IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 1 June 2022, accepted 23 July 2022, date of publication 27 July 2022, date of current version 1 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3194123

RESEARCH ARTICLE

CI-Net: Appearance-Based Gaze Estimation via Cooperative Network

YUAN LUO, JIANGTAO CHEN, AND JIAN CHEN

Key Laboratory of Optoelectronic Information Sensing and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Corresponding author: Jiangtao Chen (1185007137@qq.com)

This work was supported by the National Nature Science Foundation of China under Grant 61801061.

ABSTRACT Facial occlusion and different appearances of both eyes in natural scenes can affect the accuracy of gaze estimation based on appearance. Therefore, this paper proposes a gaze estimation model based on cooperative network: CI-Net, including a consistency estimation network (C-Net) and inconsistency estimation network (I-Net). C-Net is used to estimate the Main gaze of the true gaze, and an attention mechanism is added to adaptively assign the weight between eyes and face features. The I-Net is used to estimate the Residual residuals based on true gaze. In addition, Cross attention module is designed in this paper, through which I-Net can selectively obtain information from C-Net, to obtain more accurate eyes directions. The experimental results in this paper show that the CI-Net gain lower angle errors than the current mainstream CNN methods under the condition of different appearance of both eyes and facial occlusion.

INDEX TERMS Gaze estimation, deep learning, main gaze, residual residuals.

I. INTRODUCTION

Gaze estimation is the process of predicting gaze direction and locating gaze points. As an important research topic in the computer vision field, gaze estimation plays an important role in human-computer interaction, education, business application, and other fields. Appearance-based gaze estimation gets great breakthroughs thanks to the development and application of deep learning and CNN (convolutional neural network) in recent years. Appearance-based methods can directly learn the mapping function from eye appearance to eye direction using a normal RGB camera. Recently, with the development of deep learning, the method based on CNN has greatly improved its accuracy, but it still cannot meet the actual needs.

We found a lot of low-quality images in several widely used datasets, which ultimately affected the gaze estimation results. However, in the process of gaze estimation and data collection, it is inevitable to collect low-quality images due to facial occlusion and lighting conditions. Using full-face images as input may mitigate this effect, but most researchers have ignored it.

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

Through relevant research, we found that the gaze direction of two pairs of eyes is consistent to some extent, and if this consistency is used properly, better performance can be achieved. If high-quality consistency can be extracted, the prediction result of low-quality eye image can be improved by consistency. Therefore, the Main gaze derived from consistency are particularly important. To acquire high-quality consistency, we no longer treat both eyes equally but tends to suppress low-quality eyes. Then the gaze direction of a low-quality eye image is corrected by consistency. Therefore, this paper proposes a gaze estimation model based on cooperative network: CI-Net. The main contributions can be divided into two parts:

1) Aiming at the problem of poor robustness of existing gaze estimation models. First, we introduce a Main gaze derived from consistency (the concept of Main gaze and Residual residuals will be described in detail in a later section). To extract high-quality consistency, this paper makes the same assumption as ARE-Net [1]: The user should roughly fixate on the same targets with both eyes, which is usually the case in practice. Based on this assumption, when C-Net predicts the Main gaze, Dilated-CNN is used as the backbone network in this paper to expand the receptive field. Secondly, this paper introduces a new attention mechanism:

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

FCA-block (frequency channel attention networks), FCA-block can reassign the weight of high-level features of eyes image and face image to obtain high-quality consistency.

2) Aiming at the problem that the gaze direction of low-quality eye images is difficult to estimate accurately; an additional network I-Net is constructed to obtain inconsistencies (Residual residuals) in single eye images. The Cross attention module is embedded in the I-Net. The Residual residuals is corrected by the Main gaze according to Cross attention, and get the gaze direction of both eyes at last. Many experimental results in this paper show that CI-Net achieves advanced performance on three common datasets. The remaining content of the paper is arranged as follows: 2) Related work. 3) Method. 4) Experimental results. 5) conclusion.

II. RELATED WORK

A. METHODS BASED ON APPEARANCE

Methods based on appearance require relatively few conditions. Just set up a mapping function from the image to the gaze direction. This method can be used to estimate the gaze direction in a variety of environments while maintaining a certain accuracy. However, learning general mapping functions is still challenging due to their high nonlinearity. Up to now, many methods have been put forward to learn mapping function, and neural network is more efficient and accurate to learn mapping function. GazeNet [2] fed eye images into a 16-layer architecture. The head poses information is connected to the first fully connected layer after the convolutional layer. After this work, spatial-weighs CNN [3] is proposed. This method uses the full-face image as input and gives different weights to each area of the face image. iTracker [4] is the first network using multi-channel architecture, which takes left eye image, right eye image, face cropping image, and face grid information as input. Later, scholar Ali and Kim [5] also adopted multi-path framework for gaze estimation. The difference is that this method combines multiple sets of gaze estimation data and uses a data fusion method to take advantage of every possible information related to gaze estimation. In the research of Chen and Shi [6], Biswas *et al.* [7], Chen and Shi [8], and Murthy *et al.* [9], dilated convolution layers was used to replace ordinary convolution layers and pooling layers to enlarge the receptive field. The network architecture of gaze estimation based on dilated convolution is shown in Fig. 1. In the research of ARE-Net [1], it was proposed for the first time that extracting the binocular consistency could effectively improve the accuracy of gaze estimation. This method used additional evaluation networks to coordinate the coefficients of each loss function. The results show that the method is effective. Later, FARE-Net [10] added a face module based on ARE-Net.

