore, the method is based on an end-to-end connection, which has obvious advantages for low quality images. The appearance-based method commonly includes adaptive linear regression [10], support vector machine [11] and convolution neural network (CNN). Among of them, CNN simulates the neural network of human brain, which has certain adaptability to the environment due to its learning ability. Moreover, the same neural network can achieve different functions due to the different learning contents. In addition, CNN has a certain nonlinear mapping capability, and the relationship between its input and output can be determined without explicit modeling, which can simplify the design of the gaze tracking system in a sense. In view of the above advantages of CNN, its application has become more and more extensive in recent years, and many CNN-based methods have emerged in gaze tracking.

The methods of gaze tracking with CNNs can be divided into three categories. One is to capture the face image with the built-in camera on a phone or a tablet. The output of the network model is a 2D gaze point on the screen. The second type is to use the webcam on a laptop or a desktop computer to capture the face image. The output is the 3D gaze direction converted from the gaze point on the screen. The third category is to use a mobile phone to take the image in the natural scenes. The output is the position of the gaze point in the image.

The methods in [12] and [13] belong to the first category. Krafka, K. et al. [12] adopt AlexNet network framework [14], in which left and right eyes images, face grid, and face image are taken as input, and the gaze point as output, to carry out gaze tracking on mobile phones. Without image augmentation or calibration, the test error on the cellphones is 2.04 cm, and that on the tablets computer is 3.32 cm. In the entire GazeCapture database, 85% of the images are captured with the cellphones, while the remaining 15% are taken with the tablets. The test errors on the cellphones and on the tablets are multiplied by their respective percentages to obtain the average test error in the whole database. By computing, the average test error on the whole GazeCapture database [12] is 2.23 cm. When the training and test samples are augmented, the test errors on the mobile and tablet are 1.77 cm and 2.83 cm respectively, and the average test error on the whole database is 1.93 cm. Similar to [12], in [13], AlexNet is replaced by three ResNet-18 [15], in which a higher accuracy is obtained on the 15,000 samples of the GazeCapture database [12] .

The approaches in [16], [17], [18] and [19] belong to the second type. In [16], two CNNs are employed to learn the head pose and the gaze direction in the head coordinate system respectively, and then they are fused by a gaze transformation layer to predict the gaze in the camera coordinate system. On the MPIIGaze database [20], the prediction error reached 5.6° on the lightweight AlexNet network and 4.3° on a deeper BN-Inception network [21]. In [17], left and right eye images and head image are used as input, and three VGG-16 networks are used for learning respectively. On this basis, binocular feature spectrum weight learning module is added to model in order to solve the problem of binocular asymmetry. Cross-validation test is conducted on the MPIIGaze database, and the average angle error is 4.742°. In [18], two networks of AR-net and E-Net are designed to predict the 3D gaze direction. The input of the former network is left eye and right eye images and head pose vector, and that of the latter is left eye and right eye images. The output of the former is 3D gaze direction, and that of the latter is the probability of the influence of left and right eyes on gaze. The loss functions in the two networks are coupled and correlated to reduce the influence of binocular asymmetry on gaze accuracy. Training is performed on the UT Multiview database, and cross-database cross-validation is performed on the MPIIGaze and EYEDIAP databases. In addition, multiple cameras are adopted in the appearance-based method in [19], in which multi-task learning method is used. Eye images on MPIIGaze and ShanghaiTechGaze databases are taken as the input, ResNet-34 residual neural network and feature fusion network are used for learning. Meanwhile, gaze direction and gaze point are taken as the output. The gaze estimation error on the MPIIGAZE database is 4.55°.

The [22], [23] and [24] can be classified as the third one, which are suitable for gaze tracking in natural scenes. In [24], the main branch is the cascade one of face image and head position, the original image branch is the weight one. In the two branches, the CNNs with residual blocks are adopted. Then they are multiplied element by element to give prominence to more important information affecting gaze. In the end, the fully connected layer (FC) is used to obtain the 2D gaze point. The description of the methods in [22] and [23] can be found in [24].

The proposed method belongs to the first class. The image are captured with the mobile device. Furthermore, the image is that of the frontal face with left and right eyes. The output of the model is the position of the gaze point on the screen of the mobile device.

B. Attention Mechanism Related Work

Because context information is easily discovered by human attention mechanisms, and important features can be extracted by paying more attention to useful information, attention mechanism is now widely used.

In the aspect of detection, Li et al. propose a HAR-Net for object detection, in which a mixed attention learning method with three attention mechanisms (channel attention, spatial attention and aligned attention) is used [25]. Zhao et al. apply

attention mechanism into saliency detection [26]. The channel attention mechanism is adopted in the high-level feature branch, and the spatial attention mechanism is employed in the low-level one. Finally, the two branches are integrated to achieve saliency detection.

On the recognition side, Yan et al. add two extra branches on the basis of the original main branch, in one of which the scene-level attention is used to obtain the global context information; in the other of which the region-level attention is adopted to obtain the local context information. Finally, the three branches are combined to identify the action [27]. For the recognition of facial expression, Li et al. apply attention mechanism to extract more important features, so as to improve the accuracy of expression recognition [28]. Jiao et al. apply the attention mechanism into pedestrian re-identification, and the whole network is divided into three branches: the main one which is the global characteristic branch, the local feature one in which a region-level attention mechanism is adopted to obtain local features, the fine feature one in which the pixel-level attention mechanism is used to achieve more fine-grained information and reduce the error of regional attention [29]. Furthermore, attention mechanism is also used in target recognition in visual navigation of industrial robots [30].

In terms of segmentation, Pei et al. apply attention mechanism into the segmentation of colorectal tumor, and add a dual attention module between encoder and decoder to obtain more context information from the deep network, thus solving the difficulties of segmenting irregular colorectal tumor [31]. Zhou et al. take four multimodal MRI images as input, and obtain the characteristic representation of each modal by four encoders respectively. Then, the most important and useful features are extracted by the spatial and channel attention modules, and then the fusion is carried out. Finally, the decoder module is adopted to obtain the segmentation of brain tumor [32].

In other aspects, Chu et al. apply attention mechanism to human pose estimation, and adopt the holistic attention module which pays attention to the overall consistency of human body and the local attention module which pays attention to different body parts [33]. Zhang et al. employ the face image as the input of network, and the spatial weight module is adopted after the AlexNet network, and finally the FC layers are used to estimate the gaze [34]. In the spatial weight module, the spatial attention mechanism is adopted. Because different regions of the face image are of distinct importance to the gaze, the spatial weight module is used to distinguish them, which improves the accuracy of the gaze tracking. In addition, attention mechanism is also applied to sentiment analysis [35], semantic matching in community question and answering systems [36], and video captioning [37].

