Gaze Estimation with Eye Region Segmentation and Self-Supervised Multistream Learning

Zunayed Mahmud[1,2]**, Paul Hungler**[2]**, Ali Etemad**[1,2]

[1]Department of Electrical and Computer Engineering
[2]Ingenuity Labs Research Institute
Queen's University, Kingston, Canada
{zunayed.mahmud, paul.hungler, ali.etemad}@queensu.ca

Abstract

We present a novel multistream network that learns robust eye representations for gaze estimation. We first create a synthetic dataset containing eye region masks detailing the visible eyeball and iris using a simulator. We then perform eye region segmentation with a U-Net type model which we later use to generate eye region masks for real-world eye images. Next, we pretrain an eye image encoder in the real domain with self-supervised contrastive learning to learn generalized eye representations. Finally, this pretrained eye encoder, along with two additional encoders for visible eyeball region and iris, are used in parallel in our multistream framework to extract salient features for gaze estimation from real world images. We demonstrate the performance of our method on the EYE-DIAP dataset in two different evaluation settings and achieve state-of-the-art results, outperforming all the existing benchmarks on this dataset. We also conduct additional experiments to validate the robustness of our self-supervised network with respect to different amounts of labeled data used for training.

1 Introduction

Eye gaze acts as a non-verbal behavioral sign that can indicate human intention, attention, and interests. In addition to helping us perceive the environment, gaze has various applications in human behaviour analysis (Ishii et al. 2016), human computer interaction (HCI) (Andrist et al. 2014), human robot interaction (HRI) (Moon et al. 2014), cognitive science (Amadori et al. 2021), and virtual reality (Konrad, Angelopoulos, and Wetzstein 2020). Thus, accurate estimation of gaze has gained attention over the years, making gaze estimation a well-established area of research.

Many prior works in gaze estimation relied on lightweight machine learning algorithms (Lu et al. 2011; Wood et al. 2016b). They were mostly biased with constrained environments and fixed head-pose. Recent methods (Zhang et al. 2015, 2017; Park et al. 2018) have leveraged deep neural networks to perform gaze estimation and shown performance improvement. Despite the progress, there still exists a number of underlying open problems:

- **Eye Region Registration:** The geometric shape and orientation of *particular regions* of the eye can provide strong information cues for gaze estimation. Thus, accurate segmentation of eye regions such as the iris could

enhance gaze estimation. Most existing gaze datasets do not provide ground truths that could be readily used for eye region segmentation.

- **Labeled Data:** As is the case with any machine learning problem, obtaining labeled data for gaze estimation can be difficult, time-consuming, and expensive. This is particularly challenging for gaze estimation as eye images are mainly cropped from larger facial images, making them small and with low resolutions, which in turn makes labeling more difficult and prone to errors.

In this work, we tackle these challenges by proposing a model for self-supervised gaze estimation. Our proposed solution first performs eye region segmentation using a U-Net type network (Ronneberger, Fischer, and Brox 2015) exclusively trained on a synthetic dataset created using UnityEyes (Wood et al. 2016b). This eye segmenter learns to create eye masks detailing the visible eyeball and the iris based on the synthetic dataset, which we later use to generate eye masks for the *real* dataset. This domain transfer is a necessary step in our pipeline as large gaze datasets with ground-truth eyeball and iris segment labels do not exist. Hence the synthetic dataset was created with these labels. We then propose a multistream gaze estimator network which we first pretrain using self-supervised contrastive learning and then fine-tune for gaze estimation. The multistream gaze estimator network takes a single eye image and corresponding eye masks (created by the eye segmenter) as input and regresses the 3D gaze vectors.

In a nutshell, we make the following contributions: (**1**) We are the first to propose the usage of eye region masks as a feature input for gaze estimation. To effectively train a model for eye region segmentation on real images, we generate and use a synthetic dataset, which we then use for segmentation in the *real* domain. (**2**) We introduce a novel multistream deep neural network that utilizes both appearance and geometric factors by extracting features from raw eye image and eye masks independently to perform gaze estimation. We pretrain this model in a self-supervised manner for better generalization and reduced reliance on labels. Our model outperforms the existing supervised methods on a benchmark dataset.

