LightGazeNet: A Lightweight GNN-based Architecture for Gaze Estimation

Heena Patel, Anirban Chowdhury, Pooja Jigar Choksy, Samiksha Pradeep Pachade, Ajinkya Puar
Akeso Eyecare, Beijing, China
{heena.patel, anirban.chowdhury, pooja.choksy, samiksha.pachade, ajinkyapuar}@akesoeyecare.com

Abstract

Gaze estimation remains a fundamental yet challenging task, requiring a careful balance between accuracy and efficiency for real-world deployment. We introduce LightGazeNet, a lightweight Graph Neural Network (GNN) framework designed for appearance-based gaze estimation. LightGazeNet effectively integrates multi-modal inputs—including facial features, eye cues, 3D eye centers, head pose, and calibration data—within a compact graph-based architecture. To enhance feature fusion across heterogeneous inputs, the framework leverages multi-head attention to model complex spatial dependencies. Extensive evaluations on multiple benchmark datasets show that LightGazeNet achieves competitive or superior accuracy with significantly fewer parameters than existing methods. Furthermore, it demonstrates strong cross-dataset generalization, with calibration-based adaptation improving robustness under domain shift. By combining accuracy, efficiency, and adaptability, LightGazeNet offers a practical solution for gaze estimation in real-world settings while advancing graph-based modeling in computer vision.

1. Introduction

Gaze tracking is a transformative advancement in human-computer interaction (HCI), utilizing visual data from a user's face and eyes to infer attention and intent [8, 15]. It plays a pivotal role in medical diagnostics, intelligent transportation, driver monitoring and safety, educational engagement assessment, biometric authentication, accessibility assistance, consumer behavior studies, gaming interfaces, and cognitive science [2, 18, 30, 31]. Specialized applications such as fatigue detection [32] and mental health assessment [33] further highlight its academic and industrial significance.

Early gaze tracking systems achieved high accuracy but required intrusive hardware with direct eye contact, limiting user comfort and adoption [34]. Recent advances in computer vision have enabled non-intrusive approaches that maintain accuracy while improving usability. Modern systems fall into two categories: head-mounted devices (e.g., smart glasses) that offer stability during head motion due to eye proximity, and remote systems using external cameras for facial and eye analysis, providing greater comfort in mobile and unconstrained settings [23].

Recent gaze estimation methods handle variations in illumination, head pose, and demographics, enabling applications across cognitive studies [10], social behavior analysis [21], healthcare [11], human-computer interaction [34], and commercial systems [4]. However, widespread adoption requires computationally efficient solutions suitable for resource-constrained devices. We address this challenge by introducing a lightweight graph-based approach that achieves competitive accuracy while maintaining real-time performance on affordable hardware.

Gaze estimation methods are broadly categorized as model-based or appearance-based. Model-based approaches rely on eyeball geometry and often require specialized hardware (e.g., NIR cameras) to estimate 3D gaze. While robust to head pose and effective in controlled settings, they typically require user-specific calibration and are less suited to real-world variability.

In contrast, appearance-based methods operate with standard hardware (e.g., low-resolution webcams) and perform well in unconstrained environments. These approaches learn a mapping from facial or eye appearance to gaze direction using either traditional machine learning—such as linear regression [24], support vector machines [25], or Gaussian processes [30]—or deep learning models based on convolutional neural networks (CNNs). Traditional methods require less data and train faster but struggle to generalize across varying head poses, lighting, and user differences. In contrast, deep CNNs extract high-level gaze features, scale to large datasets, and model complex non-linear mappings, making them well-suited for real-world applications.

State-of-the-art advancements in gaze estimation integrate face and eye inputs to improve accuracy and reliability, leveraging features such as iris and pupil positions. However, simply concatenating eye feature vectors or adjusting their weights fails to fully capture the complex re-

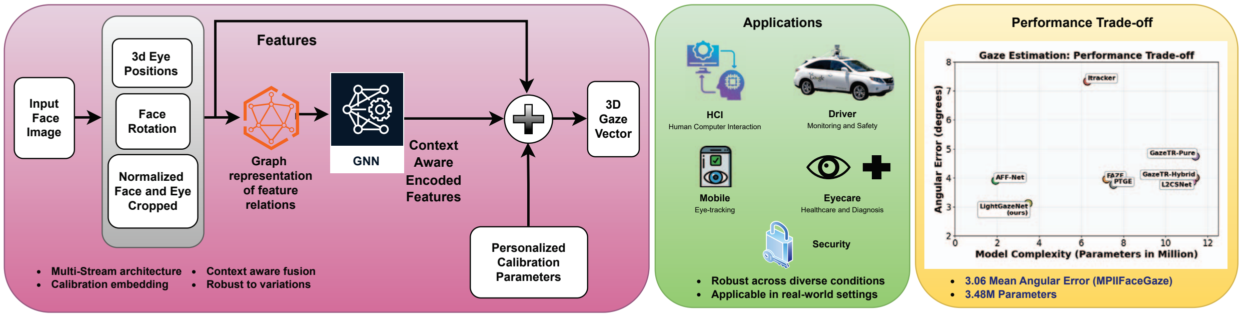


Figure 1. Overview of our proposed LightGazeNet framework. Multi-stream inputs (3D eye positions, face rotation, normalized regions) are processed through graph neural networks with personalized calibration, achieving 3.06° error with 3.48M parameters.

lationships between these inputs, especially in challenging scenarios where reliable eye features are difficult to extract [3]. Most existing methods process face and eye features independently, limiting their ability to model interactions that could enhance gaze estimation.

To overcome this, we propose a graph-based approach that explicitly models the spatial and semantic relationships between facial and eye features. Graph Neural Networks (GNNs) are well-suited for learning on structured data represented as graphs [28], and have shown effectiveness in gaze and gesture recognition [29], facial expression analysis [19], and 3D facial morphology [12]. These successes demonstrate GNNs' ability to capture spatial and structural dependencies, making them promising for modeling facial–ocular interactions in gaze estimation.

