HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 1 (2024)

Multi-View Gaze: Leveraging Multi-View Images to Disentangle Features for Accurate Gaze Estimation

Yireh Ban

Troy High School, USA

ABSTRACT

Gaze estimation is a prominent field within artificial intelligence and machine learning, rapidly developing due to the practical uses and possibilities. However, this development brings challenges, such as inaccuracy with different facial features and external factors such as lighting or camera quality. In the past, research has led to a cross encoder, or a swapping mechanism of the disentangled data from an image. The proposed method takes this one step further and incorporates multi-view images to leverage this disentanglement. Multi-view images allow for more data pairs within images to be swapped, resulting in more maximized and fine-tuned accuracy in gaze detection. Another added detail was transfer learning, or the carry over of a pre-optimized encoder to make the training process much more efficient. This can be incorporated into the real world, for example, by using it to control a computer mouse without physical movement or to detect patterns to diagnose neurodevelopmental disorders such as Attention Deficit Hyperactivity Disorder (ADHD) which can be difficult to detect in young children otherwise. The results of this newly proposed method produced more accurate results than state-of-the-art mechanisms, only having an angular error of 7.4 when trained and tested within the EVE dataset.

Introduction

Problem Definition

For many, the eyes serve more purpose than merely to see. It may serve as someone's only form of communication, or for others a vital way to convey emotion or other social cues. Many of these people may have difficulty in accessing what is considered to be a commonly used application, such as computer mice or anything needing physical activity. Using gaze estimation, this form of movement can be expanded and exponentiated to allow them to express their thoughts and interact with the world around them.

Gaze estimation, or the process of determining the direction in which a person's eyes are focused also prove to be a challenge. Despite the necessity of the development of this field and the ongoing research in result, factors such as variance in camera quality, lighting, or facial features prove to be a challenge. To refine and enhance gaze estimation, further research and development is crucial to unlock the full potential and incorporate it into the real world.

Previous Method

Previously, new methods of gaze estimation were developed. FAZE (Park et al. 2019) proposed a method consisting of a rotational representation of gaze following disentanglement of data to create an adaptable gaze estimator. This method was the first to propose a gaze representation learning technique, and it relied on the assumption that the disentanglement of data contained information about appearance, gaze direction, and head

rotation. However, this proposal was a supervised approach, requiring extensive human work, leading to difficulty in producing a larger scale dataset.

U-LinFT (Yu et al. 2020) provided an unsupervised approach. In this proposal, a network extracted gaze representations from input images to recreate and generate a redirected eye, with a calculated difference to measure error. While this method was successful in eliminating the previous problem of creating a suitable large-scale dataset, it suffered difficulties with dataset bias and data annotations due to a variety of distractions.

Cross Encoder (Sun et al. 2021) was a new approach which leveraged unsupervised learning by Sun et al. In this proposal, a cross-encoding method was discussed, involving the isolation of the gaze and appearance features and swapping them between two images, one with the same gaze and the other, the same appearance. The loss was then calculated as angular error and was used to minimize the difference between the input and output images.

Proposed Method

Following the various studies conducted and inspired by the previous gaze representation learning methods proposed in these studies, I propose a multi-view based method that leverages the current state of gaze representation. With the additional data provided by the use of multi-view images, more image pairs and data could be swapped. This additional swapping allowed for more leverage and therefore demonstrated an increased accuracy. However, this also meant that a more reliable disentangling method was essential. For this, a transfer learning process was incorporated, where a pre-trained neural network would be transferred over instead of training from scratch. This would allow for overall accuracy improvements and allow for leveraging of the already-trained encoder for knowledge and insight.

The primary contributions of this research are trifold:

- I employ a novel multi-view approach for gaze estimation, which, to the best of my knowledge, has not been previously explored.
- Through extensive experimentation, I have demonstrated that the proposed method outperforms existing approaches, showcasing its superiority.
- I also propose a practical application for the proposed method, which can be utilized for example in the diagnosis of ADHD or an eye-mouse interface.

