KD-Eye: Lightweight Pupil Segmentation for Eye Tracking on VR Headsets via Knowledge Distillation

Yanlin Li, Ning Chen, Guangrong Zhao, and Yiran Shen*

School of Software, Shandong University, Jinan, China
{yanlinli, ningchen}@mail.sdu.edu.cn
{guangrong.zhao, yiran.shen}@sdu.edu.cn

Abstract. Pupil segmentation extracts the pupil area of human eyes to track the eye movement. It is the first step for most of the vision-based gaze tracking systems which have attracted significant interests from both academic and industrial communities. When incorporated in commercial VR/AR devices, the accuracy, robustness and efficiency of the pupil segmentation is the fundamental for the successful usage of gaze tracking in human computer interaction, foveated rendering and etc. In this paper, we propose, **KD-Eye**, a lightweight vision-based pupil segmentation approach to realize accurate and efficient estimation of pupil regions. We introduce a coarse-to-fine strategy to significantly reduce the computation of the algorithm and complexity of the network. Then knowledge distillation is applied to guide the training of a lightweight student network with a large teacher network trained by a public dataset. According to our evaluation on real-world dataset and implementation on a VR platform, the coarse-to-fine strategy and lightweight network can speed up the segmentation process by over 240 times and the knowledge distillation can improve the accuracy of the student network especially when the number of training samples are limited.

Keywords: Pupil segmentation · Eye tracking · Knowledge distillation

1 Introduction

In recent years, the emerging of metaverse has attracted significant attentions from medical, scientific and industrial communities on the relevant technologies, e.g., eye tracking. There have been a number of successful applications of eye tracking technologies on different key scenarios. For examples, in the field of consumer electronics, eye tracking technology is a new modality for human-computer interaction [1] and identity authentication [2] for VR/AR headsets. Moreover, eye tracking technology is pervasively used in human visual attention analysis for market research [3], psychological research and diagnosis along with EEG signals [4] or fMRI [5].

* Corresponding author

The vision-based approaches are currently the major solutions for eye tracking systems [6–10]. They utilize CCD/CMOS cameras to achieve high tracking accuracy and low design cost. The vision-based approaches normally start with a pupil segmentation component to localize the center of eye balls and track the movement of eyes in image plane. Then a polynomial regression model [11] can be applied to map the movement of pupil in image domain to the point of gaze, i.e., what the user is looking at. In this paper, we focus on the pupil segmentation component, which is the fundamental of accurate eye tracking.

The vision-based pupil segmentation approaches can be vastly categorized as model-based [12, 13] and appearance-based [8, 14] approaches. The model-based approaches fit the observed eye images into parametric pupil shape with regression models. However, the accuracy of model-based approaches is highly dependent on the quality of the eye images and is not robust to noises, partial occlusion and challenging light condition. While appearance-based approaches [8, 14] leverage the advanced deep learning techniques to segment out the pupil region with high accuracy and they are normally robust to the variant light and appearance conditions of eye images.

When it comes to VR/AR devices, the eye tracking can be new modality for human-computer interaction [15, 16] and activities recognition [17, 18] or an enabler for adaptive scene rendering [19]. However, constrained by the limited resources on computing, energy and running memory, it is challenging to realize real-time eye tracking with common resource-consuming appearance-based approach and the pupil segmentation with deep learning model is the bottleneck. For example, according to our evaluation, a UNet-based segmentation model [20] takes over 1.5 seconds to extract the pupil region and is far from the requirement of real-time eye tracking to provide immersive user experience for applications on VR.