Motivated by this, we introduce LightGazeNet (Figure 1), a novel graph-based gaze estimation framework that employs GNNs for efficient multi-modal feature integration. Our approach processes 3D eye positions, face rotation parameters, and normalized facial and eye regions through a unified graph representation, enabling context-aware feature fusion via Graph Neural Networks (GNNs). The framework incorporates personalized calibration parameters and multi-head attention mechanisms to maintain consistent performance across diverse subjects and varying conditions. LightGazeNet achieves 3.06° mean angular error on MPIIFaceGaze with only 3.48M parameters, demonstrating superior efficiency compared to baseline methods. The lightweight design enables deployment across diverse platforms while maintaining competitive accuracy.

Our key contributions include: (1) a novel application of GNNs to gaze estimation for modeling complex inter-feature relationships through structured graph representations; (2) a multi-head attention mechanism for adaptive feature weighting and improved interpretability; and (3) a lightweight architecture achieving competitive performance with minimal computational overhead, suitable for real-world deployment scenarios.

The remainder of this paper is organized as follows. Section 2 reviews related work in gaze estimation. Section 3 presents the LightGazeNet methodology, including feature extraction, graph construction, multi-head attention GNN, and 3D gaze vector reconstruction. Section 4 presents the experimental evaluation, including ablation studies and attention analysis. Section 5 concludes the paper.

2. Related Works

Early appearance-based methods introduced multi-modal deep learning architectures for gaze estimation. **iTracker** [22] processes eye crops, face images, and face grids through separate CNN branches, demonstrating multi-input effectiveness but with limited cross-modal integration. **GazeNet** [36] incorporates head pose estimation with deeper ResNet architectures for improved generalization. To enhance spatial awareness, **Dilated-Net** [5] employs dilated convolutions to expand receptive fields without increasing parameter. In parallel, **RT-Gene** [13] focuses on real-time applications by combining appearance and geometric cues through an efficient fusion pipeline.

To enhance performance in unconstrained environments, **Gaze360** [20] employed a 360-degree spherical representation and temporal supervision to estimate gaze direction across extreme head poses and outdoor lighting conditions, enabling calibration-free estimation. **FullFace** [35] demonstrated that peripheral facial regions provide valuable cues for gaze estimation. It employed full-face input to capture subtle features such as muscular tension and facial asymmetry, which are often overlooked by eye-centric methods. **R-CNN** [26] adapted region-based CNN strategies to extract informative sub-regions from the face, allowing the model to ignore irrelevant background and focus on gaze-relevant features. Furthermore, **CA-Net** [7] introduced a channel-wise attention mechanism that dynamically weights the contributions of different feature maps, significantly improving accuracy by enhancing the discriminative power of deep features.

With the advance of transformer-based architectures, **GazeTR-Pure** and **GazeTR-Hybrid** [6] brought attention mechanisms to the forefront. GazeTR-Pure relies solely on transformer blocks for capturing long-range dependencies, while GazeTR-Hybrid combines CNN backbones with transformers to balance local and global feature learning. These methods outperform conventional CNN-based models in capturing inter-modal interactions and complex spatial patterns. Attention-based fusion was further explored in **AFF-Net** [3], which introduces a cross-modal attention mechanism to learn optimal feature integration from the face and eye modalities. AFF-Net effectively mitigates the challenges posed by occlusions and noisy eye features through adaptive feature weighting. Similarly, **L2CSNet** [1] reformulates gaze estimation as a classification task by discretizing pitch and yaw angles into bins and applying classification with soft regression, demonstrating competitive performance with simplified training pipelines.

Ongoing works have explored few-shot and meta-learning strategies, such as **FAZE**[27], which adapts to new subjects using minimal data through parameter-efficient learning and task-specific adaptation layers. FAZE achieves excellent performance even under subject variability, making it suitable for personalized gaze estimation. **PTGE** [9] utilizes progressive transformer-based modules to iteratively refine gaze predictions through intermediate supervision. Its attention mechanism and hierarchical structure enable robust gaze estimation under variable lighting and head pose conditions.

Despite significant progress, existing methods face key limitations: (1) simple concatenation for multi-modal fusion, limiting complex dependency modeling; (2) lack of explicit relational structure between facial components; (3) high computational costs preventing real-time deployment; and (4) limited cross-dataset generalization. Our **LightGazeNet** addresses these through a novel graph-based architecture that explicitly models spatial relationships between heterogeneous features via lightweight GNNs with multi-head attention, achieving superior accuracy-efficiency trade-offs with strong generalization capabilities.

3. Methodology

We propose LightGazeNet, a lightweight graph-based framework for appearance-based gaze estimation that models spatial relationships between heterogeneous input features. Unlike existing methods that process multi-modal inputs through separate branches [9, 22, 27], our approach constructs a unified graph representation where nodes encode different feature types and edges capture inter-feature dependencies, enabling joint reasoning over spatial and appearance information within a compact architecture.

As illustrated in Figure 2, LightGazeNet consists of three key components: (1) **Feature Encoding and Projection** using lightweight MobileNetV3-based CNN for image features and linear layers for geometric data, mapping heterogeneous features to a shared embedding space; (2) **Graph Construction and Reasoning** with fully connected graph formation and multi-head attention-based GNN for contextual information aggregation; and (3) **Gaze Prediction** via regression head outputting pitch and yaw angles.

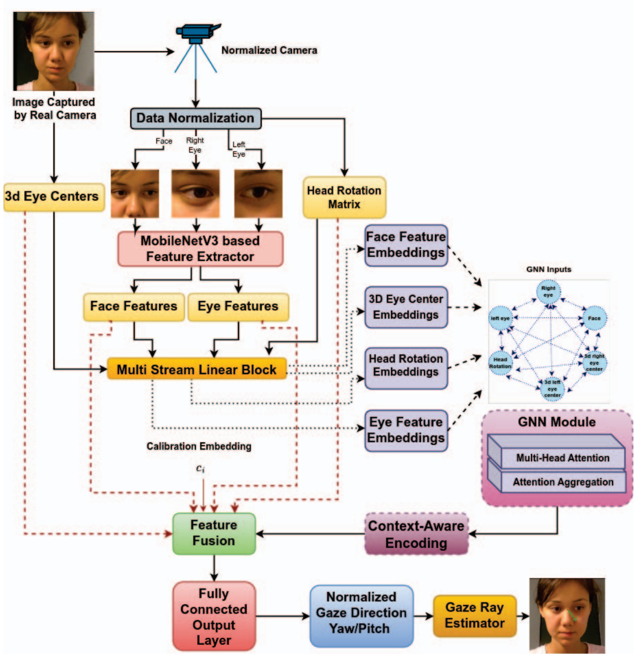


Figure 2. LightGazeNet architecture overview. The framework processes heterogeneous inputs through lightweight encoders, constructs a fully connected graph, applies multi-head attention-based GNN reasoning, and outputs gaze predictions via regression.

