PerimetryNet: A multi-scale fine grained deep network for 3D eye gaze estimation using visual field analysis

Shuqing Yu
Zhejiang University
shuqingyu@zju.edu.cn

Zhihao Wang
Zhejiang University
zhihao_wang@zju.edu.cn

Shuowen Zhou
Zhejiang University
3150103275@zju.edu.cn

Xiaosong Yang
Bournemouth University
xyang@bournemouth.ac.uk

Chao Wu
Zhejiang University
hao.wu@zju.edu.cn

Zhao Wang*
Zhejiang University
zhao_wang@zju.edu.cn

Abstract

3D gaze estimation aims to reveal where a person is looking, which plays an important role in identifying users' point-of-interest in terms of the direction, attention and interactions. Appearance-based gaze estimation methods could provide relatively unconstrained gaze tracking from commodity hardware. Inspired by medical perimetry test, we have proposed a multi-scale framework with visual field analysis branch to improve estimation accuracy. The model is based on the feature pyramids and predicts vision field to help gaze estimation. In particular, we analysis the effect of the multi-scale component and the visual field branch on challenging benchmark datasets: MPIIGaze and EYEDIAP. Based on these studies, our proposed PerimetryNet significantly outperforms state-of-the-art methods. In addition, the multi-scale mechanism and visual field branch can be easily applied to existing network architecture for gaze estimation. Related code would be available at public repository `https://github.com/gazeEs/PerimetryNet`.

Keywords: Gaze Estimation, Multi-Scale, Fine Grained, Visual Field, MPIIGaze, EYEDIAP

1 Introduction

Eye gaze estimation has been an attractive research area since its numerous application areas such as human–computer interaction, saliency detection and virtual reality by identifying the users' point-of-interest. It provides important cues for human cognition understanding [1, 2], automotive [3, 4], aviation [5], accessibility [6, 7] and visual scan path analysis [8].

Gaze estimation methods could be generally divided into feature & model-based methods and appearance-based methods. Early feature & model-based methods have employed infrared (IR) imaging or high-resolution cameras techniques and achieved commercial realm. However, such kind of solutions are mostly limited to laboratory environment due to short working distance and lack of robustness in the wild. A series of appearance-based gaze estimation methods have been proposed since they could provide relatively unconstrained gaze tracking

from commodity hardware. The development of appearance-based methods has facilitated the widespread of gaze tracking without additional cost on many platforms, such as mobile devices [9]. In addition, the application of deep convolutional neural networks (CNNs) has reduced estimation error of appearance-based system significantly [10].

In order to conduct high quality gaze estimation, previous works have tried to use multi-stream input [11, 12, 13, 14, 15],dilated convolution [13, 14, 15], multi-task techniques [16, 17], unsupervised representation learning[18] and zero-shot learning[19]. Gaze information is directly conducted as a regression result from estimation. In this work, the gaze estimation is modified as a dual task including a regression branch and a support classification branch to improve the per-angel prediction accuracy. The idea of support classification branch is inspired by medical perimetry test, where the visual field is categorized into location grids describe maps of light sensitivity. Therefore, a perimetry branch is designed in our work that calculate gaze vision field classification loss.

To verify the superiority of the proposal idea, extensive experiments are conducted on the challenging benchmark dataset MPIIGaze[20]. Our model significantly outperforms state-of-the-art works in extensive evaluations. The contributions of this work are summarized as follows:

- For the first time, we investigate the strategy of visual field analysis in-depth and propose a novel perimetry test branch for the gaze estimation problem.
- We design a multi-scale 3D gaze estimation framework with dual tasks. Our experiment demonstrates that visual field information could offer complementary information for gaze estimation.
- We conduct comprehensive evaluations on benchmark datasets and achieve a significant improvement. Notably, the proposed perimetry branch is generic to seamlessly work with existing multi-stream appearance-based model.

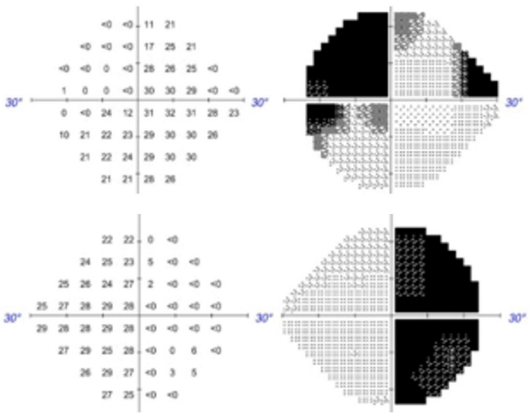

Figure 1: An example of Perimetry test result with ZEISS system: left eye (top),right eye (bottom). The numbers in the left column and grey in right column stands for the light sensitivity of visual field.