Overview

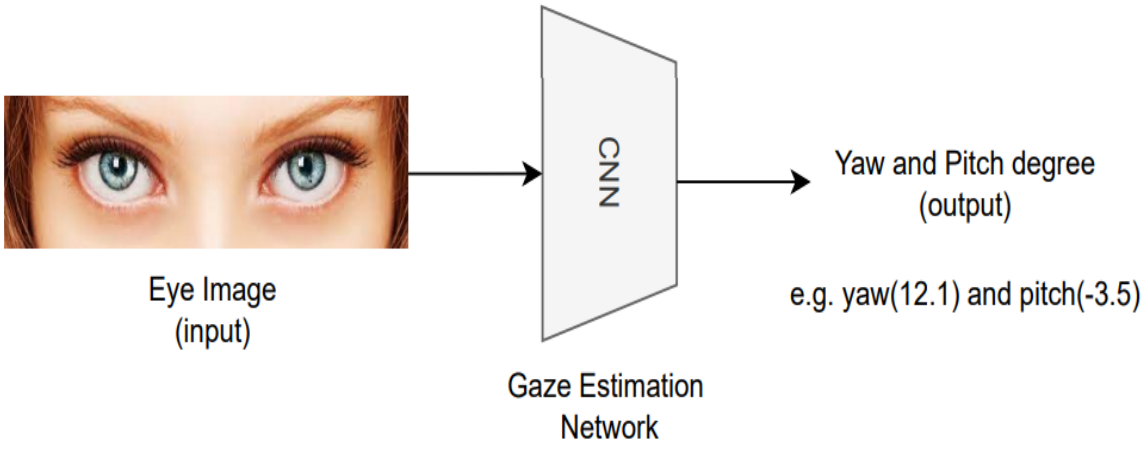
The following chapters are composed as follows: chapter 2 provides a comprehensive overview of the related work in the field, establishing the necessary background knowledge to understand my research. Chapter 3 presents a detailed explanation of the proposed method for gaze estimation, outlining its key components and techniques. Subsequently, chapter 4 focuses on experimental validation, demonstrating through rigorous experimentation that the proposed method outperforms existing approaches. Finally, in chapter 5, the paper is concluded by summarizing the main findings and implications of the research while also discussing potential future investigation in this field.

Related Work

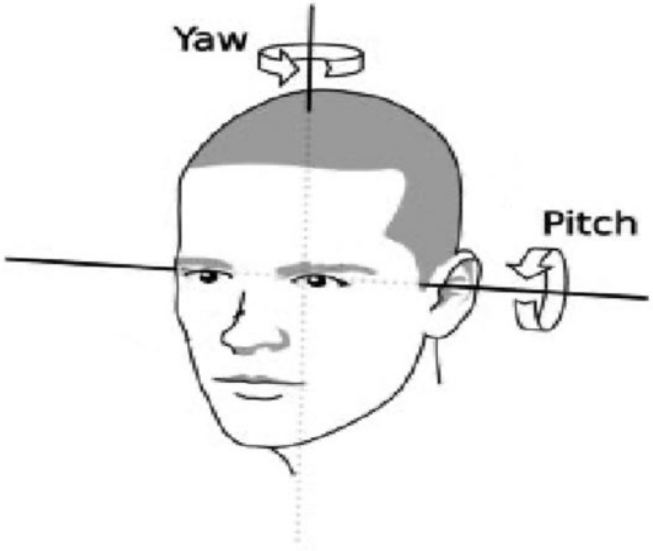
Gaze Estimation

Gaze estimation is a computer vision task that aims to determine the direction of a person's gaze based on input images, typically of the face or eyes. The output of gaze estimation is typically represented by the yaw and pitch

angles, which indicate the horizontal and vertical direction of the gaze, respectively. For example, if the predicted yaw degree was 10.5, the general gaze would be facing right, and if the pitch was 1.3, the gaze would be tilted upwards. Similarly, a yaw of -5.4 and a pitch of -2.8 would indicate a left and downward pointed gaze.



(a)



(b)

Figure 1. Gaze Estimation Concept Visualization. (a): Example of Gaze Estimation Network, (b): Yaw and Pitch Demonstration

Figure 1 (a). depicts an example of what the method uses as an input and what process it goes through to produce the output, in this particular example, using an eye image as an input and using convolutional neural network (CNN) as the network used to estimate gaze, producing a yaw and pitch degree with the provided details. Figure 1 (b). gives a visual representation of the output and what yaw and pitch looks like.

Gaze estimation has emerged as an extremely powerful tool with a multitude of possible applications. By incorporating gaze estimation into many pre-existing fields, human-computer-interactions have strengthened and improved. For example, gaze estimation enables hands-free navigation and control, improving accessibility for individuals with physical limitations and enhancing virtual reality experiences. Gaze estimation has

also proved invaluable in assessing the condition of drivers based on their eye movements. This technology has also affected industries such as advertisement, using eye tracking to keep track of customer behavior and improve user experiences. The versatility and potential gaze estimation holds will continue to shape and better technology.

