Institute for Visualization and Interactive Systems

University of Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Gaze Estimation Error Prediction of Information Visualizations

Sichun Zeng

Course of Study:	Informatik
Examiner:	Prof. Dr. Andreas Bulling
Supervisor:	Yao Wang, M.Sc.
Commenced:	2023-11-06
Completed:	2024-5-06

Abstract

As eye-tracking technology gains prominent attention in Information Visualizations (InfoVis) research, the need for high accuracy and precision in eye-tracking data becomes increasingly critical. Gaze estimation error, in the context of eye-tracking, refers to the natural discrepancy between the predicted gaze location and the actual position. Unlike interactions involving physical contact, the visual focus can only be determined through estimation, making gaze estimation errors inevitable. To minimize these errors, calibration is typically performed, where users are asked to look at five or more points on the screen to establish baseline data for ground truth. All following eye-tracking experiments are then based on this calibration data. However, there are situations where calibration is not practical, such as in some remote or online studies or during dynamic activities, where the calibration process cannot be reliably controlled. To deal with this challenge, this paper proposes VisCaiNet, a deep-learning model that predicts gaze estimation error through Area of Interest (AOI) based patterns, using duration, scanpath, scanpath length, and Hit Any AOI Rate (HAAR) as the input features, with calibration error as the output. It can effectively discern between high and low-quality gaze data based on the predefined calibration criteria, offering a solution to the challenges posed by variable conditions and unfeasible calibration situations.

Contents

List of Abbreviations

Adam Adaptive Moment Estimation. 30

AOI Area of Interest. 3

BCE Binary Cross-Entropy. 30

BERT Bidirectional Encoder Representations from Transformers. 25

CNN Convolutional Neural Network. 29

FCR Flipping Candidate Rate. 15

FOV Field of View. 13

HAAR Hit Any AOI Rate. 3

InfoVis Information Visualizations. 3

MLP Multilayer Perceptron. 27

MSE Mean Squared Error. 30

ReLU Rectified Linear Unit. 26

RNN Recurrent Neural Networks. 29

SGD Stochastic Gradient Descent. 30

1 Introduction

Eye-tracking involves using cameras or other sensing devices to detect eye movements[HNA+11]. Initially, the technology was limited to approximately identifying the direction and position of gaze[Bus35]. With the progress of sensor technology, subtle differences in the pupil can now be observed, making it possible to develop high-performance eye-tracking devices. As a result, the implementation of eye-tracking has become prevalent in the field of human-computer interaction[JK03] [MB14], particularly in the study of human interaction with information visualizations (InfoVis)[GH11].

Nevertheless, eye-tracking in information visualization studies requires precise and accurate measurement[FWT+17] of eye movements. Gaze estimation error, in the context of eye-tracking, indicates the natural discrepancy between the predicted gaze location and its actual position [BDB16]. Constrained by the quality of the eye-tracker, the different physical conditions of test subjects, and the varying experimental conditions[EGIK19], the gaze estimation error can not be eliminated.

Currently, researchers are solving the inherent limitations of gaze estimation across two dimensions: geometrical or cognitive, one-time or real-time[WZZ+21]. The geometrical dimension focuses on predicting the errors during the process of gaze estimation, considering factors that may cause the deviation, for instance, the offset accuracy or the algorithms used for mapping positions to coordinates on the screen. On the cognitive side, there is an emphasis on understanding how human perception influences gaze estimation. One-time means that calibration is performed once before the experiment starts. In contrast, real-time methods entail continuous measurement and correction throughout the experiment, sometimes even calibration-free models[SB15] are implemented.

However, most researches concentrate on dealing with the objective factors by simulating environmental conditions, regardless of whether the calibration is employed one-time, or continuously. Nevertheless, they neglect the influence of objective human cognition on gaze estimation. Indeed, there is a limited amount of research in this domain, particularly the studies that integrate AOIs.

In practice, InfoVis are designed for information retrieval. Human attention is generally not attracted to blank areas of InfoVis unless influenced by daydreaming or calibration failure. Therefore, analyzing the AOIs within InfoVis proves to be a valuable endeavor.

This paper is grounded on the foundational hypothesis that human attention is naturally captured by the most informative segments of visualization, as proposed by Itti, Koch, and Niebur [IKN98]. Conversely, areas with low saliency are often unnoticed [MHD+17]. This research aims to explore the phenomenon by asking whether gaze estimation errors can be predicted using metrics that integrate AOI data through post-hoc analysis.

The metrics employed are commonly observed gaze patterns in eye tracking, including duration, scanpath, and scanpath length. The HAAR, which measures the proportion of fixations that are included by one or more AOI [WKB+22], is additionally incorporated as the input data. This metric serves as an indicator of the relationship between gaze data and AOIs, helping to identify acceptable and unacceptable calibration errors. The threshold used to classify calibration errors is $0.5°$ to measure the performance of the model with small calibration errors[FGK+22].

The methodology begins with processing the original data from the VisRecall++ Dataset[WJH+24], extracting the input features and output calibration errors. Then a detailed data analysis is performed to measure the relationships between AOI related metrics and calibration error, confirming that human gaze interests influence the final gaze estimation. Subsequently, a neural network (VisCaliNet) that combines regression and classification tasks to predict acceptable and unacceptable calibration errors is introduced. In the evaluation phase, ablated versions of VisCaliNet are implemented. The comparison among these models uses metrics such as accuracy, precision, recall, and F1 score.

The significance of the study is highlighted in several aspects, including the identification of data quality in open-source datasets, the exploration of the impact of human cognitive factors on gaze estimation, and the development of an efficient model.

The thesis is structured as follows:

Chapter 2 – Literature Review offers a summary of related works and establishes the theoretical foundation for the research.

Chapter 3 – Data Processing and Analysis processes the raw data, extracts relevant input features, and analyzes the relationships between gaze estimation and AOIs.

Chapter 4 – VisCaliNet describes the design intentions and the architecture of the VisCaliNet model and explains the functionality and structure of each component within the neural network in detail.

Chapter 5 – Experiment Results presents the outcomes of the study and evaluates the performance of different ablated models, providing elaborate analysis and interpretation of the results.

Chapter 6 – Discussion and Future Work discusses the significance of the findings, addresses the limitations of the current study, and outlines potential future research.

Chapter 7 – Conclusion concludes the thesis, summarize the findings and contributions.

2 Literature Review

This research is based on the following theoretical areas: 1)the formation and prediction of gaze estimation error, 2) the evaluation of AOIs within InfoVis through gaze data, and 3) the application of AOIs in predicting gaze estimation error. This chapter presents a review of the literature on these research domains.

2.1 Gaze Estimation Error

Gaze estimation refers to the scientific process of predicting the focal point of the human gaze. This process requires two cameras: the scene camera and the eye camera. The scene camera, usually located in the head-mount device, captures the Field of View (FOV) of the viewer, while the eye camera tracks the viewer's pupil movements [KPB14]. Gaze estimation aims to integrate the data from these two cameras to estimate an approximate location of the gaze point.

Gaze estimation error, in this context, denotes the intrinsic difference between the estimated gaze location and the actual gaze position [BDB16]. The gaze estimation error is typically measured from the perspective of accuracy and precision [HNA+11] of gaze direction and gaze point[PSMJ22]. Because of subjective and objective limitations, it is impossible to eliminate gaze estimation errors[EGIK19]. Therefore, the optimal strategy is to identify and understand the causes of these errors and apply compensatory measures to reduce them to the lowest extent.

Researchers deal with this problem from two perspectives: whether the data originates from geometrical or cognitive sources, and whether the calibration phase is active or passive. Geometrical sources of eye-tracking data include isocentric patterns [VG11], head pose[KKL+16] and 3D face structures [XHL16], while cognitive sources involve the similarity of human gaze patterns[AGVG17]and saliency map[SMS10][SMS12]. The calibration phase is considered active when conducting methods like nine-point calibration before the experiment[CS20][LSP19], whereas passive calibration uses comprehensive techniques to estimate calibration errors[AO14].

The most widely used and straightforward approach for simulating gaze estimation error is physiological data resources with active calibration. Drewes, Masson, and Montagnini [DMM12] studies how pupil size affects gaze estimation error and introduces compensational methods for pupil's dilation and constriction situations. Nyström et al. [NAHV13] dedicate to the influence of eye physiology of the viewer on the accuracy of eye-tracking data. Barz, Bulling, and Daiber [BBD15] concentrates on the gaze mapping algorithm, presenting a model that simulates the process of correlating the position of the pupil position with the coordinate system on the screen and predicts the gaze estimation error.

Some approaches use physiological data resources with passive calibration to predict and correct gaze estimation errors. These methods either calculate the disparity between the optical and visual axes as measured from the different camera angles [ME10a][ME10b] to infer gaze position and direction, or integrate human eye data with other sensory information, such as speech [SVS03] or head poses [SYFN13], for an extensive estimation.

Other approaches focus on the gaze patterns related to human attention. Nakano and Ishii [NI10] integrate gaze direction with conversational agents to ascertain whether the user is focused on the subject being discussed. Huang et al. [HKN+16] introduce a technique that uses user interaction behaviors for automatic calibration. Bulling, Weichel, and Gellersen [BWG13] passively monitor the user's eye movements and gathers environmental factors to determine the user's status. Currently, such studies are somewhat limited in number and mainly serve as supplementary methods for predicting gaze estimation errors.