3.1. Feature Encoding and Projection

The model takes as input a face image $\mathbf{I}$, left and right eye crops $\mathbf{L}, \mathbf{R} \in \mathbb{R}^{224\times 224\times 3}$, a head rotation vector $\mathbf{r} \in \mathbb{R}^{9}$, and the 3D positions of both eyes $\mathbf{p}_L, \mathbf{p}_R \in \mathbb{R}^{3}$. We employ MobileNetV3-Small [17] with depthwise-separable convolutions and Squeeze-and-Excitation (SE) modules for efficient feature extraction. The network extracts 576-dimensional feature vectors from the face image ($\mathbf{f}_I$), left eye ($\mathbf{f}_L$), and right eye ($\mathbf{f}_R$) regions through global average pooling. All features are projected to a unified 32-dimensional space:

$$\mathbf{h}_i = \sigma(\mathbf{W}_i \mathbf{f}_i + \mathbf{b}_i), \quad i \in I, L, R \tag{1}$$

where $\mathbf{W}_i \in \mathbb{R}^{32\times 576}$, $\mathbf{b}_i \in \mathbb{R}^{32}$, and $\sigma(\cdot)$ is GELU activation [16]. Geometric features are similarly projected:

$$\mathbf{h}_j = \sigma(\mathbf{W}_j \mathbf{x}_j + \mathbf{b}_j), \quad j \in r, p_L, p_R \tag{2}$$

with $\mathbf{W}r \in \mathbb{R}^{32\times 9}$ and $\mathbf{W}_{p_L}, \mathbf{W}_{p_R} \in \mathbb{R}^{32\times 3}$.

3.2. Graph Construction and Reasoning

We construct a fully connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $|\mathcal{V}| = 6$ nodes, each encoding a 32-dimensional modality feature. The initial node feature matrix is:

$$\mathbf{X} = [\mathbf{h}_L; \mathbf{h}_R; \mathbf{h}_I; \mathbf{h}_r; \mathbf{h}_{p_L}; \mathbf{h}_{p_R}] \in \mathbb{R}^{6\times 32} \tag{3}$$

Node-wise Feature Aggregation. Let $\mathbf{x}_i$ be the i-th row of $\mathbf{X}$. Interaction scores between nodes i and j are computed as:

$$e_{ij} = \text{LeakyReLU}(\mathbf{a}^T[\mathbf{W}\mathbf{x}_i \,||\, \mathbf{W}\mathbf{x}_j]) \tag{4}$$

where $\mathbf{a}, \mathbf{W}$ are learnable parameters, and $||$ denotes concatenation.

Normalized coefficients are:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k\in\mathcal{N}_i} \exp(e_{ik})} \tag{5}$$

where k indexes neighbors of i, i.e., all other nodes in the fully connected graph.

Each node is updated as:

$$\mathbf{x}'_i = \sigma\left(\sum_{j\in\mathcal{N}_i} \alpha_{ij}\mathbf{W}\mathbf{x}_j\right) \tag{6}$$

where $\sigma(\cdot)$ is a non-linear activation. To enrich representations, P parallel projections are applied, each outputting a d'-dimensional feature. The concatenated output is:

$$\mathbf{x}'_i = \big\|_{p=1}^{P} \mathbf{x}_i'^{(p)} \in \mathbb{R}^{P\cdot d'} \tag{7}$$

This process is applied in two layers:

$$\mathbf{H}^{(1)} = \mathcal{F}(\mathbf{X}) \in \mathbb{R}^{6\times 256} \tag{8}$$

$$\mathbf{H}^{(2)} = \mathcal{F}(\mathbf{H}^{(1)}) \in \mathbb{R}^{6\times 32} \tag{9}$$

where $\mathcal{F}(\cdot)$ denotes one full pass of Equations 4–7.

The final graph-level feature is:

$$\mathbf{Z} = \text{Flatten}(\mathbf{H}^{(2)}) \in \mathbb{R}^{384} \tag{10}$$

3.3. Gaze Direction Estimation

The gaze prediction integrates GNN output $\mathbf{h}'_i \in \mathbb{R}^{384}$ with original CNN features and subject-specific calibration embeddings. We concatenate multi-modal features:

$$\mathbf{h} = [\mathbf{Z}; \mathbf{f}_R; \mathbf{f}_L; \mathbf{f}_I; \mathbf{r}; \mathbf{p}_L; \mathbf{p}_R; \mathbf{c}_i] \in \mathbb{R}^{1947+d_c} \tag{11}$$

where $\mathbf{c}_i \in \mathbb{R}^{d_c}$ is the subject-specific calibration embedding with $d_c = 6$ dimensions, defined as:

$$\mathbf{c}_i = \text{Embedding}(i, d_c) \tag{12}$$

Here, i denotes the subject identifier from the training set.

Two fully-connected layers with GELU activation produce gaze direction estimates:

$$\mathbf{z}_1 = \text{GELU}(\mathbf{W}_1\mathbf{h} + \mathbf{b}_1) \in \mathbb{R}^{1280} \tag{13}$$

$$\mathbf{g} = (\theta, \phi) = \mathbf{W}_2\mathbf{z}_1 + \mathbf{b}_2 \in \mathbb{R}^2 \tag{14}$$

where $\mathbf{W}_1 \in \mathbb{R}^{512\times(1947+d_c)}$, $\mathbf{b}_1 \in \mathbb{R}^{512}$, $\mathbf{W}_2 \in \mathbb{R}^{2\times 512}$, and $\mathbf{b}_2 \in \mathbb{R}^2$ are learnable parameters. The output (θ, ϕ) represents pitch and yaw angles in radians.

