



Unlocking the head: unleashing deep learning and depth camera for free head movement gaze estimation

Zihao Wang^{1,2}, Jiangtao Wang^{1,2}, Anna Ching Mei Ng³, Tit Ng³, Wei Qian^{1,2}, Mbazingwa Elirehema⁴, Patrice Monkam^{1,2} , Shouliang Qi^{1,2} 

Keywords:

Gaze estimation, eye movement, deep learning, depth camera, free head movement

Citation:

Wang, Z.; Wang, J.; Ng, A. C. M.; Ng, T.; Qian, W.; Elirehema, M.; Monkam, P.; Qi, S. Unlocking the head: unleashing deep learning and depth camera for free head movement gaze estimation. *Intell. Robot.* 2026, 6(3), 621-42. <https://dx.doi.org/10.20517/ir.2026.29>

Received: 17 Apr 2026

First Decision: 8 Jul 2026

Revised: 28 Jul 2026

Accepted: 1 Sep 2026

Published: 23 Sep 2026

Academic Editor:

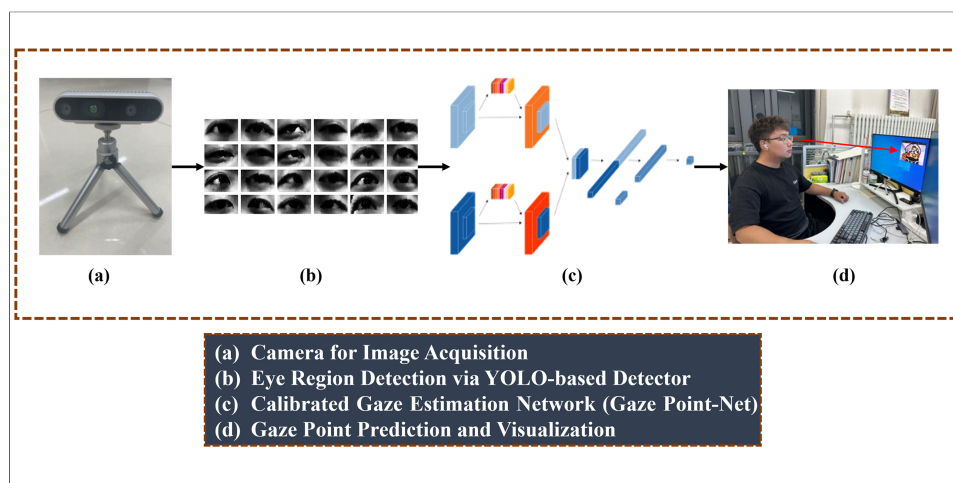
Jinhai Li

Copy Editor:

Pei-Yun Wang

Production Editor:

Pei-Yun Wang



Abstract

Gaze estimation has applications such as visual attention analysis and human-computer interaction. However, performance under free head movement conditions can be further enhanced. This study aims to improve the accuracy and robustness of appearance-based gaze estimation without head fixation. We develop a gaze-tracking system using a consumer depth camera and a deep-learning-based gaze estimation model [Gaze-Point-Net (GPN)]. A customized YOLOv5-based detector is developed to simultaneously localize the eyes and mouth, providing reliable facial landmarks for gaze estimation. The detected regions and depth image were used to calculate a vector of location and posture. In GPN, a double-channel convolutional neural network with a Squeeze-and-Excitation module extracts features from binocular images, and these features are concatenated with the location-and-posture vector to complete the multimodal gaze-position prediction task. Our GPN presents competitive performance in four experiments: (1) The error distribution in the sequential point test ranges from 4 to 10 degrees; (2) In the random point test, the



¹College of Medicine and Biological Information Engineering, Northeastern University, Shenyang 110819, Liaoning, China.

²Key Laboratory of Intelligent Computing in Medical Image, Ministry of Education, Northeastern University, Shenyang 110169, Liaoning, China.

³Shenzhen Jingmei Health Technology Co., Ltd., Shenzhen 518000, Guangdong, China.

⁴Department of Electronics and Telecommunication Engineering, Dar es Salaam Institute of Technology, Dar es Salaam 2958, Tanzania.

Correspondence to: Dr. Patrice Monkam, Prof. Shouliang Qi, College of Medicine and Biological Information Engineering, Northeastern University, Shenyang 110819, Liaoning, China. E-mail: patrice123china1@gmail.com; qisl@bmie.neu.edu.cn

calibrated and filtered GPN achieved a pixel error of 185.19 ± 116.57 pixels and an angular error of $4.52^\circ \pm 2.87^\circ$, demonstrating superior performance over counterpart methods; (3) In the trajectory tracking test, approximately 81.91% of gaze points were located within a 200-pixel tolerance radius of the target trajectory; (4) In the browsing test, the generated gaze heat and trajectory maps showed satisfactory results. The developed eye-tracking system, integrating a depth camera and deep learning models, demonstrates competitive performance and strong potential for several eye-tracking applications.

1. INTRODUCTION

Gaze estimation is an important research direction in computer vision. It can utilize devices such as color cameras and infrared equipment to capture features such as faces and eyes, and then calculate the position that the user is looking at on a screen or in a natural environment. This facilitates interaction between the user and the device or can provide data support for operations or tests by collecting physiological information from the user. Gaze estimation technology has a wide range of applications^[1,2], and can be used in education^[3,4], scientific research^[5,6], medicine^[7,8], entertainment^[9,10], automatic driving^[11,12], business^[13], sport^[14], and so on. As shown in Figure 1, gaze estimation results can be represented by a gaze-trajectory plot and a gaze heat map.

Gaze estimation systems primarily operate in the following three modes. (1) Head-mounted eye tracking. In this approach, gaze estimation is performed while wearing devices such as glasses or headsets. This method achieves high-precision gaze estimation, especially when equipped with gaze estimation hardware. As documented in^[15,16], this method has been refined. The advantage of this approach is that the device's relative position to the head remains relatively stable, eliminating the need to account for variations in head movement. It allows for a focus on eye image information alone, resulting in higher accuracy within appearance-based methods; (2) Desktop-based gaze estimation. This method employs sampling devices, such as cameras, to collect information and utilizes model-based or appearance-based techniques for gaze estimation. It is primarily used to predict a user's point of focus while using a computer or looking at a monitor^[17,18]. Notably, the camera's position is fixed relative to the monitor. It caters to scenarios where the user's head position remains relatively stable or changes slightly. Given the possibility of head movement and rotations, the method must account for these variables during gaze estimation; (3) Pose-based gaze estimation. This approach is often applied in long-range prediction scenarios, where the camera is positioned at a significant distance from the subject. It relies on the subject's overall body posture and head orientation to determine gaze direction and identify the object of interest. As shown in^[19], substantial progress has been made with this approach. This method involves numerous variables, and the task does not require extremely high precision. We used the second method, which builds a desktop line-of-sight tracking system.

There are currently three common methods for gaze estimation. (1) 2D mapping method. These approaches involve personalized calibration and utilize a mapping model to estimate the focal point corresponding to new eye features^[20,21]; (2) 3D model method. These methods employ eye structure and geometric imaging models to estimate the three-dimensional line of sight. It has higher requirements for cameras and lighting but offers greater flexibility in handling individual differences and head movements^[22,23]; (3) Appearance-based method. The appearance-based method uses eye or facial images as input and determines gaze position through the training of a mapping model^[24,25]. It is robust because it uses a large number of statistical samples, does not require specific eye features to be extracted, and requires less equipment. However, adapting it to individual activities remains challenging. It offers simplicity, efficiency, high accuracy, a small number of model parameters, and low computational requirements. Because of its minimal hardware requirements, it is suitable for deployment in scenarios involving lightweight imaging devices. Therefore, appearance-based methods have become a mainstream approach for gaze estimation.



Figure 1. Illustration of gaze estimation results, including the predicted gaze trajectory (left) and gaze distribution heat map (right). These images are used solely for visualization and illustrative purposes. They were obtained from the Kodak Lossless True Color Image Suite, originally released by Eastman Kodak Company (1999); the publicly accessible mirror at <https://rOk.us/graphics/kodak/> (accessed: 22.2.2026) was used for retrieval.

With advances in deep learning for computer vision, neural networks have shown state-of-the-art performance in gaze estimation. Appearance-based gaze estimation methods use mathematical models and take inputs such as binocular images to infer gaze direction. Consequently, neural network models in deep learning are widely applied in appearance-based gaze estimation methods. LeNet and VGG models are widely used in gaze estimation tasks^[2].

Appearance-based gaze estimation methods have been widely used on home computers^[2]. When the subject's head is significantly tilted, or farther away, the system may have difficulty accurately detecting eye position and posture, resulting in inaccurate or even failed gaze estimation. Appearance-based gaze estimation methods use traditional cameras for sampling, which is largely influenced by environmental lighting conditions. This article aims to address these issues to some extent.

In this study, we develop a desktop-based gaze estimation system capable of accurate gaze estimation under free head movement and varying lighting conditions. The main contributions are summarized as follows: (1) A desktop-based gaze estimation system is developed to support a wide range of gaze-related research through flexible experimental configurations while providing reliable gaze estimation; (2) A curated gaze estimation dataset is collected and compiled to support system development and comprehensive evaluation, with plans to make the dataset available following completion of the necessary data-sharing procedures and receipt of the required permissions; (3) Unlike conventional approaches that first estimate the gaze direction in the camera coordinate system and subsequently transform it into the world coordinate system to determine the screen gaze point^[26], the proposed system directly regresses the gaze position on the screen, simplifying the estimation pipeline; (4) Depth information is incorporated to estimate the user's spatial position and head pose, enabling robust gaze estimation under unrestricted head movement.

2. MATERIALS AND METHODS

2.1. Overview of the unlocking-the-head gaze estimation system

The procedure of our gaze estimation system is as follows [Figure 2A]: Firstly, we employ an improved You Only Look Once (YOLO) model^[27] to locate and crop eye images. We utilize an Intel RealSense depth camera^[28] to calculate the three-dimensional spatial positions of the eyes and mouth. Next, we utilize a multimodal gaze point network to predict screen gaze points. To accomplish this, we need to create and collect two datasets: one for training the YOLO object detection model and another for training the Gaze-Point-Net (GPN) gaze regression model. In our implementation, the resolution of the feature maps is increased in the detection head to improve the recognition of small-sized objects (eyes and mouth regions in the facial images). Moreover, the left-eye and right-eye classes are merged into a single “eye” class, since the two eyes exhibit highly similar appearance characteristics. The eyes and mouth are detected simultaneously during facial landmark detection. While the eye images are used in the subsequent gaze estimation process, the detected mouth position provides an additional facial landmark for characterizing the facial geometry. Combined with the binocular eye positions captured by the Intel RealSense depth camera, the three landmarks (left eye, right eye, and mouth) define a facial plane whose normal vector provides two angular degrees of freedom corresponding to head pitch and yaw in the camera coordinate system. The 3D spatial coordinates of the left and right eyes, together with these two angular components, constitute the 8-dimensional position/posture vector used in this study. Roll is not explicitly represented as an independent angular variable but is implicitly encoded in the relative 3D spatial configuration of the two eyes.

