# Monocular 3D gaze estimation using feature discretization and attention mechanism*

**SHA Tong[1], SUN Jinglin[1], PUN Siohang[2], and LIU Yu[1]****

*1. School of Microelectronics, Tianjin University, Tianjin 300072, China*

*2. Institute of Microelectronics, University of Macau, Macau 999078, China*

(Received 29 November 2022; Revised 21 January 2023)

Gaze estimation has become an important field of image and information processing. Estimating gaze from full-face images using convolutional neural network (CNN) has achieved fine accuracy. However, estimating gaze from eye images is very challenging due to the less information contained in eye images than in full-face images, and it's still vital since eye-image-based methods have wider applications. In this paper, we propose the discretization-gaze network (DGaze-Net) to optimize monocular three-dimensional (3D) gaze estimation accuracy by feature discretization and attention mechanism. The gaze predictor of DGaze-Net is optimized based on feature discretization. By discretizing the gaze angle into $K$ bins, a classification constraint is added to the gaze predictor. In the gaze predictor, the gaze angle is pre-applied with a binned classification before regressing with the real gaze angle to improve gaze estimation accuracy. In addition, the attention mechanism is applied to the backbone to enhance the ability to extract eye features related to gaze. The proposed method is validated on three gaze datasets and achieves encouraging gaze estimation accuracy.

**Document code:** A **Article ID:** 1673-1905(2023)05-0301-6

**DOI** https://doi.org/10.1007/s11801-023-2203-1

Three-dimensional (3D) gaze estimation refers to estimating the 3D gaze direction of human eyes through the eye images obtained by the digital camera. Gaze estimation has become a research hotspot in the field of image and information processing, for it has been widely used in many areas such as psychological diagnosis[1], virtual reality[2], and human-machine interaction[3].

Current gaze estimation methods can be divided into feature-based and appearance-based methods. Feature-based methods[4-6] use specific eye features such as pupil center and corneal reflection to estimate gaze. Their reliability depends on the accuracy of feature detection. Moreover, feature-based methods usually require specific hardware for illumination and capture. By contrast, appearance-based methods don't focus on specific eye features. They directly learn the mapping function from eye appearance to gaze. The methods take images from off-the-shelf cameras as input and don't require additional devices. Appearance-based methods[7-14] have mostly been used with convolutional neural network (CNN) in recent years. CNN extracts eye appearance features from high-dimensional images and learns a highly non-linear mapping relationship from eye appearance to gaze. ZHANG et al[7] proposed a method based on LeNet to estimate gaze from single eye images. ZHANG et al[8] further extended their work and proposed the GazeNet, which is a 13-convolutional-layer neural network inherited from a 16-layer visual geometry group (VGG)[15] network. This GazeNet is proved to be more effective than the LeNet-based method. CHENG et al[9] used the asymmetric information of both eyes for gaze estimation. CHEN et al[10] proposed a multistream CNN that took an full-face image, images of both eyes, and a face grid as input to estimate gaze. On the contrary, ZHANG et al[11] proposed a network, which only took a full-face image as input and used a spatial weighting mechanism to increase the weight of the eye region adaptively. Similarly, LIU et al[12] used the multi-scale channel and spatial information to select the important facial area adaptively. MURTHY et al[13] proposed a full-face-based gaze estimation method using attention and difference mechanism. ABDELRAHMAN et al[14] proposed a fine-grained gaze estimation network, which took a full-face image as input and optimized the network with two identical losses.

Recent studies mainly focus on estimating gaze from full-face images. ZHANG et al[11] showed that other facial regions, besides the eye region, also contributed to gaze estimation. By analyzing the region importance maps, they found that the eye region is the most important if the gaze direction is straight ahead, while the model puts higher importance on other regions if the gaze direction becomes

* This work has been supported by the Major Science and Technology Special Plan Projects of Yunnan (No.202002AD080001).

** E-mail: liuyu@tju.edu.cn

more extreme. However, full-face-based methods in some scenes or applications (such as virtual reality) are inapplicable, where full-face images can't be obtained. Therefore, how to improve gaze estimation accuracy of eye-based methods without other facial regions is a problem that needs to be addressed.