Multi-View Image

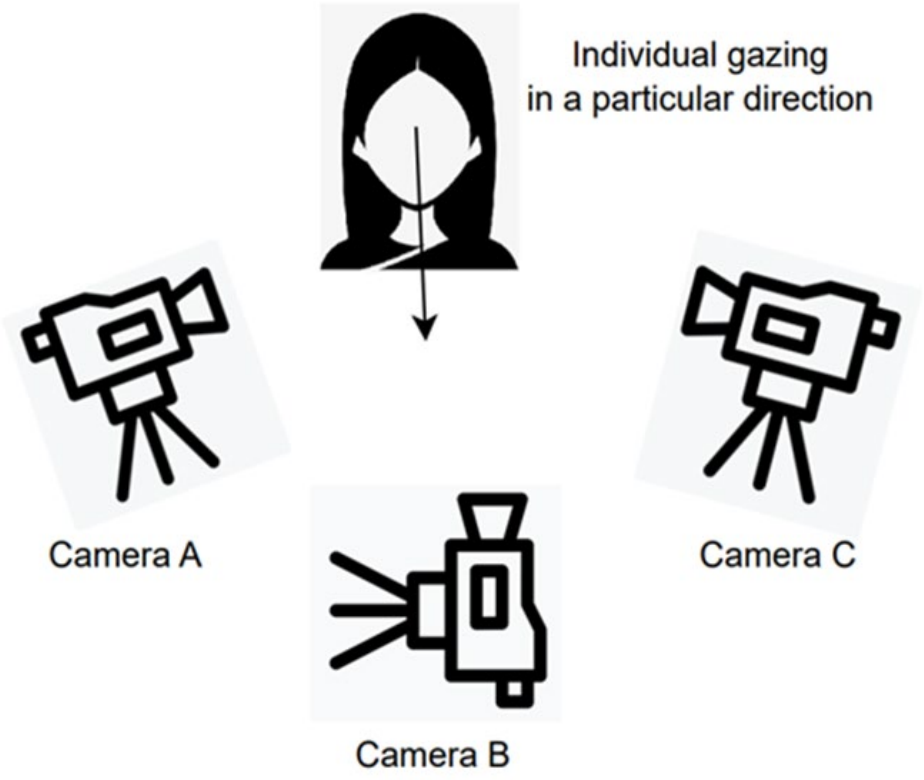


Figure 2. Multi-view image acquisition setting

(a) (b) (c)

Figure 3. Examples of acquired multi-view images. (a): Left View, (b): Front View, and (c): Right View

Multi-view images are a broad description categorizing an object with images from multiple perspectives simultaneously. By capturing multiple views of the same scene, it becomes possible to gather more information about the object, in this case, about the appearance and gaze direction. This is because as more images are taken, more swappable features are exposed which allows for increased accuracy.

Figure 2 provides a visual representation of what a multiview image entails. The image depicts an individual gazing in one direction. This scene is then captured by multiple cameras at the same time so that the images all display the same gaze direction but have different appearance-related features. For example, one such scenario is shown with Figure 3, where the same individual is shown in three images: Camera A shows the scene taken from the left of the person, B from the front, and C from the right. These images are all taken simultaneously, and therefore show the same gaze or direction of the person's eyes, but are at different angles, showcasing varying appearance-based features and visually representing multiview images.

The primary objective of this research is to establish a gaze-related correlation among these multi-view images, which we hypothesize will enhance the accuracy of the gaze estimation network. The EVE (Park et al. 2020) dataset is a widely recognized collection of samples for gaze estimation, acquired using the previously mentioned multi-view imaging settings. Using the EVE dataset, this research can be enhanced and the gaze-related features will be more prominent, allowing for improved accuracy in gaze estimation overall. Details regarding results and statistics are further explained in Chapter 4.

Gaze Representation Learning

Representation Learning

Representation learning is a crucial aspect of machine learning that involves the automatic extraction of meaningful and useful features from raw input data. This process aims to transform this raw data into a more abstract and compact representation that captures important underlying patterns, structures, or features of the data. In the context of gaze estimation, representation learning has a significant role in extracting gaze-related features from the input data.

The ultimate goal of representation learning is to discover features that are meaningful towards a task. For gaze estimation, this involves the representations that capture the essential characteristics and patterns of gaze behavior. By finding better representations the accuracy of estimation models can be enhanced.

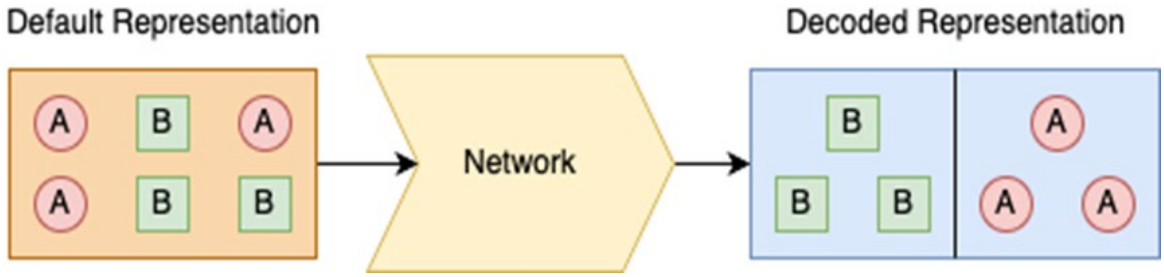


Figure 4. Diagram of Representation Learning

Gaze Representation Learning

To find better representations, the disentanglement and separation of data is necessary. This information can then be put through a cross encoder, or a feature swapping mechanism. By leveraging multiple views or perspectives, the model can capture different aspects of gaze behavior to create an improved representation.