2 Related Work

Gaze Estimation. Gaze estimation can be categorized into three different approaches: feature-based, model-based, and appearance-based. Feature-based methods (Huang et al. 2014; Xiong et al. 2014) generally rely on hand-crafted features such as pupil centers, eye corners, and iris edges extracted from eye images. Model-based approaches (Wood et al. 2016a; Wood and Bulling 2014; Park et al. 2018) on the other hand, aim to fit 3D eye models to eye images. Finally, appearance-based gaze estimation methods (Baluja and Pomerleau 1994; Tan, Kriegman, and Ahuja 2002) take raw eye images and learn to map gaze as a 2D point or a 3D angular vector.

Recent works on gaze estimation (Zhang et al. 2015, 2017; Park, Spurr, and Hilliges 2018) rely on Convolutional Neural Networks (CNNs) and have shown significant performance improvements. A multitask learning network was developed in (Yu, Liu, and Odobez 2018) where eye landmark detection was performed as an auxiliary task. In (Krafka et al. 2016), a multistream strategy was proposed where separate convolutional feature extractors were used for both eye and face images. The approach improved on the results of (Zhang et al. 2015) by a large margin. In (Wood et al. 2016b) a real-time synthetic eye image simulator, UnityEyes, was developed to generate millions of synthetic images to create scope for domain transfer learning in gaze estimation. This work shows competitive performance in cross-dataset experiment. In (Shrivastava et al. 2017) the domain gap between synthetic and real eye images was addressed and reduced by domain adversarial learning. In a subsequent work by (Lee, Kim, and Suh 2018), the domain adaptation results were further improved by leveraging bidirectional mapping between synthetic and real domains.

Contrastive Learning. Contrastive learning is primarily a self-supervised learning (SSL) method which aids deep neural networks to learn a better latent representation by utilizing augmented unlabeled data. A recent contrastive learning method, SimCLR (Chen et al. 2020), demonstrates that by applying contrastive loss and maximizing the agreement between two augmented versions of the same image, it can facilitate neural networks to learn more useful visual representations. In (Grill et al. 2020) two neural networks were used which are referred to as online and target networks, where the online network predicts the output representation of the target network given two augmented versions of the same image as input. Recent applications of contrastive learning include medical image segmentation (Chaitanya et al. 2020; Zhao et al. 2021), facial expression detection (Roy and Etemad 2021), pose estimation (Spurr et al. 2021), and others. Nonetheless, despite the high likelihood that contrastive learning could aid in better generalization of gaze estimators in dealing with challenging scenarios and low quality inputs, it has not yet been explored in this area.

3 Method

Model Structure and Overview. We aim to develop a gaze estimator that can focus on salient regions of the eye, which we hypothesize contain important gaze-related information.

Table 1: Description of transformations.

Transformations	Parameters
Gaussian noise	$\sigma = (0 - 10.0), \mu = 0$
Gaussian blur	$\sigma = (0 - 2.0)$, filter size $= (3 \times 3)$
Cutout	$h, w = (0 - 10.0)$ px
Downscale	$(1x - 2x)$
Random lines	$(0 - 2)$
Contrast change	-

While one approach would be to rely on eye landmarks (Park et al. 2018), landmark detection is time-consuming and sensitive to noise. Hence, to focus the learning of the gaze estimator network without the need to detect individual and fine-grained landmarks, we rely on the iris and visible eyeball masks, in addition to the whole image of the eye. Accordingly we design a multistream pipeline to learn representations from each of these three inputs independently, which can then be combined for gaze estimation. First, an eye segmentation module is trained to detect the visible eyeball and the iris masks. As gaze datasets with ground-truth labels for eyeball and iris regions are very rare, we train the eye segmenter on an auxiliary synthetic dataset, which we discuss in detail later in Section 4. The output masks are then used to train two encoders for learning representations from the iris and visible eyeball segments as two of the three streams of our multistream pipeline. Next, we pretrain an encoder in a contrastive self-supervised manner to learn generalized representations from real input images of the whole eye. Given the binary nature of the iris and visible eyeball masks, we do not use self-supervised learning for pretraining the two mask encoders as our empirical experiments demonstrate no performance boost with such pretraining. This encoder, along with the other two encoders are used in parallel in our multistream setup to extract relevant features. The output representations from these three encoders are then fused in the feature dimension and utilized for final gaze estimation. Below we describe the details of each of the modules of our pipeline.