To bridge the gap above, we propose, **KD-Eye**, a lightweight appearance-based pupil region segmentation approach via coarse-to-fine strategy and knowledge distillation to realize efficient and accurate segmentation. The contributions of this paper can be summarized as follows,

- We propose **KD-Eye**, an efficient and accurate approach for eye segmentation. The efficiency is achieved by designing a simple fully-convolution model, **KD-EffNet** as the backbone to reduce the model complexity and a coarse-to-fine strategy to reduce the dimensionality of input eye images.
- Knowledge distillation is introduced to guide the training process of **KD-Eye** to guarantee the segmentation accuracy of the student model, i.e., **KD-EffNet** with limited number of training samples.
- Finally, according to our evaluation on real-world dataset and implementation on resource-constrained VR platform, **KD-Eye** achieves comparable segmentation accuracy with large model while it is 240× faster than the competing approach.

The organizing of the rest of paper is as follows. We first review the related work on eye tracking in Section 2. Then the new lightweight pupil segmentation approach **KD-Eye** is described in Section 3. Section 4 presents extensive

evaluations on eye tracking datasets, and, finally, Section 5 concludes the whole paper.

2 Related Work

In this section, we will discuss the related work on invasive eye tracking and vision-based pupil segmentation.

The research on automatic eye tracking or gaze tracking systems start with invasive approaches which utilizes electro-oculography [21] or scleral search coil [22] to monitor the electrooculographic signals or change of magnetic field caused by the movement of eye-balls. However, both of the two methods require specially-designed hardware makes them hard to be used pervasively. Then the other issue is their invasive property makes them not user-friendly. For example, the scleral search coil approach [22] requires the user to wear a scleral contact lens with embedded coil of wire which makes it uncomfortable to wear for many people.

Then vision-based approaches are proposed to track the eye movement in a non-invasive way [12, 13, 8, 14]. Because it only requires CCD/CMOS cameras to capture eyes of users, its utility can be significantly extended to different application scenarios. To track the eye movement, it normally start with finding the region of pupil in image domain, or termed as pupil segmentation. The accurate pupil segmentation is the basis for the performance of following eye tracking, therefore, attracts significant research efforts.

The approaches of vision-based pupil segmentation can be vastly categorized as model-based [12, 13] and appearance-based [8, 14]. The model-based approach normally utilize the assumption that the pupil edge is conspicuous and can be represented as a ellipse. Then the ellipse shape is fitted with limited number of parametric features via polynomial regression. However, the model-based approach requires high quality and high resolution eye images to achieve accurate segmentation and are vulnerable under challenging lighting conditions. With the success of deep learning on image processing, pupil segmentation can be solved in a data-driven approach, i.e., appearance-based [14, 8]. These approaches normally train a complex deep neural network using large number of labeled eye images, i.e., whose pupil region is explicitly labeled with manual work. Then the trained neural network can be employed for the pupil segmentation tasks for the unlabeled images. To promote the research on eye tracking, there are a number of relevant datasets being published, such as EV-Eye [23], TEyeD [24], [10], ETH-XGaze [25] and [26].

3 Method

In this paper, we propose **KD-Eye**, a lightweight appearance-based approach for pupil segmentation. As shown in Figure 1, **KD-Eye** adopts a coarse-to-fine strategy to achieve both high segmentation accuracy and running efficiency. It starts with extracting the area of interest of the near-eye images to avoid processing the irrelevant pixels, such as the skin, eye bows, frames of glasses

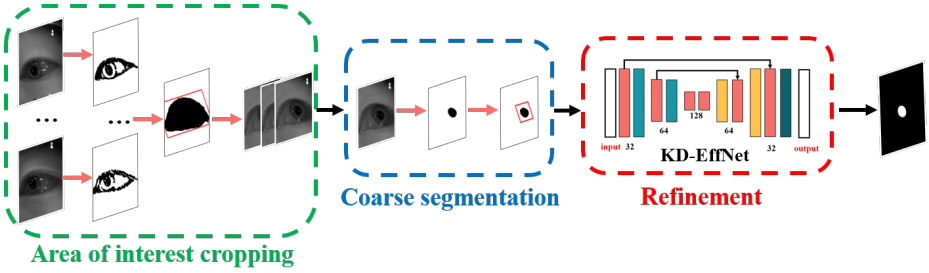

Fig. 1. Overview of KD-Eye.

and etc. Then an efficient model-based segmentation method is introduced to extract the coarse area of pupil. At last, **KD-EffNet** with only few convolutional layers is designed to refine the segmentation results. To promote the accuracy of **KD-EffNet**, knowledge distillation is adopted to train **KD-EffNet** with only limited number of training samples.