2.2 Gaze-based AOIs evaluation

AOI refers to Area of Interest. It is typically demonstrated as a rectangle or polygon within a visualization[DD17], which often includes elements like titles, legends, or any area containing semantic information. Traditionally, AOIs were predefined by skilled data analysts. However, with the development of AI algorithms, this field has seen progress in automatic AOI detection. For instance, Fuhl et al. [FKSK18] try to automatically generate AOIs based on saliency and Fichtel et al. [FLP+19] attempt to dynamically identify AOIs from videos.

Gaze-based AOI evaluation combines gaze patterns with AOIs to understand users' gaze behavior. Blascheck et al. [BKR+14] provide a comprehensive summary of AOI-based visualization studies, classifying the methods of integrating gaze patterns with AOI. Additionally, Blascheck, Raschke, and Ertl [BRE13] use heatmaps to illustrate the transitions of gaze data between AOIs, intending to understand the shifts of user attention. Drusch, Bastien, and Paris [DBP14] have introduced an innovative approach for analyzing the dynamics of AOIs on web pages.

A growing number of researchers are focusing on the AOI in terms of their memorability. Borkin et al. [BVB+13] conduct an experiment containing 2070 visualizations, comparing their memorability to identify features of InfoVis that are easy to remember. Borkin et al. [BBK+15] analyze the eye movements of 33 participants and perform follow-up interviews, identifying the AOIs that attract the most attention and are most memorable. Similarly, Wang et al. [WJBB22]introduce VisRecall and VisRecall++[WJH+24], a model and advanced model designed to predict the recallability of various types of visualizations, such as bar charts, scatter plots, pie charts, and line graphs. These models use AOIs and correspondent transferred scanpaths to enhance understanding of how the type and layout of visualizations influence participants' memories.

AOIs also provide instructions for visualization design. Blascheck et al. [BKR+17] summarize the development of AOI-based visualizations, discussing their implications and applications. Orquin, Ashby, and Clarke [OAC16] investigate the problem of overlapping AOIs, and propose an optimal size for AOIs in visualization design to enhance clarity and usability. Additionally, Byelas and Telea [BT06] concentrate on the distribution of AOIs within the field of software architecture representation. Wang et al. [WHZ+18] introduce an algorithm capable of automatically selecting between line graphs and scatter plots for data visualization. Hu et al. [HBL+19] study further, they propose a model that can directly recommend visualizations with well-organized layouts of AOI according to the characteristics of the data.

2.3 AOIs in Gaze Estimation Error Prediction

As mentioned in the previous section, gaze estimation is primarily based on eye physiology, and some approaches integrate human cognitive factors such as attention [Bis+21][HZZ+22]. Few methods incorporate AOI into predicting gaze estimation errors.

However, some studies have paved the way for subsequent research in this area. For instance, Wang et al. [WKB+22] introduce the Hit Any AOI Rate(HAAR) and Flipping Candidate Rate (FCR) to examine gaze uncertainty from the perspective of AOIs. HAAR indicates the proportion of fixations that are included by at least one AOIs, while FCR refers to the likelihood of fixations landing on overlapping AOIs. Both metrics have proven effective in quantifying gaze uncertainty. In this research, HAAR will be used as an input feature in the neural network.

Other researchers do not specifically study the use of AOIs in predicting gaze estimation errors in InfoVis, but they employ similar concepts of human gaze attention, such as the saliency of images, to estimate gaze direction and points. Sugano, Matsushita, and Sato [SMS12] extract saliencies from various users and transform them into a saliency map. They allocate probabilities to the original visualization based on saliency distributions to estimate the gaze point. Valenti, Sebe, and Gevers [VSG12] develop a system that can adjust calibration errors by analyzing AOIs. It can improve the performance of both head pose and eye gaze trackers. Chang et al. [CMQ+19]introduce SalGaze, which focuses on personalized gaze estimation using saliency. These studies provide valuable experiences of using AOIs to estimate gaze points effectively.

3 Data Processing and Analysis

In this chapter, a detailed overview of the dataset will be provided, along with an explanation of the data collection methods used in the original dataset. Subsequently, data processing techniques such as ID map construction, fixation mapping, HAAR calculation, calibration error extraction, and data preparation for the neural network will be introduced to ensure the data is optimally prepared for modeling. Following these preparations, a comprehensive statistical analysis of essential metrics will be performed and the validation of the distribution across both the training and testing datasets will be examined.

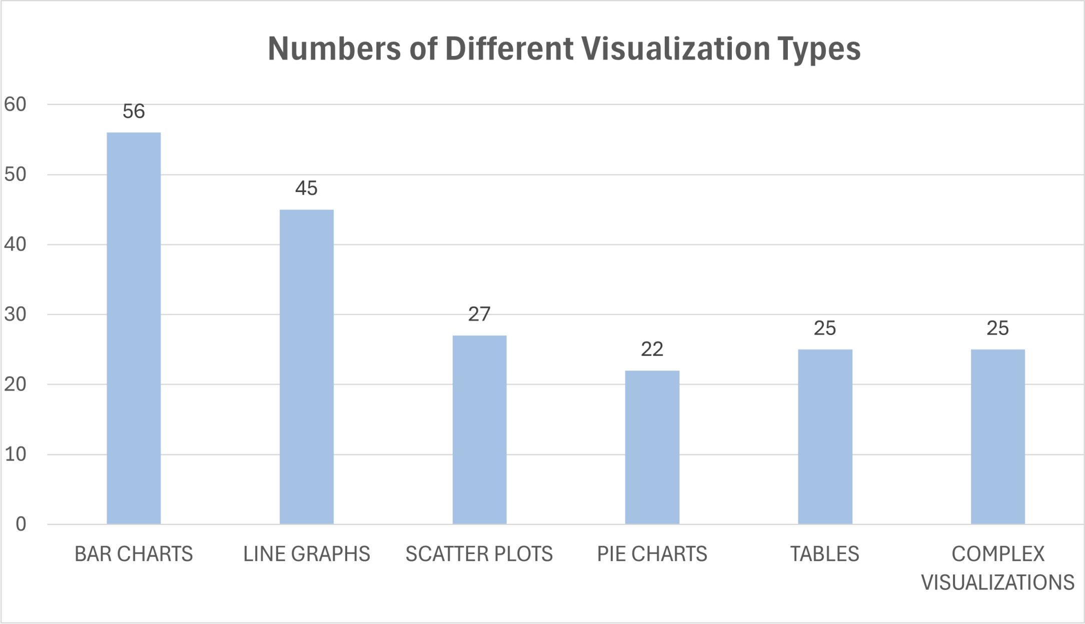


Figure 3.1: Numbers of Different Visualization Types

3.1 Dataset Overview

In this research, the VisRecall++ dataset [WJH+24] serves as the foundational dataset due to its inclusion of gaze patterns associated with AOI, which are crucial for the subsequent analysis and modeling. This dataset was collected from forty participants (15 females and 25 males) from a local university. Each participant was equipped with an Eyelink-1000 Plus eye tracker and tasked with viewing 200 visualizations.

Visualizations as Stimuli. To emphasize the task-oriented attributes of gaze patterns and enhance the impact of human attention on gaze estimation, the VisRecall++ dataset adopts various visualizations as stimuli. The dataset contains 200 information visualizations in total, including 56 bar charts, 45 line graphs, 27 scatter plots, 22 pie charts, 25 tables, and 25 other complex visualizations.

AOIs Identification. The Visrecall++ dataset includes AOIs that have been pre-identified by experienced data analysts for each visualization. These AOIs are divided into eleven categories: annotation, axis, graphical elements, legend, object, title, paragraph, source, other texts, data, and additional unspecified categories. This detailed categorization helps to analyze gaze patterns across different visual elements more effectively.

Eye Tracking Data Collection. The visualizations are randomly divided into 10 groups, each containing 20 visualizations. Participants are shown 2 to 6 groups of visualizations in total. The eye-tracking data collection begins with a calibration phase to ensure precise tracking of eye movements. Once calibrated, participants are assigned to view the visualizations while engaging in tasks specifically designed to provoke cognitive and visual responses. Throughout these sessions, fixation points and viewing durations are recorded, enabling the caption of detailed gaze patterns across all tasks and visualizations.

3.2 Data Processing

ID map construction. The raw fixation data includes the x and y coordinates and the duration of each fixation. Experienced data analysts have identified visualizations with AOIs, recording the coordinates of the AOI corners. Initially, a matrix termed "ID map" is constructed for each visualization. The size of the matrix is equal to the pixel dimensions of the image. The value within the matrix indicates the AOI to which that pixel belongs. For instance, a value of "`idmap[6][6] = 2`" indicates that the pixel at position (6,6) of the matrix is part of an AOI that is a graphical element.