Figure 2B illustrates the training process of GPN. We collected training data from volunteers, consisting of standardized binocular images and corresponding three-dimensional information computed from image depth. These datasets were assembled and utilized to train a universal model.

We trained our model on a large dataset, and to further improve its fit to each user, we designed a calibration function [Figure 2C]. With this function, we can enhance the accuracy of the model’s predictions and deploy the same model on screens of different sizes to maintain consistent task accuracy when the camera position may not be stable.

2.2. Construction and training of the GPN model

Traditional gaze estimation tasks are usually performed under the condition of a fixed head, so traditional methods do not need to consider the head’s spatial position and pose information. Since the six degrees of freedom of the head are fixed, neural network models in traditional methods accept only binocular images or use a monocular model and compute binocular images through a mirror-flipping method.

To address the asymmetry between pupil direction and eyelid features in binocular images, we designed GPN based on LeNet, similar to existing studies^[2]. Since binocular images are different, the feature extractors cannot use the same parameters for feature extraction. Therefore, we adopted a double-channel convolutional neural network feature extractor with non-shared parameters to extract features from binocular images and used multiple fully connected layers to compress and extract features. Additionally, we connected (concatenated) head spatial position and pose vectors in the hidden layer to construct a multimodal neural network for completing the multimodal gaze position prediction task. Below is the network structure of GPN [Figure 3].

Our neural network architecture mainly consists of two parts: the SELayer^[29] and GAZE attention. The SELayer is a module used to enhance the neural network’s feature expression ability. It takes the input feature map and performs adaptive average pooling to obtain a global feature vector. After processing through two fully connected layers and an activation function, the SELayer generates a weight vector used to

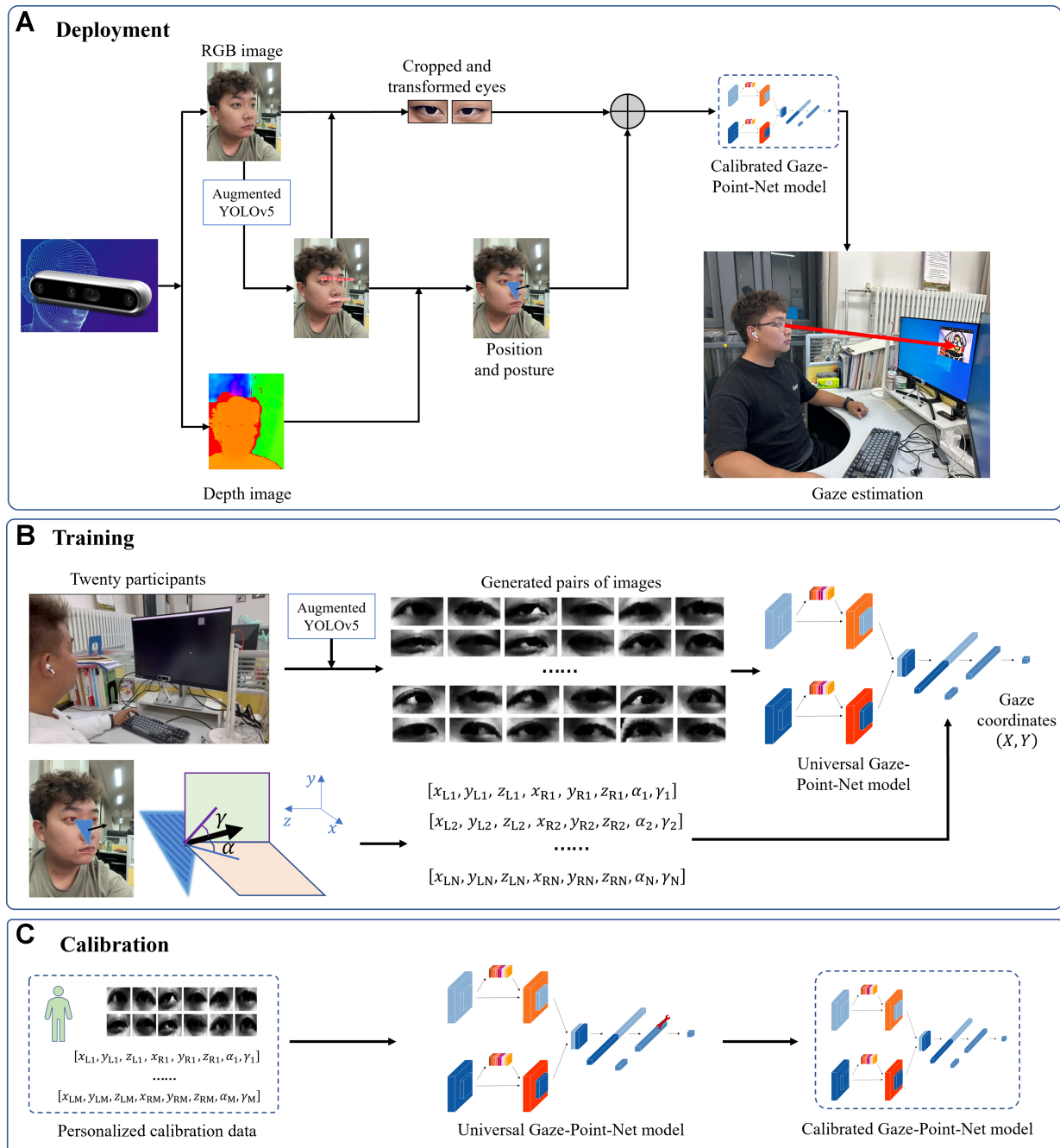


Figure 2. Overview of the unlocking-the-head gaze estimation system. (A) The deployment of the system; (B) The training of the universal GPN model; (C) The calibration of the GPN model. GPN: Gaze-Point-Net; RGB: Red, Green, Blue; YOLO: You Only Look Once.

scale the input feature map. The SELayer includes an adaptive average pooling layer, two fully connected layers, a rectified linear unit (ReLU) activation function to construct nonlinear combinations, and a Sigmoid activation function to limit the parameter range.

The GAZE attention is a neural network consisting of convolutional and fully connected layers. Its input includes two monochrome eye images (36×60 pixels each) and head position and pose information. The output is a 2D tensor representing the gaze point coordinates.

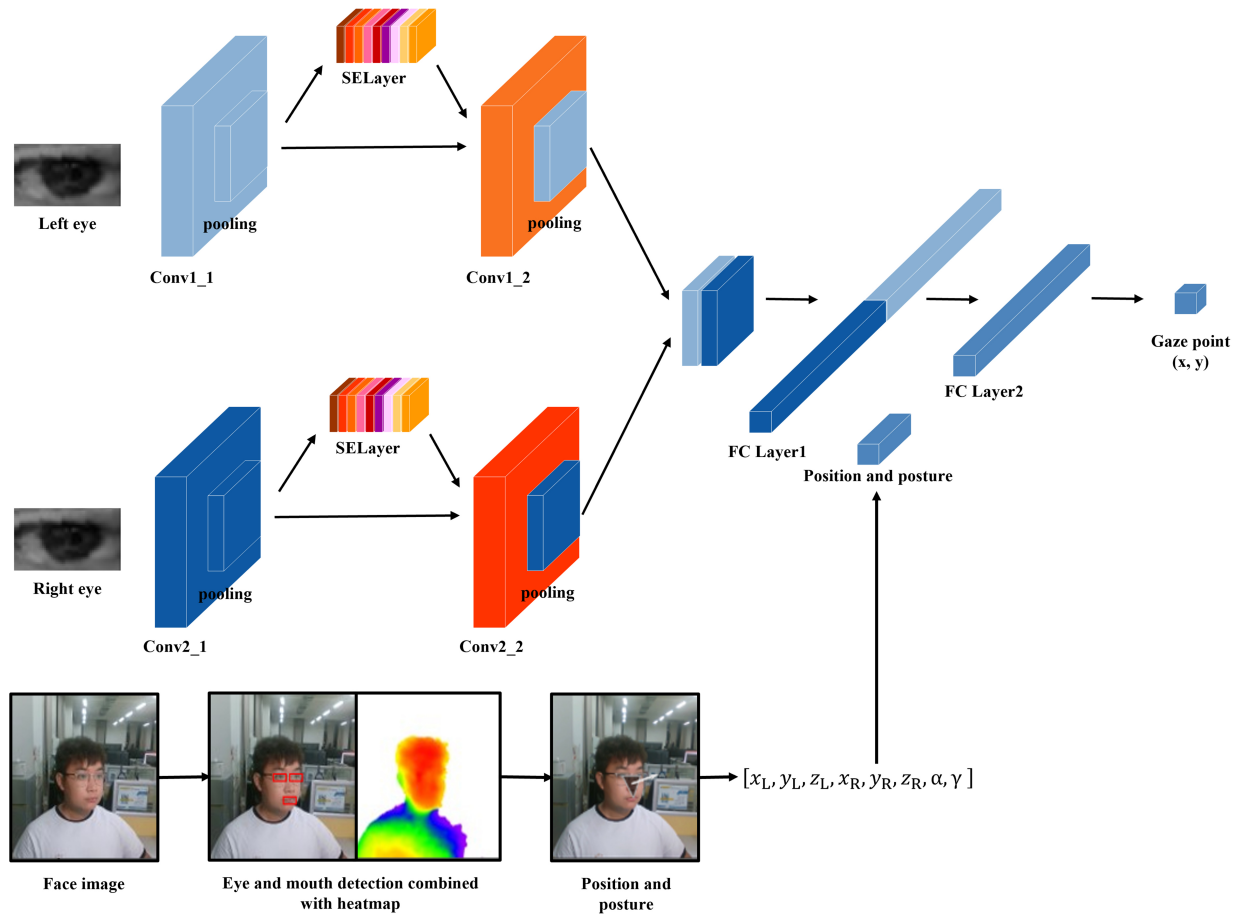


Figure 3. GPN structure. The 8-dimensional position/posture representation comprises the 3D spatial coordinates of the left and right eyes (6 dimensions) and two angular components of the facial-plane normal representing head pitch and yaw (2 dimensions). GPN: Gaze-Point-Net; FC: fully connected.