2 Related Work

In this section, 3D Gaze estimation methods including two categories: feature & model-based approaches and appearance-based approaches, would be reviewed respectively.

2.1 Feature & model-based gaze estimation methods

Commercial gaze-trackers are usually conducted by feature based gaze estimation methods using infra-red (IR) imaging techniques, where such approaches are based on well established theory that using the pupil centers and corneal reflections. For instance, Guestrin and Eizenman have presented a point of gaze estimation system in a desk top settings with an evaluation on 4 subjects [21]. A mobile device based gaze estimation has been reported by Brousseau et al that compensates for the relative roll between the system and subject's eyes [22]. Recent Tobii [1] is a predominantly feature-based gaze estimation system that claim to provide gaze accuracy of less than 1.9° error under real-world usage conditions. In addition to the above desktop setting, many real-time pupil detection in-the-wild approaches have been designed for gaze estimation [23, 24, 25, 26, 27].

[1] www.tobii.com/

However, few work has been reported about the performance comparison of these approaches in term of gaze estimation via pupil detection.

Model based gaze estimation methods extract visual features such as pupil, eyeball center and eye corners that aim to fit a geometric 3D eyeball model to conduct gaze estimation. Early model-based approaches mainly rely on high resolution cameras for feature extraction but suffers from illumination variation problem [28, 29, 30]. Later approaches have tried to overcome such requirements that only commodity web cameras are adopted, while machine learning techniques are used to empower feature extraction [31, 32].

2.2 Appearance-based gaze estimation methods

Compare with aforementioned feature & model based approaches, appearance-based methods aim to estimate gaze directions directly from images that captured using commodity cameras without handcrafted feature [16, 20]. Considering the time consuming personal calibration for every participant required by feature & model-based gaze estimation methods, these appearance-based methods have significant advantages and are benefited with the rapid growth of datasets and advancements in deep learning techniques. Considering the input type, appearance-based gaze estimation method could be generally classified into single-stream and multi-stream methods.

GAZENet [20] is a well known appearance-based gaze estimation method, which is a single-stream network that using single eye image as input. In the following work, a single-stream Spatial-Weights CNN model is proposed with full face image as input [16] where additional layers that learn spatial weights for last convolutional layer is introduced to use face information effectively. A head pose branch is designed in a CNN model which involves signal eye image and head pose [33]. A multi-task CNN based approach is proposed to extract eyeball feature, head pose and 3D eye position for gaze prediction with eye images and RGBD head image as input [34].

Multi-stream methods are proposed alternative to single-stream methods. A multi-stream approach called iTracker is designed that using left eye image, right eye image, face image and face grid as input[11]. Dilated convolution mechanism has also been applied with eye images and face image as input, where dilated convolutions are used to instead of several max pooling layers in their model [13, 14]. An extended dilated convolutional approach has introduced a gaze decomposition method that decomposes the gaze angle into the sum of a subject-independent gaze estimate from the image and a subject-dependent bias [15]. AGE-Net has tried to use an attention mechanism with the dilated convolution model [35]. FAR-Net has proposed a face-based asymmetric regression-evaluation network that utilize the asymmetry between subject's two eyes, where a confidence score is obtained from each eye[36].

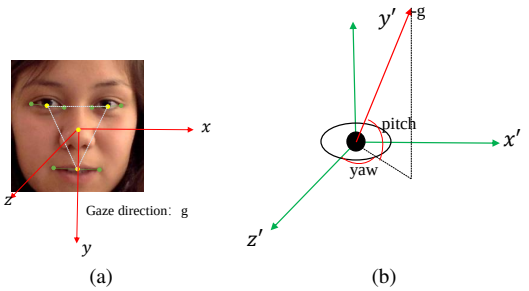

Figure 2: Related concepts of gaze direction. (a).The head coordinate system represented by red lines is defined based on a triangle connecting three midpoints of the eyes and mouth. (b). Pitch and yaw of gaze direction. Green lines represent the three-dimensional coordinate system, and red lines represent the gaze direction.