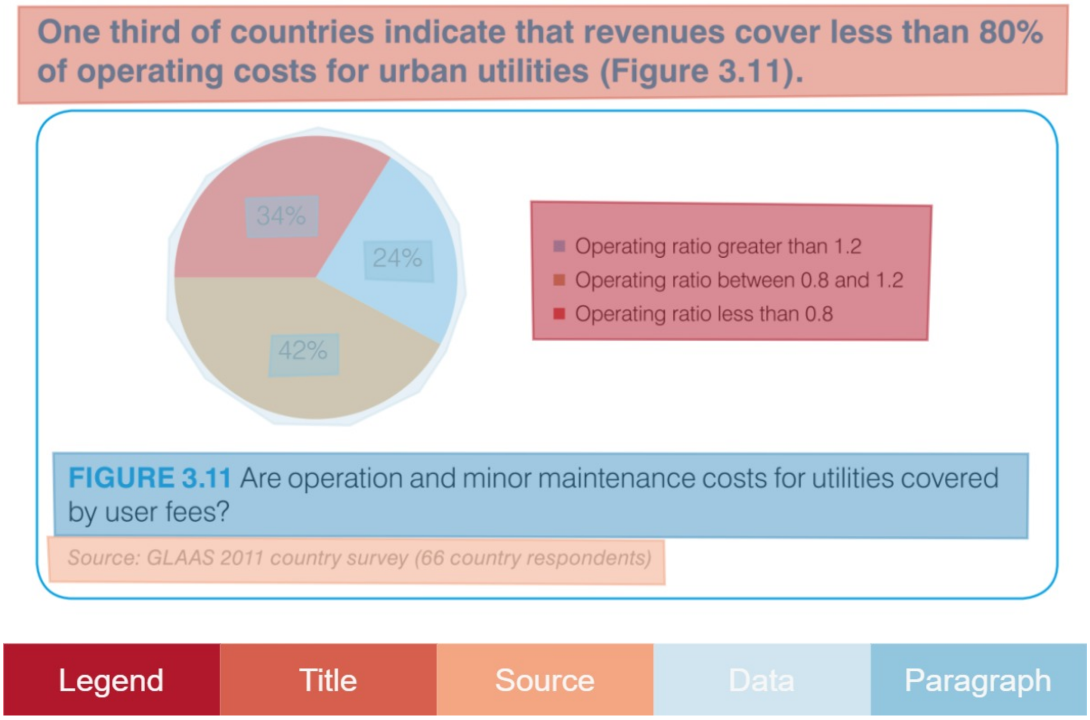


Figure 3.2: Example of a visualization annotated with AOIs which includes legend, title, source, data, and descriptive paragraph. Each AOI is marked with different colors.

Fixation Mapping. Using the coordinates of each fixation in the raw data, the fixations are mapped to the ID map, generating a sequence of numbers corresponding to the AOIs. These numbers are then converted to a sequential string that represents the participant's scanpath across the AOIs. For instance, the number 1 is converted to "X", symbolizing "axis", the number 2 to "G", representing "graphical elements", and the number 6 to "P" for "paragraph". Each token within the string indicates an AOI, with the total length of the string equal to the scanpath length.

Calculating HAAR. HIT[1] and OFF[2] are initially extracted from the scanpaths string. Since "Z" represented the blank areas within the visualization, counting the occurrence of "Z" in the scanpath string is necessary to calculate the final OFF. The HIT value corresponds to the total length of the string minus the OFF value. Following the formula provided by Wang et al. [WKB+22], the HAAR equals HIT divided by OFF plus HIT, therefore the final HAAR value is calculated.

[1] the count of fixations detected within at least one AOI[WKB+22]

[2] the count of fixations located in the white spaces [WKB+22]

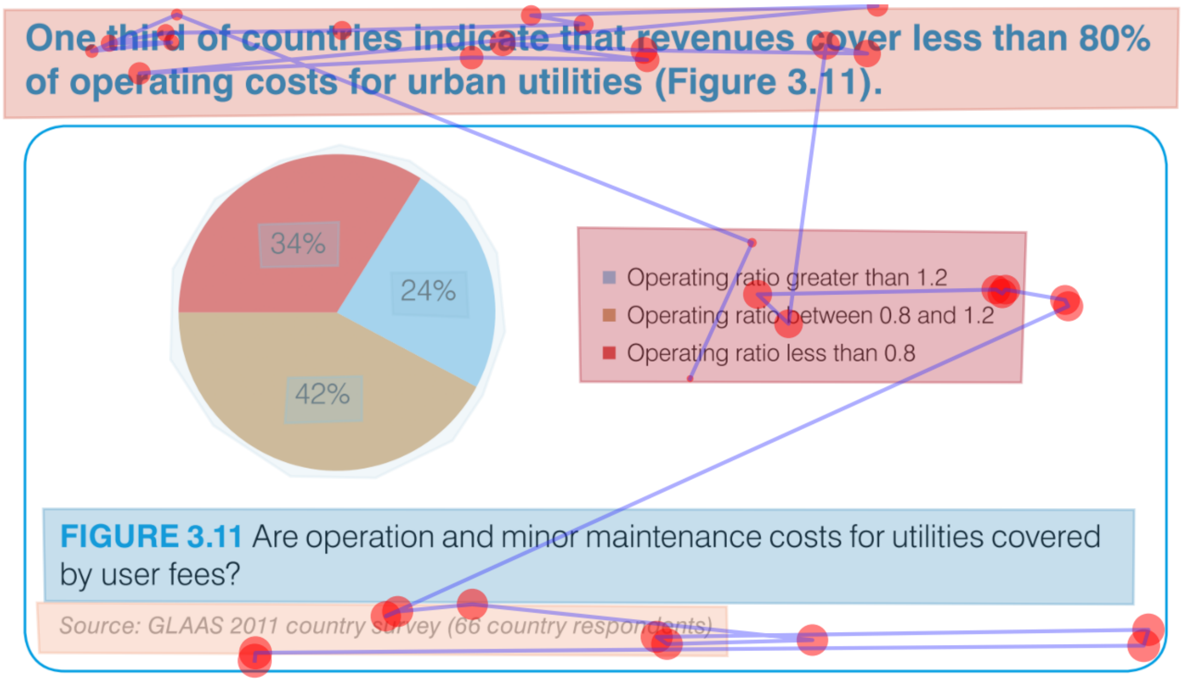


Figure 3.3: An example of information visualization marked with scanpath and duration. The purple lines represent the scanpaths: the fixation begins at the legend, moves up to the title, then returns to the legend, and finally shifts to the paragraph and source. The data in this situation is unnoticed. The corresponding transformed scanpath string was "LLTTTTTTTTZTTTTTTTLL-LLLZZSSZZSSZZSZ". In this visualization, the duration of each fixation was represented with red circles, with larger circles indicating a longer duration.

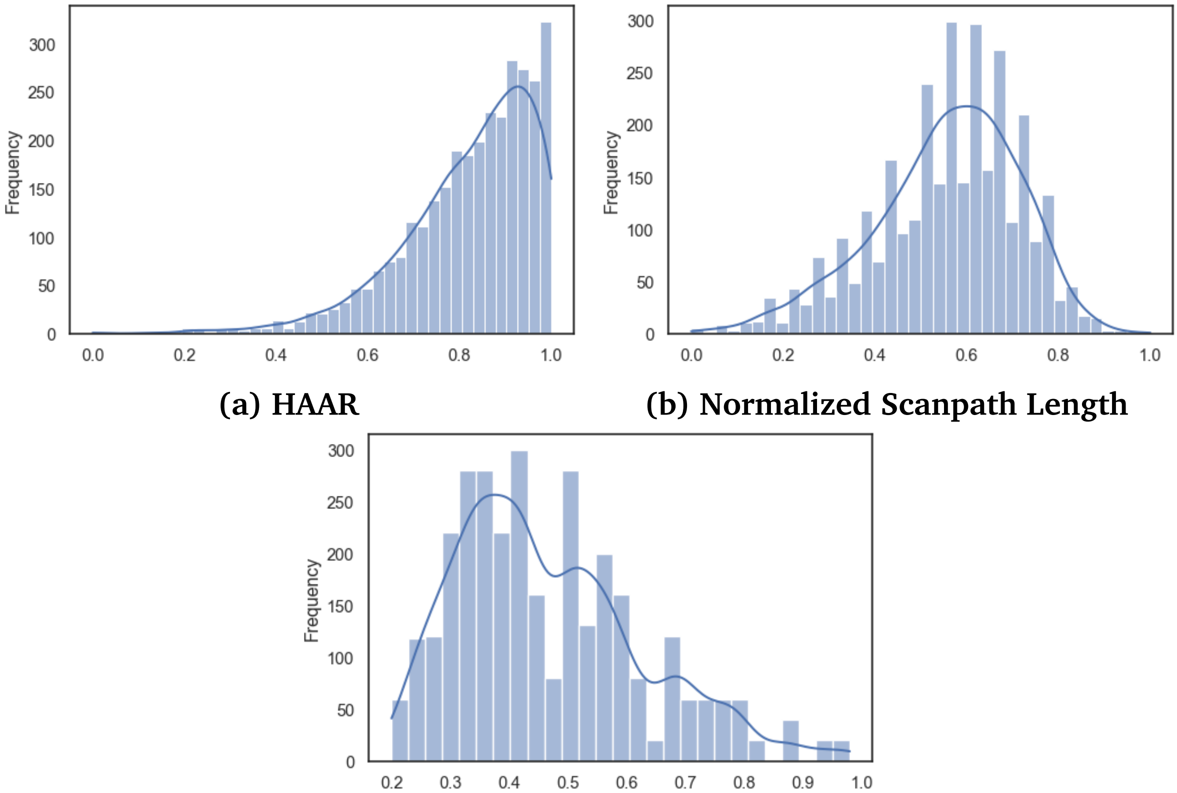
$$HAAR = \frac{HIT}{OFF + HIT}$$

Equation 3.2:Equation of Calculating HAAR [WKB+22]

***Extracting Calibration Error*.** Calibration error data is stored within the ".asc" files, which contain all eye-tracking data for each participant. Typically, each participant took calibration once, viewing multiple calibration points. From this data, the average calibration error for each eye is extracted, and the lower value from the left and right eyes is selected to represent the participant's ultimate calibration error. This method ensures that the data for the later analysis is accurate.

