HIGH SCHOOL EDITION
Journal of Student Research
Volume 14 Issue 1 (2025)

Unsupervised Gaze Representation Learning with Conjugate Gaze Consistency Loss for Enhanced Gaze Estimation

Junho Kee[1] and Giyoung Yang[#]

[1]Bergen County Academies, USA
[#]Advisor

ABSTRACT

Patients suffering from quadriplegia and other paralysis that interfere with the ability to communicate have been increasing for the past decade, and in turn, the need for better communication boards have been growing. Current communication boards, both physical and digital communication boards, all have flaws starting from the need of an assistant to the sole expense of the device. However, gaze estimation techniques have also been gaining attention to enhance the quality of communication boards by tracking the movement of the eye via a camera to assess what the user is trying to communicate. Previous studies on gaze estimation algorithms have shown that collecting data for accurate gaze values is an arduous task, and that the accuracy of the gaze estimation models has been unsatisfactory for practical use. Thus, I propose a gaze estimation-based digital communication board system that combines gaze representation learning with transfer learning. In the representation learning phase, I introduce a random sign-reversal module to efficiently isolate gaze-related features. In the transfer learning phase, I implement a medically driven loss function to enhance the system's accuracy. The proposed system achieved an angular error of 9.42 degrees which represents state-of-the-art performance compared to previous studies.

Introduction

Paralysis is the limitation in voluntary muscle movement caused by problems in the nervous system; about 1 in 50 Americans have some sort of paralysis, including about one third of stroke patients experiencing permanent disability. The range of which part of the body paralysis effects may vary, but quadriplegia, the paralysis of the body from neck down, has a detrimental effect on the patient's life, limiting the ability to move or even communicate. The patients of quadriplegia have been increasing exponentially, leaving many people to live with inconveniences in life that are often untreatable.

The traditional methods to alleviate this problem was to use physical or digital communication boards for the patients, yet these are difficult to create desirable effects (Patak et al. 2006). Primarily, for patients who cannot control their slightest movement, there must be an assistant to use the communication board. Thus, the patients can only express immediate needs for about 8 to 12 words a minute. Not only is this inconvenient and ineffective, but having an assistant to aid a patient is impractical. An alternative to this is a digital communication board, which often uses gaze estimation or slight muscle movement and predictive text software that allows patients to communicate (Kar and Corcoran 2017). However, these assisting technologies are very costly and highly specific to individuals, making it a generally unapproachable method for many patients.

Out of these methods, gaze estimation technology has been a field of interest as eyes are able to be controlled freely by patients with quadriplegia since they are controlled by the oculomotor nerves which are directly connected to the brain, not the spinal cord. However, accuracy has always been a crucial issue to all these real-time detecting techniques, as the enhancement of the accuracy of the model may hinder the algorithm to run expeditiously. Another

problem specific to gaze-estimation technique is the variation of different features on the image-based technique: skin color, eye shape, and other external features pose difficulties in solely detecting the gaze of the eye unless made specific for one individual. To tackle this problem without decelerating the model, an unsupervised gaze representation learning is proposed, encoding the image and inputting it into a random sign reverse module to create a new image with a random direction of the eye. Then with the pre-trained encoder, a transfer learning is conducted to separate the yaw and the pitch (coordinate systems of gaze estimation) from the appearances and a conjugate gaze consistency loss is applied for each left and right eye.

Background Knowledge

Gaze Estimation

Gaze estimation is a regression task which outputs the continuous movement of the eye. It takes in an image of an eye and outputs the gaze vector consisting of yaw and pitch. Yaw is the equivalent of horizontal movement of the eye and pitch is equivalent to the vertical movement of the eye, which is converted into coordinates when made into a singular vector. The basic construction of the gaze estimation algorithm is to input an image of an eye into a convolutional neural network which outputs a map of gaze related features. With these vectors, the algorithms can combine the loss function with the euclidean distance and the angular distance between the correct gaze vector and the output gaze vector to calculate the total loss. However, crucial limitations exist for this method; primarily, for an accurate gaze estimation, the model must solely focus on the movement of the eye, but eye images have excess variables such as skin color and eye shape which may impair the efficiency of the model. Secondly, neural networks are affected critically by image noises, so supervised gaze estimation algorithms require very clear training samples which are very difficult to get unless in a controlled setting (Kim and Jeong 2020). To tackle both of these problems, using an autoencoder to separate the appearance features and the gaze related features and rotating them to calculate the difference between the gaze features can effectively train the gaze estimation algorithm unsupervised.