Apart from those CNN based model, a hybrid transformer approach has been developed where a CNN is employed to learn local feature from face images and a transformer is used as encoder to estimate gaze [37].

3 Methodology

3.1 Pre-processing

Gaze vector g is usually defined as a unit vector from a reference point in the head coordinate system. Figure 2(a) shows the head coordinate

system, which is defined by a triangle connecting the three midpoints of the eyes and mouth. The x axis of the head coordinate system is defined as the direction from the center of the right eye to the center of the left eye. The y axis is defined as the direction perpendicular to the x axis in the triangle plane, and the z axis is defined as the vertical direction from the triangle plane to the back of the subject[38].In the model training, 3d gaze vector is converted into 2D form represented by pitch and yaw, as shown in Figure 2(b).

Different from earlier constrained gaze estimation, appearance-based gaze estimation needs to be implemented in unconstrained environment. Unconstrained gaze estimation complicates the problem by introducing new factors such as head posture, camera distance and illumination. The purpose of data pre-processing is to eliminate the influence of these factors by mapping input images and gaze labels to a standardized space. Sugano et al. proposed a method of data normalization for 3D appearance-based gaze estimation[38]. The basic idea is to first rotate the camera to eliminate the freedom of head rotation and then translate the camera to keep the same distance between the camera and the reference point. In our work, we follow the procedure in [38] to normalize the MPIIGaze dataset [20] and the procedure in [39] to normalize the EYEDIAP dataset [40]. Furthermore, to analyse the strategy of vision field, we split up the continuous gaze target in each dataset into bins with binary labels based on gaze labels. At last, both datasets have bins labels and continuous labels for the estimation of vision field and gaze direction.

3.2 Perimetry: vision field branch

Perimetry test is a systematic measurement of visual field function, where the visual fields are mapped to lights of different sizes and brightness [41]. It is accomplished by keeping the size and location of a target constant and varying the brightness until the dimmest target the patient can see at each of the test locations is found. Such results are essential in diagnosing diseases of the visual system since different patterns of visual loss are found with diseases of the eye, optic nerve central nervous system. An example of Perimetry test results from ZEISS system [2] is shown in Figure 1, where the numbers in left column indicate the light sensitivity threshold and right column reflect the visual field function area that dark means visual function loss. Hence, we divided the visual field into grids by pitch and yaw, where a example of the data distribution is illustrated in Figure 3. Hence, a classification task would be conducted once the feature extracted from backbone model to identify the corresponding grid. The details of architecture and loss function would be presented in the following paragraph.

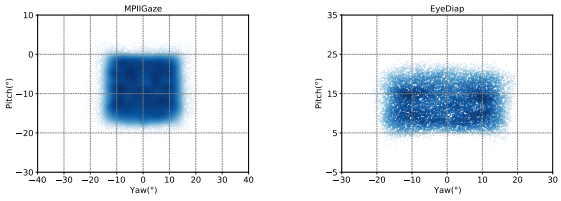

Figure 3: Distribution of gaze on MpiiGaze(left) and EYEDIAP(right).

3.3 Architecture

We propose PerimetryNet, a network for Gaze Estimation with visual field analysis branch. The PerimetryNet architecture is presented in Figure 4. It is a single-stream model that only use the face images as input since our primary focus has been to investigate the factor of visual field analysis affecting the performance.

The input images, collectively referred to by $X_{i,j}$, are first fed separately to the backbone CNN model. The features are extracted by the backbone model Resnet50 and five feature maps C1 to C5 of different scales can be get. We select feature maps C3, C4 and C5 for subsequent processing. In feature pyramid network, channel numbers will be adjusted by 1×1 convolution. Then, up-sampling and feature fusion are carried out for enhanced feature extraction. Multi-scale feature maps P3 to P5 can be get. Inspired by Retinaface[42], we also apply independent SSH modules on three feature pyramid levels to increase the receptive field. The SSH module uses a stack of 3×3 convolution replace