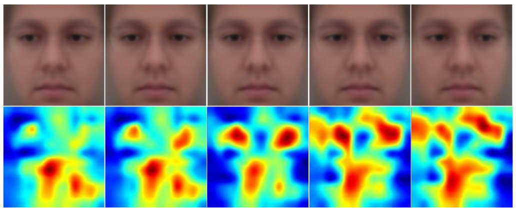

**Fig.1 Region importance map[11] indicating the facial region importance with different gaze angles**

In this paper, we propose the discretization-gaze network (DGaze-Net) to improve monocular 3D gaze estimation accuracy from two aspects. On the one hand, the gaze predictor is optimized based on discretization. Unlike other methods that directly regress gaze from eye features or facial features, the proposed method converts gaze prediction into a joint process of gaze angle classification and regression by discretizing the gaze angle into $K$ bins. In this way, the gaze angle is applied with a binned classification before regressing with the real gaze angle to improve gaze estimation accuracy. On the other hand, the attention mechanism is applied to the backbone to enhance its performance. The channel and spatial attention module in the backbone enable the model to select a set of features most relevant to gaze, thus further improving gaze estimation accuracy.

The feature extraction module of the proposed network is illustrated in Fig.2. We use a ResNet-50[16] architecture combined with the convolutional block attention module (CBAM)[17] to extract eye features. CBAM is a plug-and-play attention module that focuses on the channel and spatial axes. The channel sub-module utilizes both max-pooling outputs and average-pooling outputs with a shared network. The spatial sub-module utilizes two similar outputs pooled along the channel axis and feeds them to a convolution layer. We apply CBAM on the convolution outputs of each residual block. The input image $\boldsymbol{I}\in\boldsymbol{D}^{W\times H\times C}$, where ($W$, $H$, $C$)=(60, 36, 3), is put into the feature extraction module. Considering the resolution of $\boldsymbol{I}$ is small, to avoid over-downsampling, we change the stride of some convolutional layers. As we know, ResNet-50 can be divided into two parts. The first part is the preprocessing of input, and it contains structures from Conv7×7 to MaxPool3×3. The second part is Layer$X$ ($X$=1, 2, 3, 4), which includes (3, 4, 6, 3) residual blocks. In the original version of ResNet-50, Layer2, Layer3, and Layer4 perform downsampling directly by convolutional layers with a stride of 2. We change the stride of convolutional layers from 2 to 1 in Layer2 and Layer4, so Layer2 and Layer4 won't change the resolution of the feature map. The last block outputs the feature map $f_{\text{block}}\left(\boldsymbol{I}\right)\in\boldsymbol{F}^{8\times5\times2\,048}$. Then an adaptive average pooling is applied to downsample the feature map to $f\left(\boldsymbol{I}\right)\in\boldsymbol{F}^{1\times1\times2\,048}$.