C. Main Contributions of the Paper

In order to effectively and accurately track the gaze in low-quality images, a gaze tracking method based on CNN is proposed, in which binocular feature fusion and spatial attention mechanism (SAM) are adopted. The main contributions of the paper are summarized as follows.

1) The asymmetry of binocular image quality caused by illumination and occlusion affects the accuracy of gaze tracking. To this end, the local binocular spatial attention mechanism (LBSAM) is adopted, which is used to distinguish the importance of different regions of the two eyes, in order to improve the accuracy of gaze tracking.
2) The method of global binocular spatial attention mechanism (GBSAM) is proposed, and the cascaded branch of face image and head position is spatially weighted by the binocular images branch, which effectively improves the accuracy of gaze tracking.
3) The mechanism of binocular feature fusion is adopted. The comparison experiment between the models with binocular feature fusion and with binocular data fusion is conducted, which verifies the former is superior to the latter.

In addition, a comparative experiment is conducted on the LBSAM and the local binocular channel attention mechanism (LBCAM), which verifies the advantages of the LBSAM.

The remaining sections are arranged as follows. In Section 2, the proposed gaze tracking method is introduced, including binocular feature fusion and data fusion mechanism, LBSAM and GBSAM. In the third section, the database and evaluation metric, implementation details, comparative experimental results of various models (including binocular fusion mechanism, local and global binocular attention mechanism). The discussions are presented in Section 4. The conclusion and future work are given in Section 5.

II. Proposed Gaze Tracking Method

In the proposed gaze tracking method, the ResNet-50 CNN is used as the basis, and the feature fusion mechanism is used to fuse the features of both eyes. In addition, the SAM is integrated into the network model. The LBSAM and the GBSAM are used to improve the accuracy of gaze tracking.

A. Proposed Gaze Tracking Approach

The overall framework of the proposed gaze tracking approach is shown in Fig. 1.

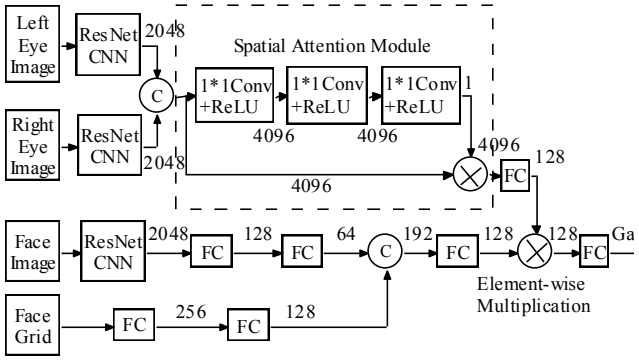

Fig. 1. The overall framework of the proposed gaze tracking approach

The "C" in the circle in Fig. 1 represents cascade. The numbers at the arrows indicate the number of output feature

layers. The whole gaze tracking system can be divided into two branches, the one of which located in the lower part is the main branch, and the other of which in the upper part is the binocular weight branch. In the main branch, firstly the ResNet CNN is used to extract the feature from the face image, and then the FC layers are adopted to reduce dimension. Meanwhile, face grid representing head position is also taken as the input, and then the dimension is reduced by the FC layers. After that, they are cascaded to obtain the output of the main branch. While, in the binocular weight branch, the left and right eyes are first cascaded by the ResNet CNN to realize the binocular feature fusion. Then the LBSAM is adopted to solve the problem of binocular asymmetry caused by varying illumination, specular reflection and occlusion, so as to enhance the accuracy of gaze tracking. Furthermore, when the image quality is poor, some useful information of the face image will be lost, resulting in the decline of the gaze tracking accuracy. However, the binocular images contain more useful information related to the gaze and can better show the details of the gaze. Therefore, the GBSAM is proposed, in which the two outputs from the main branch and the binocular weight branch are multiplied element by element. In doing so, the important regions that can reflect gaze in the main branch are highlighted, so as to effectively improve the performance of gaze tracking. Finally, the gaze point is obtained after dimensionality reduction by the FC layer. Next, the fusion mechanism, local binocular attention mechanism and GBSAM are elaborated in detail.

B. Binocular Fusion Mechanism

In CNN-based methods, there are two common fusion methods: one is data fusion, the other is feature fusion. In this paper, the binocular feature fusion method is adopted, in which the position and pose features of double eyes are extracted by two ResNet CNNs respectively. Then they are fused, as shown in a) of Fig. 2. In the binocular data fusion method, left eye and right eye images are cascaded, and then feature extraction is carried out by the ResNet network, as shown in b) of Fig. 2.

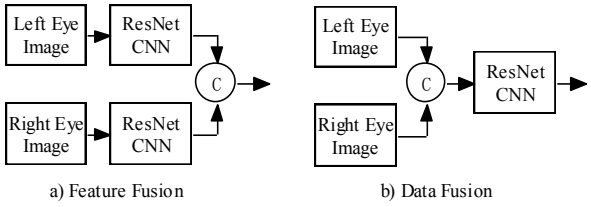

Fig. 2. Binocular fusion method

In the method based on binocular data fusion, the binocular data are cascaded together for processing, which ignores the difference between the two eyes. By contrast, in the method based on binocular feature fusion, the left and right eye images are processed separately, which highlights the difference in the details of the left and right eyes, so that the accuracy of gaze tracking is higher. In the latter section, the comparative experiments between the two fusion methods are carried out to verify the advantages of the feature fusion mechanism.

C. Local Binocular Attention Mechanism

Affected by illumination, occlusion, specular reflection and other factors, the quality of binocular images may be different, leading the problem of binocular asymmetry, which makes the influence of both eyes on gaze tracking different. Besides, due to the influence of noise and interference, different regions of eyes have different effects on gaze tracking. In order to reflect the different importance of both eyes for gaze tracking, local binocular attention mechanism is adopted. There are two attention mechanisms that can be adopted, one is spatial attention mechanism, and the other is channel attention mechanism.