Eye Region Segmentation. As mentioned above, we aim to adopt iris and visible eyeball masks in our multistream pipeline to provide useful gaze-related information in addition to the standard way of learning gaze from the entire eye image. To this end, a network is required to segment eye images into the two desired masks. We thus adopt a U-Net type model (Ronneberger, Fischer, and Brox 2015) with two output channels, one for the iris mask and the other for the visible eyeball mask. The architecture details of this model are presented in Fig. 1(A). To train this network, we use UnityEyes simulator (Wood et al. 2016b) and generate a synthetic image dataset given the scarcity of detailed and accurate datasets with labeled iris and visible eyeball regions. The network is trained from scratch using MSE loss.

Self-Supervised Learning of Eye Images. To utilize a model for effective learning of real eye images, we rely on self-supervised pretraining. This design choice is made to enable the encoder to be able to learn strong representations in the presence of different variations and challenging con-
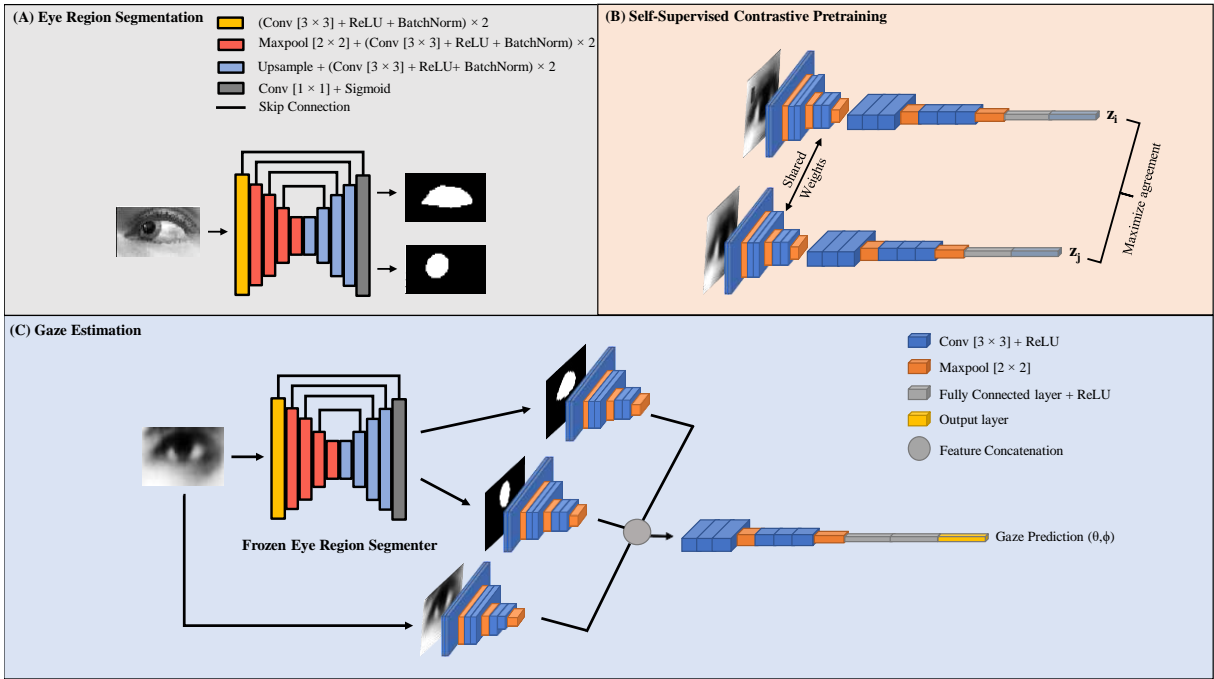
Figure 1: The proposed framework for our method. First, we train an eye region segmenter using a synthetic database which we create using a simulator (Wood et al. 2016b). Next, we pretrain the eye image encoder by self-supervised contrastive learning. Finally, we fine-tune the multistream network for gaze estimation.