Feature swapping is a novel approach that works by leveraging deep learning techniques, particularly convolutional neural networks (CNN). It involves the swapping of features extracted from two different sources,

and through the combination of the sources, the model improves its accuracy in the extraction and categorization of data. In the context of gaze estimation, the feature swapping mechanism is applied by training CNNs to extract relevant features about the gaze and the facial features. These extracted features are then swapped to minimize error and distinguish the gaze and facial features, making it more resilient to variations and challenging conditions. The usage of feature swapping heavily leverages the current gaze estimation methods and addresses many of the limitations traditional models face to produce more accurate results.

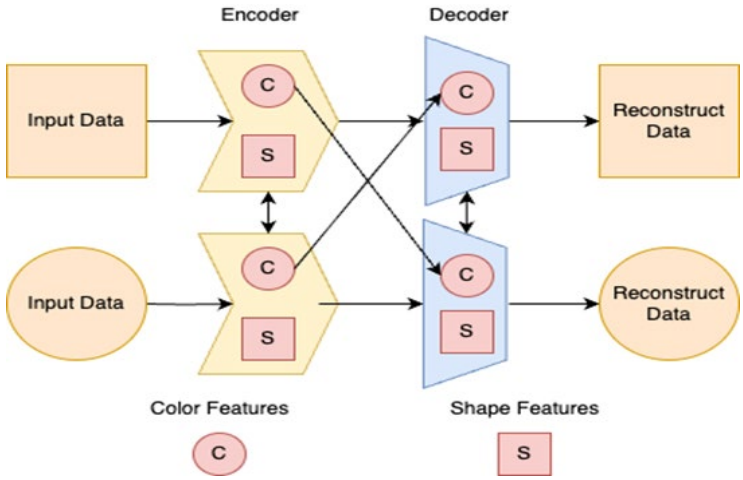
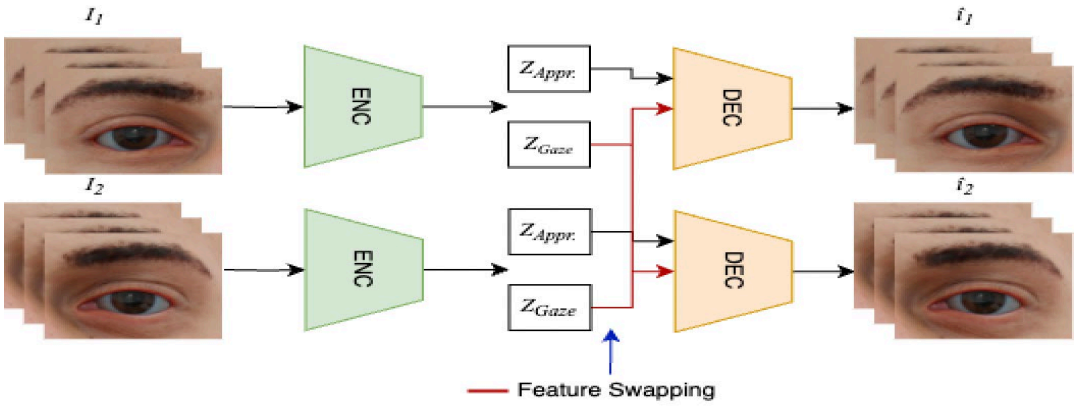


Figure 5. Diagram of Gaze Representation Learning

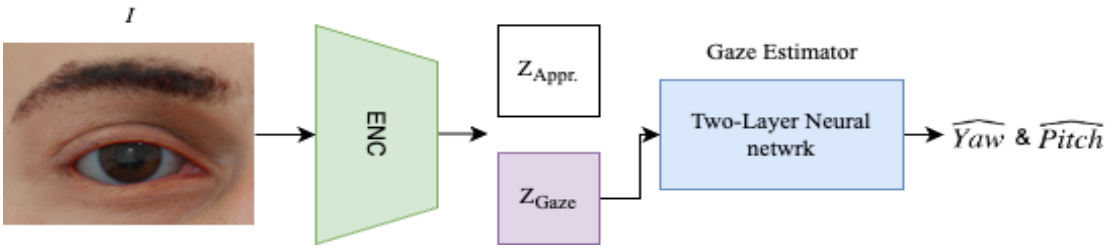
By combining the feature swapping mechanism with multi-view images, the proposed method aims to find better representations for gaze estimation. These representations effectively capture the underlying structure, patterns, and characteristics of gaze behavior, leading to more accurate and reliable gaze estimation models.

Multi-View based Gaze Representation Learning

System Overview



(a)



(b)

Figure 6. Overview Architecture of the Proposed Method. (a): Gaze Representation Learning and (b): Transfer learning (Gaze Estimation)

Figure 6 provides an overview of the architecture employed in the proposed method for gaze estimation, consisting of two key components, displayed by each of the (a): Gaze Representation Learning and (b): Transfer learning (Gaze Estimation). In the first component, the model focuses on extracting features and reconstructing the original image by swapping image features from multiple images, or multiview images, to enhance the disentanglement of features, using deep learning including convolutional neural networks (CNNs). The second component works to estimate the gaze vector which is composed of yaw and pitch by using the pretrained encoder. Among the extracted features, only the gaze-related features are further placed into a two-layer neural network and used to calculate the result.