***Preparing Data for Neuron Network*.** Additional features for the neural network are then extracted, including the scanpath length and duration. These features, along with the scanpath string, HAAR, and calibration error, are mapped to a ".csv" file according to the participant ID and the name of the visualization. Given the different ranges of calibration error, HAAR, and scanpath length, the scanpath lengths are normalized to a range from 0 to 1. Calibration error and HAAR are not normalized because their values are within the range of 0 to 1. Additionally, the duration for each participant on each visualization is stored as a list of numbers. To perform matrix operations in subsequent networks, the longest list of durations is identified and the length of all other lists is resized to the length 57 through zero padding method[ASR+22].

3.3 Data Analysis



(a) HAAR

(b) Normalized Scanpath Length

(C) Calibration Error

Figure 3.4: Distribution of HAAR, normalized scanpath length, and calibration error

HAAR and calibration error. In the VisRecall++ dataset, calibration errors range from 0.200 to 0.980, with an average of 0.468 and a variance of 0.025. This distribution shows a left-skewed tendency, indicating that most data points are concentrated at the

	Ground Truth	Train Dataset	Test Dataset
Mean	0.468	0.468	0.468
Median	0.43	0.43	0.43
Variance	0.0255	0.0257	0.0249

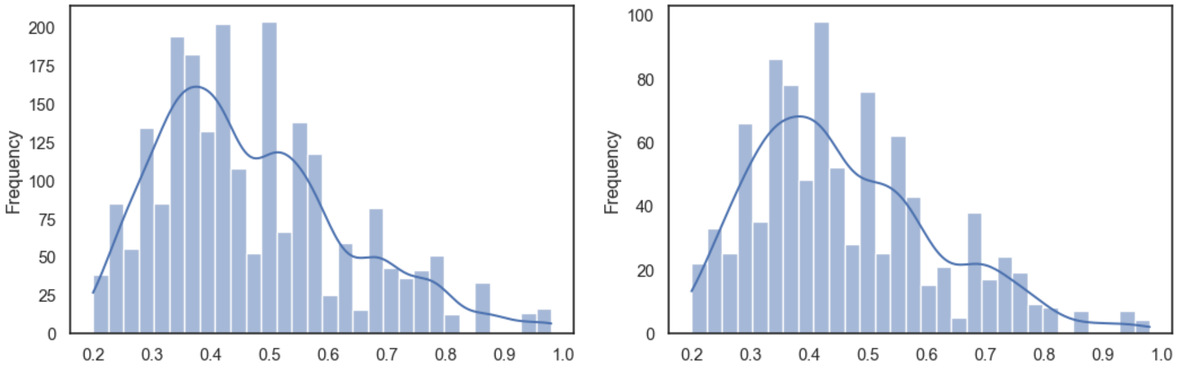
Table 3.1: Mean, Median, Variance of the train and test dataset

lower end of the range. Conversely, the HAAR exhibit a right-skewed distribution with a mean of 0.823, suggesting that most fixations occur within the AOIs. This observation aligns with the initial hypothesis that human attention is drawn to the most informative parts of a visualization Itti, Koch, and Niebur [IKN98]. The minimum HAAR value is 0.000, potentially indicating the instances of gaze estimation failure. The variance of HAAR is 0.021, showing consistency in distribution comparable to that of calibration errors.

Scanpath Length. Additionally, the normalized scanpath length has a right-skewed distribution, with a mean of 0.556 and a variance of 0.025, identical to that of the calibration error. Notably, there are two instances where the minimum normalized scanpath length is 0. The zero instances correspond to a value of 1 when transforming back to unnormalized scanpath length. These instances have a duration of 190 and 136 seconds respectively. The scanpath string related to these two instances is labeled "Z", referring to non-informative areas. The observation suggests potential calibration failures, as extended durations in such areas are unusual for actively engaged users.

Duration. As shown in Figure 3.3, the size of each red circle visually represents the fixation duration, with larger circles indicating longer fixation durations. Empirical observations reveal that the total duration of fixations within AOIs significantly exceeds that in non-informative blank spaces. This phenomenon illustrates that the engagement of participants is led to areas with information, highlighting the effectiveness of visual stimuli in guiding the attention of viewers.

Training and Testing Dataset Statistics. The entire dataset is split into training and testing subsets at proportions of 80% and 20%. To ensure the reproducibility of the results, a manual seed of 1 is set in this research. The mean and variance of the training data are 0.468 and 0.0257, while for the test data, these figures are 0.468 and 0.0249, both closely identical to the statistics of the overall dataset. Notably, the median for the training data, testing data, and the entire dataset is consistently 0.43, indicating a left-skewed distribution across all subsets.



(a) Training Dataset

(b) Testing Dataset

Figure 3.5: Distribution of Calibration Errors for the Training and Testing Datasets

4 VisCaliNet

Based on the data analysis from the earlier chapter, it can be concluded that both HAAR and scanpath length have a measurable relationship with calibration error according to their statistical metrics and distributions. Although the effects of duration and scanpath are not observable because of their complex format representations, their impacts are nonetheless apparent. This chapter aims to explore these relationships further using deep learning methods. A model named VisCaliNet will be introduced for predicting gaze estimation error. The model takes inspiration from the RecallNet [WJBB22] and incorporates several important features associated with human visual attention and cognitive responses.

4.1 Input Features

Four key inputs are used in the training process: scanpath, HAAR, duration, and scanpath length. The scanpath records the path of fixation moving across various AOIs, illustrating the shifts in the participant's focus. HAAR measures the interaction between the participants and the AOIs, providing deep understanding of attention distribution. Duration evaluates how long a viewer looks at different areas, with the longer value indicating greater interest. In addition, scanpath length countes the total fixations of a participant on one visualization, revealing the participant's engagement on the task. To ensure the normalized scanpath lengths accord with the left-skewed distribution of the calibration errors, the transformation formula `sqlnorm = 1 - sqlnorm` is applied to the scanpath length.

As shown in the earlier chapter, the scanpath in this study is transformed into a sequence of string tokens. To efficiently process the strings, the Bidirectional Encoder Representations from Transformers (BERT) model[LT18] is adopted due to its excellent performance in understanding context and distinguishing ambiguous expressions. And the BERT model had the functionality of generating embeddings from text, allowing for detailed and sophisticated interpretations of the scanpaths. To enhance efficiency and clarity during processing, separating tokens `[SEP]` is inserted between each token within the scanpath strings. This method guarantees that each character is recognized as an individual AOI, and the entire string is treated as a cohesive scanpath sequence, rather than as a single word. The tokenized scanpaths are then fed into the pre-trained BERT model, which converts the tokens into a 768-dimensional vector.

4.2 Model Design

VisCaliNet innovatively combines regression and classification techniques to address the complex task of predicting calibration errors from eye-tracking data. Initially, the model employs a regression approach to accurately predict these errors, providing a preliminary overview of the model's performance in estimating calibration errors. By analyzing the outcomes of the regression model, the tendencies and accuracy of the model are evaluated, giving instructions for subsequent studies.

Following the regression phase, the model classifies the outputs as "acceptable" and "unacceptable" based on predefined criteria. The threshold of $0.5°$ is adopted, indicating a good performance even with small calibration errors[FGK+22]. If a significant deviation in the estimated calibration error is observed during the experiment, recalibration is needed. In addition, this step assures the reliability of the data, especially when the proportion of unacceptable calibration errors is high, which might suggest that the dataset could be untrustworthy.

4.3 Model Architecture

The VisCaliNet model is based on MLP framework because of its efficiency and advancement in dealing with regression tasks[JHBB23]. For the input features HAAR and scanpath length, represented as single numerical values, the model configuration includes an input channel of size 1, a hidden channel of 5, and an output channel of 1. Furthermore, Rectified Linear Unit (ReLU) layers are intersected to add non-linearity and enhance model performance.

Duration data, represented as uniform vectors with size 57, is initially processed through a linear layer that reduces its dimension to 32. It is followed by a ReLU activation layer to include non-linear dynamics, then another linear layer is implemented to decrease the output to 10 channels. A subsequent ReLU layer is applied, followed by a final linear layer, ultimately reducing the output to a single channel.