3D Gaze Vector Reconstruction. The predicted angular parameters (θ, ϕ) are converted to a 3D unit gaze direction vector using spherical-to-Cartesian transformation:

$$\hat{\mathbf{g}}_n = \frac{1}{|\mathbf{g}_n|_2}\begin{bmatrix}\cos\theta\sin\phi \\ \sin\theta \\ \cos\theta\cos\phi\end{bmatrix} \tag{15}$$

where $|\cdot|_2$ denotes the L2 norm.

Given the initial gaze direction vector $\hat{\boldsymbol{z}} = \hat{\boldsymbol{g}}_0$ as the primary axis (typically the forward-looking direction in normalized camera coordinates), we compute the orthonormal basis vectors as follows:First, we establish the Y-axis component through cross-product normalization:

$$\hat{\mathbf{y}} = \frac{\hat{\mathbf{g}}_n \times \mathbf{x}_{init}}{|\hat{\mathbf{g}}_n \times \mathbf{x}_{init}|_2} \tag{16}$$

$$\hat{\mathbf{x}} = \frac{\hat{\mathbf{y}} \times \hat{\mathbf{g}}_n}{|\hat{\mathbf{y}} \times \hat{\mathbf{g}}_n|_2} \tag{17}$$

where, $\mathbf{x}_{init}$ represents an initial x-axis estimate derived from head pose parameters.

The complete conversation matrix $\mathbf{R}_h = [\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}]^T \in \mathbb{R}^{3\times 3}$ satisfies orthogonality constraint $\mathbf{R}_h^T\mathbf{R}_h = \mathbf{I}$ and $\det(\mathbf{R}_h) = 1$, ensuring proper rotation properties. The transformation to camera coordinates is:

$$\hat{\mathbf{g}}_c = \mathbf{R}_h^{-1}\hat{\mathbf{g}}_n \tag{18}$$

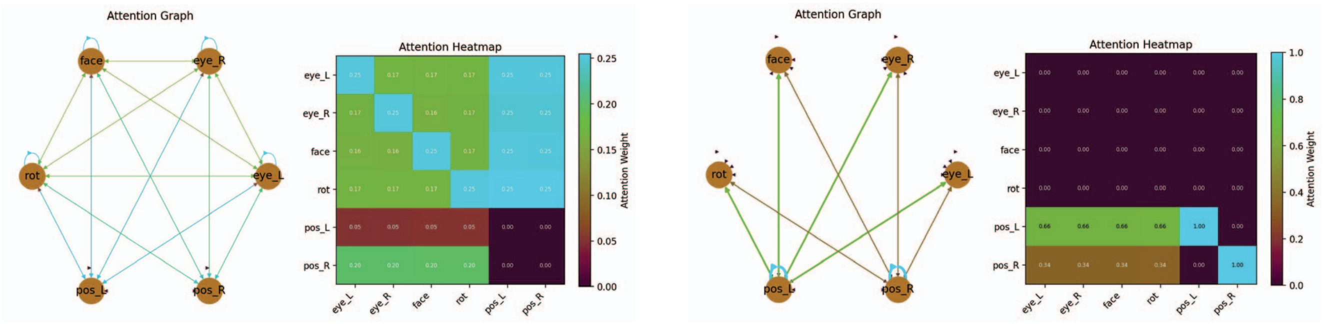
The final gaze ray in camera coordinates is formulated as:

$$\mathcal{G}_R(t) = \mathbf{o}_e + t \cdot \hat{\mathbf{g}}_c, \quad t \geq 0 \tag{19}$$

where $\mathbf{o}_e \in \mathbb{R}^3$ is the estimated eye center position and t is the ray parameter.

4. Experiments and Results

Dataset. We evaluate our method on three widely-used gaze estimation datasets: MPIIFaceGaze [35], EyeDiap [14], and GazeCapture [22]. MPIIFaceGaze contains 37,667 images from 15 participants with eye crops, facial landmarks, head pose, and 3D gaze vectors in camera coordinates. EyeDiap provides 94 video sequences from 16 participants with three protocols: continuously moving targets, discretely moving targets, and floating ball tracking.



(a) Layer 1 Attention Representation

(b) Layer 2 Attention Representation

Figure 3. Attention weights visualization for Subject p14

We use the continuously moving targets protocol following standard evaluation practices. GazeCapture is a large-scale dataset with over 2,445 participants recorded on mobile devices, providing RGB face images, eye crops, and screen-coordinate gaze points across diverse conditions.

Data processing. The gaze normalization preprocessing pipeline standardizes face and eye images along with their corresponding head poses and gaze vectors to maintain consistency across samples [37]. Face and eye images are cropped using the Python face-detection library as part of the dataset preprocessing. The cropped face and eye images are resized to dimensions of $224 \times 224 \times 3$. Pixel values are normalized to a range of $[0, 1]$. Six facial landmarks from the dataset labels define the bounding boxes for the face and eyes. The eye center is determined by averaging the eye coordinates. The pre-processing approach accounts for variations in head poses, distances, and camera parameters, ensuring robust, structured input for training gaze estimation models.

Implementation Details. The model is trained for 30 epochs using Huber loss with $\delta = 1.5$, chosen for its stability compared to L2 loss. Training employs Adam optimizer with learning rate 3×10^{-4} and batch size 64. Cross-subject evaluation uses leave-one-out validation, excluding each subject's data from training for testing. Following prior work, the final 500 images per subject form the validation set. Implementation uses PyTorch with Xavier uniform initialization.

4.1. Attention Dynamics Across Layers

Two-Stage Edge-Level Attention Dynamics. The visualizations in Figure 3 show the evolving attention dynamics across the two GNN layers for a sample of subject `p14` from the MPIIFaceGaze dataset [35]. In the first layer, all perceptual nodes (eye_L, eye_R, `face`, `rot`) exhibit strong self-attention and moderate mutual connections, with a shared focus on pos_R. This pattern reflects early fusion of appearance and pose cues, with positional information from the right eye playing a dominant role. In contrast, the second layer suppresses all perceptual node interactions, shifting attention exclusively toward spatial nodes. The positional nodes (pos_L, pos_R) become central to the graph's reasoning, highlighting a transition from appearance-level integration to fine-grained geometric refinement for accurate gaze prediction.