[2] www.zeiss.com/meditec/us/product-portfolio/perimetry.html

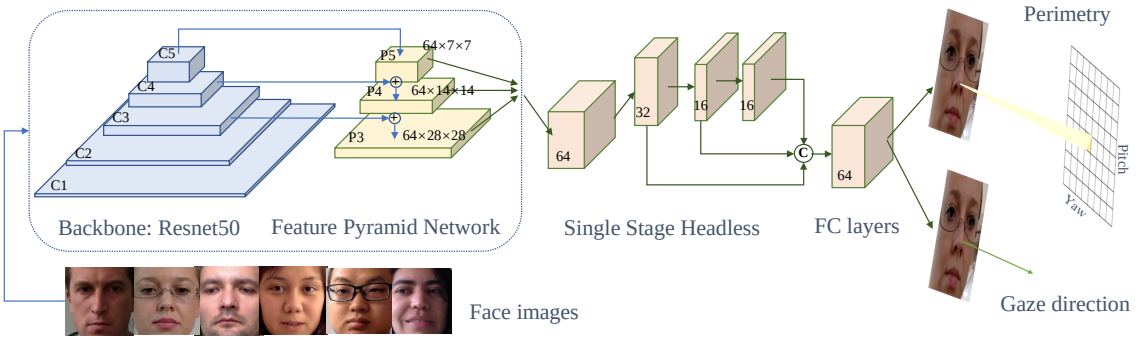

Figure 4: Network Architecture of proposed approach. PerimetryNet is designed based on the feature pyramids with three scales. For each scale of feature maps, we apply SSH modules. At last, fully connected layers are used to predict the field of view and gaze direction.

the 5×5 convolution and the 7×7 convolution. At last, each of feature maps is connected to two fully connected layers. One of the fully connected layers is used to predict the the field of view and the other is used to predict gaze direction. More specifically, In the vision field branch, we divide the field into pitch and yaw. In the gaze direction branch, the gaze vector can also be converted into pitch and yaw. We propose to use two identical losses for pitch and yaw. Each loss consists of a combined focal loss and mean-squared error.

$$L = L_{gaze} + \alpha L_{perimetry} \tag{1}$$

The focal loss is used to increase the weight of hard samples. The focal loss and MSE loss are defined as:

$$L_{perimetry} = -\left(1 - p_i\right)^{\gamma} \log\left(p_i\right) \tag{2}$$

$$L_{gaze} = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - p_i\right)^2 \tag{3}$$

Where $\gamma = 2$, p_i is the predicted value, y_i is the ground-truth and α is the weight of direct prediction of gaze direction.

The backbone model has been initialized from an ImageNet-pretrained model. Transfer learning is appealing that the low level features generate from fine-tuned pretrained model perform well, since gaze datasets contain much fewer samples than large-scale image classification datasets like ImageNet. For instance, we have found that networks with the same structure but trained from random initial weights achieve higher errors compared to PerimetryNet reported in Table 2.

The visual field analysis branch only works in training stage, the gaze information would be directly predicted during inference.

4 Experimental Results and Discussions

In this section, we carried out extensive comparisons and ablation experiments to demonstrate the effectiveness and robustness of our proposed idea. The details of our network configuration are given out in the section. Comprehensive quality and quantity results are reported to show the superiority of the perimetry branch as well as multi-scale strategy against the state-of-the-art models. Discussions of these experiments point out the interesting findings of our work.

4.1 Dataset and Metric

The performance of proposed method is evaluated on two challenging benchmark datasets: MPIIGaze [20] and EyeDiap [40].

Evaluation Metric Following most gaze estimation methods, we use gaze angular error (°) as our evaluation metric. Assuming the ground-truth gaze vector is $g \in \mathbb{R}^3$ and the predicted gaze vector is $\hat{\mathrm{g}} \in \mathbb{R}^3$, the gaze angular error(°) can be computed as:

$$\mathcal{L}_{angular} = \frac{\mathrm{g} \cdot \hat{\mathrm{g}}}{\|\mathrm{g}\| \|\hat{\mathrm{g}}\|} \tag{4}$$

Table 1: Comparison with SOTA methods for 3D gaze estimation on MPIIGaze and EYEDIAP.