1) Spatial Attention Mechanism (SAM): As shown in Fig. 1, the part in the dotted box is the spatial attention module. The input of the module is the feature map obtained by binocular feature fusion. Then it is divided into two branches, the lower branch is the input feature maps. The above branch, which is composed of three 1 × 1 convolution and ReLU activation functions, produces the weight maps. The number of feature maps output by the first two convolution remains unchanged, and that output by the last convolution is reduced to one. After that, the number of weight maps is extended to be the same as that of the input feature maps, and the weight of each output channel is the same, so that each channel of output weight maps corresponds to that of the input feature maps. Only the weights corresponding to different regions in each channel of the input feature maps are different. Then the feature maps and the weight maps are multiplied element by element to produce the output, which makes the regions greatly affecting the gaze direction given higher weights in order to differentiate the importance of different regions. In addition, the ReLU activation function increases the nonlinear expression ability of the model.

It can be seen that the binocular local spatial attention model has the following advantages. On the one hand, different regions of binocular feature maps are enhanced or suppressed by the generated spatial weight, which weakens the influence of noise and interference in binocular image on gaze tracking. On the other hand, the importance of two eyes is implicitly distinguished by the spatial weight module, which alleviates the problem of binocular asymmetry and boosts the accuracy of gaze tracking.

2) Channel Attention Mechanism (CAM): The channel attention mechanism is adopted in the SENet module, as shown in Fig. 3.

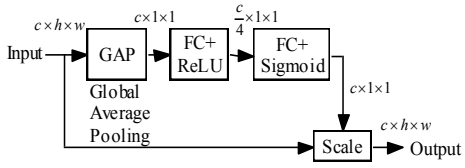

Fig. 3. Channel attention module

Similarly, the input of this module is still the feature maps after binocular feature fusion. The difference is that the

weight channel consists of global average pooling (GAP), FC layer, ReLU, FC layer, and Sigmoid. Firstly, the GAP operation is used to compress the features, and each feature channel is compressed into a real number, as shown in the Fig. 3. Then the dimension is reduced to 1 / 4 of the original by the FC layer. After that, the dimension is increased by a FC layer and transformed into the original one. The ReLU nonlinear activation function is added in the middle of the two FC layers, which can properly fit the complicated correlation between different channels and improve the nonlinear adaptability of the model. Then, the weights of *c* channels are generated by Sigmoid activation function, which are normalized weights between 0 and 1. Finally, by the scale operation, the input binocular feature maps are weighted channel by channel, and different weights are given to different feature channels, so as to improve the performance of gaze tracking.

D. Global Binocular Spatial Attention Mechanism (GBSAM)

In addition to the local binocular attention mechanism, the GBSAM is also used. In order to explain the strategy more clearly, the model without GBSAM is shown in Fig. 4. The upper part is still the binocular feature branch, and the lower part is still the main branch with the face grid and face image as input. The difference is that in the model without GBSAM, the main branch and binocular feature branch are connected in a cascade way. In the model with GBSAM (shown in Fig. 1), the two branches are connected by multiplying element by element.

The face image contains the information of face pose, eye position and eye pose. Moreover, the face grid contains the information of head position and head size. Thus, the main branch after cascading these two parts contains the complete gaze information. While, the left eye and right eye images contain the more detailed information of the eye position and eye pose, and are an important factor to reflect the gaze. With GBSAM, the main branch is weighted by binocular feature branch, which makes the important regions affecting the gaze direction given more weights, so as to effectively enhance the accuracy of gaze tracking.

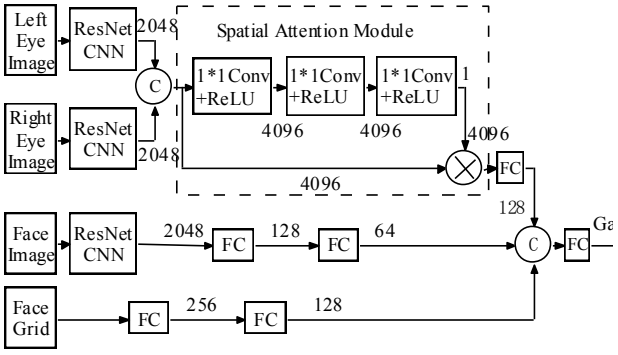

Fig. 4. Model without global binocular spatial attention mechanism

III. DATABASE AND PERFORMANCE EVALUATION

A. Database

In order to better verify the algorithm, we need to select the appropriate database. The selected database should be as close as possible to the real world with a variety of illuminations and various poses. Here we choose the large public database of GazeCapture. Besides, since we aim at the samples with face and eye images, we extract the samples in which the face and eye can be effectively detected from GazeCapture database, with a total of 1471 folders and 1490959 frames face images. Because the more samples, the longer it takes to train and test. The GazeCapture database contains a large number of samples. In order to save time, some samples are extracted to conduct comparative experiments on different models. As we know, the more samples, the better the performance of the model. Thus, reducing the number of samples generally makes the performance of the model degraded. However, because their changing trends are consistent, it does not affect the comparison and verification of various models. Here, the first 150,000 samples on GazeCapture database are selected for comparison experiments on different models, including 127,757 training samples, 15,742 test ones, and 6,501 validation ones. Finally, all 1490959 valid samples are adopted to test the proposed approach, including 1251983 training samples and 179496 test samples. The model is trained with the training samples, and tested with the test samples to get the test error. Furthermore, the advantages and effectiveness of the proposed method are verified by comparing with the other existing advanced methods.

B. Implementation Details and Evaluation Metrics

The algorithm is based on Pytorch framework. The Adam optimizer [38] with L2 regular term is adopted. The samples are fed into the network model in batches for training, so as to improve the efficiency of the algorithm with less loss of accuracy. In addition, limited by computer hardware resources, the batch size used here is set to 60.

1) loss Function: The loss function is the mean square error between the estimated gaze points and the target ones, which can be expressed as

$$Loss(w)=\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\left\|h_w(x_i)-g_i\right\|^2\right) \tag{1}$$

where w is the weight, m is the number of training samples, x is the input, $h_w(x_i)$ is estimated gaze point by the neural network, and g is the target gaze point. Then the updated weight can be calculated by

$$w_{t+1}=w_t-\eta\frac{\partial Loss(w)}{\partial w_t} \tag{2}$$

where η is the learning rate, which is set as 0.0001.