The convolutional layers of this neural network adopt a non-shared-parameter dual-channel convolutional neural network structure. Each channel contains two convolutional modules. The first convolutional module consists of 20 convolutional kernels with a size of 5×5 , a stride of 1, and a padding of 2 to keep the feature size the same. This is followed by a batch normalization layer, a ReLU activation function, and a 2×2 max-pooling layer. After the first pooling operation, the attention mechanism introduced earlier enhances the network's focus on the input data. This mechanism adaptively weights each channel, allowing the neural network to focus more on extracting important features while ignoring irrelevant ones. The second convolutional module consists of 50 convolutional kernels with a size of 5×5 , a stride of 1, and a padding of 2×2 to ensure that the feature size remains the same. This is followed by a batch normalization layer, a ReLU activation function, and a maximum pooling layer with a size of 2.

2.3. Training of the augmented GPN

During training, we found that the model underestimated the y-coordinate of the predicted gaze point. Therefore, we redesigned the loss function. Our data and labels are both represented as two-dimensional coordinates, denoted by $c_1 = (x_1, y_1)$ and $c_2 = (x_2, y_2)$, respectively, where c_1 and c_2 represent the predicted and ground-truth gaze points. Here, x_1 and y_1 denote the predicted horizontal and vertical coordinates, while x_2 and y_2 denote the corresponding ground-truth coordinates. The proposed loss function consists of the absolute errors of the x- and y-coordinates, the squared error of the y-coordinate, and the Euclidean distance between the predicted and ground-truth gaze points, weighted by 2:2:1:4, respectively [Equation

(1)]. These weighting coefficients were selected empirically through preliminary experiments to balance the contributions of the individual loss terms and promote stable model convergence. The squared y -coordinate error term further emphasizes deviations in the vertical coordinate, particularly at larger y -coordinate values.

$$L(c1, c2) = 2|x1 - x2| + 2|y1 - y2| + |y1 - y2|^2 + 4\|(x1, y1) - (x2, y2)\|^2 \quad (1)$$

A total of 20 participants were randomly divided at the participant level into training, validation, and test sets at an 8:1:1 ratio, comprising 16, 2, and 2 participants, respectively. All samples from each participant were assigned exclusively to the corresponding subset, ensuring that no participant appeared in more than one subset.

During training, we set the number of epochs to 50, the batch size to 16, and used the redesigned loss function as the objective function. To optimize the objective function, we used the stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.0001 and a dynamic learning rate strategy. Specifically, we gradually reduced the learning rate by a factor of 0.1 at epochs 20 and 30. After each training batch, we used the validation set to calculate the mean absolute error (MAE) and determine whether to update the best model.

2.4. Calibration and filtering process

The GPN model was trained using only the data from participants in the training subset, while data from the validation and test subsets were reserved for their respective evaluation purposes. Therefore, we can further improve the model's accuracy and performance under specific users' usage conditions through calibration. To this end, we designed the following calibration steps, following a similar approach to that described in^[30].

First, we collected calibration data using a method similar to that used in collecting training data. The calibration data were collected by presenting nine gray boxes arranged in two rectangles on the display screen in sequence. The collected data include standardized binocular images, head position and posture vectors, and corresponding label data. Next, the collected data were assembled into a batch of calibration datasets, and the model was fine-tuned using the same loss function and optimizer as in training. Importantly, model fine-tuning at this stage used exclusively the calibration data from volunteers included in the training set.

During model calibration, to prevent overfitting at the nine calibration points, we employed an approach similar to transfer learning by adjusting only the parameters of the last fully connected layer during fine-tuning. This method enables us to further improve the model's accuracy and performance under specific user conditions.

Moreover, a mean filter is applied to the calibrated gaze-point predictions. Specifically, the current gaze location is estimated as the mean of the five most recent predicted gaze locations. The filtered and unfiltered predictions are then compared to evaluate the effect of filtering on gaze estimation performance.

2.5. Experimental methods and measurements

2.5.1. Experiments

After the completion of the entire system's construction, we designed the following four experiments with reference to the method in^[31]: sequential and random point testing, trajectory tracking testing, and browsing mode analysis. These experiments comprehensively analyze the accuracy, robustness, and real-time performance of the system.

In the sequential point location test, we utilized our system design to track a gray box moving in a zigzag pattern from the upper left corner of the screen. Real-time gaze prediction was achieved through the use of a gaze estimation system. With participants' consent, we recorded binocular images, facial images, head position, head pose information, and normalized screen position information of the gray box for data analysis.

Similarly, in the random point location test, we used a gray block as a fixation target that jumped randomly between specific position and remained fixed for three seconds at each positions. To eliminate the influence of saccadic eye movements, we recorded data after the block had been stationary for one second.

The fixation task with a stationary target in these experiments closely resembled the primary task of current eye-tracking devices, allowing us to assess the eye-tracking system's performance accurately.

To evaluate the system's ability to track moving targets, we designed a circular trajectory tracking test. A circular target trajectory with a radius of 500 pixels was drawn at the center of the screen, and a moving box served as the gaze target. Gaze points were predicted and collected throughout the test. To quantify tracking accuracy, confidence radii of 300 and 200 pixels relative to the target trajectory were used as tolerance thresholds for determining whether the predicted gaze points were valid tracks.

Finally, we demonstrated the gaze estimation system using a browsing mode analysis task. We selected five images containing highly distinctive objects, each displayed in full screen for five seconds. Simultaneously, we captured the sequence of gaze positions of the viewers and visualized them as gaze point trajectories and heat maps to analyze and present the performance of our model. By combining the captured gaze points with the positions of the target objects in the images, we analyzed the reliability, real-time capability, and accuracy of our model.

2.5.2. Measurement

After computing the normalized gaze coordinates, they are mapped back to the $1,920 \times 1,080$ screen coordinate system to obtain the final projected gaze location, referred to as the fixation point (yellow arrow in Figure 4). Simultaneously, the corresponding ground-truth location is retained as the centroid of the target square displayed during the experiment (orange arrow in Figure 4). The predicted and ground-truth screen coordinates are then used to calculate the Euclidean distance between the two points on the screen.

To further evaluate gaze estimation accuracy in three-dimensional space, the spatial coordinates of the eyes in the camera coordinate system, together with the known geometric relationship between the camera and the display, are used to determine the distances from the midpoint of the two eye rays to both the predicted and ground-truth fixation points. Based on these distances, the cosine theorem is applied to compute the angular error between the predicted and ground-truth gaze directions (red angle in Figure 4). The resulting angular error is subsequently analyzed under different experimental conditions.

It is important to note that this study was approved by the Ethics Committee of Northeastern University (approval no. NEU-EC-2024B036S). All procedures were conducted in accordance with institutional guidelines and the Declaration of Helsinki (2024). Written informed consent was obtained from all participants prior to data collection. The experimental setup consisted of a $1,920 \times 1,080$ display and the Intel RealSense depth camera mounted directly below the screen. Data were collected from a single participant per session at a viewing distance ranging from 0.5 to 0.85 m.

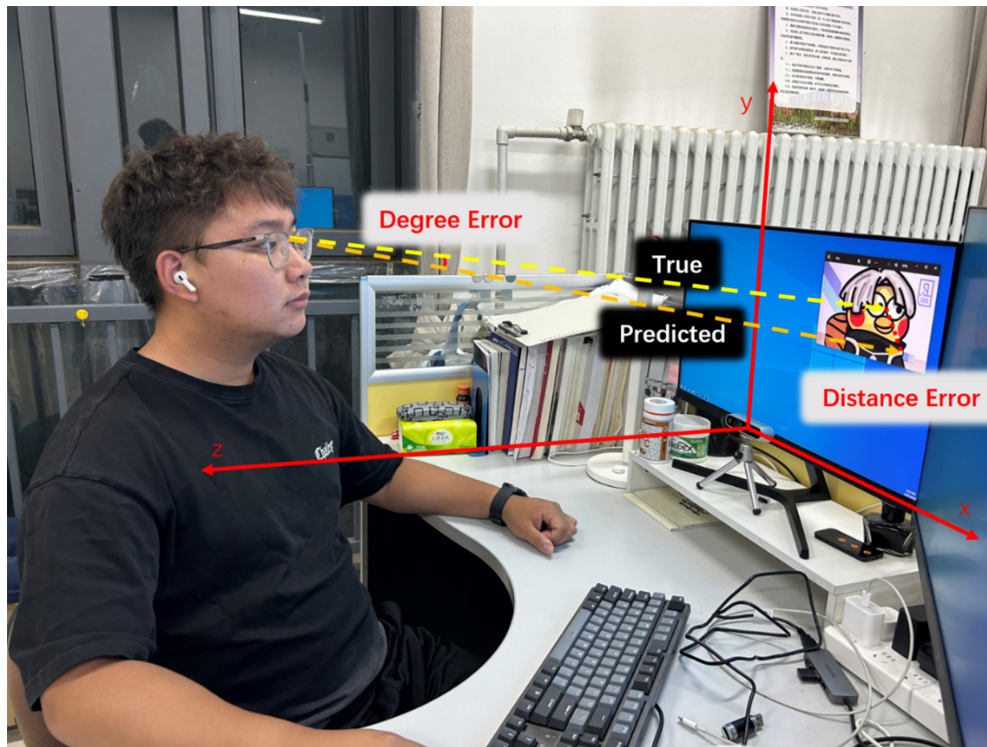


Figure 4. Camera coordinate system and gaze estimation. The image displayed on the screen was obtained from a publicly accessible online source and is included for illustrative purposes only. It is not part of the experimental dataset or quantitative evaluation.

3. RESULTS

3.1. Performance at sequential point test

We conducted sequential point testing and collected 6,572 sets of experimental data. The results of this testing, along with error analysis at multiple scales, are presented in [Figures 5](#) and [6](#). Moreover, the performance of the eyes-mouth detection model is evaluated.

During the error analysis, we considered various variables such as screen gaze position, head spatial position, head spatial pose, and lighting conditions. This comprehensive analysis allowed us to identify the factors that contribute to errors in our experimental data. By combining sequential point testing with thorough error analysis, we obtained a comprehensive understanding of our system's performance and limitations under different conditions.