B. PROBLEMS WITH THE METHODS BASED ON APPEARANCE

1) The problem of head pose can be solved to some extent by using full face and eyes images as input. However, in the case of facial occlusion and different appearances of eyes, these methods use input information indiscriminately and ignore the effect of low-quality images on gaze estimation.

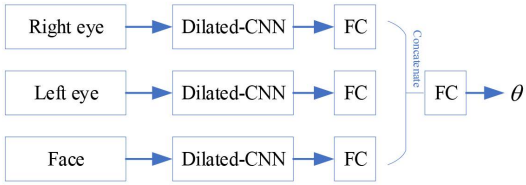

FIGURE 1. Network architecture based on dilated-CNN.

2) These methods are hard to estimate the accurate gaze direction of both eyes when two eyes are exposed to different light conditions. When one eye is in dark light, this affects the overall gaze estimation and the accuracy of that eye's gaze direction.

III. METHOD

A. CONSISTENCY AND INCONSISTENCY

Consistency: According to the visual characteristics of the human eyes, when both eyes focus on the same object, they look in roughly the same direction. To cater to this characteristic, we introduce a direction called the Main gaze. Where θ_l,θ_r is the monocular direction of the left and right eyes, and we formulate the Main gaze θ_m as follows:

$$\theta_m = \frac{\theta_l + \theta_r}{\sqrt{||\theta_l + \theta_r||_2}}. \tag{1}$$

Inconsistency: We refer to the difference between the Main gaze and the gaze direction of two eyes as inconsistency. To cater to this characteristic, we introduce a parameter called Residual residuals, and we formulate Residual residuals $\theta_{\Delta r}$,$\theta_{\Delta l}$ as follows:

$$\theta_r = \theta_m + \theta_{\Delta r}, \theta_l = \theta_m + \theta_{\Delta l} \tag{2}$$

B. CI-NET

Aiming at the problem that high-quality Main gaze cannot be obtained due to the existence of inferior images. We designed a consistency estimation network, which took two eyes and facial images as input, and added the FCA-block [11] module for weight redistribution of facial features and eyes features while expanding the receptive field with Dilated-CNN [7]. Enhance and suppress input features selectively.

Aiming at the problem that the gaze direction of low-quality eye images is difficult to estimate accurately. While ensuring high-quality consistency, we designed an inconsistency estimation network to correct the Residual residuals through Corss attention modual, to achieve an accurate estimation for low-quality eye images.

An overview of the CI-Net network is shown in Fig. 3. The network consists of two parts, a consistency network (C-Net) to estimate the Main gaze and an inconsistency network (I-Net) to estimate the Residual residuals. The Eye channel and Face channel are used to extract features from the eyes and face. These two channels are inspired by iTracker, but the