2) Regular Term to Prevent Overfitting: In order to prevent overfitting, L2 regular term is added to the loss function. Then the new loss function can be written as

$$L(w) = Loss(w) + \frac{\lambda}{2}\sum_{i=1}^{m} w_i^2 \tag{3}$$

where λ is the weight attenuation coefficient, which is set as 0.0001. Then the updated weight can be calculated by

$$w_{t+1} = w_t - \eta\left(\frac{\partial Loss(w)}{\partial w_t} + \lambda w_t\right) \tag{4}$$

where the descending gradient is given by

$$g_t = \frac{\partial Loss(w)}{\partial w_t} + \lambda w_t . \tag{5}$$

When the weight is small, the complexity of the model is low and the fitting effect is good. When the weight value increases, the complexity of the model increases, which easily leads to overfitting. By means of weight attenuation coefficient, the weight is limited, thus preventing overfitting.

3) First and Second Order Momentum Terms to Adaptively Adjust the Learning Process: The first order momentum term can be expressed by

$$g_t = g_t + \beta g_{t-1} \tag{6}$$

where β is the momentum coefficient. The added momentum term is the product of the momentum coefficient and the last updated gradient descent result. When the current gradient descent direction is the same as the last one, the momentum term accelerates the gradient descent. On the contrary, it slows down the gradient descent. On this basis, in order to reduce the impact of data fluctuations, a smooth momentum term is used here, which can be written as

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \tag{7}$$

where β_1 is the momentum coefficient, which is set to 0.9 here.

Furthermore, the second momentum term is adopted, which can be expressed as

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \tag{8}$$

where the coefficient β_2 is set to 0.999. Then the weight update formula can be rewritten as

$$w_{t+1} = w_t - \eta\left(\frac{m_t}{\sqrt{v_t} + \varepsilon}\right). \tag{9}$$

In order to prevent the denominator from being zero, a small value ε is added to the denominator. The value of ε is set to 1e-8 here. It can be seen from formula (9) that the new learning rate is equivalent to $\frac{\eta}{\sqrt{v_t} + \varepsilon}$. After introducing the second order momentum term, the faster the parameter changes, the smaller the learning rate is, and vice versa. It is obvious that the second order momentum term can adaptively regulate the learning rate. In addition, because the initial values of first and second momentum terms are 0, they are very small and close to 0 at the beginning of training. In order to overcome the problem, they are modified as

$$m_t^* = \frac{m_t}{1-\beta_1^t}, \tag{10}$$

$$v_t^* = \frac{v_t}{1-\beta_2^t}. \tag{11}$$

The final gradient update formula is

$$w_{t+1} = w_t - \eta\left(\frac{m_t^*}{\sqrt{v_t^*} + \varepsilon}\right). \tag{12}$$

To sum up, according to the six formulas of (5), (7), (8) and (10) - (12), the Adam algorithm with L2 regular term is obtained, which is the optimization method adopted here.

4) Evaluation Metric: The test error is taken as the evaluation metric. It is the mean value after the root of the square sum of error between the estimated gaze point and the target one, which can be expressed by

$$Error = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(\hat{g}_i - g_i)^2} \tag{13}$$

where $\hat{g}_i$ is the estimated gaze point. The gaze point here is the position in the camera coordinate system. The test samples are used to verify, and the test error is calculated to evaluate the performance of the models.

C. Performance Comparison of Different Models

1) Comparison of Binocular Data Fusion and Feature Fusion Models: Binocular data fusion and feature fusion models are compared. The GBSAM and LBSAM are not adopted here, that is to say, the binocular feature fusion model is to the one that remove the spatial weight module in the dotted box in Fig. 4. The binocular data fusion model is obtained by replacing a) with b) in Fig. 2. In order to save training and testing time, the first 150,000 samples in GazeCapture database are selected for the experiment, and the comparison results of binocular data fusion and feature fusion models are shown in Table I. The initials are usually used as abbreviations. However, the first letter of Feature Fusion is repeated, which is easy to be confused. In addition, Fusion and Integration have the similar meaning, so the initials of Integration are chosen instead of Fusion as the abbreviation here. Thus, Data Fusion is abbreviated as DI and Feature Fusion as FI.

TABLE I
COMPARISON OF BINOCULAR DATA FUSION AND FEATURE FUSION MODELS

Fusion method	Test Error (cm)
Data Fusion (DI)	2.5864
Feature Fusion (FI)	2.4255

It can be seen that the test error of the model with binocular feature fusion is smaller than that with binocular data fusion, which indicates the advantages of the feature fusion.

2) Comparison of Different Local Binocular Attention Models: In order to compare the local binocular attention models, comparative experiments are carried out on the models without local binocular attention mechanism, those with LBSAM, and those with LBCAM. At the same time, the two methods of binocular data fusion and feature fusion are adopted. The error comparison of different local binocular

attention models is shown in Table II, and the test error histogram is shown in Fig. 5. In the second column of Table II, N represents the model without local binocular attention mechanism, CA represents the one with LBCAM, and SA represents the one with LBSAM. The first three rows in the table are the models using data fusion, and the last three rows are those using feature fusion. The test errors of the models without GBSAM and with GBSAM are shown in the third and the forth columns, respectively.

TABLE II
ERROR COMPARISON OF MODELS WITH DIFFERENT FUSION METHODS AND LOCAL BINOCULAR ATTENTION MECHANISM

Fusion method	Local binocular attention mechanism	Test Error (cm) without GBSAM	Test Error (cm) with GBSAM
DI	N	2.5864	2.5347
DI	CA	2.5577	2.5255
DI	SA	2.5249	2.4869
FI	N	2.4255	2.3839
FI	CA	2.3920	2.3522
FI	SA	2.3831	2.3306

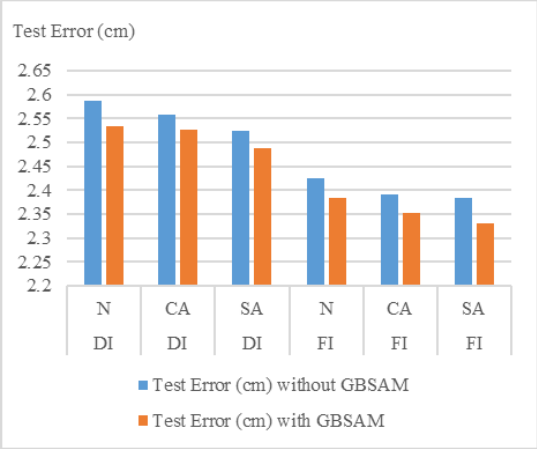

Fig. 5. Test Error comparison of models with different fusion methods and local binocular attention mechanism

By comparing the results of the last three rows in the table, it is obvious that the model without local binocular attention has the worst performance, followed by the one with LBCAM, and the one with LBSAM has the best performance. The same conclusion can be drawn by comparing the first three rows in the table. These results verify the advantages of LBSAM. Furthermore, in the first three rows of data fusion and the last three rows of feature fusion models, those with the same local binocular attention mechanism are compared. It is clear that the performance of the models with binocular feature fusion is better than that of those with data fusion, which is the same as the conclusion in the previous section. In conclusion, the model with binocular feature fusion and LBSAM has a better performance, which shows their superiority.