Representation Learning

Representation learning in the process of machine learning is the step that converts higher dimensional data, such as images, to a lower dimensional feature map with wanted features extracted from the original data (Zhang et al. 2018). By utilizing an effective representation learning technique, the algorithm is able to perform much more accurately and efficiently. Gaze estimation is also a machine learning technique using images as inputs, so multiple representation learning methods have been developed. In 2019, a method called FAZE used rotation matrices with supervised data in order to rotate the gaze features to a desirable location and added it back to the face (Park et al. 2019). Similarly, the following research utilized rotation matrices with unsupervised data, putting two images of the eyes each looking at different directions into the convolutional neural network to take the difference between the two gaze features and using an autoencoder to rotate an eye looking from one direction to another. Another effective representation learning technique was the cross-encoding technique which gathered two images of eyes looking at the same direction, isolated the gaze features to the appearance features, and swapped the gaze features so they can still appear to be viewing the same direction (Yu et al. 2020). Other studies involving cross-encoders not only swapped the gaze features but also classified data into more specific features such as the location of the camera and the left or right eye, increasing the accuracy of the model (Sun et al. 2021). This paper proposes a novel way to incorporate these methods with the introduction of a medical term, conjugate gaze ability, which is the ability of the brain to control the left and the right eye in the same vertical and horizontal direction. Therefore, on top of using rotation matrices for unsupervised gaze data, incorporating the ideas of conjugate gaze ability allows the left and the right eyes to be always available for

cross-encoding even without supervised data. The specific proposed gaze estimation algorithm will be introduced in chapter 3.

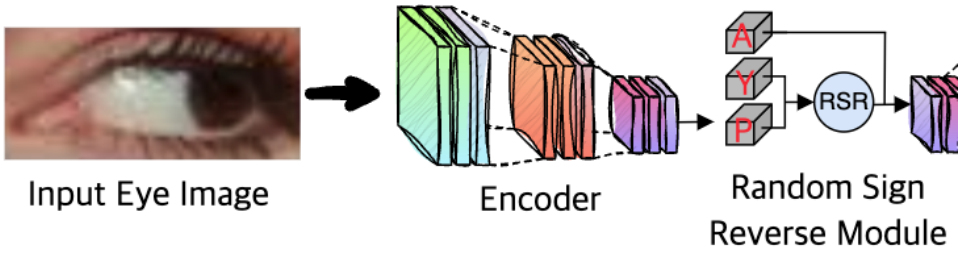
Proposed Gaze Estimation Network

Unsupervised Gaze Representation Learning

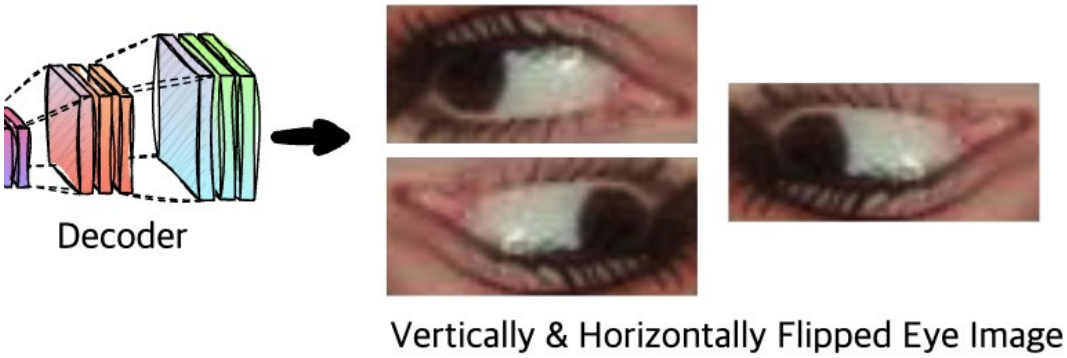
As shown in figure 1, the representation learning of this system incorporates two major steps: encoding the original eye image with a random sign reverse module, and decoding the image to create a new image of the eye looking in a different direction. This representation learning technique will aid us disentangle the necessary features needed to estimate gaze without supervised data.

Figure 1 (a) describes the encoding of the original image. The image is inputted in an encoder which separates the features of the image into three major categories: appearance, yaw, and pitch. The appearance describes the features of the human eye that are unrelated to gaze estimation such as the eye shape or the eye color. This unnecessary information will be pulled apart from the yaw and pitch features: the coordinate system for gaze estimation. After these features are separated, the yaw and pitch vector will go into a random sign reverse module which will invert the gaze either horizontally, vertically, or both horizontally and vertically.