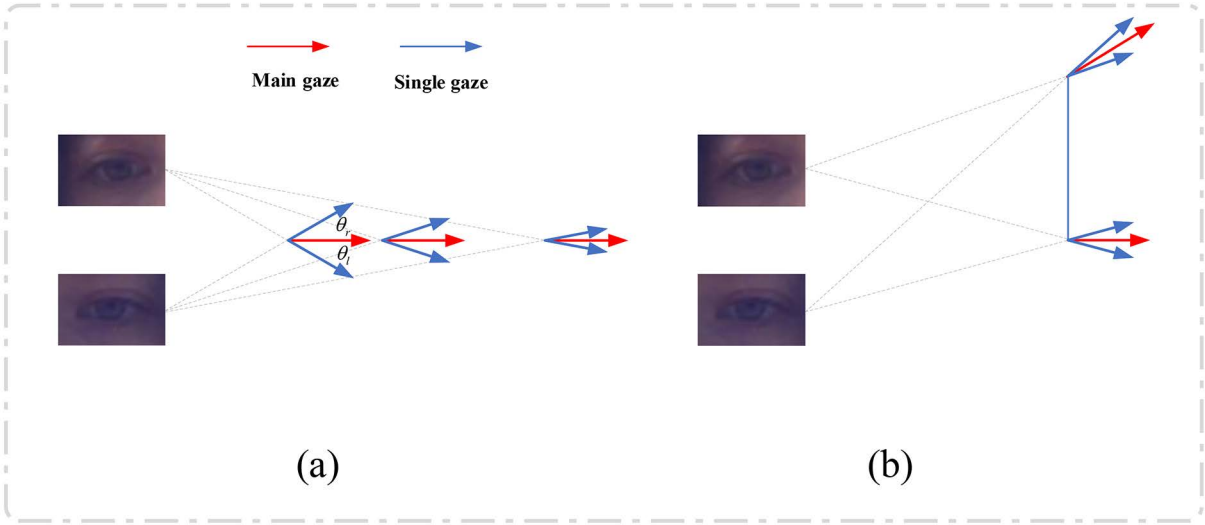

FIGURE 2. Two characteristics of human binocular vision. The blue arrow Single gaze is the gaze direction of the left and right eyes, and the red arrow is the Main gaze. (a) The gaze points are located at different distances. As the gaze point gets further away, the difference between the direction of the eyes and the main direction (Residual residuals) becomes smaller. (b) The gaze points are in different horizontal positions. As the gaze point moves horizontally in one direction, the Residual residuals of the eye in the same direction changes greatly (compared with the Residual residuals in the other direction). The presence of these two conditions would make Residual residuals inaccurate, so we will introduce a cross attention mechanism to prevent it.

difference is that we replace some convolution and pooling layers with dilated convolution layers while not using the face grid information as input. Then we added the FCA-block module in C-Net for reweighting the facial features and eyes features. We need to use consistency to guide the acquisition of the Residual residuals because the Main gaze derived from consistency plays a decisive role in gaze estimation; so, we have added the Cross attention module [12] in I-Net. Finally, we combine the Residual residuals with the Main gaze to obtain the gaze direction for each eye. Next, we will introduce the dilated convolutions module and FCA-block module in C-Net as well as the Cross attention module used to connect I-Net and C-Net.

1) DILATED CONVOLUTIONS

To obtain a larger receptive field, most current architectures use pooling layers and convolutional layers with large strides for downsampling. Both are used to obtain a larger receptive field at the cost of reducing the resolution feature map, and this loses the subtle variations in small pixel ranges. This has little impact on classification tasks such as object detection [13], which can accommodate subtle changes in position. However, for the regression task of gaze estimation, this can greatly affect the accuracy of gaze direction. Furthermore, document [14] showed that the use of dilated convolution in place of normal convolutions improves gaze estimation using a multi-channel architecture. In our network, we use dilated convolution not only for the eye region but also for the face region.

Dilated convolution increases the size of the receptive field without significantly increasing the number of parameters while maintaining spatial resolution. Given an input feature mapping μ, the size of the convolution kernel is N × M × K, where the weight is w, the bias is b, and the dilation rate is (r_1, r_1). The output feature mapping v resulting from the dilated convolution operation can be formulated as follows:

$$v(x, y) = \sum_{k=1}^{K} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} u(x + nr_1, y + mr_2, k) w_{nmk} + b \quad (3)$$

2) FCA-BLOCK

To obtain a high-quality Main gaze, we add the attention module FCA-block in C-Net. It helps to reduce the weight of low-quality eye images and suppress the information which irrelevant to gaze estimation. FCA-block is a new attention mechanism based on the squeeze and excitation block (SE-Block) [15]. Details of the FCA-block are shown in Fig. 4. Different from SE-block which uses global average pooling (GAP) to compress feature maps, FCA-block applies a two-dimensional discrete cosine transform (2DDCT) to compress the feature maps (DCT can be viewed as a weighted sum of input). And GAP just corresponds to the lowest frequency component of 2DDCT, which means that the use of GAP alone will lead to a loss of other frequency components in the feature channels. Suppose that the input is X, we divide X into multiple parts along the channel dimension.Denoted as $[X^1, X^2, \cdots, X^{n-1}], X^i \in R^{C \times H \times W}, i \in 0, 1, \cdots, n-1, C' = \frac{C}{n}$. The corresponding 2DDCT frequency component is assigned to each part and the result of the 2DDCT is used as the compression result of the channel attention. The formula for 2DDCT is as follows:

$$Freq^i = 2DDCT^{u_i, v_i}(X^i) = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} X^i_{:,h,w} B^{u_i, v_i}_{h,w} \quad (4)$$