[Figure 5](#) presents the error analysis for various variables. The horizontal axis represents the variables, while the vertical axis represents the average angular error within the corresponding interval. [Figure 5A](#) and [B](#) display the relationship between the horizontal and vertical positions of the target point on the screen and the angular error. [Figure 5C](#) presents a three-dimensional plot showing the relationship between the target point's position on the screen and the angular error. The x and y axes represent the horizontal and vertical positions of the target point on the screen, respectively, while the z-axis represents the magnitude of the angular error. The error distribution ranges from 6 to 10 degrees.

[Figure 5D](#) illustrates the relationship between the angular error and the horizontal distance between the observer's gaze and the camera. The independent variable is the horizontal position of the midpoint between the eyes in the camera coordinate system, which represents the observer's horizontal displacement. The horizontal distance is distributed from -0.1 to 0.15 meters, and the error distribution ranges from 4 to 10

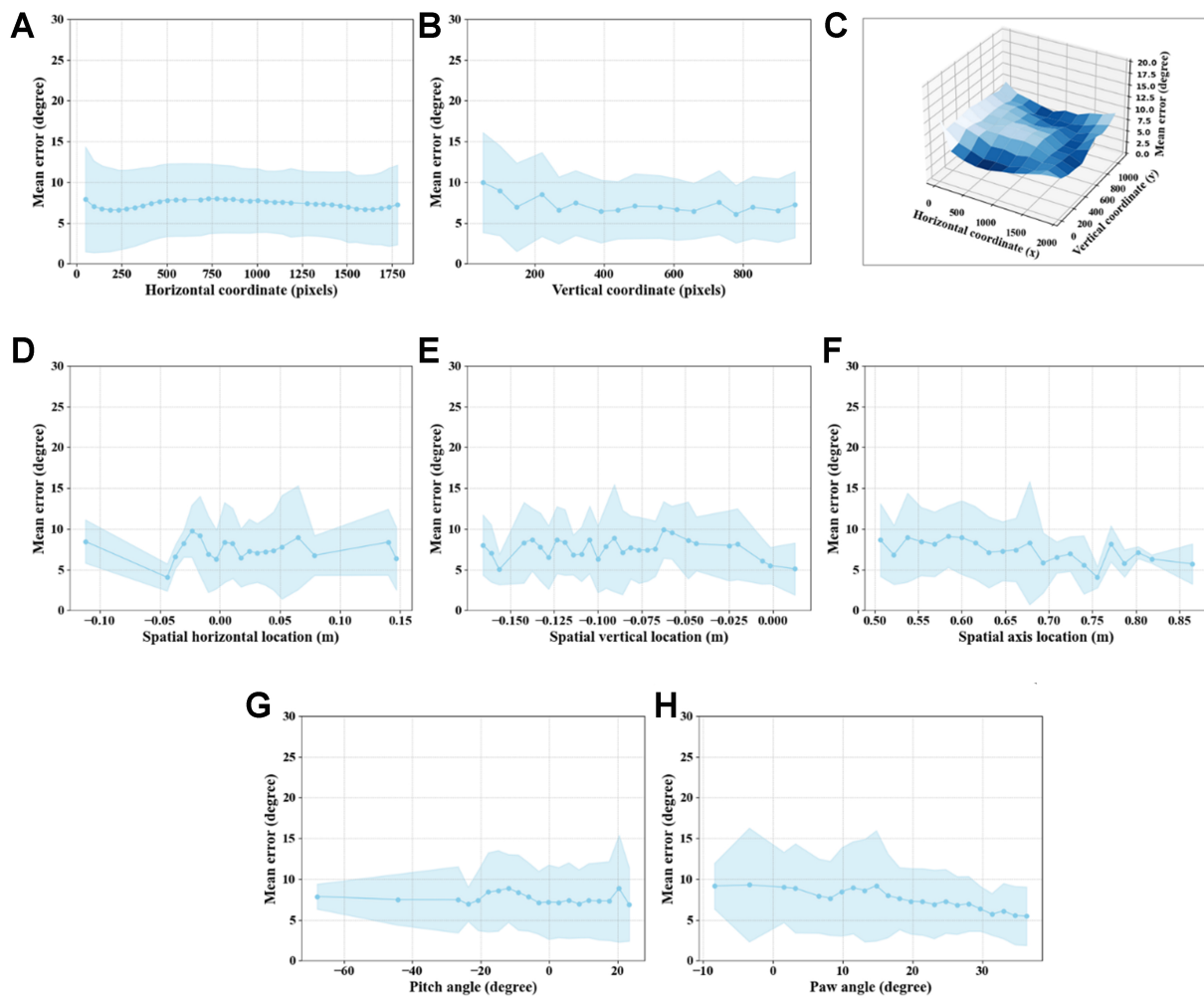


Figure 5. Performance analysis of GPN with different variables. (A) error analysis of the horizontal coordinates of the target points; (B) Error analysis of the vertical coordinates of the target points; (C) Error analysis of the horizontal and vertical coordinates of the target points; (D) Error analysis of the spatial horizontal location of volunteers' heads; (E) Error analysis of the spatial vertical location of volunteers' heads; (F) Error analysis of the spatial axis location of volunteers' heads; (G) Error analysis of the pitch angle of volunteers' heads; (H) Error analysis of the paw angle of volunteers' heads. The central points indicate the mean values, and the blue shaded regions represent the corresponding variance. GPN: Gaze-Point-Net.

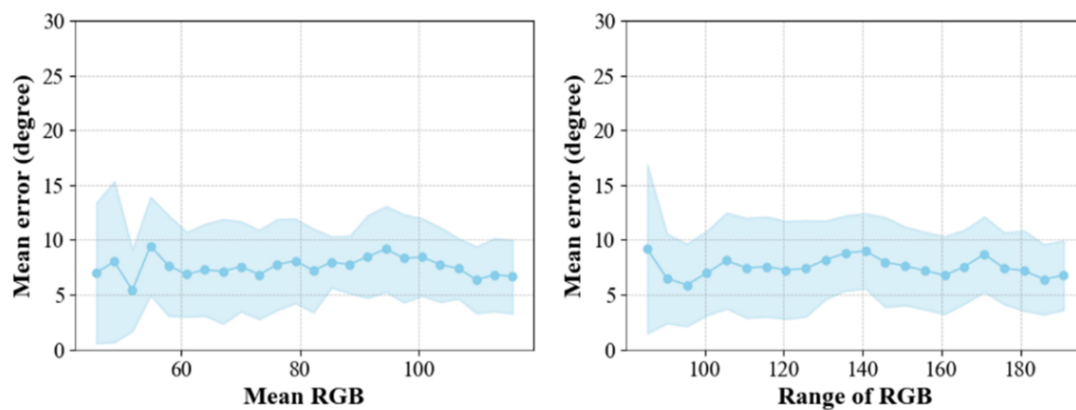


Figure 6. Performance analysis of GPN with different light conditions. The central points indicate the mean values, and the blue shaded regions represent the corresponding variance. GPN: Gaze-Point-Net; RGB: Red, Green, Blue.

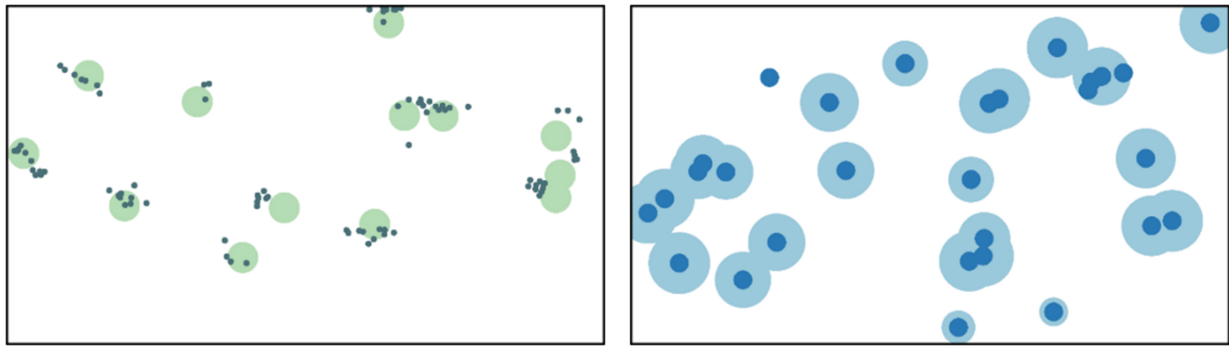


Figure 7. Random point test results.

degrees. [Figure 5E](#) demonstrates the relationship between the angular error and the vertical distance. The independent variable is the vertical position of the midpoint between the eyes in the camera coordinate system, representing the observer's vertical displacement. The vertical distance values range from -0.15 to 0 meters, while the angular error distribution ranges from 5 to 10 degrees. [Figure 5F](#) presents the relationship between the angular error and the axial distance between the observer's gaze and the camera. The independent variable is the axial position of the midpoint between the eyes in the camera coordinate system, which represents the observer's forward and backward displacement and the distance between the observer and the display. The axial distance ranges from 0.5 to 0.85 meters, and the angular error distribution ranges from 4 to 8 degrees.

Finally, [Figure 5G](#) displays the relationship between the angular error and the pitch angle of the observer's head, representing the vertical swing amplitude. The pitch angle ranges from -70 to 20 degrees, and the angular error distribution ranges from 7 to 9 degrees. [Figure 5H](#) shows the relationship between the angular error and the yaw angle of the observer's head, representing the horizontal swing amplitude. The yaw angle ranges from -10 to 38 degrees, and the angular error distribution ranges from 5 to 9 degrees.

[Figure 6](#) illustrates the relationship between the mean and range of Red, Green, Blue (RGB) values and the angular error. The mean RGB values represent the intensity of the experimental ambient light, which is distributed in the range of 40–120. The range of RGB values represents the angle of the laboratory light source, where a larger light angle corresponds to a larger RGB range, distributed in the range of 80–200. In both cases, the errors are distributed within 5–10 degrees.

3.2. Performance at random point test

We collected the results of random point tests for analysis. First, we plotted the distribution of points and the predicted standard deviation using the actual values and the predicted values. These are shown in [Figure 7](#). In the left panel, the light-colored large circle represents the fixed target point in the test, with its center at the center of the target block and its diameter equal to the target block's diagonal length. The dark-colored small circles represent the scattered positions of the predicted fixation points by the model. In the right part, the dark small circles represent the actual fixed target point locations, while the light large circles surrounding them indicate the range of predicted points corresponding to the labels. This figure shows the distribution of predicted points around the ground-truth values and the predictive standard deviation.