The scanpath data in this study is transformed into BERT embeddings, requiring a more complex architecture. To process these embeddings, MLP is applied. The neuron network begins with a linear layer that takes an input size of 768 and decreases the output to the

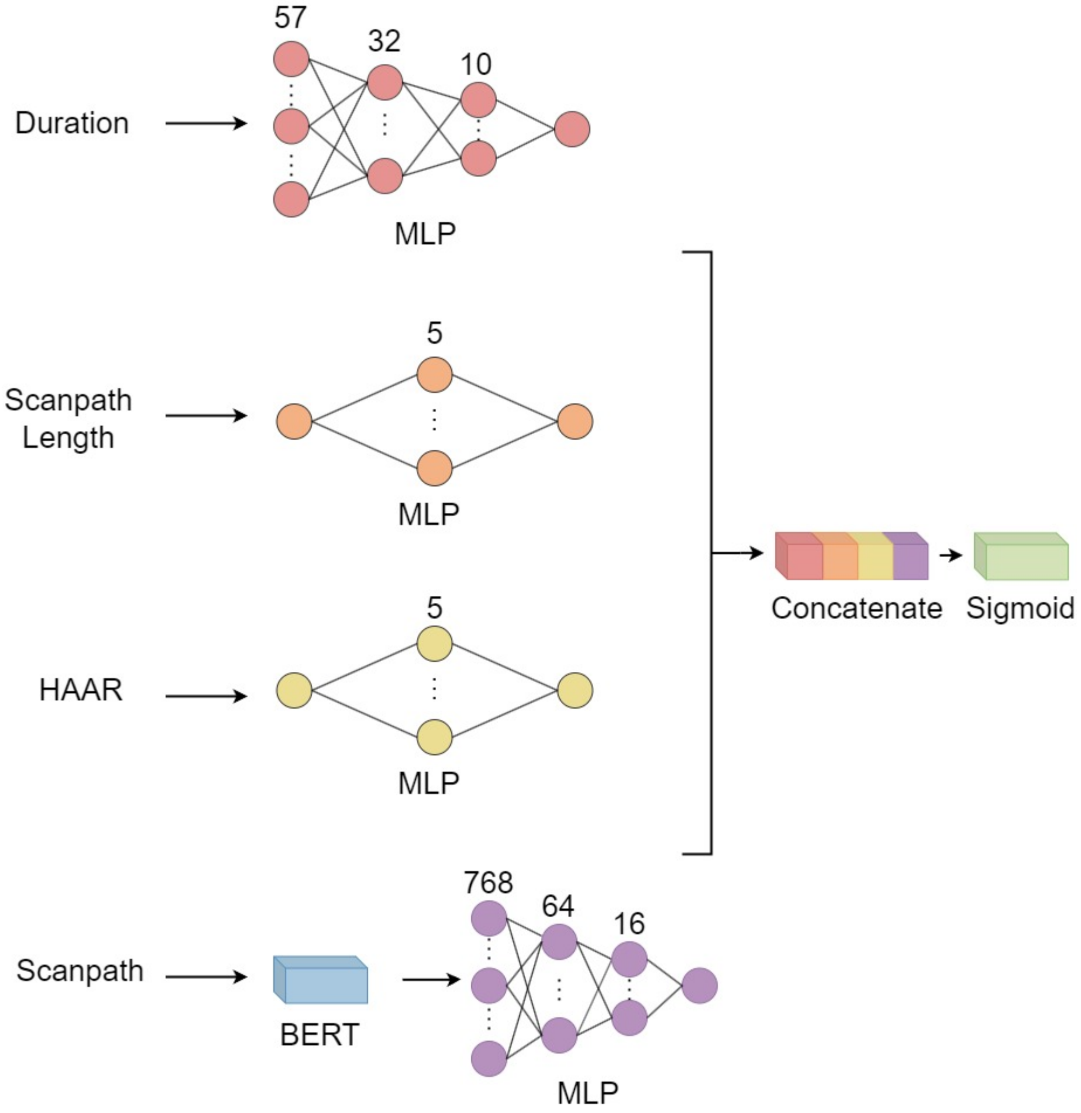


Figure 4.1: Architecture of the VisCaliNet: duration, scanpath length, and haar are fed into separate mlp. String-encoded scanpath is transformed into 768-dimensional embeddings using a pre-trained BERT model, then the results are passed into an Multilayer Perceptron (MLP). The outputs of each MLP are concatenated into a tensor of size 4 and are processed through a sigmoid function to generate the final binary output of 1 and 0.

size of 64. A subsequent linear layer further compresses the BERT embeddings down to a size of 16, and the final layer outputs a single value. ReLU activation layers are inserted between each linear layer to introduce non-linearity and enhance the learning capabilities of the model.

After training each feature independently, the four input features are concatenated into a single vector. The unified vector is then processed through a linear layer to predict the exact calibration error. The final adjustment in this model is the classification of the output from the previous networks, enabling effective handling of binary classifications. The VisCaliNet model has been transformed to handle a classification task using a sigmoid layer. Initially, calibration errors are categorized into binary classes(1 for acceptable and 0 for unacceptable) based on the threshold of 0.5°. Subsequently, a sigmoid function is applied to the categorized outputs to facilitate the classification process.

5 Experiment Results

In this chapter, numerous experiments will be conducted to evaluate the performance of VisCaliNet on the VisRecall++ dataset. Additionally, different ablated versions of VisCaliNet are implemented and evaluated to understand the impact of different model components on overall performance.

5.1 Ablated Models

Three ablated versions based on VisCaliNet are applied: MLPBERT, CNNBERT, and RNNBERT. Each was initially designed for regression tasks and later adapted to classify the output into corresponding classes based on predefined criteria.

MLPBERT. The MLPBERT model shares the same architecture as VisCaliNet, with one significant alteration: instead of employing a sigmoid function for classification, direct classification method is used, where outputs less than or equal to $0.5°$ are directly classified as 1. Outputs that exceed the threshold are classified as 0. This approach simplifies the classification process.

CNNBERT. Since Convolutional Neural Network (CNN)[LLY+21] are known for their effectiveness at processing context-aware features, one-dimensional convolutional layers are included in the training approach. The architecture of CNNBERT is identical to that of the MLPBERT model for training the duration, HAAR, and scanpath length. But for scanpath strings, which are transformed into BERT embeddings, three convolutional layers are introduced. The whole structure is designed with an input channel of 768, a hidden channel of 64, and an output channel of 1. Each convolutional layer is followed by a Batch Normalization layer, and a ReLU is included for adding non-linearity.

RNNBERT. The basic architecture of RNNBERT is adapted from CNNBERT, but with a key modification: instead of using MLP layers to process BERT embeddings, three Recurrent Neural Networks (RNN) layers are used due to their proficiency in handling natural language inputs[SM19]. The first RNN layer receives an input of size 768 and produces an output of size 64, followed by a ReLU activation layer to introduce non-linearity. The second RNN layer takes 64-sized input and reduces it to a 16-sized output, again followed by a ReLU activation layer to maintain non-linear processing capabilities. The final RNN layer decreases the final output to one single channel.

5.2 Training Details

In the testing experiments, a dataset comprising 3,169 data points is employed, each characterized by four input features and one output. The models implemented include MLPBERT, CNNBERT, and RNNBERT for regression tasks, and VisCaliNet for classification. Each model is executed with specific configurations:

- **Regression Models:** The MLPBERT adopts Mean Squared Error (MSE) as the loss function and Stochastic Gradient Descent (SGD) for optimization, and is trained for 300 epochs. In contrast, the CNNBERT and RNNBERT models use MSE as the loss function and are optimized using the Adaptive Moment Estimation (Adam) algorithm. Additionally, CNNBERT is trained for 1000 epochs, and RNNBERT is trained for 600 epochs.
- **Classification Model:** For binary classification, the VisCaliNet uses Binary Cross-Entropy (BCE) as the loss function, with Adam remaining as the optimizer. The classification model is trained for 100 epochs to balance an optimized performance and prevent overfitting.

The dataset is trained with batches with a batch size of 32. All models are trained with a consistent learning rate of 5E-6 and a weight decay of 1E-6. The experiments are conducted on a single NVIDIA GeForce RTX 3070 Ti GPU.

5.3 Evaluation

The evaluation is executed by comparing various metrics. During the model training phase, the training loss and testing loss are observed to confirm stable and effective performance. For the evaluation of classification results, the precision-recall curve is used to identify the most effective threshold for classification. Then, metrics like accuracy, precision, recall, and F1 score are used to assess the models' classification performance.