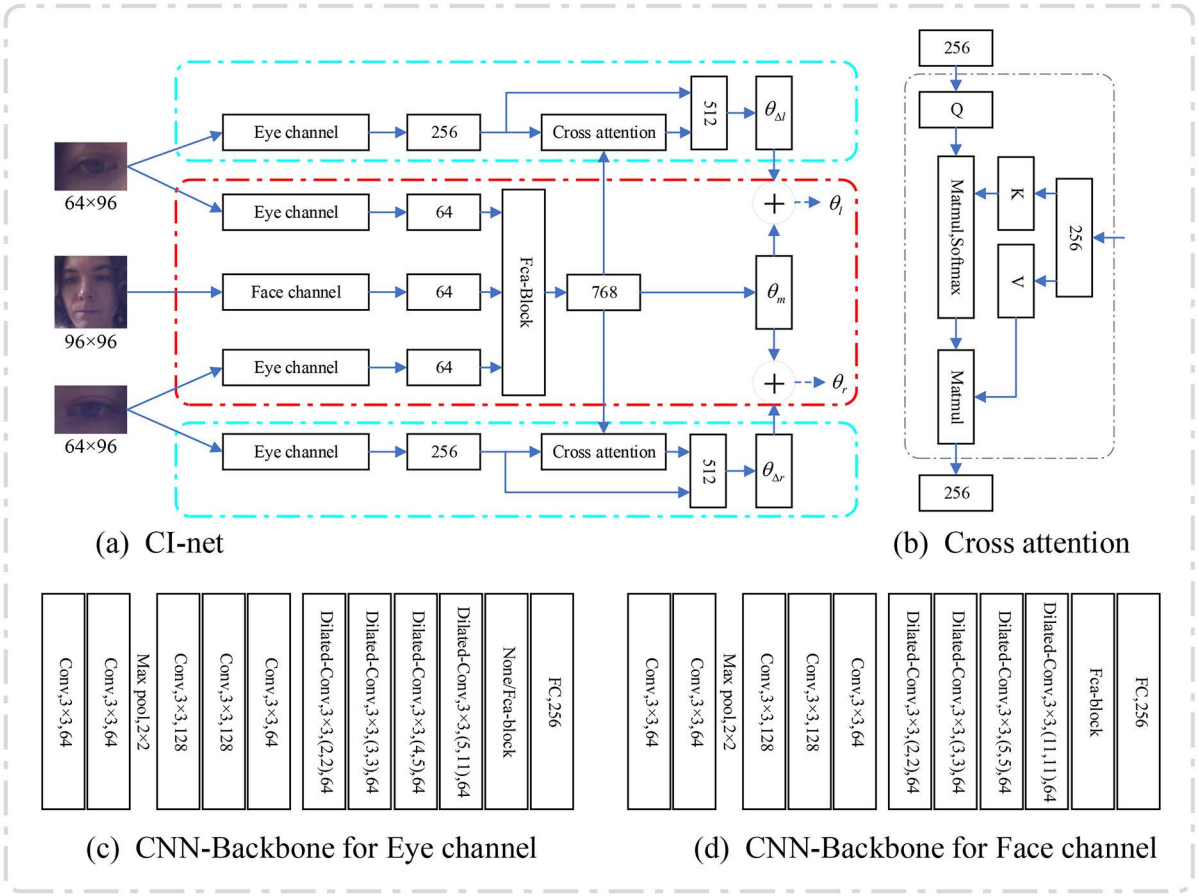

FIGURE 3. Architecture of the proposed networks. The red border is C-Net, and the blue border is I-Net. (a) the architecture of CI-Net. (b) Structure of Cross attention. (c) CNN-Backbone for Eye channel. (d) CNN-Backbone for Face channel. For the middle three branches, we need to pass FCA-block first and then stretch them into a fully connected layer.

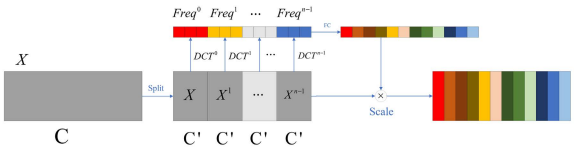

FIGURE 4. Details of FCA-block in CI-Net.

$$B_{h,w}^{u_i,v_i} = \cos\left(\frac{\pi h}{H}(u_i + \frac{1}{2})\right)\cos\left(\frac{\pi w}{W}(v_i + \frac{1}{2})\right) \tag{5}$$

where H denotes the height of X^i, W the width of X^i and (u_i, v_i) the 2D index corresponds to X^i. The entire compression vector can then be obtained by concatenating:

$$\begin{aligned} Freq &= compress(X) \\ &= cat([Freq^0, Freq^1, \cdots, Freq^{n-1}]) \end{aligned} \tag{6}$$

The entire FCA-block framework can be written as:

$$ms.att = sigmoid(fc(Freq)) \tag{7}$$

3) CROSS ATTENTION

We set up I-Net to estimate the Residual residuals. As shown in Fig. 2, our method will face two situations when estimating inconsistency. A) When the gaze points are located at different distances. As the gaze point gets further away, the difference between the direction of the eyes and the main direction (Residual residuals) becomes smaller. B) When the gaze points are in different horizontal positions. As the gaze point moves horizontally in one direction, the Residual residuals of the eye in the same direction changes greatly (compared with the Residual residuals in the other direction). To reduce the impact of the two cases above, we introduced the Cross attention module in I-Net, the module consists of a self-attention mechanism. We use the features extracted from the C-Net as Key,Value,and features extracted from the I-Net as Query. Due to the nature of the attention mechanism, the I-Net takes information from the C-Net on its own, thus improving the two situations. The output formula of Cross attention is as follows:

$$Output(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{8}$$

4) LOSS FUNCTION

We define the ground truth for the Main gaze and the monocular direction of the left and right eyes as g_t^*, g_l^* and g_r^*, thus, the labels of the Residual residuals can be defined

as shown below:

$$g^*_{\theta_l} = g^*_l - g^*_t, g^*_{\theta_r} = g^*_r - g^*_t \tag{9}$$

We use $L_2 - norm$ as the Evaluation Metrics and we train our CI-Net with the following loss function:

$$L(g, g^*) = w_t \left|\left|g_t - g^*_t\right|\right|^2 + w_l \left|\left|g_{\theta_l} - g^*_{\theta_l}\right|\right|^2 + w_r \left|\left|g_{\theta_r} - g^*_{\theta_r}\right|\right|^2 \tag{10}$$

$g = (g_t, g_{\theta_l}, g_{\theta_r})$ and w_t, w_l, w_r is used to balance the weights between the Main gaze and the Residual residuals. In this paper, $w_t = 0.5$,$w_l = 0.25$,$w_r = 0.25$.

IV. EXPERIMENTAL RESULTS

A. DATASET

MPIIGaze: MPIIGaze dataset [2] is a commonly used dataset in gaze estimation and it includes 213659 images from 15 subjects, which contain a large number of images with different illumination, head posture, and eye appearance. It also provides a standardized assessment subset of 1500 left-eye images and 1500 right-eye images from different subjects. Our method requires the capture of paired eye images, so we find one of the missing paired eye images in the dataset. In addition, we need to input the full-face image, and MPIIGaze only provides the eye image, so we obtain the corresponding face image from the MPIIGaze extended dataset MPIIFaceGaze.

EyeDiap: EyeDiap dataset [16] contains a set of video clips of 16 subjects, filmed in three different scenarios: discrete screen target, continuous screen target, and 3D floating target. As the EyeDiap dataset does not provide a standard evaluation subset, we sampled images from each video to construct the evaluation dataset. We constructed the evaluation dataset by capturing one image every 10 frames from the video clips of 14 participants.

RT-Gene: The RT-Gene dataset [17] contains 122531 images taken by 15 subjects using wearable eye-tracking glasses. Different from the MPIIGaze dataset, the subjects' distance from the camera is constantly changing, so this dataset is subject to greater variation in head posture and gaze direction.

B. IMPLEMENTATION DETAILS

The Adam optimizer is adopted with a learning rate 0.001 and a batch size of 256. All experiments are performed on a PC with an Intel Xeon E5 CPU and NVIDIA GeForce RTX 2080Ti GPU;

C. COMPARED METHODS

Single Eye [18]: Single Eye is a LeNet-based gaze estimation method. Compared with CI-Net, it only uses monocular images as input and cannot take advantage of consistency. GazeNet: GazeNet uses a network based on Single Eye that replaces the shallow network LeNet with a more powerful network VGG, which also fails to take advantage of consistency. FARE-Net: The FARE-Net is the first method to propose the use of consistency to predict the gaze direction.

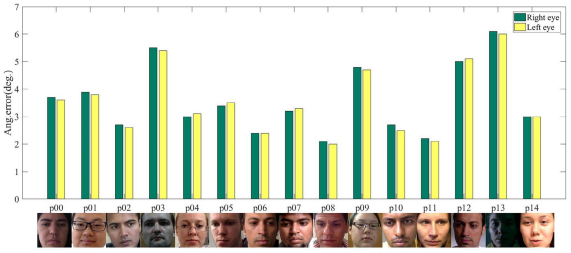

FIGURE 5. Angular error of different subjects on MPIIGaze dataset.

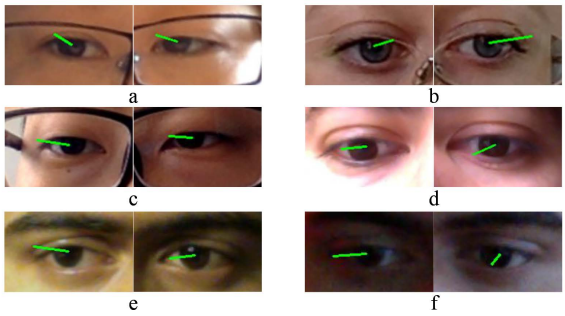

FIGURE 6. Visualization results of the evaluation datasets. (a)-(d) are the visualization results for different subjects' eye directions. (e) and (f) are the visualization results of the same subject under different eye appearances.

The model takes a pair of eyes images as input, uses AlexNet as the basic network, and uses E-Net to evaluate the eyes' gaze direction.