3) Comparison of the Models with GBSAM: In order to further enhance the accuracy of gaze tracking, on the basis of binocular feature fusion and LBSAM, the GBSAM is adopted. The model is shown in Fig. 1. The bottom branch, in which the face grid and the face image are taken as the input, is the main one of gaze tracking. The upper branch, in which the binocular images are taken as the input, is the weight one, which contains finer gaze information. The importance of the main branch is distinguished by the weight branch to obtain more accurate gaze direction. In order to verify the effect of the GBSAM on the performance of the model, the GBSAM is added to all the models in the previous section. Their test errors are shown in the last column in Table II, and the histogram is shown in orange color in Fig. 5.

It can be seen that the performance of binocular feature fusion is better than that of binocular data fusion. At the same time, the performance of the model with local LBSAM is better than that of the one with LBCAM, which is completely consistent with the conclusions in the previous section. In addition, it can be seen from Fig. 5 that the performance of the model with GBSAM is better than that of the one without GBSAM, which verifies the superiority of the GBSAM. The proposed model is shown in the last cell of Table II. It can be seen that the model with binocular feature fusion, LBSAM and GBSAM has the smallest test error and the best performance.

D. Performance Comparison with Other Gaze Tracking Methods

In order to further verify the effectiveness of the proposed approach, it is compared with other advanced approaches. The methods in [34] and [13] are re-implemented, and some relevant other comparative experiments are also conducted. They are all trained and tested on the first 150,000 samples of GazeCapture, and their test errors are shown in Table III.

TABLE III
COMPARISON WITH THE ADVANCED METHODS ON THE 150000 SAMPLES OF GAZECAPTURE DATABASE

Method	Test Error (cm)
Face Spatial Weight (Alex) [34]	3.2740
Face Spatial Weight (ResNet-50)	2.6630
Face Spatial Weight with Face Grid (ResNet-50)	2.5491
Face Image, Face Grid, Eye Images (ResNet-18) [13]	2.4402
Face Image, Face Grid, Eye Images (ResNet-50)	2.4255
Proposed Method	2.3306

In [34], only face images are used as input, and there is no face grid or eye images branch. Moreover, the spatial attention mechanism is adopted in the face branch, and Alex network is used. The test error is shown in the first row of Table III. The network model in the second row of the table is similar to it, except that ResNet-50 is used instead of Alex network. The results of the first two rows show that compared with the Alex network, because of the advantage of the residual block in ResNet-50, the network can reach a deeper layer, and then has a stronger learning ability, so the test error of the second row is dramatically reduced and its performance is significantly improved. On the basis of the model in the second row, the face grid branch is added to the model in the third row. It is clear that the performance of the model is further improved. For the model in the second row, only the face image is used to estimate the gaze, which will inevitably cause error, because of the lack of head position information. In contrast, because the head position information is implied in the face grid, adding the face grid into the network model will improve the performance in the third row. The model

adopted in [13] is similar to the one without GBSAM or LBSAM in the paper. The difference is that ResNet-18 is used in the former, while ResNet -50 is used in the latter. Both results are shown in the forth and the fifth of Table III. It can be seen that because the bottleneck residual block is used in the ResNet-50, the network can reach a deeper layer while improving the real-time performance, thus improving the performance of the network. Next, the third and fifth rows are compared, in which ResNet-50 is all used as the base network. It can be seen that the performance of the former is worse, which shows that although the SAM contributes to improving the performance of the model, it is slightly inferior due to the lack of the eye images branches. While, because eye images contain more important details related to the gaze, especially when the distance between human and camera is far or the quality of face images is poor, eye images play a vital role in improving network performance. Furthermore, by comparing the last two rows in the table, it is clear that the performance of the proposed method is significantly improved, which fully verifies the superiority of GBSAM and LBSAM.

In addition, the methods on the whole GazeCapture database are compared, shown in Table IV. The results show that the performance of the proposed method has outperformed that of the other state-of-the-art approaches. In addition, the method in [12] is compared with that in the paper. In [12] , Alex is used as the basic network, and face images, face grid and left and right eye images are used as the input. In the paper, we replace the Alex network with ResNet-50, add the LBSAM module, and adopt the GBSAM. The accuracy of the proposed method without data augmentation is higher than that of iTracker without or even with augmentation, which has further verified the advantages of SAM.

TABLE IV
COMPARISON WITH THE STATE-OF-THE-ART APPROACHES ON THE WHOLE GAZECAPTURE DATABASE

Method	Test Error (cm)
iTracker without augmentation [12]	2.23
iTracker with augmentation [12]	1.93
SD [39]	1.97
TAT [40]	1.95
Proposed Method	1.86

IV. DISCUSSION

A. The Advantages of Binocular Feature Fusion

There are two main fusion methods: data fusion and feature fusion. For binocular data fusion, the binocular images data are fused together and then fed into the ResNet network, in which the difference between the two eyes is ignored, resulting in poor accuracy of gaze tracking. For binocular feature fusion, the left and right eye images are respectively fed into the ResNet network and then fused. The difference between the two eyes is distinguished, which contains more useful information, and the accuracy of gaze tracking is higher. A large number of experiments on the GazeCapture database also verify that the performance of binocular feature fusion model is exceedingly better than that of binocular data fusion one.

B. The Difference between SAM and CAM

The attention mechanism here is not the focus of attention in the usual sense. The usual focus of attention represents the target of the user's gaze. While the attention mechanism refers to the degree of attention, which can be understood as a weighting mechanism. The SAM and CAM are used in the binocular cascade branch, which are called LBSAM and LBCAM, respectively. Both SAM and CAM are attention mechanism, in which important information can be paid more attention. Thus, the performance of the model with LBSAM or LBCAM will be improved compared with that without attention mechanism.