3.1 Area-of-interest (AoI) Extraction

As we focus on the application of eye segmentation for eye tracking in VR/AR devices, the locations of eyes of different users are similar in the collected near-eye images. Therefore, there are a large number of redundant pixels only recording the irrelevant parts around eyes, such as skin of face, eye bows, frames of glasses and etc. Those redundant pixels account for significant computation of the following processing as we aim for pixel-wise segmentation. The area-of-interest (AoI) can be obtained by analyzing the occupied pixels of eyes in the near-eye images collected from the same device. As shown in the left-most component of Figure 1, each of the near-eye image is first binarized with adaptive thresholding, i.e., the dark pixels with values less than the threshold are set to be "0" (shown as pure black in the image) and those with values larger than threshold are set to be "1" (shown as pure white in the image). Then morphological erosion and dilatation is applied to remove isolated noises, and largest connected component of the binary image is selected to remove the irrelevant dark parts including frames and eye bows. Finally, the AoI is retained by finding a bounding box covers all the pixels occupied by eyes of different users. In this paper, the bounding box accounts for 60% of the whole image. As the above process can be offline, the input size of KD-Eye can be reduced significantly without introducing extra computing burden during the online segmentation stage.

3.2 Coarse Segmentation

Then we propose an efficient model-based pupil segmentation method to obtain a coarse estimation of pupil area. Different from the offline AoI extraction, the

coarse pupil segmentation must be conducted online as the pupil moves significantly within the AoI. The coarse segmentation is based on the fact that the pupil region is darker than the iris region. Therefore, a simple thresholding method with morphological erosion and dilatation can be applied to extract the region of pupil. As the interference of black eye lashes, the segmented region is quite coarse according to the evaluations in Section 4. After the coarse pupil region is obtained, we calculate the diameter d of the minimum outer circle of the region and apply an adaptive bounding box centering at the centroid of the outer circle. The edge length (l) of the bounding box slightly larger than the circle (we set $l = 1.5d$ in our paper) to accommodate the inaccuracy of the coarse segmentation. It is worth noting that the model-based coarse pupil segmentation cannot be applied directly on original near-eye images as the dark eye bows and black frames of glasses make the model-based segmentation challenging and introduces extra online computation.

3.3 Refined segmentation with KD-EffNet

Though the coarse segmentation is efficient and effective to find the pupil area with near-eye images, the boundary of the pupil region cannot be extracted precisely sometimes. It can only provide coarse segmentation results and fails at some pixels due to the influence of lighting conditions, change of appearance, irrelevant movement of eyes, e.g., blinks. To overcome the issues, we propose, **KD-EffNet** an efficient pupil region refinement approach via knowledge distillation. **KD-EffNet** adopts the power of non-linearity of deep learning (DL) and knowledge distillation to improve the accuracy of pupil segmentation compared with model-based approach. To guarantee the efficiency of DL-based approach, the coarse segmentation component above is able to reduce the dimensionality of the inputs significantly. At the same time, **KD-EffNet** adopts a very simple network design to save the computation during pupil segmentation. We will introduce how **KD-EffNet** is designed and trained in the rest of the section.

KD-EffNet training via Knowledge Distillation A number of challenging design issues need to be addressed when considering the implementation of the eye segmentation or tracking systems on resource-constrained devices. First, the resource-constrained devices call for a lightweight approach with simple model as backbone. Second, it is challenging and requires significant efforts to collect and manually-label large-scale dataset in near-eye setting for training the segmentation model in an appearance-based approach.