Gaze Representation Learning

The ultimate goal of the proposed gaze representation learning is to leverage gaze estimation using improved representation and disentanglement with the use of multiview images. To achieve this goal, first, a set of multiview images must be obtained, which provide different perspectives and angles, which can help improve the accuracy of gaze estimation by providing an increased amount of data to swap. Next, the gaze estimation algorithm processes the images, extracting and disentangling the data and swapping them using neural networks. As a result of swapping a greater number of eye image pairs, the trained network becomes capable of extracting more consistent features, leading to improved accuracy in gaze estimation.

This swapping is demonstrated in figure 6(a). Once the multiview images (I_1, I_2) go into the encoder (*ENC*), they are extracted from their data. This data is then disentangled into $Z_{Appr.}$ and Z_{Gaze}, representing the appearance related data and representing the gaze related data, respectively. The encoder can be more concisely defined as follows: $ENC: I_k \rightarrow Z; \quad Z=\{Z_{Appr.}, Z_{Gaze}\}$, where $k \in \{1,2\}$. The Z_{Gaze} data is then swapped between the two images before being placed in the decoder. The decoder can be more concisely defined as follows: $DEC: Z=\{Z_{Appr.}, Z_{Gaze}\} \rightarrow \hat{\imath}_k$, where $k \in \{1,2\}$. The results produced, should, in theory, be an identical replica of the original input image, therefore increasing the accuracy of extracting the gaze-related data.

To train the proposed architecture, a loss function, denoted as L_1, is defined and used during the training process. This loss function plays a crucial role in optimizing the model's performance. The L_1 loss function measures the average pixel intensity difference between the prediction and ground truth. The absolute difference is computed for each sample then averaged over all the samples. By minimizing the loss, calculated by L_1, the model is ultimately improved due to the reduction of discrepancies between the predicted and ground truth. By utilizing this loss function, the model trains to estimate gaze directions closer to the ground truth. Theoretically, when the loss value reaches 0, the model will produce 100% accuracy and will be the best possible.

Equation 1: L1 Loss Function

$$L_1 = \sum_{x}^{X}\sum_{y}^{Y}|I_1(x,y) - \hat{I}_1(x,y)| + \sum_{x}^{X}\sum_{y}^{Y}|I_2(x,y) - \hat{I}_2(x,y)|$$

Here, *X* and *Y* denote the width and height of the inputted image, respectively. *I(x,y)* represents the pixel intensity at the location *(x,y)*.

Transfer Learning (Gaze Estimation)

The ultimate goal of transfer learning in the scope of gaze estimation is to carry over a previously trained neural network. The network used in the transfer learning step is the same as previously mentioned in Chapter 3.2. The gaze-related feature, Z_{Gaze}, is run through the network, producing the gaze vector *V={yaw, pitch}* of the gaze. More simply put, using this network, the input consists of multiview images, and the output is the gaze direction, or *GazeEST:* $Z_{Gaze} \rightarrow V=\{yaw, pitch\}$.

The use of transfer learning is beneficial due to the already-optimized encoder, allowing for leveraged knowledge and insight, instead of starting from scratch. Transfer learning also enables building upon a pre-trained model which significantly speeds up the learning process and improves performance.

Mean squared error will be used to train the proposed *ENC* and *GazeEST*. The mean squared error loss function is calculated as follows:

Equation 2: Mean Squared Error Loss Function

$$L_{MSE} = \frac{1}{N}\sum_{n}^{N}(v_n - \hat{v}_n)^2$$

Here, *N* represents the total number of samples. Similarly, v_n represents the nth ground truth gaze vector while $\hat{v}_n$ represents the predicted gaze vector.

Implementation Details

In the architectural design of the gaze estimation implementation, a two-fold approach involving Gaze Representation and Transfer Learning is employed. For Gaze Representation, we utilize an encoder-decoder structure with Reset-50 as the encoder, allowing for feature extraction. In the decoder, downsampling layers were transitioned into upsampling layers to enhance the preservation of data. Following, in the transfer learning phase, we retain the Resnet-50 encoder and introduce a two-layer neural network for gaze estimation. During training, the learning rate is 0.0001, conducting 200 epochs, and the learning rate decays at epochs 120 and 160 with decay factor 0.1. The batch size is set to 512 for sufficient training. In contrast, for transfer learning, we maintain the same learning rate but adopt a shorter training duration of 60 epochs with a learning rate decay at epoch 40 with 0.1. The batch size is kept constant at 512.

Experimental Result

Dataset

The EVE (Park et al. 2020) dataset provided multi-view and was used to train and test for the within-dataset and few-shot experiments. EVE is a dataset that holds about 105 hours of footage, providing over 12 million frames and 1,327 unique visual stimuli from fifty-four participants at four different camera views.

The GazeCapture dataset was used as the test dataset for the cross-dataset experiment. This dataset had about 2.5 million frames of data from over 1,450 participants.