However, they focus on different perspectives. In CAM, the focus is on feature channels, and different weights are given to different feature channels. In SAM, the focus is on different regions. The weights of different channels are the same, but those of different regions in each channel are different, which makes the importance of different regions distinguished.

Due to the influence of illumination, occlusion and specular reflection, the image quality of some eye regions will degrade. With LBSAM, the weights in these eye regions can be reduced, while other eye regions can be given higher weights and attention. Therefore, the influence of not only the above interference but also binocular asymmetry can be reduced, so as to improve the accuracy of gaze tracking. The experimental results also prove that the model with LBSAM has the best performance, so SAM is more suitable for the practical application here.

Furthermore, though the influence of binocular asymmetry can also be suppressed in [18], that of the above interference can not be reduced. The reason is that the attention mechanism is not adopted in [18] and the design idea is also different from the proposed method. The details can be found in the introduction section on Page 2.

C. The Importance of Attention Mechanism and the Model Input

The importance of attention mechanism and model input can be seen in comparison to other approaches.

The proposed method is compared with that in [24]. Both are suitable for different applications, and the input and output of the models are also different. The method in [24] is used to track the gaze in natural scenes, so the captured image may be a back or a side face. The output of the model is the position of the gaze in the image in natural scenes. The input of the model is the original complete image, face image and head position. In contrast, the method proposed in this paper aims at the situation where the frontal face can be captured. The output of the model is the position of the gaze point on the screen of a mobile device. The input of the model is left and right eye images, face image and face grid, as shown on the right side of Fig. 6.

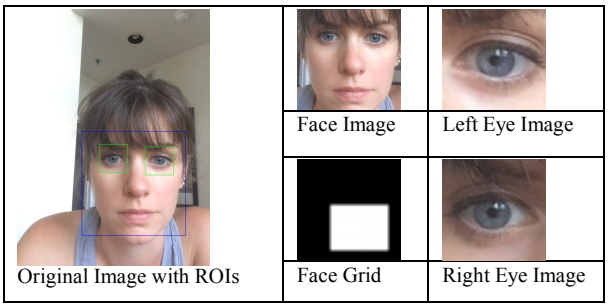

Fig. 6. A visual input image example of the model in the proposed method

The image on the left side of Fig. 6 is the original one with ROIs (regions of interest) of face and eyes. The size of the original input image is 480 × 640. The face, left eye and right eye images are extracted from the input image. Then they are resized to 224 × 224 and fed into the network model. The face grid size is 25 × 25, which represents the size and position of the head relative to the original image. There are also similarities between the two methods. The GBSAM in the proposed method is similar to the attention mechanism in [24]. The cascade branch of face image and head position is the main one. Its output is multiplied element by element with the weight, so that the important information that determines the position of the gaze point is extracted to obtain more accurate gaze point. The difference is that the weights are designed differently. In [24], since the original image contains the position information of gaze point, the output of the original image branch is designed as the weight. In the proposed method, the left and right eye images can be extracted from the frontal face images, so the output of binocular cascade branch is taken as the weight. Because the left and right eye images contain more important details, which directly reflects the position of the gaze point, the accuracy of the gaze tracking is bound to be improved. Furthermore, in the proposed method, LBSAM is used. The advantages are shown in the previous section.

In order to visually represent the spatial weights, the feature maps of local binocular spatial weights and global binocular spatial weights for the image in Fig. 6 are drawn here, as shown in Fig. 7 and Fig. 8 respectively.

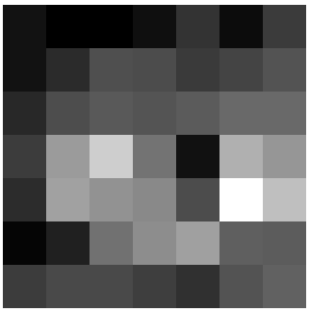

Fig. 7. Feature map of local binocular spatial weights

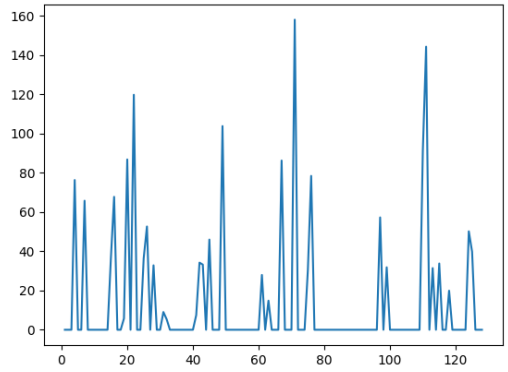

Fig. 8. Feature map of global binocular spatial weights

As shown in Fig. 1 of Section 2, after the features of the left and right eye images are extracted by ResNet-50 respectively, the corresponding features with a size of 7 × 7 and 2048 channels are generated, and then the binocular combined features with 4096 channels are obtained after cascade. Then, by the operation of three 1 × 1 convolution and ReLU, a gray image with a size of 7 × 7 and 1 channel is generated, which is the feature map of local binocular spatial weights, as shown in Fig. 7. After that, the number of the spatial weights is extended to 4096. Then, the binocular combined features are multiplied element by element with the local binocular spatial weights to extract useful information. As can be seen from Fig. 7, the regions with larger gray values correspond to the larger weights. The weight in the white region is the largest, and the binocular combined feature of the corresponding region is extracted to the greatest extent, while the weights in the black regions are zero, and the corresponding regions contain useless information and are ignored. Then, after dimensionality reduction by the FC layer, the 1D features with 128 channels are generated, which are the global binocular spatial weights, as shown in Fig. 8. The horizontal axis represents each channel and the vertical axis represents the global spatial weight. After that, the main features (the combined features of face image and face grid) are multiplied element by element by the global binocular spatial weights to filter the information. It can be seen that the main features corresponding to the channels (such as channel 80 and channel 90) with zero weight are filtered out, while those corresponding to the other channels with different weights are extracted. Finally, after dimensionality reduction by the FC layer, the gaze position is obtained. In conclusion, due to the weighting effect of LBSAM and GBSAM, information in features is filtered and extracted, which effectively improves the accuracy of gaze tracking.

In addition, The proposed method is compared with that in [34]. In [34], due to the lack of head position information in the face image, it causes a large error to estimate the gaze by only a face image. In contrast, in the proposed method, face grid with head position information is also used as the input of the model to make up for the above deficiencies. Moreover, the left and right images with more gaze detail information

are also adopted as the input so as to further improve the performance.