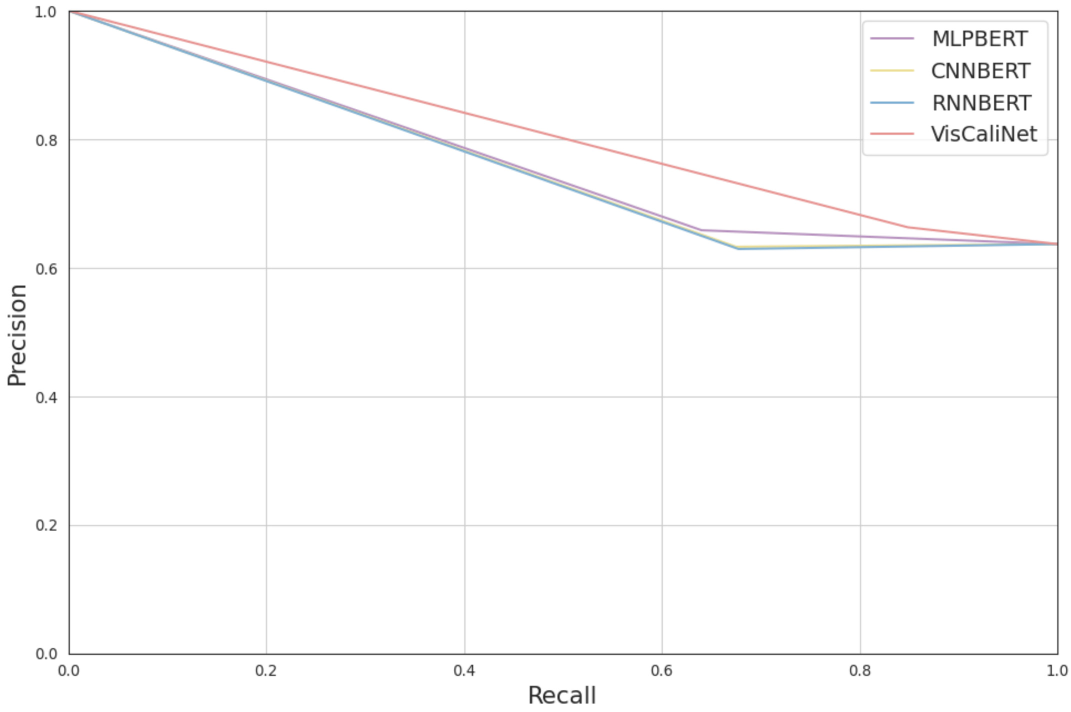


Figure 5.1: Precision-Recall Curve for all ablated models

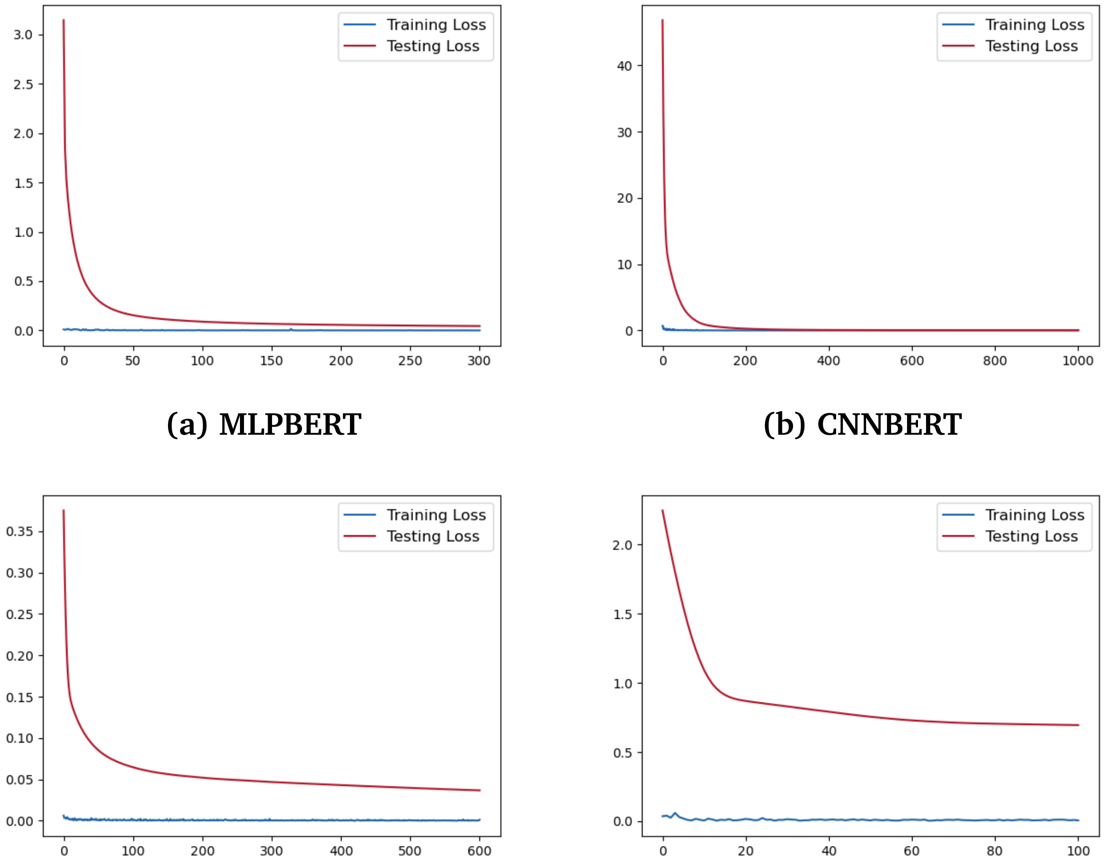
Precision-Recall Curve

The precision-recall curve is utilized to present the relation between recall and precision under different thresholds. The most significant threshold is then chosen based on the curve's results. Figure 5.1 shows that the overall precision and recall range from 0.6 to 1.0. As recall increases, precision correspondingly decreases. Notably, VisCaliNet demonstrated the highest precision compared to other models. MLPBERT performs second best, while RNNBERT and CNNBERT show the lowest precision overall.

A relatively higher precision [FWT+17] is then chosen for the VisCaliNet model to ensure that when calibration error is predicted as acceptable, it is indeed correct. The specific values are shown in Table 5.2. This method is important for maintaining the overall precision of the eye-tracking data. Whenever unacceptable calibration data is identified, recalibration should be conducted.

Metrics

Test Loss. As illustrated in Figure 5.2, the training loss for all regression models remains consistently low, with a significant reduction in test loss observed within the first 100 epochs. Beyond this point, the test loss stabilizes, consistently remaining at a relatively low level. The final recorded losses for MLPBERT, CNNBERT, and RNNBERT models are 0.029, 0.037, and 0.029. The loss of VisCaliNet is higher compared to other models because it is specifically designed for a classification task. However, it stabilizes after 50 epochs. The final loss is 0.695.



(a) MLPBERT

(b) CNNBERT

(C) RNNBERT

(d) VisCaliNet

Figure 5.2: Changes in training and testing loss with the increase of epochs

Accuracy. Accuracy represents the proportion of correct predictions the models make. VisCaliNet achieves the highest accuracy, recording a score of 0.618. Close behind, MLPBERT scores 0.607, while RNNBERT and CNNBERT have values of 0.582 and 0.571, respectively, indicating a noticeable spread in performance with a difference of approximately 0.057 between the highest and the lowest scores.

Precision. Precision is an indicator showing the correctness of positive predictions. VisCaliNet and MLPBERT gain the best performance, scoring a precision of 0.663 and

	Accuracy	Precision	Recall	F1 Score
MLPBERT	0.607	0.660	0.789	0.719
CNNBERT	0.571	0.655	0.691	0.672
RNNBERT	0.582	0.644	0.770	0.701
VisCaliNet	0.628	0.663	0.847	0.744

Table 5.1: Accuracy, Precision, Recall, and F1 Score of Different Models

0.660, closely followed by CNNBERT at 0.655. RNNBERT showes a slightly lower precision of 0.644.

Recall. The difference in recall among the models is significant, with values ranging from 0.691 to 0.847. VisCaliNet has the highest recall at 0.847, showing its superior performance, while CNNBERT records the lowest at 0.691. MLPBERT and RNNBERT have scores of recall of 0.789 and 0.770 respectively.

F1 Score. VisCaliNet significantly surpasses other models with an F1 score of 0.744, exhibiting a good performance when combining precision and recall. In contrast, due to its weaker recall performance, CNNBERT has the lowest F1 score at only 0.672. Meanwhile, MLPBERT and RNNBERT obtain 0.719 and 0.701 respectively.

6 Discussion and Future Work

Among all the ablated versions, VisCaliNet demonstrates the best performance and successfully meets the initial goals of this research. This chapter explores the insights gained from the evaluation phase, highlights the significance of the VisCaliNet, and discusses its limitations. At the end of the chapter, the potential future work built on the foundation laid by the current research will be outlined.

Discussion

Adding more input features has proven to be beneficial. Initially, only the linear relationship between HAAR and calibration error was considered, which was directly related to the AOIs. However, this approach resulted in less ideal performance. Subsequently, the model was enhanced by incorporating the scanpath, transformed into BERT embeddings, and durations, listed as an array, which led to acceptable outcomes. The further incorporation of scanpath length resulted in a modest performance improvement, highlighting the benefits of expanding the feature set.

Regarding network architectures, MLPs is excellent in processing numerical values, making them highly suitable for predicting gaze estimation error. Conversely, while CNNs are typically advantageous for computer vision tasks, they does not show the same level of performance in the ablated application. In some instances, a RNN architecture may serve as a feasible alternative, offering flexibility and efficiency where needed.

In terms of classification techniques, incorporating a sigmoid classification layer in the VisCaliNet lead to a slight improvement compared to the direct classification approach used in the standard MLPBERT, which suggests that sigmoid functions may be more effective at handling binary classification tasks in specific scenarios.

Significance

Exploring Human Cognitive Factors on Gaze Estimation. As mentioned above, the potential of human gaze information in predicting gaze estimation errors is notably insufficiently investigated. This research endeavors to enrich the academic conversation about how cognitive factors impact gaze estimation, filling the gap in this field.

Model Efficiency and Simplicity. VisCaliNet requires a minimal set of input features and maintains a low level of overall complexity. This simplicity minimizes the consumption of computing resources, and enhances reproducibility and user experience, enabling its application across various scenarios. Moreover, the model relies on established parameters without monitoring environmental factors, thereby improving its simplicity.

Identifying Calibration Process in Open-Source Datasets. In the current era of big data, a vast collection of open-source eye-tracking datasets is available for in-depth analysis. These datasets provide essential information on gaze behaviors to the public. However, they often lack details on the calibration error, an important metric for measuring the accuracy of gaze estimation. This absence leaves the validity of the calibration process suspicious, casting doubt on the quality of the gaze data collection. This research offers a solution that utilizes existing gaze data to evaluate the reliability of the calibration process, hence improving the credibility of further studies using open-source eye-tracking datasets.