To solve the challenges above, we design **KD-EffNet**, a lightweight neural network with only few pairs of convolution and pooling layers to achieve efficient segmentation. **KD-EffNet** is trained via knowledge-distillation to achieve accurate segmentation with limited number of self-collected training samples (target dataset). The training pipeline of **KD-EffNet** is shown in Figure 2. The large deep neural network, **Teacher-Net**, is trained by large-scale, publicly-available dataset *TEyeD* [24] (source dataset) with manually-segmented pupil regions while the simple **KD-EffNet** is trained with our small-scale, self-collected dataset under the guidence of the **Teacher-Net**. When training **KD-EffNet**,

the labeled samples from target dataset are feed to both **KD-EffNet** and **Teacher-Net**. Though the source and target datasets are different from lighting conditions, view angles, participants and etc. The common features, like the appearance of eyes, can be exploited to guide the training of **KD-EffNet** with limited samples from target dataset. The loss function L_c of the knowledge distillation is combination of hard loss L_h and soft loss L_s:

$$L_c = \alpha L_s + (1 - \alpha) L_h \tag{1}$$

The hard loss is the cross-entropy loss of the output of softmax when distillation *temperature* $T = 1$, i.e.,

$$L_h = -\frac{1}{n} \sum_{i=1}^{n} y_i \log(h_i) \tag{2}$$

where n is the number of pixels, y_i is the binary value of the i_{th} pixel (it is 1 when belonging to pupil, it is 0 otherwise). h_i is i_{th} output from softmax.

When the distillation *temperature* $T = t$, the soft loss L_s is defined as,

$$L_s = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} \log(s_{ij}) \tag{3}$$

where m is the number of classes which equals to 2 for segmentation problem. x_{ij} is the output of the **Teacher-Net** and s_{ij} is the output of the student-net, i.e., the **KD-EffNet**.

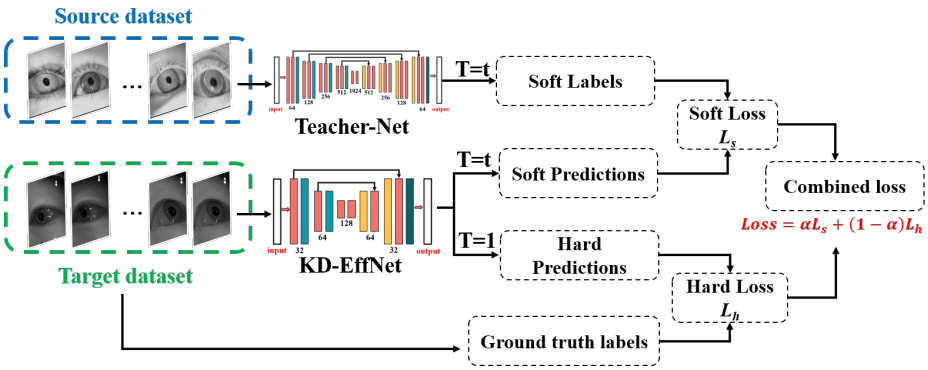

Fig. 2. The training pipeline of **KD-EffNet** via knowledge distillation.

Details of Network Design The deep and complex **Teacher-Net** is based on the architecture of U-Net [20] consisting of an encoder and decoder to extract information of the near-eye images from different scales. As shown in Figure 3, the encoder is a stack of four pairs of convolutional layers following maxpooling.

The structure of decoder is similar to the encoder with up-conv layers instead of maxpooling. At last, 1×1 convolution operation is applied to extract the final segmentation results. Skip connections are adopted to fuse the features extracted from different levels.

The **KD-EffNet** is designed by significantly reducing the number of layers of the **Teacher-Net**. As shown at the bottom of Figure 3, its encoder only employs two convolution layers with maxpooing, thus, the convolution and up-conv layers are also reduced to two pairs and the number of channels of each layer are also significantly reduced. Intuitively, the number of parameters of **KD-EffNet** is only 1/34 of **Teacher-Net** (0.9M v.s. 31M).