During the experimental process, we retained the predicted data from the uncalibrated model and the unfiltered results, resulting in four data groups. These data are illustrated in [Figure 8](#). The left graph depicts the functional relationship between angular error and the true values on the X-axis, where the scatter points represent the distribution of the data. The scatter points are plotted at a ratio of 1:100. The four curves

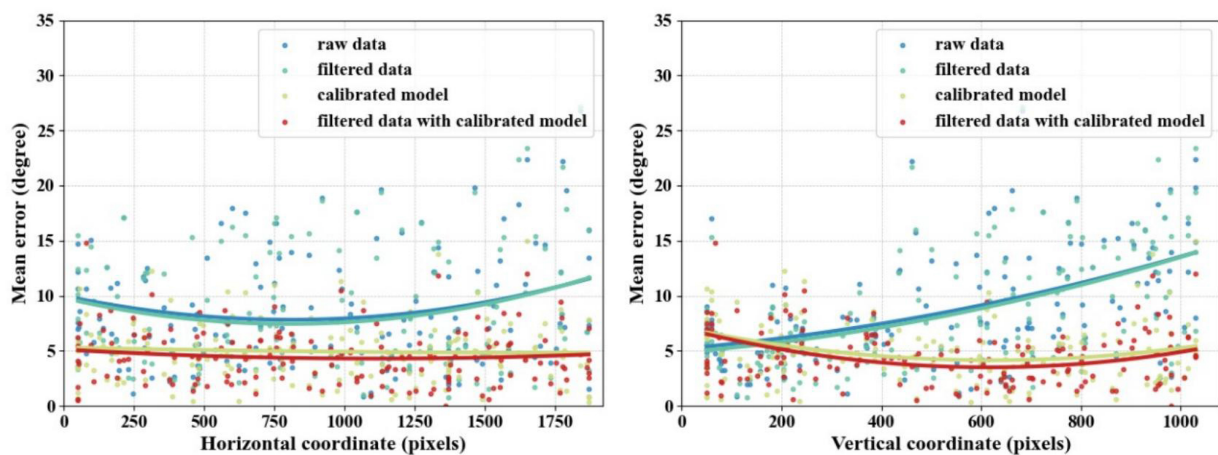


Figure 8. Data filtering and model calibration analysis.

represent the quadratic polynomial fit for each dataset, illustrating how the average angular error varies with the x-coordinate. The right graph presents a similar error analysis, with the Y-axis representing the true values. As shown in the graph, filtering the raw data has minimal effect on reducing the average error, whereas filtering after calibration substantially reduces it.

To analyze the performance of the model, we conducted a series of control experiments and tested the following mathematical models: (1) K-nearest neighbors (KNN). In machine learning, KNN is commonly used for pattern recognition tasks. We expanded the image into a one-dimensional vector and concatenated it with the head position and attitude information, then used the KNN method with a neighborhood size of 3 for testing^[30]; (2) Random forest (RF)^[32]. RF is an effective regression method in machine learning. We tested the model using the same data-processing method, with a decision tree size of 300, a maximum depth of 20, and a maximum of 65 features at each node; (3) Linear regression (LR)^[33]. LR is a simple regression method in machine learning and has been applied effectively to medical image analysis tasks^[33]. We used the same data processing method and tested the model using LR.

Finally, utilizing various models including KNN, RF, LR, GPN, and calibrated GPN, we obtained the prediction results for random-point testing. The results from different models were compared, and the error analysis results are presented in Figure 9. Figure 9A shows the error analysis graph with the true values on the x-axis, which has the same pattern as Figure 8, consisting of an error scatter plot and a fitting curve. In contrast, Figure 9B shows the error analysis graph with the true values on the y-axis.

Scattered points represent the error distribution of data sampled at a certain proportion, and a quadratic function is used to fit the scattered points into a curve. The blue, cyan, and yellow colors represent the results of KNN, LR, and RF, respectively. Pink represents the uncalibrated model results, while red represents the calibrated model results with added data-filtering operations. Our model shows a significantly smaller error distribution and lower standard deviation. Additionally, applying the filtering operations after calibration substantially improves the model's performance.

Figure 9C and D present numerical comparisons of the average angular error and average pixel error for the aforementioned models, respectively. The two figures depict the KNN, RF, LR, general GPN, and calibrated GPN models from left to right, representing progressively lighter shades. The results indicate that our model's performance is comparable to RF before calibration but improves significantly after calibration.

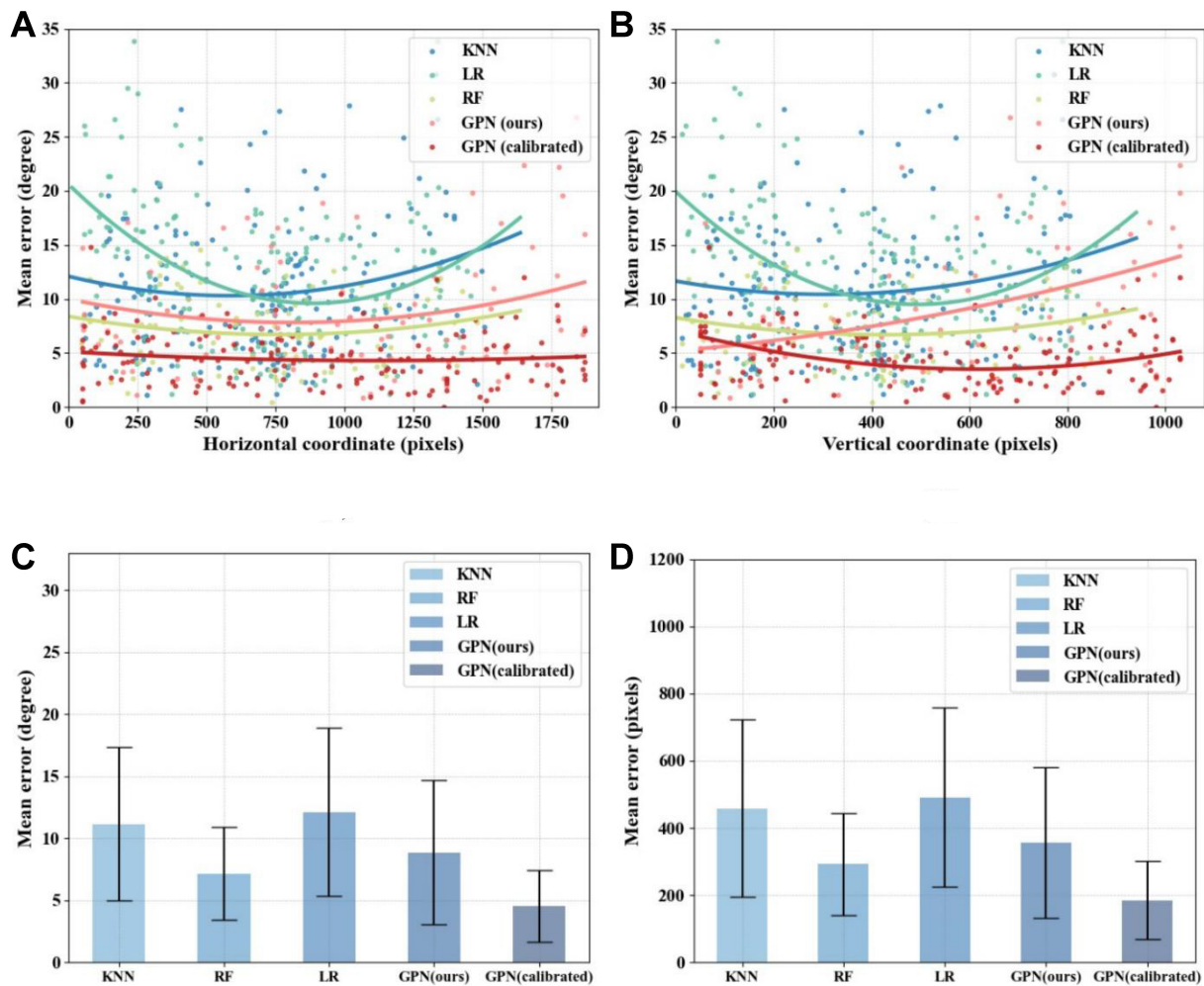


Figure 9. Performance comparison analysis of GPN and other models. (A) scatterplot and fitted curve analysis of the horizontal coordinates of the target points between different models; (B) Scatterplot and fitted curve analysis of the vertical coordinates of the target points between different models; (C) Analysis of mean angular errors for different models; error bars represent the standard deviations of angular errors across the test samples ($n = 2$); (D) Analysis of mean pixel errors for different models; error bars represent the standard deviations of pixel errors across the test samples ($n = 2$). GPN: Gaze-Point-Net; KNN: K-nearest neighbors; LR: linear regression; RF: random forest.

Table 1. Pixel and angular errors of the calibrated GPN and its counterparts

Model	Pixel error	Angular error
KNN ^[30]	458.36 ± 263.17	11.16 ± 6.19
RF ^[32]	292.75 ± 152.27	7.15 ± 3.72
LR ^[33]	491.26 ± 267.23	12.12 ± 6.75
GPN	357.03 ± 223.80	8.85 ± 5.81
Calibrated GPN	185.19 ± 116.57	4.52 ± 2.87

Bold content indicates the proposed approach, which achieved the best performance. GPN: Gaze-Point-Net; KNN: K-nearest neighbors; RF: random forest; LR: linear regression.

Table 1 presents the mean and standard deviation of pixel error and angular error for different models. The screen has a pixel range of 2202, and its visible viewing angle ranges from 60 to 70 degrees; the calibrated GPN also includes the subsequent data-filtering operation. Our calibrated GPN achieves a pixel error of 185.19 ± 116.57 and a angular error of 4.52 ± 2.87 , which is much smaller than that of KNN, RF, LR, and GPN.

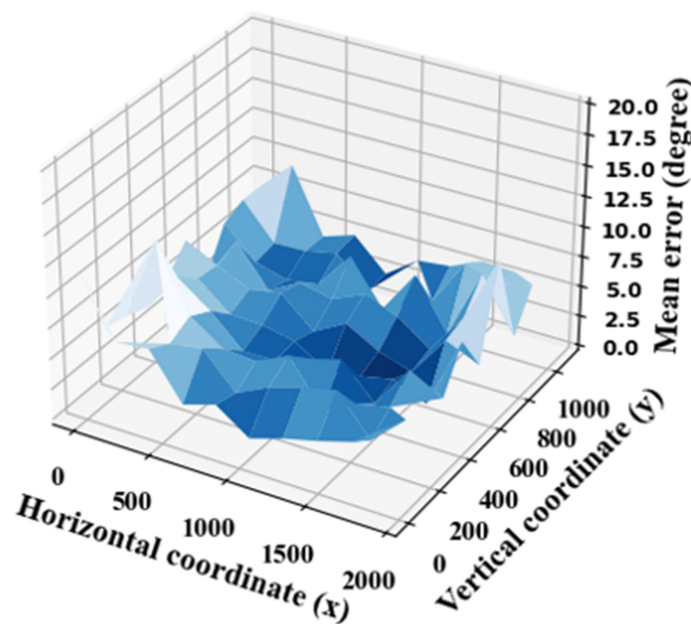


Figure 10. Three-dimensional error analysis.