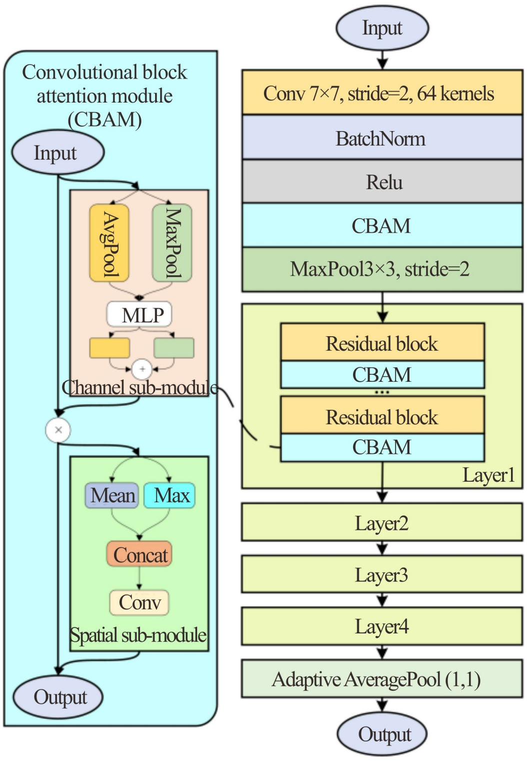

**Fig.2 Feature extraction module architecture**

The overall framework of the proposed DGaze-Net is illustrated in Fig.3. It consists of the feature extraction module and gaze angle predictors. The 3D gaze vector can be converted into a two-dimensional (2D) gaze angle vector in the spherical coordinate system. The 2D gaze angle vector is composed of two eye rotation angles: yaw $g_\phi$ and pitch $g_\theta$. Unlike the previous work that directly regress $g_\phi$ and $g_\theta$ in one predictor, we use two predictors to predict $g_\phi$ and $g_\theta$ separately. At first, we discretize the continuous gaze angle in datasets into bins for binned gaze classification. The angle values of [−99°, +99°] are discretized into 66 bins equally. Each bin $X_i$ covers a range of angles from $X_i^{\min}$ to $X_i^{\max}$ and votes for the mean of all training samples in this angle range, $x_i$. The eye image $\boldsymbol{I}\in\boldsymbol{D}^{60\times36\times3}$ is first put into the feature extraction module, which outputs $f\left(\boldsymbol{I}\right)\in\boldsymbol{F}^{1\times1\times2\,048}$. Then concatenate feature $f(\boldsymbol{I})$ with head pose $\boldsymbol{h}$, and send them into two gaze angle predictors separately. In each gaze angle predictor, the fusion feature is put into a fully-connected layer with 66 softmax-normalized neurons. We calculate the expectation $\hat{g}_\phi$ ($\hat{g}_\theta$) over the softmax-normalized output probability $p_i$ of the 66 neurons as the final output of each predictor:

$$\begin{cases}\hat{g}_\phi=\sum\limits_{i=1}^{66}x_i\cdot p_i^\phi\\ \hat{g}_\theta=\sum\limits_{i=1}^{66}x_i\cdot p_i^\theta\end{cases}. \tag{1}$$

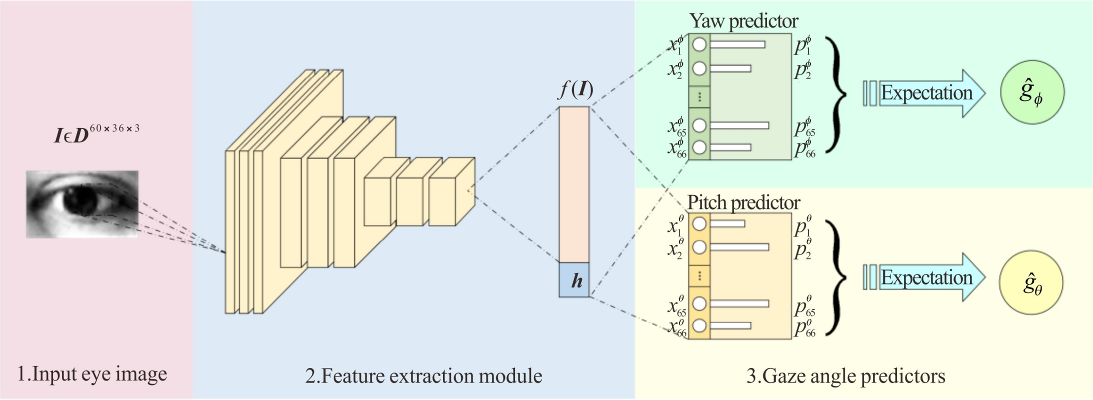

**Fig.3 Overall framework of the discretization-gaze network**

We use two separate losses to optimize the DGaze-Net, one for each predictor. Each loss consists of a cross-entropy loss and a mean-squared error loss. The idea behind this is that we use the bin classification to predict the neighborhood of the gaze angle, then apply the fine-grained regression on the expectation of the binned outputs $\hat{g}_{\phi}$ ($\hat{g}_{\theta}$).