V. Conclusions and Future Work

Gaze tracking is widely used. However, due to the influence of many factors, such as varying illuminations, various head pose, different distance between human and camera, the occlusion of hair or glasses, and low-resolution images, it is a great challenge to accurately carry out the gaze tracking. In view of this, we propose an approach based on binocular feature fusion and CNN, in which LBSAM and GBSAM are integrated into the network model to boost the accuracy of gaze tracking. The input of the model is left and right eye images, face image and face grid. The output of the model is the position of the gaze point on the screen of a mobile device.

The models with binocular feature fusion and data fusion are compared on the GazeCapture database. The experimental results show that the performance of the former is better than that of the later. The model with LBSAM, the one with LBCAM, and the one without attention mechanism are compared. The experimental results verify the superiority of the proposed LBSAM. The comparative experiments on the models with GBSAM and without GBSAM are carried out. The experimental results show that the accuracy of gaze tracking is higher with GBSAM. The test error on the whole GazeCapture database is 1.86 cm in the proposed method with binocular feature fusion, LBSAM and GBSAM. It is compared with the other existing advanced approaches. The result shows that the performance of the former has outperformed that of the latter, which has verified the effectiveness of the proposed approach.

However, since the proposed method is applicable to the case where there are face images or even binocular images, the gaze tracking of the side face or the back image cannot be carried out. Furthermore, data augmentation methods are not used in the paper, and the accuracy of gaze tracking needs to be further improved. In addition, this paper only focuses on the gaze tracking for static images, instead of using video frame images for dynamic tracking, which will be our future work. Besides, we will further optimize the algorithm to improve its speed in the future.

References

[1] R. A. Naqvi, M. Arsalan, G. Batchuluun, H. S. Yoon, and K. R. Park, "Deep Learning-Based Gaze Detection System for Automobile Drivers Using a NIR Camera Sensor," *Sensors,* vol. 18, no. 2, Feb 2018, Art. no. 456.

[2] A. P. Yow, D. Wong, H. Liu, H. Zhu, I. J. Ong, A. Laude *et al.*, "Automatic visual impairment detection system for age-related eye diseases through gaze analysis," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2017, pp. 2450-2453.

[3] J. Orlosky, Y. Itoh, M. Ranchet, K. Kiyokawa, J. Morgan, and H. Devos, "Emulation of Physician Tasks in Eye-Tracked Virtual Reality for Remote Diagnosis of Neurodegenerative Disease," *IEEE Transactions on Visualization and Computer Graphics,* vol. 23, no. 4, pp. 1302-1311, 2017.

[4] J. Guo, Y. Liu, Q. Qiu, J. Huang, C. Liu, Z. Cao *et al.*, "A Novel Robotic Guidance System With Eye-Gaze Tracking Control for Needle-Based Interventions," *IEEE Transactions on Cognitive and Developmental Systems,* vol. 13, no. 1, pp. 179-188, 2021.

[5] S. Bottos and B. Balasingam, "Tracking the Progression of Reading Using Eye-Gaze Point Measurements and Hidden Markov Models," *IEEE Transactions on Instrumentation and Measurement,* vol. 69, no. 10, pp. 7857-7868, 2020.

[6] L. Dai, J. Liu, Z. Ju, and Y. Gao, "Iris Center Localization Using Energy Map With Image Inpaint Technology and Post-Processing Correction," *IEEE Access,* vol. 8, pp. 16965-16978, 2020.

[7] L. Dai, J. Liu, Z. Ju, and Y. Gao, "Iris center localization using energy map synthesis based on gradient and isophote," *Journal of Intelligent & Fuzzy Systems,* Article vol. 38, no. 4, pp. 4511-4523, 2020.

[8] Z. Zheng, Y. Wang, J. Barnes, X. Li, C.-H. Park, and M. Jeon, "Non-invasive Gaze Direction Estimation from Head Orientation for Human-Machine Interaction," in *Human-Computer Interaction. Interaction Technologies*, Cham, 2018, pp. 380-389: Springer International Publishing.

[9] K. A. F. Mora and J. M. Odobez, "Geometric Generative Gaze Estimation (G3E) for Remote RGB-D Cameras," in *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[10] F. Lu, Y. Sugano, T. Okabe, and Y. Sato, "Adaptive Linear Regression for Appearance-Based Gaze Estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 36, no. 10, pp. 2033-2046, 2014.

[11] Y. L. Wu, C. T. Yeh, W. C. Hung, and C. Y. Tang, "Gaze direction estimation using support vector machine with active appearance model," *Multimedia Tools and Applications,* vol. 70, no. 3, pp. 2037-2062, 2014.

[12] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik *et al.*, "Eye Tracking for Everyone," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2176-2184.

[13] E. T. Wong, S. Yean, Q. Hu, B. S. Lee, J. Liu, and R. Deepu, "Gaze Estimation Using Residual Neural Network," in *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 2019, pp. 411-414.

[14] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in neural information processing systems,* vol. 25, no. 2, pp. 1-9, 2012.

[15] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, New York, 2016, pp. 770-778.

[16] H. Deng and W. Zhu, "Monocular Free-Head 3D Gaze Tracking with Deep Learning and Geometry Constraints," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3162-3171.

[17] H. Q. Yan, " Research on Gaze Estimation in Free Posture," University of Electronic Science and Technology, Chengdu, China, 2019.

[18] Y. Cheng, F. Lu, and X. Zhang, "Appearance-Based Gaze Estimation via Evaluation-Guided Asymmetric Regression," in *Computer Vision – ECCV 2018*, Cham, 2018, pp. 105-121: Springer International Publishing.

[19] D. Lian, L. Hu, W. Luo, Y. Xu, L. Duan, J. Yu *et al.*, "Multiview Multitask Gaze Estimation With Deep Convolutional Neural Networks," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 30, no. 10, pp. 3010-3023, 2019.

[20] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 41, no. 1, pp. 162-175, 2018.

[21] C. Szegedy, L. Wei, J. Yangqing, P. Sermanet, S. Reed, D. Anguelov *et al.*, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1-9.

[22] A. Continente, A. Khosla, C. Vondrick, and A. Torralba, "Where are they looking?," in *Advances in Neural Information Processing Systems 28*, 2015, pp. 199-207.

[23] D. Lian, Z. Yu, and S. Gao, "Believe It or Not, We Know What You Are Looking At!," in *Computer Vision – ACCV 2018*, Cham, 2019, pp. 35-50: Springer International Publishing.