After the necessary features are extracted and classified, this data will be decoded to create an image of the eye based on the inverted gaze direction vectors as shown in figure 1 (b). Inverting the gaze of the image of the eye is salient in calculating the difference between the gaze vectors of images to train the gaze estimation algorithm without labeled data.



(a)



(b)

Figure 1. Network architecture of the proposed unsupervised gaze representation learning
(a): Feature extraction and proposed random sign reverse module and (b): image reconstruction

Equation 1: Reconstruction loss function

$$L_{rec} = \sum_{k \in F} \sum_{j \in Y} \sum_{i \in X} |I_k(i,j) - \hat{I}_k(i,j)|_1$$

The reconstruction loss function is a vital component of the representation learning process. It calculates the loss by how much the gaze vector of the original image and the inverted image differs, allowing the model to learn gaze patterns. In the equation, X represents the yaw values of the image, Y represents the pitch values of the image, and F denotes the random sign reverse values. In this function, the image with the gaze vector (i, j) for given a yaw and pitch value is inverted in one of three directions of F and compared to the original image with the gaze vector which is not inverted. The losses calculated from these differences are added up to calculate the reconstruction loss. Based on this value, the model can capture gaze-related information which allows the model to predict the actual gaze value.

Gaze Estimation

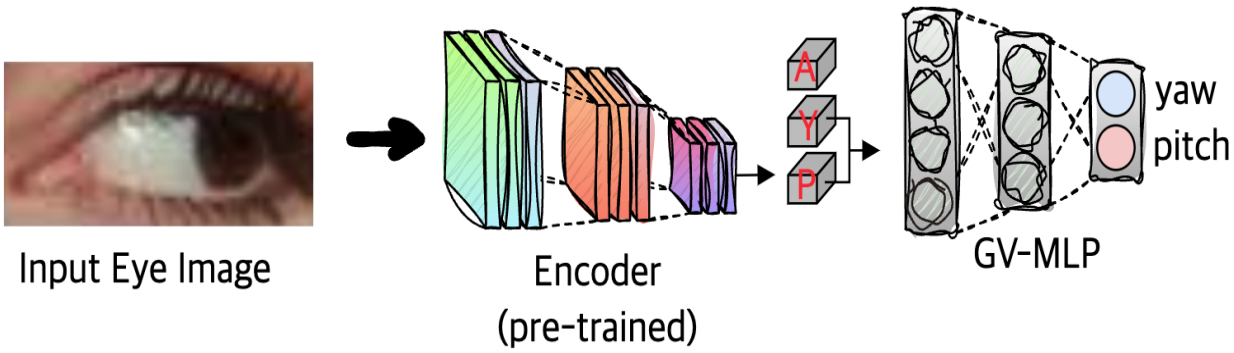


Figure 2. Architecture of the proposed network

The actual gaze estimation will use transfer learning from the pre-trained algorithm from the representation learning. The model that was trained unsupervised will be encoded to separate gaze related features from irrelevant features, and will go through a convolutional layer, GV-MLP (gaze vector multi-layer perceptron), to perform a regression task to calculate the vector from a continuous value. The GV-MLP can accurately predict yaw and pitch with the help of an aggregation of two separate loss functions: traditional gaze estimation loss function (equation 2) and the new proposed loss function based on biological principles involving the consistent gaze ability, further explained in equation 3.

Equation 2: L2 distance loss function

$$L_{L2} = \sum_{k \in [L,R]} |y - \hat{y}_k|^2 + |p - \hat{p}_k|^2$$

The L2 distance loss function is a conventional way to calculate loss for gaze estimation: it simply calculates the difference between the estimated gaze from the actual gaze. In the equation, $\hat{y}$ and p^ are both the ground truth–

or the actual gaze vector of either the left or the right eye based on k. The y and p are the predictions of the gaze, which are then put into the distance function with their respective pairs and added up. This loss function will be the basis of calculating the loss of the model.

Equation 3: Conjugate gaze consistency loss function

$$L_{cons} = |\hat{y}_L - \hat{y}_R|^2 + |\hat{p}_L - \hat{p}_R|^2$$

The conjugate gaze consistency loss function uses the principles of gaze consistency ability, a medical term for the left and the right eyes' ability to look in the same direction. Based on this, the difference of the yaws and the pitches of the left and the right eye must be minimized for the better performance of the model. To conduct this, the difference between the yaws of the left and the right eye and the difference between the pitches of the left and the right eye are each calculated, squared, and added to each other to obtain the loss.