Limitations

Range of Calibration Error. This study utilizes preprocessed and cleaned calibration errors within the range from 0 to 1. However, in real-world scenarios, particularly during online data collections where rigorous calibration processes are lacking, the calibration errors can exceed this range. This limitation may restrict the applicability of VisCaliNet to settings where calibration is more controlled.

Dataset Composition and Size. The dataset currently includes approximately 3,000 data points, all from university students. While this sample size is adequate for this research, it does not fully represent the variety of the demographic groups. It is both important and feasible to consider expanding the dataset to other groups, improving the generalizability of the results.

Input Features. This model primarily relies on scanpath-related data to represent human attention and cognition. Althought the approach appears effective, incorporating additional metrics may enhance its capabilities. For instance, the FCR, as introduced by Wang et al. [WKB+22], effectively evaluates gaze uncertainty by indicating the likelihood of fixations landing on overlapping AOIs. Furthermore, integrating saliency as another metric may significantly improve the model's ability to predict gaze estimation error.

Future Work

The primary focus of future work will be to handle the limitations identified in the current research. This includes adjusting the model to accommodate calibration errors beyond the current range of 0 to 1, enabling it to deal with a wider scope of real-life scenarios. Moreover, incorporating more complex metrics such as the FCR and saliency will perhaps provide a more accurate result in gaze estimation using human attention and cognition.

Based on the enhanced model, the next phase will involve developing an auto-correction model. This model will not only predict calibration errors but also actively correct them in real-time. It will involve developing algorithms capable of detecting the root causes of calibration errors and creating a standard correction formula that adjusts gaze estimation closer to the actual gaze position. Implementing a system that automatically applies these corrections will enhance the overall accuracy and utility of the device.

Once the improvements are implemented, the final step will be to develop an interface that integrates the auto-correction model. This interface will utilize the initially trained data as the foundational model. During user interactions, with the user's consent, new data will also be collected. The data will then be fed into the model, allowing it to integrate the new information and make adaptations accordingly. Over time, this iterative process will refine the model's accuracy and efficiency, continuously improving its performance by involving the latest data inputs.

7 Conclusion

In this research, VisCaliNet is developed. It is a model based on deep learning algorithms using MLP as the primary network component, designed to predict gaze estimation errors. Through detailed data analysis, this study discovers the relationship between AOI related gaze patterns and gaze estimation errors. Innovatively, AOI-based gaze metrics and the (HAAR) are concatenated as input features, with calibration error as the output. Multifaceted evaluations are conducted to measure the performance of the model, confirming its efficiency from different perspectives. Additionally, ablation studies are performed to compare the impact of minor modifications on the overall performance of the VisCaliNet model. The result showed that this study demonstrates strong performance in identifying acceptable and unacceptable calibration errors using the criteria of $0.5°$, offering a practical solution for situations where calibration is difficult.

Bibliography

[AGVG17] F. Alnajar, T. Gevers, R. Valenti, S. Ghebreab. “Auto-calibrated gaze estimation using human gaze patterns.” In: *International Journal of Computer Vision* 124 (2017), pp. 223–236 (cit. on p. 13).

[AO14] K. Alberto Funes Mora, J.-M. Odobez. “Geometric generative gaze estimation (g3e) for remote rgb-d cameras.” In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 1773–1780 (cit. on p. 13).

[ASR+22] I. A. Ahmed, E. M. Senan, T. H. Rassem, M. A. Ali, H. S. A. Shatnawi, S. M. Alwazer, M. Alshahrani. “Eye tracking-based diagnosis and early detection of autism spectrum disorder using machine learning and deep learning techniques.” In: *Electronics* 11.4 (2022), p. 530 (cit. on p. 21).

[BBD15] M. Barz, A. Bulling, F. Daiber. “Computational modelling and prediction of gaze estimation error for head-mounted eye trackers.” In: *DFKI ResearchReports* 1.1 (2015), pp. 1–10 (cit. on p. 14).

[BBK+15] M. A. Borkin, Z. Bylinskii, N. W. Kim, C. M. Bainbridge, C. S. Yeh, D. Borkin, H. Pfister, A. Oliva. “Beyond memorability: Visualization recognition and recall.” In: *IEEE transactions on visualization and computer graphics* 22.1 (2015), pp. 519–528 (cit. on p. 15).

[BDB16] M. Barz, F. Daiber, A. Bulling. “Prediction of gaze estimation error for error-aware gaze-based interfaces.” In: *Proceedings of the ninth biennial acm symposium on eye tracking research & applications*. 2016, pp. 275–278 (cit. on pp. 9, 13).

[Bis+21] P. Biswas et al. “Appearance-based gaze estimation using attention and difference mechanism.” In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 3143–3152 (cit. on p. 15).

[BKR+14] T. Blascheck, K. Kurzhals, M. Raschke, M. Burch, D. Weiskopf, T. Ertl. “State-of-the-art of visualization for eye tracking data.” In: *Eurovis (stars)*. 2014, p. 29 (cit. on p. 14).

[BKR+17] T. Blascheck, K. Kurzhals, M. Raschke, M. Burch, D. Weiskopf, T. Ertl. "Visualization of eye tracking data: A taxonomy and survey." In: *Computer Graphics Forum*. Vol. 36. 8. Wiley Online Library. 2017, pp. 260–284 (cit. on p. 15).

[BRE13] T. Blascheck, M. Raschke, T. Ertl. "Circular heat map transition diagram." In: *Proceedings of the 2013 Conference on Eye Tracking South Africa*. 2013, pp. 58–61 (cit. on p. 14).

[BT06] H. Byelas, A. Telea. "Visualization of areas of interest in software architecture diagrams." In: *Proceedings of the 2006 ACM symposium on Software visualization*. 2006, pp. 105–114 (cit. on p. 15).

[Bus35] G. T. Buswell. "How people look at pictures: a study of the psychology and perception in art." In: (1935) (cit. on p. 9).

[BVB+13] M. A. Borkin, A. A. Vo, Z. Bylinskii, P. Isola, S. Sunkavalli, A. Oliva, H. Pfister. "What makes a visualization memorable?" In: *IEEE transactions on visualization and computer graphics* 19.12 (2013), pp. 2306–2315 (cit. on p. 15).

[BWG13] A. Bulling, C. Weichel, H. Gellersen. "EyeContext: Recognition of high-level contextual cues from human visual behaviour." In: *Proceedings of the sigchi conference on human factors in computing systems*. 2013, pp. 305–308 (cit. on p. 14).

[CMQ+19] Z. Chang, J. Matias Di Martino, Q. Qiu, S. Espinosa, G. Sapiro. "Salgaze: Personalizing gaze estimation using visual saliency." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2019, pp. 0–0 (cit. on p. 16).

[CS20] Z. Chen, B. Shi. "Offset calibration for appearance-based gaze estimation via gaze decomposition." In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2020, pp. 270–279 (cit. on p. 13).

[DBP14] G. Drusch, J. Bastien, S. Paris. "Analysing eye-tracking data: From scanpaths and heatmaps to the dynamic visualisation of areas of interest." In: *Advances in science, technology, higher education and society in the conceptual age: STHESCA* 20.205 (2014), p. 25 (cit. on p. 15).

[DD17] A. T. Duchowski, A. T. Duchowski. *Eye tracking methodology: Theory and practice*. Springer, 2017 (cit. on p. 14).

[DMM12] J. Drewes, G. S. Masson, A. Montagnini. "Shifts in reported gaze position due to changes in pupil size: Ground truth and compensation." In: *Proceedings of the symposium on eye tracking research and applications*. 2012, pp. 209–212 (cit. on p. 14).

[EGIK19] B. V. Ehinger, K. Groß, I. Ibs, P. König. “A new comprehensive eye-tracking test battery concurrently evaluating the Pupil Labs glasses and the EyeLink 1000.” In: *PeerJ* 7 (2019), e7086 (cit. on pp. 9, 13).

[FGK+22] Y. Feng, N. Goulding-Hotta, A. Khan, H. Reyserhove, Y. Zhu. “Real-time gaze tracking with event-driven eye segmentation.” In: *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE. 2022, pp. 399–408 (cit. on pp. 10, 26).

[FKSK18] W. Fuhl, T. C. Kübler, T. Santini, E. Kasneci. “Automatic Generation of Saliency-based Areas of Interest for the Visualization and Analysis of Eye-tracking Data.” In: *VMV*. 2018, pp. 47–54 (cit. on p. 14).

[FLP+19] E. Fichtel, N. Lau, J. Park, S. Henrickson Parker, S. Ponnala, S. Fitzgibbons, S. D. Safford. “Eye tracking in surgical education: gaze-based dynamic area of interest can discriminate adverse events and expertise.” In: *Surgical endoscopy* 33 (2019), pp. 2249–2256 (cit. on p. 14).

[FWT+17] A. M. Feit, S. Williams, A. Toledo, A. Paradiso, H. Kulkarni, S. Kane, M. R. Morris. “Toward everyday gaze input: Accuracy and precision of eye tracking and implications for design.” In: *Proceedings of the 2017 Chi conference on human factors in computing systems*. 2017, pp. 1118–1130 (cit. on pp. 9, 31).