Subject-specific variations in attention arise from differences in gaze behavior, such as head pose and ocular structure, as well as the varying relevance of spatial and visual cues across samples. The GNN adaptively adjusts its focus based on feature salience and progressively refines representations across layers.

Visualizing Inter-Subject Variability in Attention Maps. In Figure 4, the t-SNE visualization of second-layer attention patterns reveals clear clustering by subject identity across all 15 subjects from the MPIIFaceGaze dataset[35]. Each subject forms a distinct cluster, indicating that the GNN effectively captures subject-specific gaze representations. This consistency within subjects and separability across subjects highlights the model's capacity to encode personalized gaze features. Such discriminative attention embeddings are well-suited for few-shot personalization,

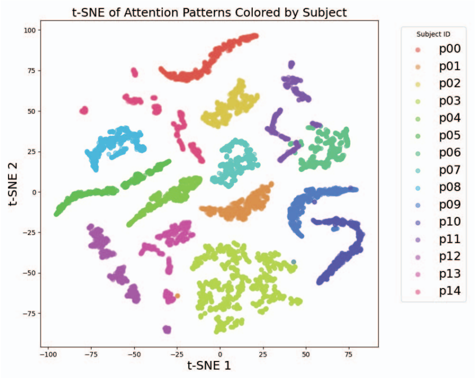


Figure 4. t-SNE visualization of Layer 2 attention patterns showing distinct subject-specific clusters across all subjects from the MPIIFaceGaze dataset[35].

enabling efficient adaptation to new users by leveraging similarity to existing clusters.

4.2. Ablation Study

Feature Extractor Comparison. We evaluate different backbone networks for feature extraction. Table 1 shows MobileNetV3-Small achieves the best accuracy-efficiency trade-off. While EfficientNet-B0 provides comparable accuracy, it significantly increases computational cost. The custom CNN reduces model size but suffers substantial accuracy degradation.

Table 1. Feature extractor comparison on model efficiency and accuracy.

Feature Extractor	Model Size (MB)	Parameters (M)	Angular Error (°)
EfficientNet-B0	37.66	9.83	3.84
CNN	6.56	1.72	5.84
MobileNetV3-Small (ours)	18.65	3.48	**3.39**

Loss Function Analysis. Table 2 compares different loss functions. Huber Loss achieves the lowest angular error (3.39°), demonstrating superior performance compared to MSE, L1, and Angular Error losses.

Table 2. Loss function comparison on MPIIFaceGaze dataset.

Loss Function	Angular Error (°)
L1 Loss	3.51
MSE	3.74
Angular Error	3.54
Huber Loss	**3.39**

GNN vs CNN Relational Module. We replace the GNN with a CNN-based relational module to assess graph-based reasoning effectiveness. Table 3 shows the CNN alternative increases angular error to 4.18°, confirming that structured graph connectivity is crucial for modeling inter-feature relationships in gaze estimation.

Table 3. GNN vs CNN relational module comparison.

Relational Module	Model Size (MB)	Parameters (M)	Angular Error (°)
CNN-based Model	17.32	4.53	4.18
LightGazeNet	18.65	3.48	**3.39**

4.3. Model Training and Error Convergence

LightGazeNet demonstrates consistent performance across all 15 subjects on MPIIFaceGaze dataset, with median angular errors ranging 3-6 degrees (Figure 5). Subjects p00, p02, p04, p05, p06, and p12-p14 show tight interquartile ranges with minimal variance, while subjects p01, p03, p07, p09, and p11 exhibit larger variance due to challenging conditions including extreme poses and illumination variations. The consistent performance validates the model's effectiveness despite subject-specific variations.

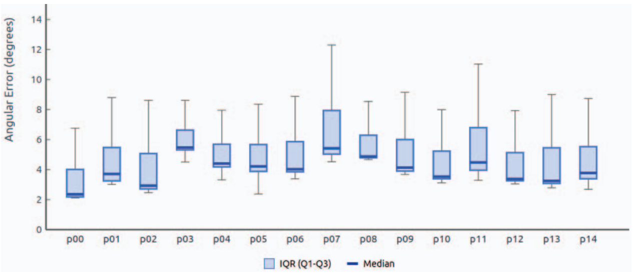


Figure 5. Gaze angular error convergence over 30 training epochs for 15 training subjects of MPIIFaceGaze dataset.

Calibration Training Strategy. During calibration, we employ selective parameter updating to preserve learned representations while enabling personalized adaptation. Only subject-specific calibration embeddings and final output layer parameters are trainable, while all backbone parameters (feature extractors, GNN components, projection layers) remain frozen. This preserves pre-trained feature extraction capabilities while allowing personalized adaptation through minimal parameter updates.

Calibration Module Evaluation. LightGazeNet incorporates a learnable calibration module that adapts to individual users through personalized embedding parameters. Unlike traditional post-processing methods, our approach integrates calibration directly into the network architecture, enabling end-to-end optimization.

Experimental Protocol. We evaluated the effectiveness of the calibration using varying sample sizes ($k \in \{1, 9, 16, 32\}$). For each subject, we reserve the last 500 samples for testing and randomly select k samples from the remaining data for calibration.

Table 4. Calibration results on MPIIFaceGaze dataset. Angular error decreases consistently with more calibration samples.

Calibration Samples (k)	Angular Error (°)	Improvement (%)
Uncalibrated	3.39	-
1	3.28	3.24
9	3.15	7.08
16	3.06	9.73
32	2.99	11.80

Table 4 shows a consistent improvement with increasing calibration samples, achieving 11. 80% error reduction (0.40° improvement) with 32 samples. The robust uncalibrated baseline (3.39°) demonstrates strong generalization, while the calibration module effectively captures user-specific variations without compromising underlying representations.