Equation 4: Total gaze loss function

$$L_{gaze} = L_{L2} + \beta L_{cons}$$

Finally, the total loss of the gaze estimation can be calculated by adding the two loss functions from equations 2 and 3 with some alterations. The loss calculated from the conjugate gaze consistency loss function will be multiplied to a hyperparameter, β which can be manually adjusted to maximize the performance of the algorithm.

Experimental Results

Dataset

The dataset utilized to test and train the performance of the gaze estimation algorithm is EVE (End-to-end Video-based Eye-tracking) dataset as shown in figure 3 (Park et al. 2020). This dataset collected data from 54 participants and consists of 4 camera views with over 12 million frames. This dataset was specifically chosen over the other available datasets such as EyeDiap or MPII-NV dataset due to its diverse distribution of yaws and pitches as shown in figure 4. Thus, the model is able to track the eye movement in a greater range using the EVE dataset.

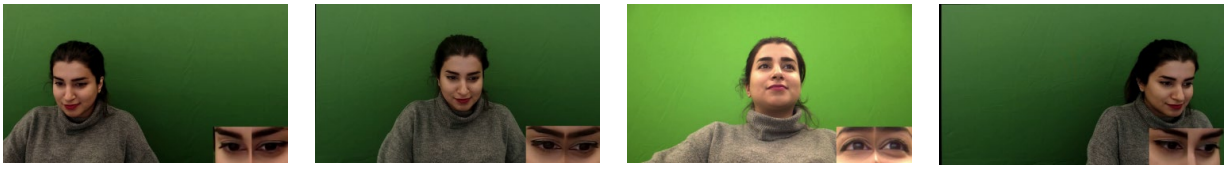
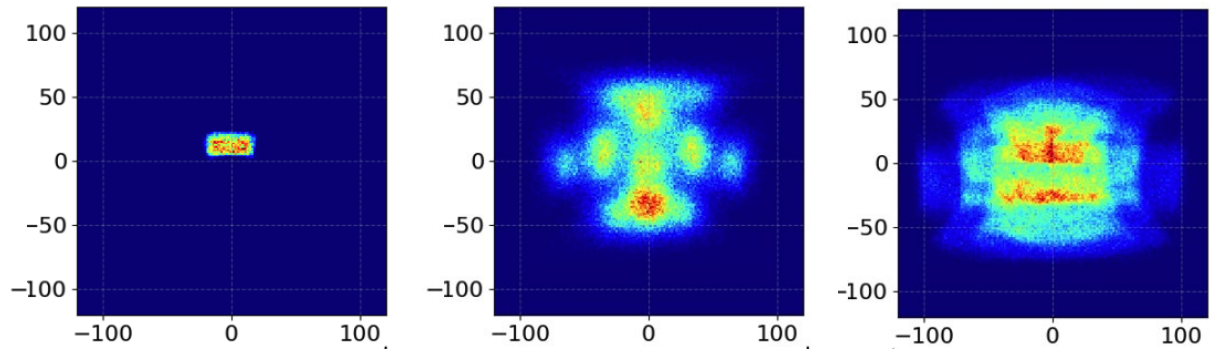


Figure 3. Samples in EVE dataset (Park et al. 2020)



(a): EyeDiap (b): MPII-NV (c): EVE

Figure 4. Yaw and pitch distribution of three datasets
(a): EyeDiap (Funes Mora et al. 2014), (b): MPII-NV (Zhang et al. 2015), and (c): EVE (Park et al. 2020)

Evaluation Metric

The evaluation of the angular error of this algorithm will utilize equation 5.

Equation 5: Angle Error Function

$$\theta=\frac{180}{\pi}\times cos^{-1}\left(\frac{\vec{gt}*\vec{pred}}{|\vec{gt}|_1\times|\vec{pred}|_1}\right)$$

This equation obtains the angular distance between the ground truth angle and the prediction angle by dot producing the two vector values of angles which will be divided by the cross product of the magnitude of the vectors. This will result in a value from 0 to 1 with 0 being the angles of the two vectors being identical to 1 being the maximum difference of the angles. Then, the inverse cosine function will convert this value into radians, then 180/π will be multiplied to convert this value into degrees.