D. RESULT

1) WITHIN DATASET

Fig. 5 and 6 show the gaze estimation results of CI-Net. In Fig. 5, 15 subjects from the dataset MPIIGaze are selected. It can be seen from the figure that, in the case of low-quality of one eye images, lower angular errors can also be obtained according to consistency. Fig. 6 shows the visualization results of our method on the dataset MPIIGaze. (a)-(d) are the visualization results for different subjects' eye directions. (e) and (f) are the visualization results of the same subject under different eye appearances. The results show that the proposed method can adapt to different eye appearances and get advanced accuracy and robustness.

To illustrate the overall performance of the model in this paper, we further calculated the accumulative proportion on the dataset MPIIGaze to illustrate the distribution of different angular errors for different methods. As can be seen from Fig. 7, the error distribution of 80% of the results of our method is within 10° for both left and right eyes. Significant advantages over all three other methods. The formula for accumulative proportion is shown below:

$$p = \frac{Num(s|s \leq s_{A.e}, s \in S)}{Num(s)} \tag{11}$$

where S denotes the set of all angular errors on the MPIIGaze and $s_{A.e}$ represents a specific angular error.

To demonstrate the performance of the proposed method on low-quality eye images. We show three appearance-based

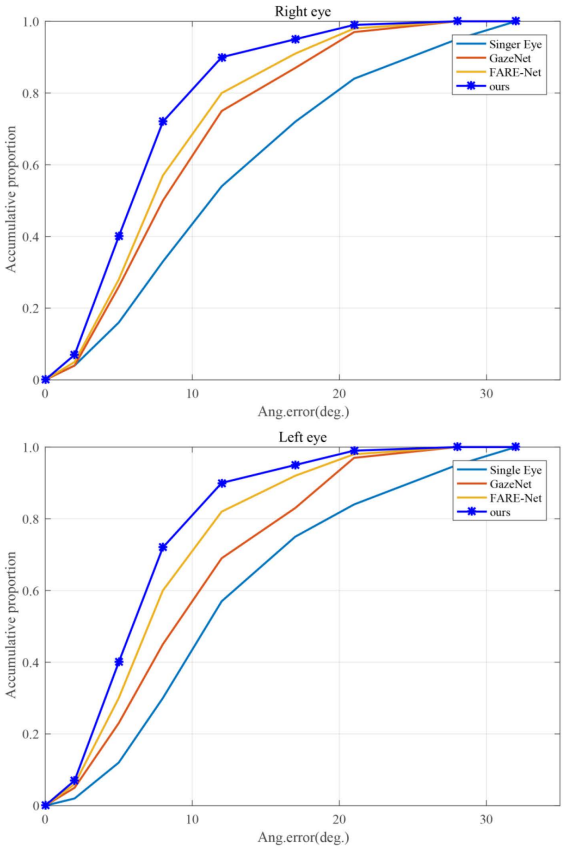

FIGURE 7. Accumulative proportion of angular error on MPIIGaze.

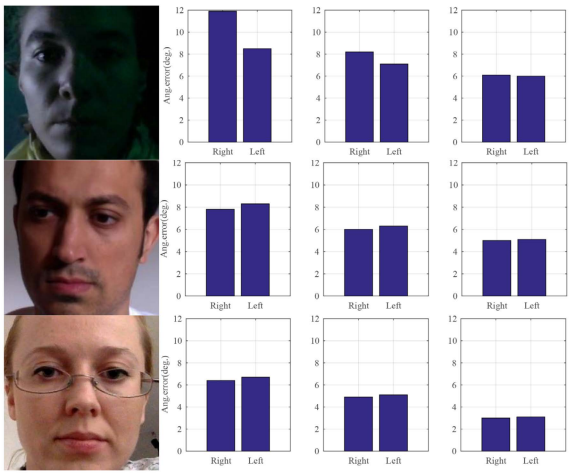

FIGURE 8. Comparison of two eyes' gaze errors. The first column of bars shows the gaze errors of Single Eye, the second column of bars shows the gaze errors of FARE-Net, and the third column of bars shows the gaze errors of CI-Net. It shows that our methods get lower angular errors and smaller differences between the two eyes.

gaze estimation methods, shown in Fig. 8, for Single Eye, FARE-Net, and CI-Net, and derive left and right eye gaze angle error for specific samples. In this case, the Single Eye method, which uses only monocular images as input, performs better for eyes with a high-quality image, but it is unable to obtain the exact gaze angle from low-quality images which will greatly affect the final accuracy. And FARE-Net, performs asymmetric optimization, improving the better eye and the worse eye through a designed evaluation and feedback strategy so that both eyes have a smaller angular error, which will obtain better overall accuracy. However, this method did not suppress useless information from low-quality images and did not explore the effect of inconsistency on the gaze directions of the two eyes. The results show that CI-Net with consistency and inconsistency gets better performance than Single eye and FARE-Net.