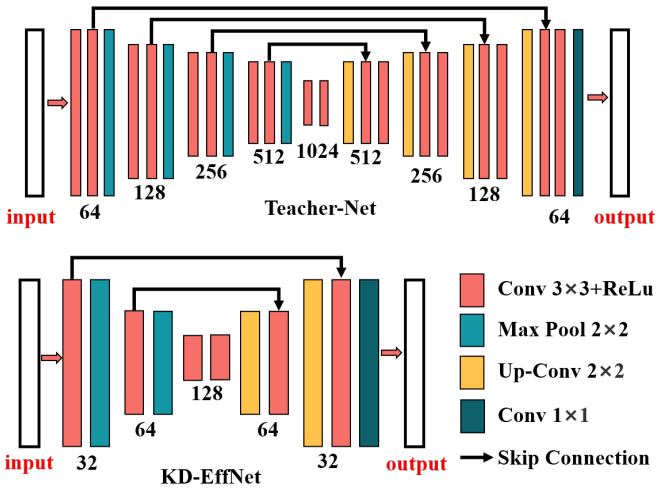

Fig. 3. The detailed parameter settings of fully-convolutional **Teacher-Net** and **KD-EffNet**. The kernel size of convolution and up-conv layers is 3 and 2 respectively, and pooling size is 2. The number of channels shown at the bottom of each layer.

4 Evaluation

In this section, we will evaluate the performance of **KD-Eye** in terms of accuracy and efficiency with a number of pupularly used metrics. **Intersection over Union (IoU)** estimates the proportion of overlap between the segmented area and ground truth over the whole pupil region [27]. **F1 Score** or dice coefficients measures the similarity between the segmented area and ground truth pupil region [28]. **Pixel error** (P_e) of the centers of the segmented area and ground truth is also considered as an evaluation metric for segmentation accuracy.

For evaluation on efficiency, we calculate the sizes of the models (number of coefficients), memory usage, FLOPs and runtime of inference on Snap-

dragon XR2, a resource-constrained platform popularly embedded in a number of cutting-edge VR devices. We implement both **Teacher-Net** and efficient **KD-EffNet** for evaluation to show the effectiveness of proposed components including KD-based training for segmentation accuracy, AoI extraction and coarse-segmentation for segmentation efficiency.

Dataset collection. We recruit 30 subjects to collect near-eye images to evaluate the performance of accuracy and efficiency of our proposed approaches. Each subject participates in the data collection sessions to follow the guidance of a stimulus shown on the screen. Videos of both eyes are recorded during the collection sessions. Then we uniformly pick a subset of 9000 frames to manually label the pupil region for training and evaluation. A public dataset *TEyeD* [24] is employed to train the **Teacher-Net** for knowledge distillation.

4.1 Evaluation on segmentation accuracy

We first evaluate and compare the segmentation accuracy of different approaches to show whether each of the proposed components takes in effect in pupil segmentation as expected. By varying the number of subjects used for evaluation, the corresponding IoUs, F1 scores and P_e of different approaches are calculated and shwon in Table 1-3. Specifically, when m subjects are adopted for evaluation, each of the subject is chosen alternatively as testing and the rest m-1 subjects are used for training to form a leave-one-out cross validation procedure. In the table, the **Coarse-only** is the approach segments the pupil region using the coarse-segmentation component only to show if the following appearance-based approach can improve the segmentation accuracy. The results of **Teacher-Net** and **Eff-Net** are obtained from training the **Teacher-Net** and **EffNet** with images of different number of subjects in our self-collected dataset only without knowledge distillation. Finally, **KD-EfNet** takes all the components proposed in this paper in which the Teacher-Net component is pre-trained with the public dataset *TEyeD* and the knowledge distillation training employs our self-collected images from different number of subjects.