ditions (low resolution, noise, occlusions, etc.) often available in eye images which are generally cropped from facial images. To this end, we first define a set of transformations as described in Table 1 which are randomly applied to augment the images. Following (Chen et al. 2020), we only draw positive pairs within a minibatch N and the other $2(N-1)$ samples are considered to be negatives. To preserve the gaze direction, we avoid applying any geometric transformation which might potentially harm the eye shape and orientation. Next, as shown in Fig. 1(B), a contrastive approach is used to learn useful representations using the positive and negative augmented samples. The contrastive loss (Chen et al. 2020) is used for pretraining the eye encoder, which is defined as:

$$L_{i,j} = -log \frac{exp(cosine(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} exp(cosine(z_i, z_k)/\tau)} \quad (1)$$

where i and j are the positive pairs, $\mathbb{1}_{[k \neq i]} \in [0, 1]$ is a function which results in 1 only if $i \neq k$, and the temperature τ is a scaling parameter for the cosine similarity, set to 0.1.

Encoder, Fusion, and Gaze Estimation Details. The details of the encoders used for each of the streams are presented in Fig. 1(C). It should be emphasized that the architecture details of all three encoders are identical with their training being the only difference. Subsequent to these three encoders, the learned representations are concatenated in the feature dimension and subsequently go through a final encoding process prior to gaze estimation. The 3 fully connected (FC) layers used for gaze estimation consist of 256, 128, and 2 units respectively. Output gaze is estimated in pitch (θ) and yaw (ϕ) angles based on common practice in

the field (Zhang et al. 2015; Park, Spurr, and Hilliges 2018). The entire pipeline, including the encoders and gaze estimator are trained using MSE loss.

4 Experiments and Results

Dataset and Preprocessing Steps. As mentioned earlier in Section 3, to perform eye region segmentation, we generate a synthetic dataset using the eye simulator, UnityEyes (Wood et al. 2016b). The simulator provides synthetic eye images and their corresponding landmarks, which we use to generate the output masks. We generate a total of 60,000 synthetic images. To bring the domain of these synthetic data closer to that of real images, we apply the same augmentations presented earlier in Table 1. It should be re-emphasized, however, that this is merely carried to as a preprocessing step, and contrastive learning is *not* used for training the U-Net with the synthetic data.

For training the eye image encoder as well as gaze estimation, we use the EYEDIAP dataset (Funes Mora, Monay, and Odobez 2014). There are multiple sessions in the dataset with different illumination conditions, continuous gaze targets, and head pose. The low resolution images and extreme occlusions make this dataset highly challenging for gaze estimation task. We evaluate our proposed solution on the screen target gaze session (CS/DS) and static and mobile (SP/MP) head-pose scenario. The input eye images are taken from the VGA video provided for this session. The raw images are first normalized to have frontal head-pose and then converted to a fixed size of 36×60. Later, these images are converted to gray-scale and histogram equalized. We fol-

Table 2: Performance comparison on EYEDIAP dataset.

Methods	LOSO	5-fold
(Zhang et al. 2015)	7.60	-
(Zhang et al. 2017)	6.3	-
(Yu, Liu, and Odobez 2018)	6.50	-
(Park, Spurr, and Hilliges 2018)	-	10.3
(Wang et al. 2019)	-	9.90
Ours (Multistream Network)	6.29	6.48
Ours (SSL + Multistream Network)	**6.15**	**6.34**

Table 3: Performance of our method using different amounts of output labels.

Labeled Data(%)	Angle Error$^\circ$
100%	6.34
75%	6.86
50%	7.01
25%	7.45

low the same person-independent or leave-one-subject-out (LOSO) protocol as in (Funes-Mora and Odobez 2016), as well as 5-fold cross-validation following (Park, Spurr, and Hilliges 2018).