The cross-entropy loss $L_c$ is defined as

$$L_c\left(\hat{g}, g\right) = -\sum_i g_i \log \hat{g}_i . \qquad (2)$$

The mean-squared error loss $L_{\text{mse}}$ is defined as

$$L_{\text{mse}}\left(\hat{g}, g\right) = \frac{1}{N}\sum_0^N \left(\hat{g} - g\right)^2 , \qquad (3)$$

where $g$ is the ground-truth value and $\hat{g}$ is the predicted value. We aim to make $L_c$ and $L_{\text{mse}}$ have equal contribution to gaze estimation. Simply combining the losses by a fixed weighted sum $L=L_c+\alpha L_{\text{mse}}$ can't adaptively adjust the weights of $L_c$ and $L_{\text{mse}}$. It may cause one of the losses to dominate the network optimization and affect the performance of gaze estimation. KENDALL et al[18] propose a principled way to adaptively adjust the weight of each loss using homoscedastic uncertainty. Based on this, we define the final loss of each predictor as

$$\begin{cases} L_{\phi} = \dfrac{1}{2\alpha_1^2} L_c\left(\hat{g}_{\phi}, g_{\phi}\right) + \dfrac{1}{2\alpha_2^2} L_{\text{mse}}\left(\hat{g}_{\phi}, g_{\phi}\right) + \log \alpha_1 \alpha_2 \\ L_{\theta} = \dfrac{1}{2\beta_1^2} L_c\left(\hat{g}_{\theta}, g_{\theta}\right) + \dfrac{1}{2\beta_2^2} L_{\text{mse}}\left(\hat{g}_{\theta}, g_{\theta}\right) + \log \beta_1 \beta_2 \end{cases} , \qquad (4)$$

where ($\alpha_1$, $\alpha_2$) and ($\beta_1$, $\beta_2$) are the learnable parameters. Large scale $\alpha_1$ will decrease the contribution of $L_c\left(\hat{g}_{\phi}, g_{\phi}\right)$, whereas small scale $\alpha_1$ will increase its contribution. The scale is regulated by the last term $\log\alpha_1\alpha_2$ to prevent setting $\alpha_i$ too large. It's similarly for $\beta_i$.

We perform experiments on MPIIGaze[7], real-time eye gaze estimation in natural environments (RT-Gene)[19], and UT-Multiview[20] datasets. The proposed method achieves encouraging gaze estimation accuracy on these datasets. Ablation studies are conducted to demonstrate the effectiveness of the proposed structure. In addition, we change the input of some full-face-based methods to single-eye images to analyze the effect on gaze estimation when other facial regions can't be obtained.

MPIIGaze is one of the most commonly used datasets in gaze estimation, which contains 3 000 eye images for each of 15 subjects. In addition, it also provides full-face images. UT-Multiview comprises 24 320 (23 040 synthesized training data and 1 280 normalized test data) eye images for each of the 50 subjects. RT-Gene contains 122 531 images of 15 participants using wearable eye-tracking glasses. This dataset provides eye images and full-face images and has higher variation in gaze angles. We follow a leave-one-subject-out evaluation for MPIIGaze dataset, a 3-fold validation for UT-Multiview dataset and a 3-fold validation for RT-Gene dataset according to the evaluation protocol provided by the dataset.

We implement the structure of our model using Pytorch and train the model with the Adam optimizer on these datasets. We use the initial learning rate of 0.001 and multiply it by 0.1 after every 5 000 iterations. For MPIIGaze, we train the model for 50 epochs. For UT-Multiview, there are 30 epochs applied. For RT-Gene, the model is trained for 40 epochs.

There are several full-face-based methods and eye-based methods tested for comparison. Results of the following methods are obtained from our implementation or published papers. To analyze the effect on gaze estimation without facial regions and to compare more fairly with our eye-based method, we modify some full-face-based methods to eye-based methods and conduct experiments. L2CS-Net[14] uses the full-face image as input to predict 3D gaze. We change the input of L2CS-Net to eye images and resize the eye images from (60×36) to (448×448) to match the input size. The modified network is called L2CS-Net*. Dilated-Net[10] is a multistream CNN that takes the full-face image and eye images as input to estimate gaze. We shrunk the Dilated-Net to one eye network (Dilated-Net*) to compare with our method. It's similarly for RT-Gaze[19] and I2D-Net[13]. Minst[7], GazeNet[8], ARE-Net[9], MeNet[21], Capsule-Net[22] and MSGazeNet[23] are the major eye-based gaze estimation methods.