Methods	Publisher	MPIIGaze(degree)	EYEDIAP(degree)
iTraker(AlexNet)[11]	CVPR 2016	5.6°	9.9°
Spatial-Weights CNN[16]	CVPRW 2017	4.8°	6.0°
RT-Gene(4 model)[43]	ECCV 2018	4.3°	/
MeNet[44]	CVPR 2019	4.9°	/
Bayesian Approach[45]	CVPR 2019	4.3°	9.9°
FAR-Net[36]	IEEE TIP 2020	4.3°	5.71°
CA-Net[12]	AAAI 2020	4.1°	5.3°
I2DNet[14]	JEMR 2021	4.3°	/
AGE-Net[35]	CVPRW 2021	4.09°	/
GazeTR[37]	Arxiv 2021	4.0°	5.17°
GEDDNet[15]	IEEE TPAMI 2022	4.5°	5.4°
L2CS-Net[46]	Arxiv 2022	3.92°	/
PrimetryNet(Ours)	/	**3.72°**	**5.10°**

MPIIGaze MPIIGaze dataset is collected in real-world conditions with 15 people from diverse ethnic backgrounds under illumination, appearance and head pose variation. We use the "Evaluation Subset", which contains 3,000 images per subject, which is 45000 samples in total. The reference point for image normalization is set to the center of the face. We follow same procedure as state-of-the-art works [11, 12, 14, 15, 16, 35, 36, 43, 44] used for cross-subject validation.

EYEDIAP EYEDIAP dataset contains videos of full face with continuous screen target, discrete screen target or floating target, and with static or dynamic head pose. We followed the evaluation protocol described in [16], which is used in most SOTA works. We use the data from screen targets that used in prior work, where 14 subjects (three female, none with glasses) are included. The reference point for image normalization is set to the midpoint of both eyes [47].

4.2 Implementation Details

The entire model has been trained on 2 NVIDIA RTX 3090 GPUs while PyTorch is used as deep learning frameworks. Furthermore, we apply Adam optimizer with a learning rate of 0.0001 with a step of 5 times decay at the 5th, the 10th and 20th epochs. For simplicity, weight α is set as 1, while a resnet-50 is employed as backbone model in the ablation study. After that, our perimetry branch and main gaze estimation branch are installed with the backbone models for the fine-tuning process to further enhance the model performance. The grid of visual field is set as 3° for MPIIGaze and 3° for EYEDIAP.

4.3 Comparison with SOTA methods

We compare our work with prior works [11, 12, 14, 15, 16, 35, 36, 37, 43, 44, 45, 46] that use face images or face plus eye images as input on the MPIIGaze and EYEDIAP datasets. The results are shown in Table 1. The proposed PerimetryNet achieves a 3.72° mean angular error on MPIIGaze and a 5.10° mean angular error on EYEDIAP, outperforming the state-of-the-art significantly. Detail performance on each subject is presented in Figure 5. Some examples from EYEDIAP with estimated gaze are illustrated in Figure 6.

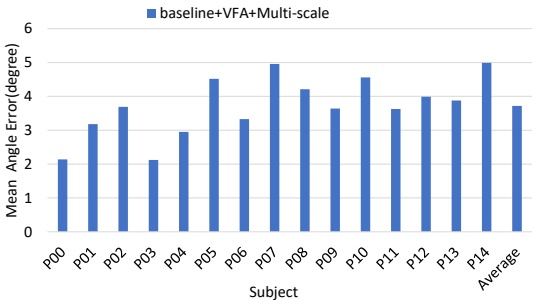

Figure 5: Mean angle error for subjects in MPIIGaze.

Table 2: Ablation study on MPIIGaze.

Methods w/ pretrain	MAE
baseline	4.29°
baseline + VFA	3.83°
baseline + VFA + Multi-scale (**PrimetryNet**)	**3.72**°
PrimetryNet w/o pretrain	3.81°

4.4 Ablation Study

Experiments in this section are conducted on MPIIGaze dataset for the purpose of analysing the different components and designs of our method. The overall test accuracy of experiments is compared using MAE while Resnet-50 is chosen as the baseline model. The training details are available in the public repository. The results shown in Table 2 have indicated that the proposed visual field branch and multi-scale mechanism are able to improve the performance of gaze estimation. The transfer learning strategy with Imagenet-pretrained model parameters could also reduce error.

5 Conclusion

To sum up, a multi-scale gaze estimation frame work is proposed in this paper, where a visual field analysis branch that inspired by medical perimetry test is designed to improve estimation accuracy. A comprehensive study has been taken to investigate the effect of the multi-scale component and the visual field branch . The experimental results have shown that the proposed PerimetryNet outperforms state-of-the-art methods on two challenging benchmarks. The ablation study has also presented the effectiveness of the multi-scale component and the visual field branch. In addition, the proposed multi-scale mechanism and visual field branch can be easily applied to existing network architecture, e.g. multi-stream framework, for gaze estimation. A further study on calibration settings like [15] will be taken in future.