Performance Comparison

The performance comparison of the proposed approach of utilizing the random sign reverse module and the incorporation of the concepts of conjugate gaze ability to the state-of-the-art gaze estimation methods is to accurately evaluate the performance of the proposed method compared to previous approaches.

Table 1. Performance comparison with state-of-the-art gaze estimation methods

Method	**Angle Error (degrees)**
(Park et al. 2019)	14.56
(Yu et al. 2020)	12.19
(Sun et al. 2021)	10.27
(Gideon et al. 2022)	10.03
Proposed	9.42

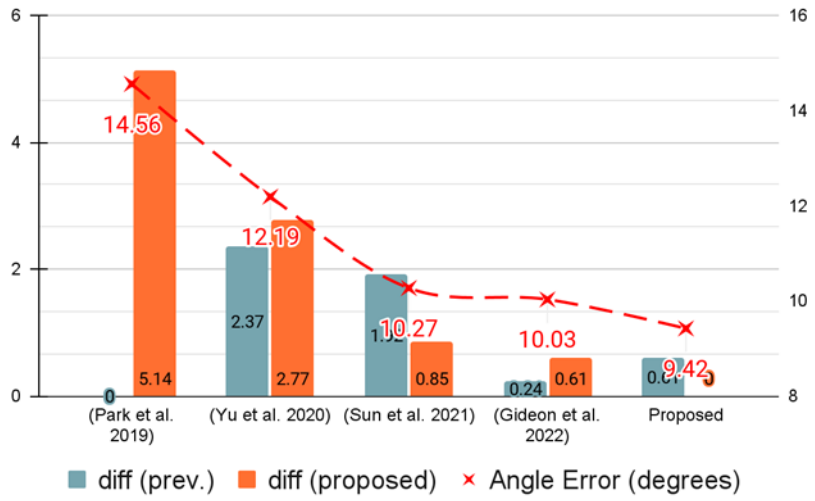


Figure 5. Performance comparison with state-of-the-art gaze estimation methods

Primarily, as shown in table 1 and figure 5, the difference between the angle error of our proposed method to the Park's proposed method differs by 5.14 degrees. This method was the first attempt to incorporate representation learning to gaze estimation, utilizing a supervised learning algorithm via rotation matrices. Then, Yu's method, which differs to our proposed method by 2.77 degrees, conducted representation learning utilizing an unsupervised approach also utilizing rotation matrices. Sun's proposed method has a greater angle error than our proposed method by 0.85 degrees in which they utilized cross-encoding techniques to swap gaze-related features from each other for another unsupervised method. Finally, the method proposed by Gideon has a greater angle error to our proposed method by 0.61 degrees, where he gathered multi-view images of the eye from different angles. Multi-dimensional cross encoder was utilized to swap significant features to each others' data to enhance the unsupervised training of the model.

Table 2. Three-component ablation study

Method	Representation Learning	RSR Module	Consistency Loss	Angle Error (degrees)
Baseline	X	X	X	17.83
Abl. Model 1	O	X	X	17.81
Abl. Model 2	O	O	X	9.96
Abl. Model 3	X	X	O	14.28
Proposed	O	O	O	9.42

The three-component ablation study of the proposed method was conducted to demonstrate the significance of each different proposed component of the algorithm. There are three components utilized: representation learning without the random sign reverse (RSR) module, representation learning with RSR module, and the consistency loss algorithm utilized in transfer learning.

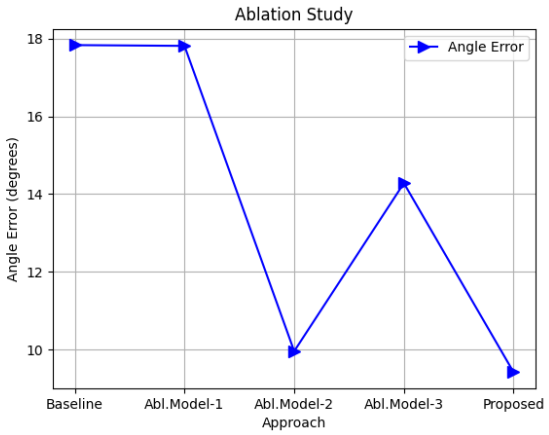


Figure 6. Three-component ablation study

Table 2 and figure 6 depicts the performance of each ablation model with certain components activated and the others deactivated. Primarily, without using any components, the ablation model had an angle error of 17.83 degrees, which did not significantly improve when representation learning was utilized without the RSR module. However, the incorporation of the RSR module to the representation learning resulted in a dramatic decrease of angle error to 9.96 degrees. Utilizing the consistency loss function but not any representation learning methods, though not as extreme as utilizing the RSR module, resulted in a significant decrease in the error to 14.28 degrees. Finally, our proposed method of utilizing all the components resulted in the angle error of 9.42 degrees, a 8.41 degree difference from the module without any components, showing the significance of each component.