The 2D gaze angle vector $\left(\hat{g}_{\phi}, \hat{g}_{\theta}\right)$ predicted by the network is converted into a 3D vector $\hat{\boldsymbol{g}} \in \mathbb{R}^3$ in the Cartesian coordinate system. The angular error $L_{\text{angular}}$ between predicted gaze vector $\hat{\boldsymbol{g}} \in \mathbb{R}^3$ and ground truth $\boldsymbol{g} \in \mathbb{R}^3$ is employed to measure the accuracy of gaze estimation. For training and testing, we use the left eye of the subject as input.

$$L_{\text{angular}} = \arccos\frac{\boldsymbol{g}\cdot\hat{\boldsymbol{g}}}{\|\boldsymbol{g}\|\cdot\|\hat{\boldsymbol{g}}\|} . \qquad (5)$$

Tab.1 shows the results of $L_{\text{angular}}$ within full-face-based methods on three datasets. We can see that switching the input from full-face images to eye images will cause a significant decline in gaze estimation accuracy, which also stress the importance of research on eye-based methods. As can be seen from Tab.1, the proposed DGaze-Net achieves encouraging gaze estimation accuracy with limited input conditions. It can be noticed that our method shows less good accuracy than the initial implementations of L2CS-Net[14] and I2D-Net[13]. Considering the difference of input size (face image resolution (448×448) vs. eye image resolution (60×36)), the accuracy of our method is acceptable.

**Tab.1 Comparison with full-face-based methods (The performance of our method is shown in bold)**

| Method | MPIIGaze | UT-Multiview | RT-Gene |
|---|---|---|---|
| Face-Net[11] | 4.8° | - | 10.0° |
| RT-Gaze[19] | 4.8° | - | 8.6° |
| RT-Gaze* | 5.1° | 5.5° | 11.5° |
| Dilated-Net[10] | 4.8° | - | 8.5° |
| Dilated-Net* | 5.2° | 5.8° | 13.1° |
| L2CS-Net[14] | 3.9° | - | - |
| L2CS-Net* | 5.8° | 6.7° | 12.3° |
| I2D-Net[13] | 4.3° | - | 8.4° |
| I2D-Net* | 5.1° | - | - |
| **DGaze-Net** | **4.8°** | **5.4°** | **10.4°** |

Tab.2 shows the results of gaze estimation within eye-based methods. MSGazeNet[23] is a study that decomposes the structure of eye regions and combines them to estimate gaze using a multistream network. Therefore, the structure of MSGazeNet is quite complex, containing a U-net style network for anatomical eye region isolation and a multistream gaze estimation network. The proposed DGaze-Net shows similar gaze estimation accuracy to state-of-the-art methods with a relatively simple structure.

We also present the gaze accuracy of our method and ARE-Net[9] for each subject on MPIIGaze in Fig.4. It can be observed from Fig.4 that the performance of our method is more stable for different subjects.

**Tab.2 Comparison with eye-based methods (The performance of our method is shown in bold)**

| Method | MPIIGaze | UT-Multiview | RT-Gene |
|---|---|---|---|
| Mnist[7] | 6.3° | 6.2° | 14.9° |
| GazeNet[8] | 5.5° | 5.9° | 12.8° |
| ARE-Net[9] | 5.0° | - | - |
| MeNet[21] | 4.9° | 5.6° | - |
| CapsuleNet[22] | 5.7° | - | - |
| MSGazeNet[23] | 4.8° | 5.3° | - |
| **DGaze-Net** | **4.8°** | **5.4°** | **10.4°** |