Acknowledgements

References

[1] Rima-Maria Rahal and Susann Fiedler. Understanding cognitive and affective mechanisms in social psychology through eye-tracking. *Journal of Experimental Social Psychology*, 85:103842, 2019.

[2] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5396–5406, 2020.

[3] Tony Poitschke, Florian Laquai, Stilyan Stamboliev, and Gerhard Rigoll. Gaze-based interaction on multiple displays in an automotive environment. In *2011 IEEE International Conference on Systems, Man, and Cybernetics*, pages 543–548. IEEE, 2011.

[4] Gowdham Prabhakar, Aparna Ramakrishnan, Modiksha Madan, LRD Murthy, Vinay Krishna Sharma, Sachin Deshmukh, and Pradipta Biswas. Interactive gaze and finger controlled hud for cars. *Journal on Multimodal User Interfaces*, 14(1):101–121, 2020.

[5] LRD Murthy, Abhishek Mukhopadhyay, Varshit Yellheti, Somnath Arjun, Peter Thomas, M Dilli Babu, Kamal Preet Singh Saluja, DV JeevithaShree, and Pradipta Biswas. Evaluating accuracy of eye gaze controlled interface in military aviation environment. In *2020 IEEE Aerospace Conference*, pages 1–12. IEEE, 2020.

[6] Maria Borgestig, Jan Sandqvist, Gunnar Ahlsten, Torbjörn Falkmer, and Helena Hemmingsson. Gaze-based assistive technology in daily activities in children with severe physical impairments–an intervention study. *Developmental Neurorehabilitation*, 20(3):129–141, 2017.

[7] Yu-Hsin Hsieh, Maria Borgestig, Deepika Gopalarao, Joy McGowan, Mats Granlund, Ai-Wen Hwang, and

Figure 6: Gaze estimation result visualization for subjets from EYEDIAP. Red lines show the predicted gaze direction.

Helena Hemmingsson. Communicative interaction with and without eye-gaze technology between children and youths with complex needs and their communication partners. *International journal of environmental research and public health*, 18(10):5134, 2021.

[8] Sukru Eraslan, Yeliz Yesilada, and Simon Harper. Eye tracking scanpath analysis techniques on web pages: A survey, evaluation and comparison. *Journal of Eye Movement Research*, 9(1), 2016.

[9] Nachiappan Valliappan, Na Dai, Ethan Steinberg, Junfeng He, Kantwon Rogers, Venky Ramachandran, Pingmei Xu, Mina Shojaeizadeh, Li Guo, Kai Kohlhoff, et al. Accelerating eye movement research via accurate and affordable smartphone eye tracking. *Nature communications*, 11(1):1–12, 2020.

[10] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4511–4520, 2015.

[11] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2176–2184, 2016.

[12] Yihua Cheng, Shiyao Huang, Fei Wang, Chen Qian, and Feng Lu. A coarse-to-fine adaptive network for appearance-based gaze estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10623–10630, 2020.

[13] Zhaokang Chen and Bertram E Shi. Appearance-based gaze estimation using dilated-convolutions. In *Asian Conference on Computer Vision*, pages 309–324. Springer, 2018.

[14] LRD Murthy, Siddhi Brahmbhatt, Somnath Arjun, and Pradipta Biswas. I2dnet-design and real-time evaluation of appearance-based gaze estimation system. *Journal of Eye Movement Research*, 14(4), 2021.

[15] Zhaokang Chen and Bertram Shi. Towards high performance low complexity calibration in appearance based gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[16] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written

all over your face: Full-face appearance-based gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 51–60, 2017.

[17] Yuk-Hoi Yiu, Moustafa Aboulatta, Theresa Raiser, Leoni Ophey, Virginia L Flanagin, Peter Zu Eulenburg, and Seyed-Ahmad Ahmadi. Deepvog: Open-source pupil segmentation and gaze estimation in neuroscience using deep learning. *Journal of neuroscience methods*, 324:108307, 2019.

[18] Yu Yu and Jean-Marc Odobez. Unsupervised representation learning for gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7314–7324, 2020.

[19] Yang Liu, Lei Zhou, Xiao Bai, Yifei Huang, Lin Gu, Jun Zhou, and Tatsuya Harada. Goal-oriented gaze estimation for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3794–3803, 2021.

[20] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):162–175, 2017.

[21] Elias Daniel Guestrin and Moshe Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on biomedical engineering*, 53(6):1124–1133, 2006.