Angular Error

Angular error is a metric used to provide a quantitative measure of how accurate the gaze estimation algorithm is in predicting the direction a person is looking. It provides a numerical value that represents the distance between the estimated gaze and the ground truth, represented as equation 3. By using the angular error, the model can then be optimized to decrease the error value. Theoretically, once this value hits zero, the model will predict the gaze correctly one hundred percent of the time.

Equation 3: Angular Error

$$\text{Angular Error} = cos^{-1}(cos(\alpha - \beta)) = \frac{A \cdot B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_i B}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum}}$$

Here, A and B denote the predicted gaze vector and its ground truth gaze vector, respectively, and the dot product of A and B are taken and divided by the product of the norms. In the best case scenario, the angular error is zero degrees, and the worst case is 180 degrees, due to the inverse cosine.

Experimental Protocol

In this research, five different experiments were conducted to demonstrate the efficiency and accuracy of the proposed algorithm in comparison to current state-of-the-art methods. The first experiment is a within dataset test, which allows training and testing with the same dataset to compare the performance of different algorithms. The second experiment uses a cross dataset test in order to confirm the domain generalization performance. Thirdly, a few-shot test is conducted, which is the usage of only a few training images to practice transfer learning. Fourth, latent code visualization is conducted, which uses high dimensional feature spaces using a t-SNE algorithm, which compresses high dimensional data into low dimensional data. This uses clustering, expressed in a graph, to determine accuracy, where the more definitely grouped clusters give a visual demonstration of the relationship in data. Lastly, data augmentation tests the algorithm by providing a variety of images, to identify which change may have benefited the method.

Within Dataset Evaluation

Table 1. Angular Error Comparisons on EVE (Park et al. 2020) Dataset

Method	Angular Error
FAZE (Park et al. 2019)	15.8
U-LinFT (Yu et al. 2020)	11.4
CrossEncoder (Sun et al. 2021)	9.7
Proposed Method	**7.4**

An within dataset evaluation uses the same dataset for both training and testing. Table 1 shows the angular error values between the 4 presented methods compared under the EVE (Park et al. 2020) dataset using this method of evaluation. As shown, for the EVE dataset, SOTA-1, 2, and 3 demonstrated a higher angular error value than the proposed method, exhibiting the effectiveness of the proposed feature swapping mechanism.

Cross Dataset Evaluation

Table 2. Angular Error Comparisons of Cross Dataset Evaluation

Method	Angular Error	
	GazeCapture	Gaze360
FAZE (Park et al. 2019)	23.1	22.7
U-LinFT (Yu et al. 2020)	21.8	19.5
CrossEncoder (Sun et al. 2021)	17.4	15.7
Proposed Method	**9.7**	**8.2**

The process of training and testing the methods use two different datasets: a method is trained using one dataset and tested with another. The dataset used to train and test are kept constant within a comparison, and the training and testing datasets can be varied to prove the reproducibility and increase the reliability of the results.

Table 2 shows the angular error values between the four presented methods compared under the GazeCapture and Gaze 360 datasets and trained by the EVE dataset. In both cases, the proposed method produced a lower angular error value than the state-of-the-art methods currently published, displaying the benefit of this method.

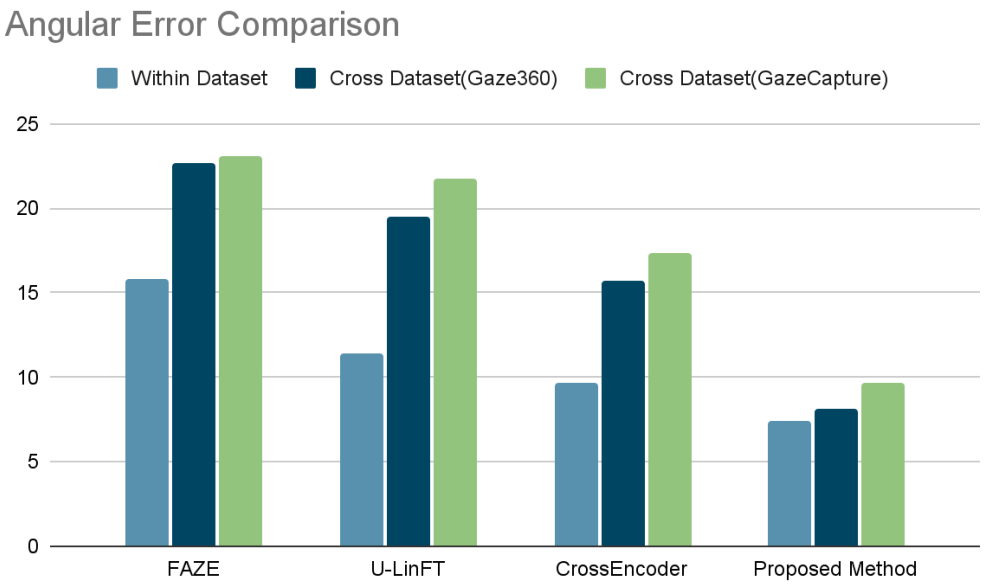


Figure 7. Angular Error Comparison Table for Within and Cross Dataset Experiments

Figure 7 shows the effectiveness of the proposed method over novel state-of-the-art methods as it maintains the lowest angular error throughout both experiments. Additionally, the proposed method also displays the lowest performance drop between within and cross dataset angular errors, showing its generalizability.