Table 1. IoUs of KD-Eye on pupil segmentation.

No. of Sub	5	10	15	20	25	30
Coarse-only	0.777	0.777	0.777	0.777	0.777	0.777
Teacher-Net	0.905	0.927	0.934	0.935	0.937	0.938
EffNet	0.672	0.851	0.907	0.909	0.921	0.926
KD-EffNet	0.765	0.884	0.912	0.917	0.923	0.927

By comparing the results of different approaches in the tables, we can make a number of observations. First, the increase of the number of subjects for training can improve the segmentation accuracy for the three different appearance-based approaches (i.e., **Teacher-Net**, **EffNet** and **KD-EffNet**). Second, with the

Table 2. F1 scores of KD-Eye on pupil segmentation.

No. of Sub	5	10	15	20	25	30
Coarse-only	0.860	0.860	0.860	0.860	0.860	0.860
Teacher-Net	0.949	0.960	0.965	0.965	0.967	0.967
EffNet	0.781	0.916	0.948	0.952	0.958	0.960
KD-EffNet	0.879	0.940	0.955	0.956	0.960	0.962

Table 3. Pixel errors (P_e) of KD-Eye on pupil segmentation.

No. of Sub	5	10	15	20	25	30
Coarse-only	14.46	14.46	14.46	14.46	14.46	14.46
Teacher-Net	2.47	0.87	0.76	0.73	0.73	0.73
EffNet	24.26	7.41	2.89	1.29	0.94	0.84
KD-EffNet	12.33	3.87	1.02	1.03	0.79	0.77

initial coarse segmentation (**coarse-only**) as the reference, except for the extremely small trainingset (No. of Sub = 5), the appearance-based approach can significantly improve the segmentation accuracy. Third, by comparing different appearance-based approaches, **KD-EffNet** with knowledge distillation training strategy can achieves the second best segmentation accuracy and is more accurate than its efficient counterpart, **EffNet**. However, the most accurate model, **Teacher-Net**, is significantly more complex than **KD-EffNet**, and it consumes significantly more resources as our following evaluation.

4.2 Evaluation on Resources Consumption

Then we evaluate the resources consumption of the pupil segmentation approaches with different settings. The resources considered in this part include number of parameters of models (**Size**), floating point operations (**FLOPs**), the memory usage when running the segmentation algorithm (**Memory**) and average computation time (**Runtime**) for segmenting pupil area of each frame. To show the effectiveness of the AoI extraction and coarse-segmentation on reducing the resources consumption, **Eff-Net(pure)** is also implemented without the AoI extraction and coarse-segmentation components.

Table 4. Resources consumption of pupil segmentation.

Models	**Size**	**FLOPs**	**Memory**	**Runtime**
Teacher-Net	31M	62.6B	553MB	1475.8ms
KD-EffNet (pure)	0.9M	16.7B	256MB	86.8ms
KD-EffNet	0.9M	0.678B	10.4MB	6.1ms

The three approaches, **Teacher-Net**, **KD-EffNet** and **KD-EffNet (pure)** are implemented on a resource-constrained SnapDragon XR2 platform. XR2 is a popularly adopted platform for a number of different off-the-shelf VR devices. It features an octa-core Kryo processor and a Adreno 650 GPU. During the evaluation, the running memory and runtime is obtained by conducting 1000 independent trials to obtain the average of the results. By comparing different approaches shown in Table 4, we can find the model of **KD-EffNet** is significantly simpler than the deep Teacher-Net: it only contains 0.9M coefficients which is approx. only 3% of **Teacher-Net**. The reduction on size of model saves approx. 73% FLOPs, 54% running memory and 94% runtime by comparing the results of **Teacher-Net** and **EffNet(pure)**. Then by introducing the preprocessing of AoI extraction and coarse-segmentation to reduce the size of input for the, though the model remains the same, the reduction on resources consumption is tremendous. By comparing results of **KD-EffNet** and **Teacher-Net**, we can observe **KD-EffNet** saves approx. 99% FLOPs, 98% running memory and 99.6% runtime. The results suggest that our efficient approach can process over 160 frames per second in realtime while the complex deep network (**Teacher-Net**) requires approx. 1.5s to handle just one frame. It indicates over 240 times speed-up with our optimization on efficiency.