4.4. Comparative Analysis with baseline Methods

To comprehensively evaluate the effectiveness of the proposed LightGazeNet architecture, we conduct an extensive comparative analysis against baseline gaze estimation methods across multiple benchmark datasets.

For fair and consistent evaluation, all comparative models are assessed on the MPIIFaceGaze[35], EyeDiap[14], and GazeCapture[22] datasets using identical training protocols. Models that employ calibration modules are evaluated using their original calibration strategies with results reported using 16 calibration samples as the standard benchmark, while models without calibration components are evaluated using their native implementations without additional modifications.

Table 5. Subject-wise angular error (degrees) on MPIIFaceGaze. **Bold** indicates average, underline indicates best per subject.

Subjects	L2CSNet [1]	FAZE [27]	PTGE [9]	LightGazeNet
P00	2.87	2.38	2.92	2.02
P01	2.97	3.99	2.74	2.86
P02	3.30	2.60	2.90	2.40
P03	3.89	5.20	5.82	4.28
P04	2.80	4.13	2.78	3.02
P05	4.11	2.74	2.30	2.30
P06	3.22	3.74	3.91	3.27
P07	4.84	5.86	4.57	4.55
P08	4.24	4.30	4.52	4.22
P09	3.32	5.31	5.72	3.51
P10	4.85	4.39	4.61	2.91
P11	4.36	3.29	3.30	2.99
P12	2.45	4.51	2.51	2.87
P13	4.65	4.13	4.04	2.83
P14	5.07	2.76	3.38	2.48
Avg	**3.86**	**3.95**	**3.76**	**3.06**

Table 6. Angular error comparison (degrees) on MPIIFaceGaze and EyeDiap. **Bold** indicates best, underline second-best.

Method	Publication	MPIIFaceGaze	EyeDiap
Mnist[35]	CVPR'15	6.39	7.37
Itracker[22]	CVPR'16	7.33	7.13
GazeNet[36]	TPAMI'17	5.76	6.79
FullFace[35]	CVPRW'17	4.93	6.53
RT-Gene[13]	ECCV'18	4.66	6.02
R-CNN[26]	BMVC'18	4.10	5.31
Dilated-Net[5]	ACCV'19	4.42	6.19
Gaze360[20]	ICCV'19	4.06	5.36
FAZE[27]	ICCV'19	3.95	4.31
AFF-Net[3]	ICPR'20	3.90	6.41
CA-Net[7]	AAAI'20	4.27	5.27
GazeTR-Pure[6]	ICPR'22	4.74	5.72
GazeTR-Hybrid[6]	ICPR'22	4.00	5.17
L2CSNet[1]	ICFSP'23	3.86	3.05
PTGE[9]	JCSE'23	3.76	3.34
LightGazeNet (ours)	-	**3.06**	**2.91**

Table 5 presents subject-wise angular errors on the MPIIFaceGaze dataset, comparing the proposed method with three recent baselines: L2CSNet, FAZE, and PTGE. The proposed model achieves the lowest mean error of 3.06°, outperforming others on 9 out of 15 subjects. Notably, it delivers significant improvements for individual subjects (e.g., P00, P10, and P14) while maintaining competitive performance on others, indicating consistent accuracy across diverse individuals and gaze patterns. Table 6 extends the evaluation across MPIIFaceGaze and EyeDiap datasets, where the method establishes new state-of-the-art results with mean angular errors of 3.06° and 2.91° respectively. These results surpass recent competitive approaches such as PTGE and L2CSNet, demonstrating superior cross-dataset performance and adaptability to diverse experimental conditions and dataset characteristics.

Table 7. Distance error comparison on GazeCapture dataset (cm). **Bold** indicates best, underline second-best.

Method	Publication	Phone	Tablet	Full Dataset
Itracker[22]	CVPR'16	1.86	2.81	2.34
FAZE[27]	ICCV'19	1.96	2.84	1.73
AFF-Net[3]	ICPR'20	1.62	**2.30**	1.96
GazeTR-Hybrid[6]	ICPR'22	1.87	2.74	2.31
PTGE[9]	JCSE'23	1.68	2.84	1.88
L2CSNet[1]	ICFSP'23	**1.53**	2.51	1.77
LightGazeNet (ours)	-	1.59	2.41	**1.69**

For mobile device evaluation, Table 7 presents distance error results on the GazeCapture dataset, which comprises gaze data collected from both phones and tablets. The proposed approach achieves the lowest overall error of 1.69 cm across the full dataset, demonstrating superior performance compared to existing methods. It also attains a strong balance between device types, ranking second-best on both the phone subset with an error of 1.59 cm and the tablet subset with an error of 2.41 cm. This consistent accuracy across different device categories underscores the method's robustness to variations in hardware and user-device interaction, a key factor for practical deployment in real-world mobile gaze tracking applications.

Cross-Dataset Evaluation. To assess generalization capability, we train our model on GazeCapture[22] and evaluate on MPIIFaceGaze[35]. This protocol tests the model's ability to generalize across different data collection methodologies, demographic distributions, and imaging conditions.

Figure 6 shows cross-dataset performance with initial 3.71° angular error in zero-shot settings, improving to 2.76° with strategic calibration using only 32 target samples (25.7% error reduction). The adaptation follows power-law decay with most gains achieved at K=16 samples (2.97°), indicating efficient convergence with minimal target exposure.

Model Complexity vs. Accuracy Trade-off. We analyze

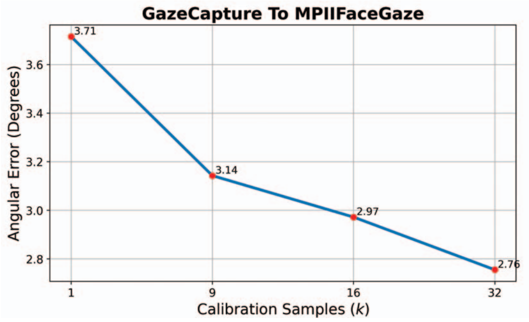


Figure 6. Cross-dataset evaluation: Angular error vs. calibration samples when training on GazeCapture and testing on MPIIFaceGaze. Performance improves with strategic calibration using minimal target samples.