Few-Shot Evaluation

Table 3. Few-shot Evaluation on EVE (Park et al. 2020) Dataset

Method	Angular Error				
	8	16	32	64	128
FAZE (Park et al. 2019)	23.0	16.6	12.9	10.8	10.5
U-LinFT (Yu et al. 2020)	22.4	15.9	12.2	9.7	9.3
CrossEncoder (Sun et al. 2021)	19.6	14.3	10.7	9.5	8.8
Proposed Method	**18.3**	**13.0**	**10.4**	**8.4**	**8.1**

The few-shot evaluation test is another method of testing the effectiveness of the systems used in the proposed method compared to the current state-of-the-art ones. Differently from the previously proposed methods, the current method uses transfer learning to carry over a pre-trained neural network. This neural network will, theoretically, lead to better results and less error even with less data to train with. Table 3 supports this theory, with the proposed error having less error than the other methods, even with smaller shot size training.
error value than the state-of-the-art methods currently published, displaying the benefit of this method.

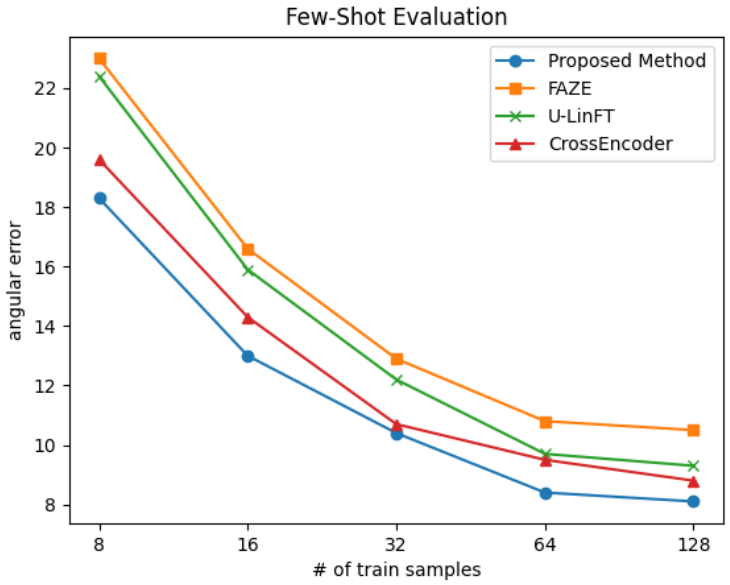
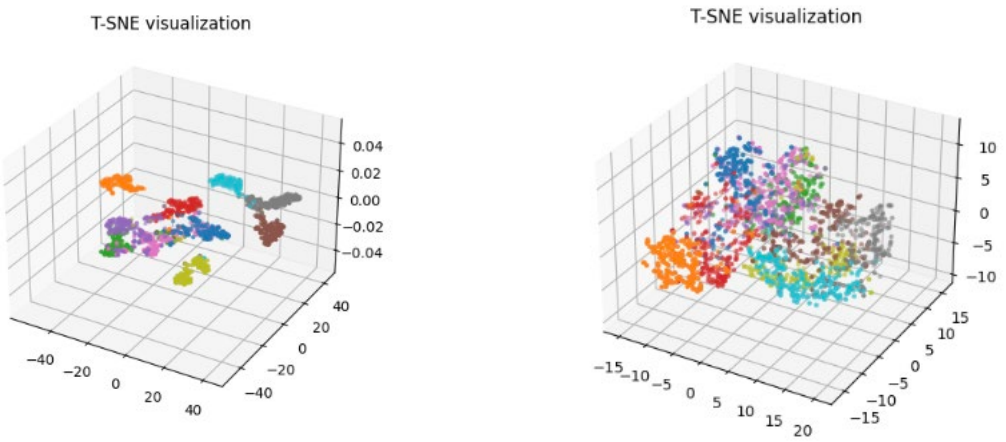


Figure 8. Few-Shot Evaluation Graph

In figure 8, it is also visible that as the number of training shots decrease, the difference in angular error between the proposed and other methods increase, displaying the effectiveness of the transfer learning technique in the proposed method.

Feature Map Evaluation



(a) (b)

Figure 9. t-SNE Evaluation. (a): with transfer learning, (b): without transfer learning

In this method of testing, a latent code visualization takes place, using a t-SNE algorithm. This algorithm takes high dimensional data and compresses it into low dimensional data, which is more handleable. Each data value is plotted (as can be seen in figure 9), and the plotted values are color coded to display the relationships between the data. Clustering or other data patterns can be seen and provides another method of visual data interpretation.