5 Conclusion

In this paper, we propose **KD-Eye**, a lightweight approach for pupil segmentation. It employes the power of knowledge distillation to improve the segmentation accuracy of **KD-EffNet**, a simple neural network with only few convolution layers. Then a coarse-to-fine strategy is proposed to speed up the processing of segmentation by significantly reducing the size of input while preserving high segmentation accuracy. According to our implementation and evaluation, **KD-EffNet** can achieve processing rate over 160fps which is over 240 times faster than competing approach and consumes only 1-2% computation and memory resources of the baseline.

Acknowledgments. This work is supported by Shandong Provincial Natural Science Foundation, China, Grant No. 2022HWYQ-040.

Data Availability Statement The datasets used in this manuscript are all publicly available. The corresponding repositories are all properly cited in the manuscript.

Disclosure of Interests. The authors declare that they have no conflict of interest.

References

1. Clay, V., König, P., Koenig, S.: Eye tracking in virtual reality. Journal of Eye Movement Research **12** (04 2019). https://doi.org/10.16910/jemr.12.1.3

2. Lohr, D., Griffith, H., Komogortsev, O.: Eye know you: Metric learning for end-to-end biometric authentication using eye movements from a longitudinal dataset (04 2021)
3. Oliveira, D., Machín, L., Deliza, R., Rosenthal, A., Walter, E.H., Giménez, A., Ares, G.: Consumers' attention to functional food labels: Insights from eye-tracking and change detection in a case study with probiotic milk. LWT - Food Science and Technology **68**, 160–167 (2016)
4. Gwizdka, J., Hosseini, R., Cole, M., Wang, S.: Temporal dynamics of eye-tracking and eeg during reading and relevance decisions. Journal of the Association for Information Science and Technology **68**(10), 2299–2312
5. Zaretskaya, N., Bause, J., Polimeni, J.R., Grassi, P.R., Scheffler, K., Bartels, A.: Eye-selective fmri activity in human primary visual cortex: Comparison between 3 t and 9.4 t, and effects across cortical depth. NeuroImage **220**, 117078 (2020), https://www.sciencedirect.com/science/article/pii/S1053811920305644
6. Valliappan, N., Dai, N., Steinberg, E., He, J., Rogers, K., Ramachandran, V., Xu, P., Shojaeizadeh, M., Guo, L., Kohlhoff, K., et al.: Accelerating eye movement research via accurate and affordable smartphone eye tracking. Nature Communications **11**(1), 1–12 (2020)
7. Cheng, Y., Wang, H., Bao, Y., Lu, F.: Appearance-based gaze estimation with deep learning: A review and benchmark. arXiv preprint arXiv:2104.12668 (2021)
8. Morimoto, C.H., Mimica, M.R.M.: Eye gaze tracking techniques for interactive applications. Comput. Vis. Image Underst. **98**(1), 4–24 (apr 2005)
9. Wang, X., Zhang, J., Zhang, H., Zhao, S., Liu, H.: Vision-based gaze estimation: a review. IEEE Transactions on Cognitive and Developmental Systems (2021)
10. Kim, J., Stengel, M., Majercik, A., De Mello, S., Dunn, D., Laine, S., McGuire, M., Luebke, D.: Nvgaze: An anatomically-informed dataset for low-latency, near-eye gaze estimation. In Proceedings of the 2019 CHI Conference, Association for Computing Machinery, New York, NY, USA (2019), https://doi.org/10.1145/3290605.3300780
11. Cherif, Z., Nait-Ali, A., Motsch, J., Krebs, M.: An adaptive calibration of an infrared light device used for gaze tracking. In: Proceedings of the 19th IEEE Instrumentation and Measurement Technology Conference. vol. 2, pp. 1029–1033 vol.2 (2002). https://doi.org/10.1109/IMTC.2002.1007096
12. Wang, K., Ji, Q.: Real time eye gaze tracking with 3d deformable eye-face model. In: Proceedings of IEEE International Conference on Computer Vision (ICCV). pp. 1003–1011 (2017). https://doi.org/10.1109/ICCV.2017.114
13. Guestrin, E., Eizenman, M.: General theory of remote gaze estimation using the pupil center and corneal reflections. IEEE Transactions on Biomedical Engineering **53**(6), 1124–1133 (2006). https://doi.org/10.1109/TBME.2005.863952
14. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4511–4520 (2015)
15. Kwok, T.C., Kiefer, P., Schinazi, V.R., Adams, B., Raubal, M.: Gaze-guided narratives: Adapting audio guide content to gaze in virtual and real environments. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 1–12 (2019)
16. Mariakakis, A., Goel, M., Aumi, M.T.I., Patel, S.N., Wobbrock, J.O.: SwitchBack: Using focus and saccade tracking to guide users' attention for mobile task resumption. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. pp. 2953–2962 (2015)