We compare the performance of our proposed method with other appearance-based gaze estimation methods on three datasets (MPIIGaze, EyeDiap, and RT-Gene). We list seven gaze estimation methods based on appearance. As shown in Table 1, our CI-Net achieves lower angle errors than the other seven methods. Specifically, compared to FARE-Net, which is currently highly accurate, Performance is improved by 12% on the MPIIGaze dataset, 5% on the EyeDiap dataset, and 6% on the RT-Gene dataset. Our method achieves better performance than FARE-Net, probably due to the extraction of binocular consistency and inconsistency. However, the effectiveness of the three modules in the extraction process is unknown, so this paper carries out a further study (ablation study).

2) ABLATION STUDY

Previous results have demonstrated that our method achieves better performance in three datasets, but the effectiveness of each module in our method is still unknown. Therefore, we performed an ablation study of the CI-Net on MPIIGaze. We replaced two of the modules, FCA-block, and Cross attention, with FC layers. As for our Dilated CNN module, we redesigned the structure to output the same size of feature maps at the last layer of the network without using dilated convolution. First, we use the normal convolutional layers instead of all the dilated convolution in the Eye channel and Face channel. And two maximum pooling layers are added behind the seventh and eighth convolutional layers of the Eye channel, and three maximum pooling layers are added behind the fourth, sixth, and eighth convolutional layers of the Face channel. As can be seen from Table 2, the angular error rises by 0.6° when replacing only the dilated convolution, which verifies the effectiveness of the dilated convolution. And as can be seen from the rest of the data in Table 2, the Fca-net and the dilated convolution modules have a greater impact on our final results of gaze estimation than the Cross attention modules, these two modules are used for estimation for the Main gaze. This validates our previous idea that the Main gaze estimated by the consistency plays a decisive role in our results.

3) CROSS-DATASET EVALUATION

This part focuses on exploring the implications across datasets. Evaluation across datasets [19] is a huge challenge for all learning-based methods. We trained the network on the MPIIGaze dataset and tested it on the EyeDiap and RT-Gene datasets. The results are shown in Table 3,

TABLE 1. Angular errors of different gaze estimation methods within datasets.

	Backbone	Input	Dataset	Ang.erroe(deg.)
SingleEye	LeNet	Eye and head pose	MPIIGaze	6.3
GazeNet	VGG-16	Eye and head pose	MPIIGaze	5.5
			UT Multiview	5.4
iTracker	VGG-16	Face	MPIIGaze	5.6
	AlexNet			6.2
Spatial-Weights	AlexNet	Face	MPIIGaze	4.8
			EyeDiap	6.0
ARE-Net	AlexNet	Eyes	MPIIGaze	5.0
			EyeDiap	6.1
Dilated-Net	Dilated-CNN	Face and eyes	MPIIGaze	4.5
			EyeDiap	5.4
FARE-Net	AlexNet	Face and eyes	MPIIGaze	4.3
			EyeDiap	5.7
			RT-Gene	8.4
Ours	Dilated-CNN	Face and eyes	MPIIGaze	3.8
			EyeDiap	5.4
			RT-Gene	7.9

TABLE 2. Ablation study of CI-Net on MPIIGaze dataset.

C-Net		I-Net		C-Net;I-Net		Ang.error(deg.)
FC	FCA-block	FC	Cross attention	CNN+Pool	Dilated-CNN	
Y		Y		Y		5.6
Y		Y			Y	4.7
Y			Y	Y		5.2
Y			Y		Y	4.2
	Y	Y		Y		4.9
	Y	Y			Y	4.0
	Y		Y	Y		4.4
	Y		Y		Y	3.8

TABLE 3. Experimental results of Cross-Dataset evaluation and comparison. Models are trained on MPIIGaze and tested on EyeDiap and RT-Gene.

	GazeNet	SingleEye	FARE-Net	Ours
EyeDiap	16.1	17.4	13.5	12.1
RT-Gene	19.5	15.1	16.2	13.6

where CI-Net outperforms the other three appearance-based methods.

4) ADDITIONAL ANALYSIS

We performed convergence analysis on our CI-Net on the MPIIGaze dataset and the results are shown in Fig. 9. The angular error decreases rapidly throughout the iteration, reaching a minimum value after approximately 70 epochs. Overall, the proposed network is shown to be able to converge quickly and robustly.

Some failure cases can also help understand the proposed method. Therefore, we have chosen a representative case and the results are shown in Fig. 10. High-quality consistency

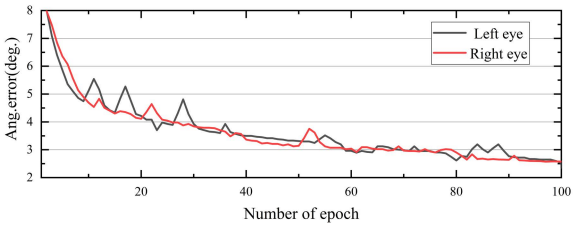

FIGURE 9. Validation on the convergence of the CI-Net.