[GH11] J. Goldberg, J. Helfman. “Eye tracking for visualization evaluation: Reading values on linear versus radial graphs.” In: *Information visualization* 10.3 (2011), pp. 182–195 (cit. on p. 9).

[HBL+19] K. Hu, M. A. Bakker, S. Li, T. Kraska, C. Hidalgo. “Vizml: A machine learning approach to visualization recommendation.” In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pp. 1–12 (cit. on p. 15).

[HKN+16] M. X. Huang, T. C. Kwok, G. Ngai, S. C. Chan, H. V. Leong. “Building a personalized, auto-calibrating eye tracker from user interactions.” In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2016, pp. 5169–5179 (cit. on p. 14).

[HNA+11] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, J. Van de Weijer. *Eye tracking: A comprehensive guide to methods and measures*. oup Oxford, 2011 (cit. on pp. 9, 13).

[HZZ+22] Z. Hu, K. Zhao, B. Zhou, H. Guo, S. Wu, Y. Yang, J. Liu. “Gaze target estimation inspired by interactive attention.” In: *IEEE Transactions on Circuits and Systems for Video Technology* 32.12 (2022), pp. 8524–8536 (cit. on p. 15).

[IKN98] L. Itti, C. Koch, E. Niebur. “A model of saliency-based visual attention for rapid scene analysis.” In: *IEEE Transactions on pattern analysis and machine intelligence* 20.11 (1998), pp. 1254–1259 (cit. on pp. 10, 22).

[JHBB23] C. Jiao, Z. Hu, M. Bâce, A. Bulling. “SUPREYES: SUPer Resolutin for EYES Using Implicit Neural Representation Learning.” In: *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 2023, pp. 1–13 (cit. on p. 26).

[JK03] R. J. Jacob, K. S. Karn. “Eye tracking in human-computer interaction and usability research: Ready to deliver the promises.” In: *The mind’s eye*. Elsevier, 2003, pp. 573–605 (cit. on p. 9).

[KKL+16] H.-I. Kim, J.-B. Kim, J.-E. Lee, T.-Y. Lee, R.-H. Park. “Gaze estimation using a webcam for region of interest detection.” In: *Signal, Image and Video Processing* 10 (2016), pp. 895–902 (cit. on p. 13).

[KPB14] M. Kassner, W. Patera, A. Bulling. “Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction.” In: *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: Adjunct publication*. 2014, pp. 1151–1160 (cit. on p. 13).

[LLY+21] Z. Li, F. Liu, W. Yang, S. Peng, J. Zhou. “A survey of convolutional neural networks: analysis, applications, and prospects.” In: *IEEE transactions on neural networks and learning systems* 33.12 (2021), pp. 6999–7019 (cit. on p. 29).

[LSP19] E. Lindén, J. Sjostrand, A. Proutiere. “Learning to personalize in appearance-based gaze tracking.” In: *Proceedings of the IEEE/CVF international conference on computer vision workshops*. 2019, pp. 0–0 (cit. on p. 13).

[LT18] J. Lee, K. Toutanova. “Pre-training of deep bidirectional transformers for language understanding.” In: *arXiv preprint arXiv:1810.04805* 3 (2018), p. 8 (cit. on p. 25).

[MB14] P. Majaranta, A. Bulling. “Eye tracking and eye-based human–computer interaction.” In: *Advances in physiological computing*. Springer, 2014, pp. 39–65 (cit. on p. 9).

[ME10a] D. Model, M. Eizenman. “An automatic personal calibration procedure for advanced gaze estimation systems.” In: *IEEE Transactions on Biomedical Engineering* 57.5 (2010), pp. 1031–1039 (cit. on p. 14).

[ME10b] D. Model, M. Eizenman. “User-calibration-free remote gaze estimation system.” In: *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*. 2010, pp. 29–36 (cit. on p. 14).

[MHD+17] L. E. Matzen, M. J. Haass, K. M. Divis, Z. Wang, A. T. Wilson. “Data visualization saliency model: A tool for evaluating abstract data visualizations.” In: *IEEE transactions on visualization and computer graphics* 24.1 (2017), pp. 563–573 (cit. on p. 10).

[NAHV13] M. Nyström, R. Andersson, K. Holmqvist, J. Van De Weijer. “The influence of calibration method and eye physiology on eyetracking data quality.” In: *Behavior research methods* 45 (2013), pp. 272–288 (cit. on p. 14).

[NI10] Y. I. Nakano, R. Ishii. “Estimating user’s engagement from eye-gaze behaviors in human-agent conversations.” In: *Proceedings of the 15th international conference on Intelligent user interfaces*. 2010, pp. 139–148 (cit. on p. 14).

[OAC16] J. L. Orquin, N. J. Ashby, A. D. Clarke. “Areas of interest as a signal detection problem in behavioral eye-tracking research.” In: *Journal of Behavioral Decision Making* 29.2-3 (2016), pp. 103–115 (cit. on p. 15).

[PSMJ22] P. Pathirana, S. Senarath, D. Meedeniya, S. Jayarathna. “Eye gaze estimation: A survey on deep learning-based approaches.” In: *Expert Systems with Applications* 199 (2022), p. 116894. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2022.116894. URL: https://www.sciencedirect.com/science/article/pii/S0957417422003347 (cit. on p. 13).

[SB15] Y. Sugano, A. Bulling. “Self-calibrating head-mounted eye trackers using egocentric visual saliency.” In: *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. 2015, pp. 363–372 (cit. on p. 9).

[SM19] R. C. Staudemeyer, E. R. Morris. “Understanding LSTM–a tutorial into long short-term memory recurrent neural networks.” In: *arXiv preprint arXiv:1909.09586* (2019) (cit. on p. 29).

[SMS10] Y. Sugano, Y. Matsushita, Y. Sato. “Calibration-free gaze sensing using saliency maps.” In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE. 2010, pp. 2667–2674 (cit. on p. 13).

[SMS12] Y. Sugano, Y. Matsushita, Y. Sato. “Appearance-based gaze estimation using visual saliency.” In: *IEEE transactions on pattern analysis and machine intelligence* 35.2 (2012), pp. 329–341 (cit. on pp. 13, 16).

[SVS03] J. S. Shell, R. Vertegaal, A. W. Skaburskis. “EyePliances: attention-seeking devices that respond to visual attention.” In: *CHI’03 extended abstracts on Human factors in computing systems*. 2003, pp. 770–771 (cit. on p. 14).

[SYFN13] B. A. Smith, Q. Yin, S. K. Feiner, S. K. Nayar. “Gaze locking: passive eye contact detection for human-object interaction.” In: *Proceedings of the 26th annual ACM symposium on User interface software and technology*. 2013, pp. 271–280 (cit. on p. 14).

[VG11] R. Valenti, T. Gevers. “Accurate eye center location through invariant isocentric patterns.” In: *IEEE transactions on pattern analysis and machine intelligence* 34.9 (2011), pp. 1785–1798 (cit. on p. 13).

[VSG12] R. Valenti, N. Sebe, T. Gevers. “What are you looking at? Improving visual gaze estimation by saliency.” In: *International journal of computer vision* 98 (2012), pp. 324–334 (cit. on p. 16).

[WHZ+18] Y. Wang, F. Han, L. Zhu, O. Deussen, B. Chen. “Line Graph or Scatter Plot? Automatic Selection of Methods for Visualizing Trends in Time Series.” In: *IEEE Transactions on Visualization and Computer Graphics* 24.2 (2018), pp. 1141–1154. DOI: 10.1109/TVCG.2017.2653106 (cit. on p. 15).

[WJBB22] Y. Wang, C. Jiao, M. Bâce, A. Bulling. “VisRecall: Quantifying information visualisation recallability via question answering.” In: *IEEE Transactions on Visualization and Computer Graphics* 28.12 (2022), pp. 4995–5005 (cit. on pp. 15, 25).

[WJH+24] Y. WANG, Y. JIANG, Z. HU, C. RUHDORFER, M. BÂCE, A. BULLING. “VisRecall++: Analysing and Predicting Visualisation Recallability from Gaze Behaviour.” In: (2024) (cit. on pp. 10, 15, 18).

[WKB+22] Y. Wang, M. Koch, M. Bâce, D. Weiskopf, A. Bulling. “Impact of gaze uncertainty on aois in information visualisations.” In: *2022 symposium on eye tracking research and applications*. 2022, pp. 1–6 (cit. on pp. 10, 15, 19, 20, 37).

[WZZ+21] X. Wang, J. Zhang, H. Zhang, S. Zhao, H. Liu. “Vision-based gaze estimation: a review.” In: *IEEE Transactions on Cognitive and Developmental Systems* 14.2 (2021), pp. 316–332 (cit. on p. 9).

[XHL16] C. Xiong, L. Huang, C. Liu. “Remote gaze estimation based on 3D face structure and iris centers under natural light.” In: *Multimedia Tools and Applications* 75 (2016), pp. 11785–11799 (cit. on p. 13).

All links were last followed on May 6, 2024.

Declaration

I hereby declare that the work presented in this thesis is entirely my own and that I did not use any other sources and references than the listed ones. I have marked all direct or indirect statements from other sources contained therein as quotations. Neither this work nor significant parts of it were part of another examination procedure. I have not published this work in whole or in part before. The electronic copy is consistent with all submitted copies.

place, date, signature