Table 2. Multiple regression task evaluation metrics for the GPN and its counterparts

Model	RMSE	MAE	R ²	MAPE	MSPE
KNN ^[30]	373.7343	291.3501	0.2357	130.6664	1,025.7574
RF ^[32]	233.3364	187.8462	0.6088	90.7369	437.8502
LR ^[33]	392.678	311.6328	0.0536	112.1944	579.4017
GPN	293.8332	232.2993	0.3423	62.4100	185.9423
Calibrated GPN	156.5305	121.8607	0.8224	46.6341	124.8420

Bold content indicates the proposed approach, which achieved the best performance. GPN: Gaze-Point-Net; RMSE: root mean squared error; MAE: mean absolute error; R²: R-squared; MAPE: mean absolute percentage error; MSPE: mean squared percentage error; KNN: K-nearest neighbors; RF: random forest; LR: linear regression.

We employed multiple regression task evaluation metrics to analyze the performance of the models used in our study, as shown in Table 2. As in Table 1, the calibrated GPN includes the subsequent data-filtering operation. The metrics include the root mean squared error (RMSE), MAE, R-squared (R²), mean absolute percentage error (MAPE), and mean squared percentage error (MSPE).

Based on the analysis of RMSE and MAE, our model exhibits significantly smaller errors compared to other models. Furthermore, for MAPE and MSPE, our model shows smaller relative errors and fewer instances of large errors. Additionally, the R² coefficient indicates a stronger correlation between our model's regression results and the reference standards.

The relationship between the horizontal and vertical coordinates of the ground truth values on the screen and the angular error is illustrated in Figure 10 using a three-dimensional plot. From the final coordinate error analysis graph, the model's prediction error was within 5° in most locations. After calibration and filtering, the average angular prediction error is substantially reduced to 4.52°, demonstrating the effectiveness of the proposed method in achieving accurate gaze estimation across different target locations.

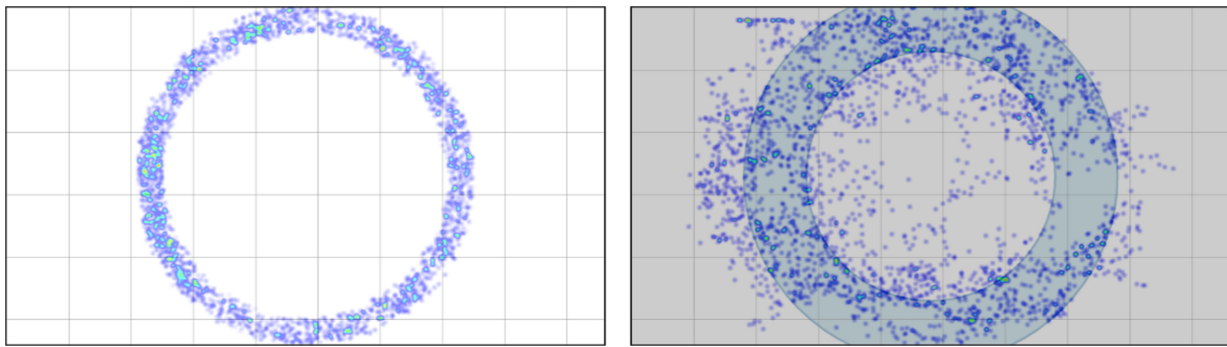


Figure 11. Circular trajectory tracking test results.

3.3. Performance at trajectory tracking test

The results of the trajectory tracking test are shown in Figure 11. The left part displays the distribution heatmap of prediction errors. The heatmap depicts the trajectory of the moving target, with brighter colors indicating larger errors and darker colors indicating smaller errors. The right part shows the distribution of scattered points within a certain confidence region. The light-colored circular rings in the background represent tolerance boundaries around the target trajectory, with the radius indicating the allowable distance from the trajectory. Scatter plots depict the distribution of predicted gaze points. A confidence radius is defined to determine whether predicted gaze points fall within a specified distance from the target trajectory and are therefore considered valid tracks. With a tolerance radius of 300 pixels, the tracking accuracy reaches 93.75%. When the tolerance radius is reduced to 200 pixels, the accuracy is 81.91%.

3.4. Analysis of browsing patterns

Figure 12 presents the browsing-mode analysis, in which the gaze positions of a participant were simultaneously recorded using a professional eye tracker (Tobii Pro) and the proposed eye-tracking system while freely viewing each image^[34]. The first column [Figure 12A] shows four representative images viewed by the participant. The second and fourth columns [Figure 12B and D] present the heat maps generated by the professional eye tracker and the proposed system, respectively, while the third and fifth columns [Figure 12C and E] show the corresponding gaze trajectory maps. The heat maps illustrate the distribution and density of gaze points, with deeper red indicating higher gaze density and thus the observer's focal areas of attention. The trajectory maps plot gaze points in chronological order, illustrating the temporal progression of visual attention across the image. Overall, the heat maps and trajectory maps generated by the proposed system closely resemble those obtained from the professional eye tracker, with only minor differences in gaze-point locations. These results demonstrate that the proposed system can provide gaze-tracking patterns comparable to those of the professional eye tracker during free-viewing tasks.

4. DISCUSSION

4.1. Principal findings

In this study, we designed a comprehensive system for gaze estimation and evaluated its performance through various experiments, drawing on the concept of explicit gaze control in neural talking-head synthesis^[35]. We demonstrated the feasibility of gaze estimation without infrared cameras and showed that a multimodal approach incorporating head spatial position and pose information can enhance the accuracy and generalization performance of gaze estimation. Furthermore, while training the object detection model, we found that increasing the resolution of the feature maps in YOLOv5 can improve the model's ability to recognize small-sized objects. Additionally, combining similar but different object categories (the left and right eyes) into the same class can significantly improve detection performance for that class.

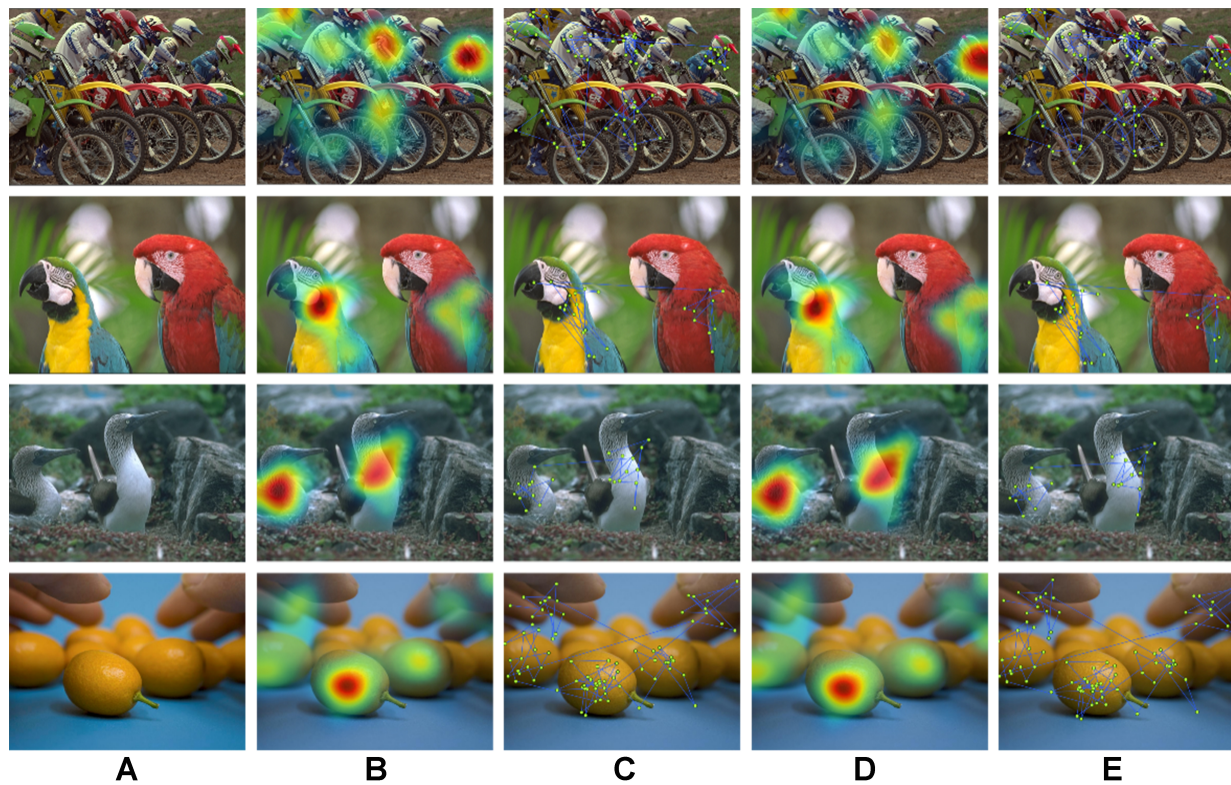


Figure 12. Representative examples of browsing-pattern analysis. (A) Original images viewed by the participant; (B) and (D) heat maps generated by the Tobii Pro eye tracker and the proposed system, respectively; (C) and (E) corresponding gaze trajectory maps. They were obtained from the Kodak Lossless True Color Image Suite, originally released by Eastman Kodak Company (1999); the publicly accessible mirror at <https://r0k.us/graphics/kodak/> (accessed: 22.2.2026) was used for retrieval, and the Berkeley Segmentation Dataset and Benchmark (BSDS300) (<https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/>)^[34].

In the computational workflow, we discovered that in the process of using neural networks for calculations, separating and then fusing features - namely, first computing the head's position and pose information while using perspective transformation to remove pose information from eye images, and then utilizing neural networks for feature fusion - improved the accuracy to some extent.

We directly regress the target fixation point, which reduces computational complexity and makes calibration more efficient. Calibrating the gaze estimation model before each prediction task greatly improves its accuracy.

Finally, our model performs much better in fixed-point gaze tasks (Random point test) compared to scanning tasks (Sequential point test) involving moving points. This is because our random-point testing procedure averages angular errors and standard deviations after the random point remains fixed for a certain period, yielding slightly lower errors than sequential-point testing involving continuous movement of the target point.