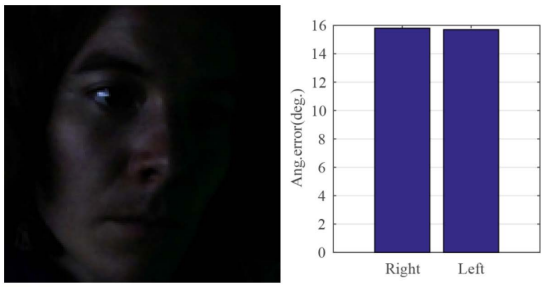

FIGURE 10. Failure cases with large angular errors.

cannot be extracted because of low-quality images of both eyes and face. Without the guarantee of high-quality consistency, the angular error of gaze direction is substantially higher.

V. CONCLUSION

Low-quality images due to different appearance of two eyes and occlusion, resulting in high nonlinearity in mapping function between the image and the gaze direction Learning such a mapping function is extremely challenging. To address this problem, we propose a network based on binocular consistency and inconsistency (CI-Net). First, C-Net obtains high-quality consistency through dilated convolution and FCA-block. Then, the Residual residuals are guided by high-quality consistency through Cross attention. Finally, obtain the gaze directions of both eyes. In addition, we have demonstrated the effectiveness of the three modules mentioned above through an ablation study. Our experiments show that the proposed CI-Net achieves leading performance on three public datasets.

REFERENCES

[1] Y. Cheng, F. Lu, and X. Zhang, "Appearance-based gaze estimation via evaluation-guided asymmetric regression," in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science, 2018, pp. 100–115.
[2] 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.
[3] 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.
[4] 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.
[5] A. Ali and Y.-G. Kim, "Deep fusion for 3D gaze estimation from natural face images using multi-stream CNNs," *IEEE Access*, vol. 8, pp. 69212–69221, 2020.
[6] Z. Chen and B. E. Shi, "Appearance-based gaze estimation using dilated-convolutions," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 309–324.
[7] L. R. D. Murthy and 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.
[8] Z. Chen and B. Shi, "Towards high performance low complexity calibration in appearance based gaze estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Feb. 7, 2022, doi: 10.1109/TPAMI.2022.3148386.
[9] L. R. D. Murthy, S. Brahmbhatt, S. Arjun, and P. Biswas, "I2DNet—Design and real-time evaluation of an appearance-based gaze estimation system," *J. Eye Movement Res.*, vol. 14, no. 4, Aug. 2021.
[10] Y. Cheng, X. Zhang, F. Lu, and Y. Sato, "Gaze estimation by exploring two-eye asymmetry," *IEEE Trans. Image Process.*, vol. 29, pp. 5259–5272, 2020.
[11] Z. Qin, P. Zhang, F. Wu, and X. Li, "FcaNet: Frequency channel attention networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 783–792.
[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[13] Y. Luo, Y. Cai, B. Wang, J. Wang, and Y. Wang, "SiamFF: Visual tracking with a Siamese network combining information fusion with rectangular window filtering," *IEEE Access*, vol. 8, pp. 119899–119910, 2020.
[14] X. Lei, H. Pan, and X. Huang, "A dilated CNN model for image classification," *IEEE Access*, vol. 7, pp. 124087–124095, 2019.
[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[16] K. A. F. Mora, F. Monay, and J.-M. Odobez, "EYEDIAP: A database for the development and evaluation of gaze estimation algorithms from RGB and RGB-D cameras," in *Proc. Symp. Eye Tracking Res. Appl. (ETRA)*, 2014, pp. 255–258.
[17] T. Fischer, H. J. Chang, and Y. Demiris, "RT-GENE: Real-time eye gaze estimation in natural environments," in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science, 2018, pp. 334–352.
[18] 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.
[19] A. A. Akinyelu and P. Blignaut, "Convolutional neural network-based methods for eye gaze estimation: A survey," *IEEE Access*, vol. 8, pp. 142581–142605, 2020.

YUAN LUO received the M.S. degree from the Chongqing University of Posts and Telecommunications (CQUPT), Chongqing, China, in 1996, and the Ph.D. degree from Chongqing University, Chongqing, in 2003. She was a Visiting Scholar with the Université de Montréal, Canada, in 2006. She is currently a Professor with CQUPT. Her research interests include computer vision, photoelectric sensing, image processing, and mobile robots.

JIANGTAO CHEN received the B.S. degree from the Chongqing University of Posts and Telecommunications (CQUPT), Chongqing, China, in 2020, where he is currently pursuing the master's degree with the College of Photoelectrics. His current research interests include pattern recognition and gaze estimation.

JIAN CHEN received the B.S. degree from the Chongqing University of Posts and Telecommunications (CQUPT), Chongqing, China, in 2020, where he is currently pursuing the master's degree with the College of Photoelectrics. His current research interests include mobile robot and semantic SLAM.