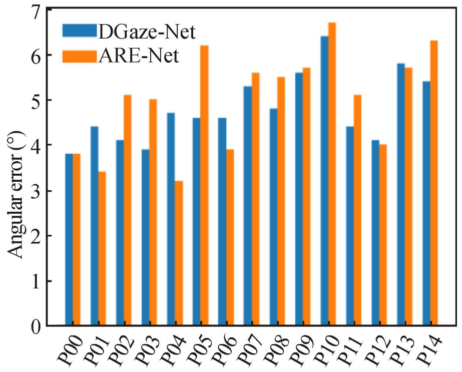

**Fig.4 Comparison of gaze accuracy of each subject between our method and ARE-Net[9] on MPIIGaze**

To further analyze the performance of the attention-based feature extraction module and the discretization-based gaze angle predictors, ablation studies are conducted on MPIIGaze. The results are presented in Tab.3. The attention modules are removed from the feature extraction module to build Var.1. We remove the gaze angle predictors and directly regress $g_{\phi}$ and $g_{\theta}$ in one fully-connected layer to build Var.2. Both attention modules and discretization-based predictors are removed to build Var.3. Besides, we conduct more ablation studies on the loss function to analyze the contribution of the binned classification and regression on gaze estimation. Var.4 and Var.5 are built to optimize the DGaze-Net with only cross-entropy loss $L_{\text{c}}$ and only mean-square error loss $L_{\text{mse}}$. Var.6 and Var.7 are built to optimize the DGaze-Net with a fixed weighted sum $L=L_{\text{c}}+\alpha L_{\text{mse}}$ ($\alpha$=1, 2). Lastly, we present the result of DGaze-Net optimized by the proposed joint loss of $L_{\text{c}}$ and $L_{\text{mse}}$ with adaptive coefficients.

Comparing Var.1 and Var.3 with DGaze-Net and Var.2, we can find that the methods using the attention-based feature extraction module as the backbone have better accuracy than those without attention modules. This proves the necessity of the attention mechanism that enables the network to focus on features most relevant to gaze. DGaze-Net and Var.1 achieve better accuracy than Var.2 and Var.3. This proves the effectiveness of our discretization-based gaze predictor. Besides, it can be noticed that DGaze-Net using the proposed joint loss with adaptive coefficients achieves better accuracy than Var.4, Var.5, Var.6, and Var.7. This proves the rationality of the proposed loss. In particular, DGaze-Net has the best performance when both the attention-based feature extraction module and the discretization-based gaze angle predictors are used simultaneously.

**Tab.3 Ablation study on MPIIGaze**

| Method | Angular error |
|---|---|
| Var.1 (ResNet-50 + Gaze angle predictors) | 5.4° |
| Var.2 (Feature extraction module + FC) | 5.3° |
| Var.3 (ResNet-50 + FC) | 5.7° |
| Var.4 (Optimized only by $L_c$) | 5.7° |
| Var.5 (Optimized only by $L_{mse}$) | 5.4° |
| Var.6 (Optimized by $L_c + L_{mse}$) | 5.2° |
| Var.7 (Optimized by $L_c + 2 \cdot L_{mse}$) | 5.1° |
| **DGaze-Net** (Adaptive coefficients) | **4.8°** |

In this work, we present the DGaze-Net, a simple architecture for monocular 3D gaze estimation. We exploit feature discretization and attention mechanisms to modify the gaze predictor and the backbone of DGaze-Net. The DGaze-Net reports competitive performance on eye-based gaze estimation on three gaze datasets. Ablation studies demonstrate the validity of the proposed structure.

## Statements and Declarations

The authors declare that there are no conflicts of interest related to this article.

## References

[1] CANIGUERAL R, HAMILTON A F D C. The role of eye gaze during natural social interactions in typical and autistic people[J]. Frontiers in psychology, 2019, 10.

[2] LI B, ZHANG Y, ZHENG X, et al. A smart eye tracking system for virtual reality[C]//2019 IEEE MTT-S International Microwave Biomedical Conference (IMBioC), May 6-8, 2019, Nanjing, China. New York：IEEE, 2019：1-3.

[3] WANG H, DONG X, CHEN Z, et al. Hybrid gaze/EEG brain computer interface for robot arm control on a pick and place task[C]//2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), August 25-29, 2015, Milan, Italy. New York：IEEE, 2015：1476-1479.

[4] SIROHEY S, ROSENFELD A, DURIC Z. A method of detecting and tracking irises and eyelids in video[J]. Pattern recognition, 2002, 35(6)：1389-1401.

[5] WU L, XU X, SHEN C. Eye detection and tracking using IR source[J]. Optoelectronics letters, 2006, 2：145-147.

[6] HIROTAKE Y, AKIRA U, TOMOKO Y, et al. Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions[C]//Proceedings of the 2008 Symposium on Eye Tracking Research & Applications (ETRA '08), March 26-28, 2008, Savannah, Georgia. New York：ACM, 2008：245-250.