4.2. Comparison with previous studies

Since existing gaze estimation studies often adopt different datasets, input modalities, and evaluation protocols, direct performance comparisons across datasets may not be fully reliable. Therefore, Table 3 provides a reference comparison with representative methods reported in the literature, rather than a strict quantitative benchmark. The results highlight the effectiveness of our integrated gaze estimation system under the experimental setting adopted in this study.

Table 3. Comparison with representative gaze estimation methods reported in the literature

Study	Key aspects	Performance
Our method	- Gaze estimation - Eye movement - Depth camera - Free head movement - Calibration	Angular error = $4.52^\circ \pm 2.87^\circ$ Pixel error: = 185.19 ± 116.57
Arar et al., 2017 ^[36]	- Eye movements - HCI - Gaze estimation - Video-based remote eye trackers - User calibration - Estimation bias - Explicit and implicit user calibration methods - Regression-based user calibration techniques - Weighted least squares regression - Real-time cross-ratio based gaze estimation - State-of-the-art user calibration methods	Angular error = $1.01^\circ \pm 1^\circ$
Bao et al., 2022 ^[37]	- Deep learning-based approaches - Appearance-based gaze estimation - Unseen environments - Generalizing gaze estimation algorithm - Rotation-consistency property - Unsupervised domain adaptation - Distribution loss - Cross-domain gaze estimation tasks - Experimental results - Baselines - Computer vision tasks - Physical constraints	Angular error = 5.70°
Zhang et al., 2020 ^[38]	- ETH-XGaze - HCI - Custom datasets - Comparison across methods - ETH-XGaze dataset - High-resolution images - Extreme head poses - Ground truth gaze targets - Robustness of Gaze estimation methods - Standardized experimental protocol - Benchmark website - Unified gaze estimation research	Angular error = 8° at extreme angle
Kim et al., 2020 ^[39]	- Smart interactive environments - User intent prediction - Gaze estimation technology - Interaction techniques - Deep learning-based approach - Low-light conditions - Eye image enhancement - MPIIGaze dataset	Angular error = 9.52° when the light is dim
Ren et al., 2023 ^[40]	- Low-light environment - Gaze estimation - Feature fusion - Multi-level information elements - Gaze conduction principle - Multi-level information element fusion model - Optimized input modes and network structures - GazeCapture dataset - Average error	Angular error = 4.38°

HCI: Human-computer interaction.

In existing appearance-based gaze estimation methods, it has been demonstrated that combining head spatial position and three-dimensional posture methods is highly effective for gaze estimation under free-viewing conditions^[36]. Our method includes a process for correcting the head roll angle. When the head undergoes angular changes along the axial rotation, we use a matrix transformation to correct the inverse rotation. This

eliminates roll-angle features in the image. We incorporate this corrected feature as an input to the model alongside other modalities, using the multimodal approach we propose.

The gaze estimation method for head rotation in Ref.^[37] directly computes on images without reverse-rotation correction. In comparison, their method relies on separate feature extraction, whereas our approach more thoroughly extracts rotation features. Our method first disentangles the features before fusing them, resulting in improved performance. Ultimately, our method achieves an error 1.18° lower than that of their approach.

Similarly, our approach incorporates perspective transformation for separating the head's yaw and pitch angles. As a result, our method demonstrates a certain degree of adaptability to both horizontal and vertical head rotations. The method described in^[38] focuses on gaze estimation for baseline angles. They trained their model on datasets with significant angle variations and achieved an error of approximately 8 degrees. Figure 5G and H illustrate the error analysis results of our method in these two dimensions. Compared with this specialized approach tailored to the specific conditions, our method achieves a performance close to its capabilities.

We acquired data under various lighting conditions by adjusting the illumination. Furthermore, during the image preprocessing phase, techniques such as histogram equalization and data standardization were applied to mitigate the impact of variations in image RGB values. In contrast to previous work^[39], which focused on dim-lighting scenarios (resulting in an error of 9.52 degrees), our approach demonstrates an average error reduction of 5.00 degrees. As illustrated in Figure 6, our method appears to be less affected by variations in light intensity and direction, thus exhibiting strong performance across different lighting conditions.

The method mentioned in^[40] is a gaze tracking approach for mobile devices. It utilizes the front camera of a smartphone to capture images, locates 68 facial landmarks using a model, and performs estimation. Our method, by contrast, is deployed on desktop platforms and benefits from higher computational power. We employ the YOLO model for target detection; in comparison, our method exhibits higher fault tolerance, faster computation speed, and lower resource consumption. On desktop platforms, we achieve performance similar to theirs with greater efficiency and resource savings.

Our approach shares some similarities with the method proposed in^[41], as both utilize facial-related information for gaze estimation. However, Ref.^[41] employs an long short-term memory (LSTM)-based framework to model temporal dependencies from continuous facial sequences, whereas our method focuses on individual-frame gaze estimation by integrating head pose information, binocular features, and calibration strategies. This design reduces the dependency on temporal information and enables a more lightweight implementation for scenarios where continuous data acquisition is unavailable or unnecessary. Although Ref.^[41] demonstrates the effectiveness of temporal modeling for gaze estimation, a direct quantitative comparison between the two methods is difficult due to differences in datasets, input settings, and experimental protocols. Nevertheless, the comparison highlights different design considerations: temporal modeling can improve performance when sequential data are available, while our framework provides a practical alternative for frame-based gaze estimation with reduced dependence on continuous recordings. Future work will further investigate this issue using unified experimental protocols and larger datasets.

Compared to existing methods, one advantage is our calibration approach. We fine-tune neural networks for calibration, enabling robust calibration under various conditions such as changes in characters, lighting, camera parameters, camera field of view, camera position, screen size, and more. This calibration

significantly enhances accuracy. During system deployment, our approach eliminates the need for fixed camera placement, head fixation, and external conditions like controlled lighting. This allows tasks to be accomplished without setting up a series of external conditions, providing a more flexible and efficient solution.

4.3. Limitations and future work

Despite the promising performance of the proposed gaze estimation system, several limitations remain.

First, the current system relies solely on images captured under visible light conditions, without incorporating infrared sensors. Consequently, its performance is sensitive to illumination variations. Although data augmentation and training strategies were employed to improve robustness, illumination effects cannot be fully eliminated. This limitation could be addressed by introducing advanced data enhancement techniques, such as diffusion-based models, to improve image quality under challenging lighting conditions and enhance robustness.

Second, the target detection module operates on a per-frame basis without enforcing temporal consistency. This leads to fluctuations in detection results across consecutive frames, causing inconsistencies in the extracted eye regions and resulting in random directional deviations in gaze predictions. To overcome this limitation, temporal consistency constraints could be incorporated, and gaze estimation models designed for continuous video streams could be developed to leverage inter-frame information and reduce prediction drift.

Third, the current pipeline lacks fine-grained preprocessing of eye images, such as explicit segmentation of the iris and eyelids, which limits the effectiveness and accuracy of feature extraction. In addition, the absence of temporal modeling prevents mitigation of randomness inherent in single-frame predictions. This issue could be alleviated by integrating eye-region segmentation methods to isolate key anatomical structures and adopting sequence-based modeling approaches to better exploit temporal dynamics.

Finally, although the current study adopts a participant-level data partitioning strategy to ensure subject-disjoint training, validation, and testing, further evaluation using more rigorous subject-independent protocols, such as leave-one-subject-out cross-validation, would provide additional evidence of the proposed method's robustness and generalizability. Moreover, a systematic ablation analysis of individual components within the proposed framework would further clarify the contribution of each module and provide deeper insights into the model design. Future studies will investigate larger and more diverse cohorts, incorporate more comprehensive cross-participant validation strategies, and conduct extensive ablation studies to further assess the model's applicability and reliability in real-world scenarios.

5. CONCLUSION

In this study, we developed a gaze estimation system that integrates direct screen position regression, binocular gaze estimation, multimodal feature fusion, and model calibration into a unified framework. By directly regressing screen positions, the system eliminates the coordinate transformation process, thereby simplifying the estimation pipeline, improving computational efficiency, reducing error accumulation, and increasing fault tolerance. A binocular model with unshared parameters is adopted to minimize the influence of binocular disparity encountered in monocular models. In addition, a multimodal convolutional neural network (CNN) model incorporating head spatial position and pose information enhances the system's adaptability to natural head movements and rotations. Perspective transformation is further applied to separate binocular image pose information from multimodal feature fusion, thereby improving the overall estimation performance. Finally, combined with the model calibration strategy, the proposed system

supports gaze estimation in a wide range of daily-life applications. The proposed system also exhibits strong calibration capability, enabling effective adaptation to variations across different operating conditions. Moreover, the model remains relatively compact, and when combined with the high computational efficiency of the YOLO-based detector, the overall inference speed does not limit the camera's sampling frequency. Consequently, the integrated system achieves both high efficiency and accuracy during inference, enabling reliable fixed-point gaze estimation in practical applications.

DECLARATIONS

Authors' contributions

Conceptualization, methodology, software, investigation, data curation, writing - original draft: Wang, Z.

Software, validation, formal analysis, visualization: Wang, J.

Resources, investigation, validation: Ng, A. C. M.

Resources, project administration, validation: Ng, T.

Validation, formal analysis, writing - review and editing: Qian, W.

Writing - review and editing, validation: Elirehema, M.

Supervision, writing - review and editing, funding acquisition: Monkam, P.

Supervision, conceptualization, writing - review and editing, funding acquisition: Qi, S.

Availability of data and materials

The dataset generated and analyzed in this study is not currently publicly available due to pending data-sharing permissions. It may be obtained from the corresponding author upon reasonable request, subject to the required ethical and data-sharing approvals.

AI and AI-assisted tools statement

During the preparation of this manuscript, the authors used ChatGPT (version 5.0, released 2025-08-07) solely for language polishing and grammar checking. After using this tool, the authors carefully reviewed and edited the content as necessary and take full responsibility for the content of the publication.

Financial support and sponsorship

This study was supported by the National Natural Science Foundation of China (82472076), the Fundamental Research Funds for the Central Universities (N26LPY002, N25BJD013), and Open Funding from Shenzhen Jingmei Health Technology Company Ltd.

Conflicts of interest

Ng, A. C. M. and Ng, T. are affiliated with Shenzhen Jingmei Health Technology Co., Ltd., while the other authors have declared no conflicts of interest.

Ethical approval and consent to participate

This study was approved by the Ethics Committee of Northeastern University (approval no. NEU-EC-2024B036S) and was conducted in accordance with institutional and ethical guidelines and the Declaration of Helsinki (2024). Written informed consent was obtained from all experimental participants.