17. Lan, G., Heit, B., Scargill, T., Gorlatova, M.: GazeGraph: Graph-based few-shot cognitive context sensing from human visual behavior. In: Proceedings of the ACM Conference on Embedded Networked Sensor Systems (SenSys). pp. 422–435 (2020)
18. Srivastava, N., Newn, J., Velloso, E.: Combining low and mid-level gaze features for desktop activity recognition. In Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **2**(4), 1–27 (2018)
19. Kim, J., Jeong, Y., Stengel, M., Akşit, K., Albert, R., Boudaoud, B., Greer, T., Kim, J., Lopes, W., Majercik, Z., et al.: Foveated AR: Dynamically-foveated augmented reality display. ACM Transactions on Graphics **38**(4), 1–15 (2019)
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241 (2015)
21. Barea, R., Boquete, L., Mazo, M., Lopez, E.: System for assisted mobility using eye movements based on electrooculography. IEEE Transactions on Neural Systems and Rehabilitation Engineering **10**(4), 209–218 (2002). https://doi.org/10.1109/TNSRE.2002.806829
22. Robinson, D.A.: A method of measuring eye movemnent using a scieral search coil in a magnetic field. IEEE Transactions on Bio-medical Electronics **10**(4), 137–145 (1963). https://doi.org/10.1109/TBMEL.1963.4322822
23. Zhao, G., Yang, Y., Liu, J., Chen, N., Shen, Y., Wen, H., Lan, G.: Ev-eye: Rethinking high-frequency eye tracking through the lenses of event cameras. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 62169–62182. Curran Associates, Inc. (2023)
24. Fuhl, W., Kasneci, G., Kasneci, E.: Teyed: Over 20 million real-world eye images with pupil, eyelid, and iris 2d and 3d segmentations, 2d and 3d landmarks, 3d eyeball, gaze vector, and eye movement types. In: 2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). pp. 367–375. IEEE (2021)
25. Zhang, X., Park, S., Beeler, T., Bradley, D., Tang, S., Hilliges, O.: Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 365–381. Springer International Publishing, Cham (2020)
26. Wang, K., Su, H., Ji, Q.: Neuro-inspired eye tracking with eye movement dynamics. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9823–9832 (2019)
27. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 658–666 (2019)
28. Chaudhary, A.K., Kothari, R., Acharya, M., Dangi, S., Nair, N., Bailey, R., Kanan, C., Diaz, G., Pelz, J.B.: RITnet: Real-time semantic segmentation of the eye for gaze tracking. In: Proceedings of IEEE/CVF International Conference on Computer Vision Workshop (ICCVW) (2019)