Figure 9(a) displays the visualization of the method with transfer learning, while figure 9(b) displays the visualization of the method without transfer learning. It is clearly visible that figure 9(a) shows very clustered and organized data points, while figure 9(b) shows very scattered and unorganized data. This is representative of the state of disentanglement, and figure 9(b) acts as a control to further display the effectiveness of transfer learning.

Application

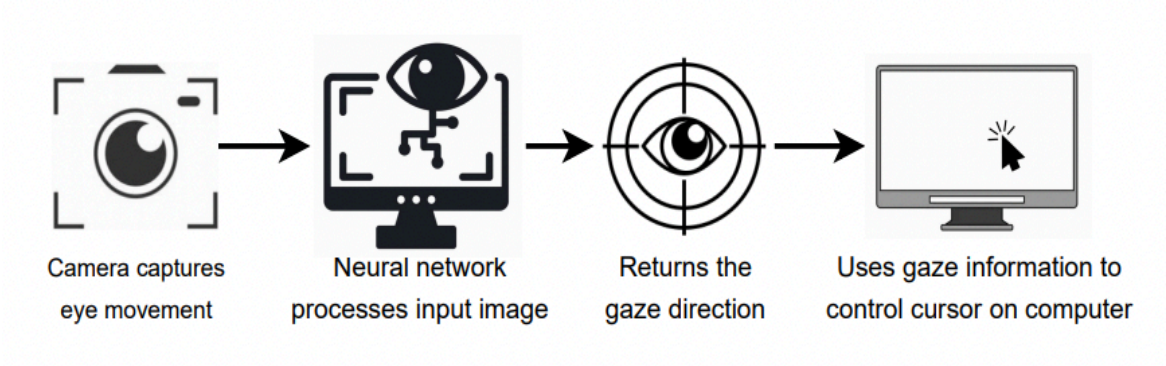


Figure 10. Diagram of the proposed human computer interaction system

The proposed method of gaze estimation has many real-life applications. For example, with human-computer interactions (HCI), gaze estimation can enhance traditional input methods of using mouse and keyboard by

allowing users to interact with computers using their eyes. This enables hand-free control of devices, which is especially useful for those with physical disabilities. Figure 10 displays how such HCI is possible, through a camera capturing eye movement, then processing with a neural network to calculate gaze, which can be used to to control the cursor of a computer.

Conclusion

With this research, the proposed method provides a leveraged solution to accurate gaze estimation, providing a new approach and applying never seen before techniques. The method uses disentanglement procedures, feature swapping, multiview images and transfer learning to create a more accurate and leveraged system. The disentanglement of features allows for separation between the data that is related to gaze direction and the data related to appearance and facial features. This is essential to the overall system as without it, the gaze-related data would be confused with other, unrelated data. The disentangled gaze features are then swapped between the two eyes using a feature swapping technique. As the gaze-related features are swapped, the output should still remain the same, as both eyes are likely looking in the same direction. Using the error calculated between the input and output images, the neural network is optimized. Theoretically, when the input and output images are identical, the network is fully optimized and the data disentanglement has no error. Instead of the usual two points of data swapping with traditional methods, multiview images allow for many more. Multiview images are images that are taken of the same thing, at the same time, but at different angles. This allows for more feature swapping as multiple more perspectives are given, which should all theoretically have the same gaze direction. Minimizing the error of this allows for a much more accurate and leveraged system. This pre-maximized system can then be carried over to be implemented, which heavily decreases the time and resources needed to train an otherwise new neural network. This system of transfer learning allows for a maximized system that can be trained with fewer data points to produce the yaw and pitch of gaze with increasing accuracy. To evaluate the effectiveness of the proposed method, five experiments were conducted, comparing this method to previous state-of-the-art methods, evaluated with angular error. The proposed method proved to have less error in all scenarios, including those with less data, more data, and varying datasets. An additional experiment was organized to visualize the results, which led to a perceptible demonstration of the outperformance of the proposed method compared to the state-of-the-art methods in both efficiency and estimation accuracy. In the future, this method could be applied medically to help detect Attention Deficit Hyperactivity Disorder (ADHD) in those who may be unable to diagnose with the traditional method. This proposed method of gaze estimation has proven to be highly effective and useful, enabling a wide range of applications and further research.

Acknowledgments

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

References

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., & Torralba, A. (2019). Gaze360: Physically unconstrained gaze estimation in the wild. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 6912-6921).

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., & Torralba, A. (2016). Eye tracking for everyone. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2176-2184).

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.

Park, S., Mello, S. D., Molchanov, P., Iqbal, U., Hilliges, O., & Kautz, J. (2019). Few-shot adaptive gaze estimation. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 9368-9377).

Sun, Y., Zeng, J., Shan, S., & Chen, X. (2021). Cross-encoder for unsupervised gaze representation learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3702-3711).

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(11).

Yu, Y., & Odobez, J. M. (2020). Unsupervised representation learning for gaze estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7314-7324).