[7] ZHANG X, SUGANO Y, FRITZ M, et al. Appearance-based gaze estimation in the wild[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, USA. New York：IEEE, 2015：4511-4520.

[8] ZHANG X, SUGANO Y, FRITZ M. MPIIGaze：real-world dataset and deep appearance-based gaze estimation[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 41(1)：162-175.

[9] CHENG Y, LU F, ZHANG X. Appearance-based gaze estimation via evaluation-guided asymmetric regression[C]//2018 European Conference on Computer Vision (ECCV), September 8-14, 2018, Munich, Germany. Berlin, Heidelberg：Springer, 2018.

[10] CHEN Z, SHI B E. Appearance-based gaze estimation using dilated-convolutions[C]//2018 Asian Conference on Computer Vision, December 2-6, 2018, Perth, Australia. Berlin, Heidelberg：Springer, 2019：309-324.

[11] ZHANG X, SUGANO Y, FRITZ M, et al. It's written all over your face：full-face appearance-based gaze estimation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, July 21-26, 2017, Honolulu, HI, USA. New York：IEEE, 2017.

[12] LIU S, LIU D, WU H. Gaze estimation with multi-scale channel and spatial attention[C]//Proceedings of the 2020 9th International Conference on Computing and Pattern Recognition (ICCPR), October 30-November 1, 2020, Xiamen, China. New York：ACM, 2020：303-309.

[13] MURTHY L R D, BISWAS P. Appearance-based gaze estimation using attention and difference mechanism[C]//2021 IEEE Conference on Computer Vision and Pattern Recognition Workshops, June 20-25, 2021, virtual. New York：IEEE, 2021：3143-3152.

[14] ABDELRAHMAN A, HEMPEL T, KHALIFA A, et al. L2CS-Net：fine-grained gaze estimation in unconstrained environments[C]//Proceedings of the 2022 29th IEEE International Conference on Image Processing (ICIP), October 16-19, 2022, Bordeaux, France. New York：IEEE, 2022.

[15] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//Proceedings of the 2015 International Conference on Learning Representations (ICLR), May 7-9, 2015, San Diego, CA, USA. Banff：Computational and Biological Learning Society, 2015.

[16] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 26-July 1, 2016, Las Vegas, USA. New York：IEEE, 2016.

[17] WOO S, PARK J, LEE J Y, et al. CBAM：convolutional block attention module[C]//Proceedings of the 2018

European Conference on Computer Vision (ECCV), September 8-14, 2018, Munich, Germany. Berlin, Heidelberg：Springer, 2018.

[18] KENDALL A, GAL Y, CIPOLLA R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 18-22, 2018, Salt Lake City, USA. New York：IEEE, 2018：7482-7491.

[19] FISCHER T, CHANG H, DEMIRIS Y. RT-GENE：real-time eye gaze estimation in natural environments[C]//Proceedings of the 2018 European Conference on Computer Vision (ECCV), September 8-14, 2018, Munich, Germany. Berlin, Heidelberg：Springer, 2018.

[20] SUGANO Y, MATSUSHITA Y, SATO Y. Learning-by-synthesis for appearance-based 3D gaze estimation[C]//Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 23-28, 2014, Columbus, USA. New York：IEEE, 2014：1821-1828.

[21] XIONG Y, KIM H, SINGH V. Mixed effects neural networks (MeNets) with applications to gaze estimation[C]//Proceedings of 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 16-20, 2019, Long Beach, USA. New York：IEEE, 2019：7735-7744.

[22] BERNARD V, WANNOUS H, VANDEBORRE J P. Eye-gaze estimation using a deep capsule-based regression network[C]//Proceedings of the 2021 International Conference on Content-Based Multimedia Indexing (CBMI), June 28-30, 2021, Lille, France. New York：IEEE, 2021：1-6.

[23] MAHMUD Z, HUNGLER P, ETEMAD A. Multistream gaze estimation with anatomical eye region isolation by synthetic to real transfer learning[EB/OL]. (2022-06-18) [2022-11-12]. https：//arxiv.org/abs/2206.09256.