Consent for publication

Written informed consent was obtained from all experimental participants for the publication of this study, including the use of their data and any facial images contained within the figures.

Copyright

© The Author(s) 2026.

REFERENCES

1. Casado-Aranda, L.; Sánchez-Fernández, J.; Ibáñez-Zapata, J. Evaluating communication effectiveness through eye tracking: benefits, state of the art, and unresolved questions. *Int. J. Bus. Commun.* **2023**, *60*, 24–61. DOI

2. Liu, J.; Chi, J.; Yang, H.; Yin, X. In the eye of the beholder: a survey of gaze tracking techniques. *Pattern. Recognit.* **2022**, *132*, 108944. DOI
3. Fan, K.; Cao, J.; Meng, Z.; et al. Predicting the reader's English level from reading fixation patterns using the siamese convolutional neural network. *IEEE. Trans. Neural. Syst. Rehabil. Eng.* **2022**, *30*, 1071–80. DOI PubMed
4. Wang, F. S.; Gianduzzo, C.; Meboldt, M.; Lohmeyer, Q. An algorithmic approach to determine expertise development using object-related gaze pattern sequences. *Behav. Res. Methods.* **2022**, *54*, 493–507. DOI PubMed PMC
5. Fan, X.; Wang, F.; Song, D.; Lu, Y.; Liu, J. GazMon: eye gazing enabled driving behavior monitoring and prediction. *IEEE. Trans. Mob. Comput.* **2021**, *20*, 1420–33. DOI
6. Čegovnik, T.; Stojmenova, K.; Jakus, G.; Sodnik, J. An analysis of the suitability of a low-cost eye tracker for assessing the cognitive load of drivers. *Appl. Ergon.* **2018**, *68*, 1–11. DOI PubMed
7. Li, P.; Hou, X.; Duan, X.; Yip, H.; Song, G.; Liu, Y. Appearance-based gaze estimator for natural interaction control of surgical robots. *IEEE. Access.* **2019**, *7*, 25095–110. DOI
8. Chhimpa, G. R.; Kumar, A.; Garhwal, S.; Dhiraj. Development of a real-time eye movement-based computer interface for communication with improved accuracy for disabled people under natural head movements. *J. Real. Time. Image. Proc.* **2023**, *20*, 81. DOI
9. Ramirez Gomez, A.; Lankes, M. Eyesthetics: making sense of the aesthetics of playing with gaze. *Proc. ACM. Hum. Comput. Interact.* **2021**, *5*, 1–24. DOI
10. Papavaslopoulou, S.; Sharma, K.; Melhart, D.; et al. Investigating gaze interaction to support children's gameplay. *Int. J. Child. Comput. Interact.* **2021**, *30*, 100349. DOI
11. Deane, O.; Toth, E.; Yeo, S. H. Deep-SAGA: a deep-learning-based system for automatic gaze annotation from eye-tracking data. *Behav. Res. Methods.* **2023**, *55*, 1372–91. DOI PubMed PMC
12. Yuan, G.; Wang, Y.; Yan, H.; Fu, X. Self-calibrated driver gaze estimation via gaze pattern learning. *Knowl. Based. Syst.* **2022**, *235*, 107630. DOI
13. Namnakani, O.; Abdrabou, Y.; Grizou, J.; et al. Comparing dwell time, pursuits and gaze gestures for gaze interaction on handheld mobile devices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery; 2023. pp. 1–17. DOI
14. Hacques, G.; Dicks, M.; Komar, J.; Seifert, L. Visual control during climbing: variability in practice fosters a proactive gaze pattern. *PLoS. One.* **2022**, *17*, e0269794. DOI PubMed PMC
15. Fischer, T.; Chang, H. J.; Demiris, Y. RT-GENE: real-time eye gaze estimation in natural environments. In *Computer Vision - ECCV 2018*. Cham: Springer International Publishing; 2018. pp. 339–57. DOI
16. Wu, Z.; Rajendran, S.; Van As, T.; Badrinarayanan, V.; Rabinovich, A. EyeNet: a multi-task deep network for off-axis eye gaze estimation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, Seoul, Korea. Oct 27–28, 2019. IEEE; 2019. pp. 3683–7. DOI
17. Wang, Y.; Jiang, Y.; Li, J.; et al. Contrastive regression for domain adaptation on gaze estimation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, USA. Jun 18–24, 2022. IEEE; 2022. pp. 19354–63. DOI
18. Nonaka, S.; Nobuhara, S.; Nishino, K. Dynamic 3D gaze from afar: deep gaze estimation from temporal eye-head-body coordination. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, USA. Jun 18–24, 2022. IEEE; 2022. pp. 2182–91. DOI
19. Fatahipour, H.; Mosavi, M. R.; Fariborz, J. Uncalibrated eye gaze estimation using SE-ResNext with unconstrained head movement and ambient light change. *Research. Square.* **2023**. DOI
20. Ma, C.; Baek, S.; Choi, K.; Ko, S. Improved remote gaze estimation using corneal reflection-adaptive geometric transforms. *Opt. Eng.* **2014**, *53*, 053112. DOI
21. Shin, Y.; Choi, K.; Kim, S.; Ko, S. A novel single IR light based gaze estimation method using virtual glints. *IEEE. Trans. Consum. Electron.* **2015**, *61*, 254–60. DOI
22. Chi, J.; Liu, J.; Wang, F.; Chi, Y.; Hou, Z. 3-D gaze-estimation method using a multi-camera-multi-light-source system. *IEEE. Trans. Instrum. Meas.* **2020**, *69*, 9695–708. DOI
23. Liu, J.; Chi, J.; Hu, W.; Wang, Z. 3D model-based gaze tracking via iris features with a single camera and a single light source. *IEEE. Trans. Human. Mach. Syst.* **2021**, *51*, 75–86. DOI
24. Qin, J.; Shimoyama, T.; Sugano, Y. Learning-by-novel-view-synthesis for full-face appearance-based 3D gaze estimation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, New Orleans, USA. Jun 19–20, 2022. IEEE; 2022. pp. 4977–87. DOI
25. Li, J.; Chen, Z.; Zhong, Y.; et al. Appearance-based gaze estimation for ASD diagnosis. *IEEE. Trans. Cybern.* **2022**, *52*, 6504–17. DOI

26. Wang, Q.; Wang, H.; Dang, R.; et al. Style transformed synthetic images for real world gaze estimation by using residual neural network with embedded personal identities. *Appl. Intell.* **2023**, *53*, 2026–41. DOI
27. Khanam, R.; Hussain, M. What is YOLOv5: a deep look into the internal features of the popular object detector. *arXiv* **2024**, arXiv:2407.20892. Available online: <https://doi.org/10.48550/arXiv.2407.20892>. (accessed 2026-09-11).
28. Keselman, L.; Iselin Woodfill, J.; Grunnet-Jepsen, A.; Bhowmik, A. Intel(R) realSense(TM) stereoscopic depth cameras. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, USA. Jul 21–26, 2017. IEEE; 2017. pp. 1267–76. DOI
29. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-excitation networks. *IEEE. Trans. Pattern. Anal. Mach. Intell.* **2020**, *42*, 2011–23. DOI PubMed
30. Park, S.; De Mello, S.; Molchanov, P.; Iqbal, U.; Hilliges, O.; Kautz, J. Few-shot adaptive gaze estimation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea. Oct 27 - Nov 02, 2019. IEEE; 2019. pp. 9367–76. DOI
31. Valliappan, N.; Dai, N.; Steinberg, E.; et al. Accelerating eye movement research via accurate and affordable smartphone eye tracking. *Nat. Commun.* **2020**, *11*, 4553. DOI PubMed PMC
32. Criminisi, A.; Shotton, J.; Konukoglu, E. Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Found. Trends. Comput. Graph. Vis.* **2012**, *7*, 81–227. DOI
33. Zhang, Y.; Cai, J.; Cui, C.; Qi, S.; Zhao, D. Predicting breast cancer response to neoadjuvant therapy by integrating radiomic and deep-learning features from early-and-peak phases of DCE-MRI. *BMC. Cancer.* **2025**, *25*, 1747. DOI PubMed PMC
34. Martin, D.; Fowlkes, C.; Tal, D.; Malik, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of 8th IEEE International Conference on Computer Vision. ICCV 2001*, Vancouver, Canada. Jul 07–14, 2001. IEEE; 2001. pp. 416–423. DOI
35. Doukas, M. C.; Ververas, E.; Sharmanska, V.; Zafeiriou, S. Free-HeadGAN: neural talking head synthesis with explicit gaze control. *IEEE. Trans. Pattern. Anal. Mach. Intell.* **2023**, *45*, 9743–56. DOI PubMed
36. Arar, N. M.; Gao, H.; Thiran, J. A regression-based user calibration framework for real-time gaze estimation. *IEEE. Trans. Circuits. Syst. Video. Technol.* **2017**, *27*, 2623–38. DOI
37. Bao, Y.; Liu, Y.; Wang, H.; Lu, F. Generalizing gaze estimation with rotation consistency. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, USA. Jun 18–24, 2022. IEEE; 2022. pp. 4197–206. DOI
38. Zhang, X.; Park, S.; Beeler, T.; Bradley, D.; Tang, S.; Hilliges, O. ETH-XGaze: a large scale dataset for gaze estimation under extreme head pose and gaze variation. In *Computer Vision - ECCV 2020*. Cham: Springer International Publishing; 2020. pp. 365–81. DOI
39. Kim, J. H.; Jeong, J. W. Gaze in the dark: gaze estimation in a low-light environment with generative adversarial networks. *Sensors* **2020**, *20*, 4935. DOI PubMed PMC
40. Ren, Z.; Fang, F.; Hou, G.; Li, Z.; Niu, R. Appearance-based gaze estimation with feature fusion of multi-level information elements. *J. Comput. Des. Eng.* **2023**, *10*, 1080–109. DOI
41. Li, Y.; Huang, L.; Chen, J.; Wang, X.; Tan, B. Appearance-based gaze estimation method using static transformer temporal differential network. *Mathematics* **2023**, *11*, 686. DOI

Disclaimer/Publisher’s Note: All statements, opinions, and data contained in this publication are solely those of the individual author(s) and contributor(s) and do not necessarily reflect those of OAE and/or the editor(s). OAE and/or the editor(s) disclaim any responsibility for harm to persons or property resulting from the use of any ideas, methods, instructions, or products mentioned in the content.



© The Author(s) 2026. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.