Implementation. We train the eye segmenter U-Net, the 3 encoders, and the gaze estimator network using Adam optimizer with an initial learning rate of 0.00001. For segmentation, we train the network for 50 epochs using a step learning rate decay with a step size of 5 and a factor of 0.1. For the gaze estimator network, first we pretrain the eye encoder with contrastive learning as discussed earlier, for 50 epochs. Here, we again use Adam optimizer and a learning rate of 0.0001 with cosine learning rate decay. We then train the multistream gaze estimator for 25 epochs where we keep the eye image encoder frozen for 5 epochs and then fine-tune the whole network for 20 epochs. At this step, we use a plateau learning rate decay with decay factor of 0.1 and patience of 3. We train our multistream network with the same hyperparameters to create a baseline without the self-supervision. We use a batch size of 32 for the segmentation task and 128 for the contrastive learning and gaze estimation task. Our networks are implemented using PyTorch with a single Nvidia 2080 Ti GPU.

Results. We perform both LOSO and 5-fold evaluation on the EYEDIAP dataset to evaluate the gaze estimation error achieved by our proposed self-supervised multistream network. We compare the performance of our proposed framework against the existing state-of-the-art solutions on this dataset. The results as shown in Table 2 are reported in mean angular errors across subjects and folds. It is observed from the table, that our method outperforms the existing benchmarks by a large margin in both LOSO and 5-fold evaluations. Specifically, our supervised multistream baseline improves the highest benchmark result by a mean angle error of $3.42°$ in the 5-fold evaluation while achieving similar performance as (Zhang et al. 2017) with LOSO. With contrastive learning, the proposed supervised baseline is further improved by $0.14°$ in both LOSO and 5-fold experiments.

Impact of Output Labels. To obtain a clear understanding of the effectiveness of the contrastive learning strategy on our multistream network beyond improvement of the results, we also conduct an experiment to examine the sensitivity of our network by varying the amounts of output labels. We choose a range of labeled data to fine-tune our network and compare the performance of gaze estimation. The analysis as reported in Table 3 reflects the performance comparison using different amounts of output labels used in the down-

stream fine-tuning. It can be observed that when we use only 75% and 50% of the labeled data, the gaze estimation error is still quite low. As we keep decreasing the percentage of labeled data to only 25%, our network can still maintain a very stable performance and outperform existing state-of-the-art methods on the 5-fold evaluation. This validates the robustness of our multistream network through the self-supervised step.

5 Conclusion and Future Work

This paper presents a novel multistream neural network for gaze estimation. Our model relies on a mask of the visible eyeball, a mask of the iris, and the entire image of the eye. To obtain the eye masks, we generate a synthetic dataset and use it to train a U-Net model for eye region segmentation. This segmenter is then frozen and used to generate the masks from real eye images. The masks, as well as the original eye image, are then passed to three separate encoders. The obtained embeddings are then fused to be used for gaze estimation. To allow the eye image encoder to learn more generalized representations, we use contrastive learning for pretraining. We validate our method on the EYEDIAP dataset and demonstrate that our method outperforms other works in the area with both 5-fold and LOSO schemes, while performing consistently and robustly even when small subsets of output labels are used for downstream gaze estimation. For future work, we plan to improve the eye region segmentation and perform more in-depth analysis of the segmentation performance. We would also like to further explore the effects of self-supervision on our gaze estimation model and experiment with additional augmentations.

References

Amadori, P. V.; Fischer, T.; Wang, R.; and Demiris, Y. 2021. Predicting Secondary Task Performance: A Directly Actionable Metric for Cognitive Overload Detection. *IEEE Transactions on Cognitive and Developmental Systems*, 1–1.

Andrist, S.; Tan, X. Z.; Gleicher, M.; and Mutlu, B. 2014. Conversational gaze aversion for humanlike robots. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction*, 25–32.

Baluja, S.; and Pomerleau, D. 1994. Non-intrusive gaze tracking using artificial neural networks. Technical report, Pittsburgh, PA, USA.

Chaitanya, K.; Erdil, E.; Karani, N.; and Konukoglu, E. 2020. Contrastive learning of global and local features for medical image segmentation with limited annotations. *arXiv preprint arXiv:2006.10511*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607.

Funes Mora, K. A.; Monay, F.; and Odobez, J.-M. 2014. 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*, 255–258.

Funes-Mora, K. A.; and Odobez, J.-M. 2016. Gaze estimation in the 3d space using rgb-d sensors. *International Journal of Computer Vision*, 118(2): 194–216.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P. H.; Buchatskaya, E.; Doersch, C.; Pires, B. A.; Guo, Z. D.; Azar, M. G.; et al. 2020. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*.