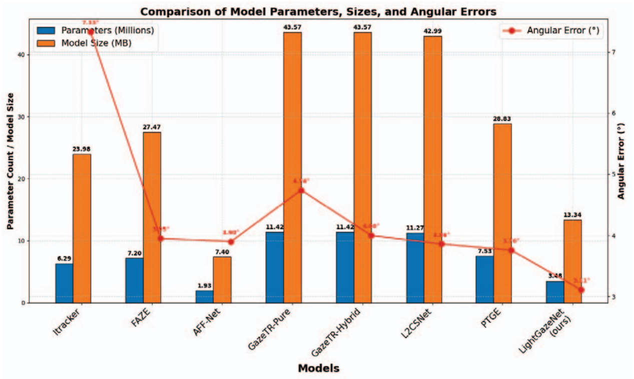


Figure 7. Parameter-accuracy trade-off analysis comparing LightGazeNet with state-of-the-art gaze estimation methods. Our approach achieves superior accuracy with significantly reduced computational complexity.

computational efficiency alongside accuracy to assess practical applicability. Figure 7 demonstrates the parameter-accuracy trade-off across methods. Transformer-based approaches (GazeTR-Pure/Hybrid) achieve 4.74°/4.00° angular error but require 11.42M parameters (43.57MB). L2CSNet attains 3.86° error with comparable complexity (11.27M parameters). PTGE reduces parameters to 7.53M (28.83MB) while maintaining 3.76° accuracy. LightGazeNet achieves optimal performance: 3.06° angular error with only 3.48M parameters (13.34MB). This represents 54% parameter reduction compared to PTGE while improving accuracy by 18.6%. The computational efficiency enables real-time inference on resource-constrained platforms without sacrificing estimation accuracy.

Qualitative Evaluation: Performance Spectrum and Robustness.

Figure 8 illustrates the effectiveness of our model under diverse challenging conditions. Despite motion blur, occlusion, and non-frontal head poses, the predicted gaze vectors remain closely aligned with ground truth, resulting in low angular errors. In EyeDiap[14] samples (a–c), large head rotations introduce asymmetry and reduce eye visibility; yet, the model maintains accuracy by leveraging head pose and facial geometry. Samples (d–e) demonstrate stable predictions under mild tilt and shadow. In GazeCapture[22] images, the model performs well even under motion blur (f), occlusion (g, j), and reduced eye visibility (h, i). MPIIFaceGaze[35] samples (k) and (m) involve low lighting, while (l) presents limited visual cues from the eyes—yet all achieve accurate gaze predictions. Sample (n) includes occlusion from accessories and spectacle shadows, while (o) presents an extreme gaze angle with facial occlusion—both still resulting in acceptable performance. These results highlight the model's ability to generalize under challenging conditions by leveraging cross-modal cues, even when individual inputs are degraded.

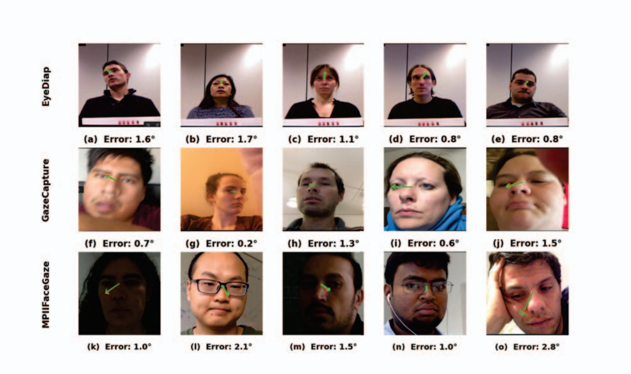


Figure 8. Qualitative results showing gaze prediction under visual uncertainties such as blur, occlusion, and head pose variation. Samples (a–e) are from EyeDiap; (f–j) from GazeCapture and (k -o) MPIIFaceGaze.

5. Conclusion

In this paper, we introduced LightGazeNet, a lightweight and efficient graph-based framework for appearance-based gaze estimation. By explicitly modeling the relationships between facial features, eye cues, and head pose using Graph Neural Networks with multi-head attention, our method achieves competitive accuracy with significantly fewer parameters. LightGazeNet is designed to run in real time on resource-constrained platforms, making it well-suited for practical applications across diverse environments. We believe LightGazeNet will serve as a valuable resource for the gaze estimation community and inspire further research into graph-based multi-modal fusion for robust, and efficient visual attention modeling. Future work will focus on incorporating temporal information and exploring self-supervised adaptation to further enhance robustness and usability in unconstrained settings.

Acknowledgments

The authors gratefully acknowledge Akeso Eyecare for providing the computational resources and support necessary for this research. Portions of the research in this paper used the EYEDIAP Dataset made available by the Idiap Research Institute, Martigny, Switzerland [14]. We are also grateful to the authors of the public gaze estimation datasets [22, 35] for their contributions to the research community.

References

[1] Ahmed A Abdelrahman, Thorsten Hempel, Aly Khalifa, Ayoub Al-Hamadi, and Laslo Dinges. L2cs-net: Fine-grained gaze estimation in unconstrained environments. In 2023 8th International Conference on Frontiers of Signal Processing (ICFSP), pages 98–102. IEEE, 2023.

[2] S Akshay, S Dhanush, G Nath Aswin, and J Amudha. Eyehelp: Predicting an aoi based on eye gaze for patient assistance. Procedia Computer Science, 258:1486–1495, 2025.

[3] Yiwei Bao, Yihua Cheng, Yunfei Liu, and Feng Lu. Adaptive feature fusion network for gaze tracking in mobile tablets. In 2020 25th International Conference on Pattern Recognition (ICPR), pages 9936–9943. IEEE, 2021.

[4] Carlos Bermejo, Dimitris Chatzopoulos, and Pan Hui. Eyeshopper: Estimating shoppers' gaze using cctv cameras. In Proceedings of the 28th ACM international conference on multimedia, pages 2765–2774, 2020.