[22] Braiden Brousseau, Jonathan Rose, and Moshe Eizenman. Accurate model-based point of gaze estimation on mobile devices. *Vision*, 2(3):35, 2018.

[23] Wolfgang Fuhl, Thomas Kübler, Katrin Sippel, Wolfgang Rosenstiel, and Enkelejda Kasneci. Excuse: Robust pupil detection in real-world scenarios. In *International conference on computer analysis of images and patterns*, pages 39–51. Springer, 2015.

[24] Wolfgang Fuhl, Thiago C Santini, Thomas Kübler, and Enkelejda Kasneci. Else: Ellipse selection for robust pupil detection in real-world environments. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pages 123–130, 2016.

[25] Thiago Santini, Wolfgang Fuhl, and Enkelejda Kasneci. Pure: Robust pupil detection for real-time pervasive eye tracking. *Computer Vision and Image Understanding*, 170:40–50, 2018.

[26] Thiago Santini, Wolfgang Fuhl, and Enkelejda Kasneci. Purest: Robust pupil tracking for real-time pervasive eye tracking. In *Proceedings of the 2018 ACM symposium on eye tracking research & applications*, pages 1–5, 2018.

[27] Shaharam Eivazi, Thiago Santini, Alireza Keshavarzi, Thomas Kübler, and Andrea Mazzei. Improving real-time cnn-based pupil detection through domain-specific data augmentation. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, pages 1–6, 2019.

[28] Kenneth Alberto Funes Mora and Jean-Marc Odobez. Geometric generative gaze estimation (g3e) for remote rgb-d cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1773–1780, 2014.

[29] Li Sun, Zicheng Liu, and Ming-Ting Sun. Real time gaze estimation with a consumer depth camera. *Information Sciences*, 320:346–360, 2015.

[30] Stefania Cristina and Kenneth P Camilleri. Model-based head pose-free gaze estimation for assistive communication. *Computer Vision and Image Understanding*, 149:157–170, 2016.

[31] Seonwook Park, Xucong Zhang, Andreas Bulling, and Otmar Hilliges. Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In *Proceedings of the 2018 ACM symposium on eye tracking research & applications*, pages 1–10, 2018.

[32] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE, 2018.

[33] Rajeev Ranjan, Shalini De Mello, and Jan Kautz. Light-weight head pose invariant gaze tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2156–2164, 2018.

[34] Dongze Lian, Ziheng Zhang, Weixin Luo, Lina Hu, Minye Wu, Zechao Li, Jingyi Yu, and Shenghua Gao. Rgbd based gaze estimation via multi-task cnn. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2488–2495, 2019.

[35] Pradipta Biswas et al. Appearance-based gaze estimation using attention and difference mechanism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3143–3152, 2021.

[36] Yihua Cheng, Xucong Zhang, Feng Lu, and Yoichi Sato. Gaze estimation by exploring two-eye asymmetry. *IEEE Transactions on Image Processing*, 29:5259–5272, 2020.

[37] Yihua Cheng and Feng Lu. Gaze estimation using transformer. *arXiv preprint arXiv:2105.14424*, 2021.

[38] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1821–1828, 2014.

[39] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Revisiting data normalization for appearance-based gaze estimation. In *Proceedings of the 2018 ACM symposium on eye tracking research & applications*, pages 1–9, 2018.

[40] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proceedings of the symposium on eye tracking research and applications*, pages 255–258, 2014.

[41] U Schiefer, J Pätzold, F Dannheim, P Artes, and W Hart. Conventional perimetry. *Ophthalmologe*, 102(6):627–646, 2005.

[42] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5203–5212, 2020.

[43] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 334–352, 2018.

[44] Yunyang Xiong, Hyunwoo J Kim, and Vikas Singh. Mixed effects neural networks (menets) with applications to gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7743–7752, 2019.

[45] Kang Wang, Rui Zhao, Hui Su, and Qiang Ji. Generalizing eye tracking with bayesian adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11907–11916, 2019.

[46] Ahmed A Abdelrahman, Thorsten Hempel, Aly Khalifa, and Ayoub Al-Hamadi. L2cs-net: Fine-grained gaze estimation in unconstrained environments. *arXiv preprint arXiv:2203.03339*, 2022.

[47] Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. Appearance-based gaze estimation with deep learning: A review and benchmark. *arXiv preprint arXiv:2104.12668*, 2021.