Huang, M. X.; Kwok, T. C.; Ngai, G.; Leong, H. V.; and Chan, S. C. 2014. Building a self-learning eye gaze model from user interaction data. In *Proceedings of the 22nd ACM International Conference on Multimedia*, 1017–1020.

Ishii, R.; Otsuka, K.; Kumano, S.; and Yamato, J. 2016. Prediction of who will be the next speaker and when using gaze behavior in multiparty meetings. *ACM Transactions on Interactive Intelligent Systems*, 6(1): 1–31.

Konrad, R.; Angelopoulos, A.; and Wetzstein, G. 2020. Gaze-contingent ocular parallax rendering for virtual reality. *ACM Transactions on Graphics*, 39(2): 1–12.

Krafka, K.; Khosla, A.; Kellnhofer, P.; Kannan, H.; Bhandarkar, S.; Matusik, W.; and Torralba, A. 2016. Eye tracking for everyone. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2176–2184.

Lee, K.; Kim, H.; and Suh, C. 2018. Simulated+ unsupervised learning with adaptive data generation and bidirectional mappings. In *International Conference on Learning Representations*.

Lu, F.; Okabe, T.; Sugano, Y.; and Sato, Y. 2011. A head pose-free approach for appearance-based gaze estimation. In *British Machine Vision Conference*, 1–11.

Moon, A.; Troniak, D. M.; Gleeson, B.; Pan, M. K.; Zheng, M.; Blumer, B. A.; MacLean, K.; and Croft, E. A. 2014. Meet me where i'm gazing: how shared attention gaze affects human-robot handover timing. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, 334–341.

Park, S.; Spurr, A.; and Hilliges, O. 2018. Deep pictorial gaze estimation. In *Proceedings of the European Conference on Computer Vision*, 721–738.

Park, S.; Zhang, X.; Bulling, A.; and Hilliges, O. 2018. 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*, 1–10.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241.

Roy, S.; and Etemad, A. 2021. Self-supervised Contrastive Learning of Multi-view Facial Expressions. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, 253–257.

Shrivastava, A.; Pfister, T.; Tuzel, O.; Susskind, J.; Wang, W.; and Webb, R. 2017. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2107–2116.

Spurr, A.; Dahiya, A.; Wang, X.; Zhang, X.; and Hilliges, O. 2021. Self-Supervised 3D Hand Pose Estimation from monocular RGB via Contrastive Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11230–11239.

Tan, K.-H.; Kriegman, D. J.; and Ahuja, N. 2002. Appearance-based eye gaze estimation. In *Sixth IEEE Workshop on Applications of Computer Vision, 2002.*, 191–195.

Wang, K.; Zhao, R.; Su, H.; and Ji, Q. 2019. Generalizing eye tracking with bayesian adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11907–11916.

Wood, E.; Baltrušaitis, T.; Morency, L.-P.; Robinson, P.; and Bulling, A. 2016a. A 3d morphable eye region model for gaze estimation. In *European Conference on Computer Vision*, 297–313.

Wood, E.; Baltrušaitis, T.; Morency, L.-P.; Robinson, P.; and Bulling, A. 2016b. Learning an appearance-based gaze estimator from one million synthesised images. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, 131–138.

Wood, E.; and Bulling, A. 2014. Eyetab: Model-based gaze estimation on unmodified tablet computers. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, 207–210.

Xiong, X.; Liu, Z.; Cai, Q.; and Zhang, Z. 2014. Eye gaze tracking using an RGBD camera: a comparison with a RGB solution. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, 1113–1121.

Yu, Y.; Liu, G.; and Odobez, J.-M. 2018. Deep multitask gaze estimation with a constrained landmark-gaze model. In *Proceedings of the European Conference on Computer Vision Workshops*.

Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2015. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4511–4520.

Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2017. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1): 162–175.

Zhao, X.; Vemulapalli, R.; Mansfield, P. A.; Gong, B.; Green, B.; Shapira, L.; and Wu, Y. 2021. Contrastive Learning for Label Efficient Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10623–10633.