Application

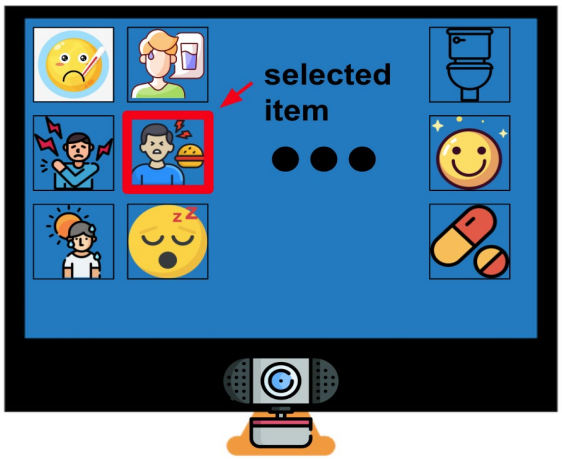


Figure 7. System setup

Finally, I present a practical setup for implementing a gaze estimation-based digital communication board system. Figure 7 illustrates the setup, which consists of a webcam and a monitor. The webcam captures the user's facial image and crops the eye area, feeding it into the proposed gaze estimation network. Eye cropping can be easily achieved using off-the-shelf facial analysis algorithms like dlib (Suwarno and Kevin 2020). The network outputs a gaze vector, represented in yaw and pitch degrees. A rule-based logic is then applied to analyze the user's gaze direction by

comparing the predicted yaw and pitch with predefined threshold values. This logic enables users to select specific items from the communication board independently which offers a faster and more efficient solution for patients without the need for an assistant.

Conclusion

In this paper, I aimed to improve the accuracy of gaze-estimation modules in order to create communication boards that are more accessible and practical for patients who have significant impaired movements–including but not limited to quadriplegia patients. The method I proposed was incorporating random sign reverse modules to effectively train the algorithm gaze prediction unsupervised and using principles of gaze consistency abilities to further enhance the performance of the algorithm. The proposed model could predict the gaze with an accuracy with the angle error of 9.42 degrees, a significant improvement from the state-of-the-art gaze estimation methods. Additionally, I separately tested different ablation modules of the algorithm by excluding certain components of the algorithm to verify its efficiency, where it validated how both the RSR module and the conjugate loss function result in a decrease in the angle error, with the RSR module creating a more remarkable difference. I successfully developed a gaze estimation algorithm by utilizing unsupervised representation learning model and biological principles involving the conjugate gaze ability for a more practical and helpful communication board for patients in need of effective communication devices.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

Funes Mora, K. A., Monay, F., & Odobez, J. M. (2014, March). 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 (pp. 255-258).

Gamper, J., Koohbanani, N. A., Benes, K., Graham, S., Jahanifar, M., Khurram, S. A., ... & Rajpoot, N. (2020). Pannuke dataset extension, insights and baselines. arXiv preprint arXiv:2003.10778.

Kar, A., & Corcoran, P. (2017). A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms. IEEE Access, 5, 16495-16519.

Kim, J. H., & Jeong, J. W. (2020). Gaze in the dark: Gaze estimation in a low-light environment with generative adversarial networks. Sensors, 20(17), 4935.

Park, S., Aksan, E., Zhang, X., & Hilliges, O. (2020). Towards end-to-end video-based eye-tracking. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16 (pp. 747-763). Springer International Publishing.

Patak, L., Gawlinski, A., Fung, N. I., Doering, L., Berg, J., & Henneman, E. A. (2006). Communication boards in critical care: patients' views. Applied nursing research, 19(4), 182-190.

Suwarno, S., & Kevin, K. (2020). Analysis of face recognition algorithm: Dlib and opencv. Journal of Informatics and Telecommunication Engineering, 4(1), 173-184.

Zhang, D., Yin, J., Zhu, X., & Zhang, C. (2018). Network representation learning: A survey. IEEE transactions on Big Data, 6(1), 3-28.

Zhang, X., Sugano, Y., Fritz, M., & Bulling, A. (2015). Appearance-based gaze estimation in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4511-4520).