[5] Zhaokang Chen and Bertram E Shi. Appearance-based gaze estimation using dilated-convolutions. In Asian Conference on Computer Vision, pages 309–324. Springer, 2018.

[6] Yihua Cheng and Feng Lu. Gaze estimation using transformer. In 2022 26th International Conference on Pattern Recognition (ICPR), pages 3341–3347. IEEE, 2022.

[7] Yihua Cheng, Shiyao Huang, Fei Wang, Chen Qian, and Feng Lu. A coarse-to-fine adaptive network for appearance-based gaze estimation. In Proceedings of the AAAI conference on artificial intelligence, pages 10623–10630, 2020.

[8] Govind R Chhimpa, Ajay Kumar, Sunita Garhwal, and Dhiraj Kumar. Real-time human–computer interface based on eye gaze estimation from low-quality webcam images: integration of convolutional neural networks, calibration, and transfer learning. Digital Scholarship in the Humanities, 40 (1):64–74, 2025.

[9] Seung Hoon Choi, Donghyun Son, Yunjong Ha, Yonggyu Kim, Seonghun Hong, and Taejung Park. Looking to personalize gaze estimation using transformers. Journal of Computing Science and Engineering, 17(2):41–50, 2023.

[10] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. Detecting attended visual targets in video. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 5396–5406, 2020.

[11] Senuri De Silva, Sanuwani Dayarathna, Gangani Ariyarathne, Dulani Meedeniya, Sampath Jayarathna, and Anne MP Michalek. Computational decision support system for adhd identification. International Journal of Automation and Computing, 18(2):233–255, 2021.

[12] Giuseppe Maurizio Facchi, Giuliano Grossi, Alessandro D'Amelio, Francesco Agnelli, Chiarella Sforza, Gianluca Martino Tartaglia, and Raffaella Lanzarotti. Graph neural networks for 3d facial morphology: Assessing the effectiveness of anthropometric and automated landmark detection. Pattern Recognition Letters, 2025.

[13] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In Proceedings of the European conference on computer vision (ECCV), pages 334–352, 2018.

[14] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In Proceedings of the symposium on eye tracking research and applications, pages 255–258, 2014.

[15] Dan Witzner Hansen and Qiang Ji. In the eye of the beholder: A survey of models for eyes and gaze. IEEE transactions on pattern analysis and machine intelligence, 32(3):478–500, 2009.

[16] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415, 2016.

[17] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF international conference on computer vision, pages 1314–1324, 2019.

[18] Tanmay Jain, Samiksha Bhatia, Chandan Sarkar, Priyanka Jain, and NK Jain. Real-time webcam-based eye tracking for gaze estimation: Applications and innovations. In 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), pages 1–7. IEEE, 2024.

[19] Hozaifa Kassab, Mohamed Bahaa, and Ali Hamdi. Gcf: Graph convolutional networks for facial expression recognition. In 2024 Intelligent Methods, Systems, and Applications (IMSA), pages 166–171. IEEE, 2024.

[20] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In Proceedings of the IEEE/CVF international conference on computer vision, pages 6912–6921, 2019.

[21] Yuki Kodama, Yasutomo Kawanishi, Takatsugu Hirayama, Daisuke Deguchi, Ichiro Ide, Hiroshi Murase, Hidehisa Nagano, and Kunio Kashino. Localizing the gaze target of a crowd of people. In Computer Vision–ACCV 2018 Workshops: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers 14, pages 15–30. Springer, 2019.

[22] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2176–2184, 2016.

[23] Jiahui Liu, Jiannan Chi, Huijie Yang, and Xucheng Yin. In the eye of the beholder: A survey of gaze tracking techniques. Pattern Recognition, 132:108944, 2022.

[24] Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato. Adaptive linear regression for appearance-based gaze esti-

mation. IEEE transactions on pattern analysis and machine intelligence, 36(10):2033–2046, 2014.

[25] Hu-Chuan Lu, Guo-Liang Fang, Chao Wang, and Yen-Wei Chen. A novel method for gaze tracking by local pattern model and support vector regressor. Signal Processing, 90 (4):1290–1299, 2010.

[26] Cristina Palmero, Javier Selva, Mohammad Ali Bagheri, and Sergio Escalera. Recurrent cnn for 3d gaze estimation using appearance and shape cues. arXiv preprint arXiv:1805.03064, 2018.

[27] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-shot adaptive gaze estimation. In Proceedings of the IEEE/CVF international conference on computer vision, pages 9368–9377, 2019.

[28] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. IEEE transactions on neural networks, 20(1): 61–80, 2008.

[29] Lei Shi, Cosmin Copot, and Steve Vanlanduit. Gaze gesture recognition by graph convolutional networks. Frontiers in Robotics and AI, 8:709952, 2021.

[30] Mohsen Shirpour, Steven S Beauchemin, and Michael A Bauer. A probabilistic model for visual driver gaze approximation from head pose estimation. In 2020 IEEE 3rd Connected and Automated Vehicles Symposium (CAVS), pages 1–6. IEEE, 2020.

[31] B Sindhu, K Ambika, M Sumalatha, and N Sindhuri. A novel multimodal biometric fusion for enhanced personal authentication. In International Conference on Artificial Intelligence and Smart Energy, pages 238–245. Springer, 2025.

[32] Hyo Sik Yoon, Na Rae Baek, Noi Quang Truong, and Kang Ryoung Park. Driver gaze detection based on deep residual networks using the combined single image of dual near-infrared cameras. IEEE Access, 7:93448–93461, 2019.

[33] Shengming Zhang, Zhenwei Zhang, Renqiong Xu, Xuetao Wei, and Jiaxin Zhang. Supporting the development of contactless mental health assessment: An explorations of the relationships between gaze patterns, depression, anxiety, and insomnia. In International Conference on Human-Computer Interaction, pages 417–426. Springer, 2025.

[34] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4511–4520, 2015.

[35] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pages 51–60, 2017.

[36] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. IEEE transactions on pattern analysis and machine intelligence, 41(1):162–175, 2017.

[37] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Revisiting data normalization for appearance-based gaze estimation. In Proceedings of the 2018 ACM symposium on eye tracking research & applications, pages 1–9, 2018.