[24] L. Dai, J. Liu, Z. Ju, and Y. Gao, "Attention Mechanism based Real Time Gaze Tracking in Natural Scenes with Residual Blocks," *IEEE Transactions on Cognitive and Developmental Systems, doi: 10.1109/TCDS.2021.3064280*, 2021.

[25] Y. Li and S. Wang, "HAR-Net: Joint Learning of Hybrid Attention for Single-Stage Object Detection," *IEEE Transactions on Image Processing,* vol. 29, pp. 3092-3103, 2020.

[26] T. Zhao and X. Wu, "Pyramid Feature Attention Network for Saliency detection," Accessed: Jan. 3, 2020. [Online]. Available: https://arxiv.org/abs/1903.00179?context=cs.CV
[27] S. Yan, J. S. Smith, W. Lu, and B. Zhang, "Multibranch Attention Networks for Action Recognition in Still Images," *IEEE Transactions on Cognitive and Developmental Systems,* vol. 10, no. 4, pp. 1116-1125, 2018.
[28] J. Li, K. Jin, D. Zhou, N. Kubota, and Z. Ju, "Attention Mechanism-based CNN for Facial Expression Recognition," *Neurocomputing,* Jun. 2020.
[29] S. Jiao, J. Wang, G. Hu, Z. Pan, L. Du, and J. Zhang, "Joint Attention Mechanism for Person Re-Identification," *IEEE Access,* vol. 7, pp. 90497-90506, 2019.
[30] H. Li and J. Li, "Recognition of Robot Based on Attention Mechanism and Convolutional Neural Network," in *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, 2019, pp. 2578-2584.
[31] Y. Pei, L. Mu, Y. Fu, K. He, H. Li, S. Guo *et al.*, "Colorectal Tumor Segmentation of CT Scans Based on a Convolutional Neural Network With an Attention Mechanism," *IEEE Access,* vol. 8, pp. 64131-64138, 2020.
[32] T. Zhou, S. Ruan, Y. Guo, and S. Canu, "A Multi-Modality Fusion Network Based on Attention Mechanism for Brain Tumor Segmentation," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 2020, pp. 377-380.
[33] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context Attention for Human Pose Estimation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5669-5678.
[34] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 2299-2308.
[35] L. Wang and P. Liu, "A Sentiment Analysis Model Based on Attention Mechanism and Compound Model," in *2019 18th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)*, 2019, pp. 203-206.
[36] Z. Hou, X. Cai, S. Chen, and B. Li, "A model based on dual-layer attention mechanism for semantic matching," in *2019 IEEE International Conference of Intelligent Applied Systems on Engineering (ICIASE)*, 2019, pp. 105-108.
[37] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang *et al.*, "STAT: Spatial-Temporal Attention Mechanism for Video Captioning," *IEEE Transactions on Multimedia,* vol. 22, no. 1, pp. 229-241, 2020.
[38] N. Shirish Keskar and R. Socher, "Improving Generalization Performance by Switching from Adam to SGD," p. arXiv:1712.07628Accessed on: December 01, 2017Available: https://ui.adsabs.harvard.edu/abs/2017arXiv171207628S
[39] C. Yang, L. Xie, C. Su, and A. L. Yuille, "Snapshot Distillation: Teacher-Student Optimization in One Generation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2854-2863.

[40] T. Guo, Y. Liu, H. Zhang, X. Liu, Y. Kwak, B. I. Yoo *et al.*, "A Generalized and Robust Method Towards Practical Gaze Estimation on Smart Phone," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 1131-1139.

LIHONG DAI received the B.S. degree in automatic control and M.S. degree in control theory and control engineering from University of Science and Technology Liaoning, China, in 2000 and 2004 respectively. She is currently pursuing the Ph.D. degree in Pattern Recognition and Intelligent System at Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang City, China.

Since 2004, she has been a teacher in school of Electronic and Information Engineering, University of Science and Technology Liaoning, China. She is currently a Senior Lecturer. Her research interests include gaze tracking, computer vision, machine learning, pattern recognition, and their applications on human-robot interaction and collaboration.

JINGUO LIU (M'07-SM'18) received the Ph.D. degree in mechatronics from Shenyang Institute of Automation (SIA), Chinese Academy of Sciences (CAS), in 2007, where he has been a Full Professor, since January 2011. He has been the Assistant Director of the State Key Laboratory of Robotics, since 2008, and has also been the Associate Director of the Center for Space Automation Technologies and Systems, since 2015. His research interests include bio-inspired robotics and space robot. He has authored or coauthored five books, over 100 articles and holds 50 patents in above areas.

He is a Member of the IEEE Technical Committee on Safety, Security, and Rescue Robotics, a Member of the IEEE Technical Committee on Marine Robotics, and the Senior Member of the Chinese Mechanical Engineering Society. He was a recipient of the T. J. TARN Best Paper Award in Robotics from the 2005 IEEE International Conference on Robotics and Biomimetics, the Best Paper Award of the Chinese Mechanical Engineering Society, in 2007, the Best Paper Nomination Award from the 2008 International Symposium on Intelligent Unmanned Systems, the Best Paper Award from the 2016 China Manned Space Academic Conference, the Outstanding Paper Award from the 2017 International Conference on Intelligent Robotics and Applications, and the Best Paper Award from the 2018 International Conference on Electrical Machines and Systems. He services as the Associate Editor or Technical Editor of several journals such as IEEE/ASME Transactions on Mechatronics, Journal of Field Robotics, Mechanical Sciences, Science China Technological Sciences, Chinese Journal of Mechanical Engineering, and Chinese Journal of Aeronautics.

ZHAOJIE JU (M'08-SM'16) received the B.S. degree in automatic control and the M.S. degree in intelligent robotics from the Huazhong University of Science and Technology, China, and the Ph.D. degree in intelligent robotics from the University of Portsmouth, U.K. He held research appointments at University College London, London, U.K., before he started his independent academic position at the University of Portsmouth, in 2012. He has authored or coauthored over 200 publications in journals, book chapters, and conference proceedings and received five Best Paper Awards and one Best AE Award in ICRA2018. His research interests include machine intelligence, pattern recognition and their applications on human motion analysis, multi-fingered robotic hand control, human–robot interaction and collaboration, and robot skill learning.

Dr. Ju is an Associate Editor of several journals, such as IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS and Neurocomputing.