INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2026, VOL. 72, NO. 1, PP. 1–8
Manuscript received October 14, 2025; revised February, 2026. doi: 10.24425/ijet.2026.157892

HybridGaze: A naturalistic dataset and multistream model for robust gaze estimation

Michał Chwesiuk, and Piotr Popis

***Abstract*—Gaze estimation plays a central role in computer vision and human-computer interaction, enabling applications in assistive systems, attention modeling, and human-robot collaboration. However, existing datasets often rely on infrared-based hardware, are collected in constrained laboratory environments, or lack precise synchronization between stimuli and gaze data, which limits model generalization to real-world conditions.**

To address these challenges, we present HybridGaze - an open-source eye tracking dataset collected using a Tobii tracker combined with a standard RGB webcam. The recordings are processed into eye images and facial landmarks, providing synchronized gaze annotations and facial information across a variety of visual tasks. By capturing gaze data in naturalistic settings, the dataset reflects real-world visual behavior and serves as a valuable benchmark for gaze estimation research.

Furthermore, we introduce GazeModalNet, a multi-stream neural network that estimates gaze direction from two complementary sources: eye images and facial landmarks. Together, the dataset and model establish a strong foundation for developing robust, multimodal gaze estimation systems beyond laboratory constraints.

***Keywords*—eye tracking; appearance-based gaze estimation; feature-based gaze estimation; eye tracking dataset; Deep learning; data synchronization**

I. Introduction

GAZE estimation is a key research direction in computer vision and human-computer interaction (HCI), with applications spanning assistive technologies, attention analysis [1], and human-robot interaction [2]. Accurate gaze estimation models rely on large-scale, high-quality datasets that capture real user behavior under diverse conditions. However, many existing datasets suffer from limitations: they are often collected in constrained laboratory environments, rely on specialized infrared hardware [3], or lack precise synchronization between visual stimuli and eye movements [4]. These constraints hinder the development of models that generalize well to naturalistic, real-world settings.

In this work, we introduce a new dataset designed to bridge this gap. It was collected using a Tobii eye tracker [5] in combination with a standard RGB webcam, providing synchronized gaze annotations and facial recordings. Participants were instructed to fixate on predefined screen targets, enabling us to establish ground-truth mappings between video frames and gaze coordinates. This setup addresses a key challenge: while most state-of-the-art research leverages expensive infrared trackers [3], real-world applications typically operate with commodity cameras. Our dataset therefore provides a valuable resource for developing and benchmarking methods that move closer to practical deployment.

Beyond the dataset itself, we release a modular software framework for experiment design and data collection. The framework integrates eye tracking hardware with webcams, extracts facial landmarks in real time, and supports calibration routines with live feedback. It is designed to be flexible and extensible, allowing other researchers to adapt it to diverse experimental protocols without rebuilding core components from scratch.

We demonstrate the utility of the dataset through a set of supervised learning experiments, training gaze estimation models to regress gaze targets from eye and face images. Our evaluation covers multiple model architectures and includes cross-participant experiments, highlighting the dataset's potential for calibration-free and person-independent gaze prediction. These results establish strong performance baselines and illustrate how the dataset supports research beyond controlled calibration tasks.

To facilitate progress in this field, we are releasing both the dataset and the collection framework to the research community. While our benchmark evaluation focuses on the calibration subset, the full dataset also contains naturalistic viewing conditions, enabling exploration of new research directions such as unconstrained gaze tracking and domain adaptation.

II. Related Works

Eye tracking research has advanced significantly in recent years, driven by improved datasets and more robust gaze estimation models that perform reliably under real-world conditions. Most approaches fall into two main categories: appearance-based and feature-based methods. Appearance-based techniques use raw RGB images of the face or eyes and employ deep neural networks to directly predict gaze direction. These methods require large-scale data but benefit from end-to-end learning without relying on explicit assumptions about eye geometry. Feature-based methods, in contrast, extract interpretable geometric or visual features-such as eye corners, pupils, or facial landmarks-and use these to infer gaze. While they typically need less data and offer better interpretability,

Authors are with Warsaw University of Technology, Poland (e-mail: Michal.Chwesiuk@pw.edu.pl, piotr.popis2.stud@pw.edu.pl).

they are more susceptible to performance degradation under varying lighting, occlusions, or anatomical differences. Recently, hybrid approaches that combine image data with geometric features have shown promising results, particularly for person-independent gaze estimation.

Before the dominance of deep learning, most gaze estimation systems relied on feature-based methods that explicitly modeled the geometry of the eye. These approaches typically detect key landmarks of the eye, such as pupil center (PC), eyelid contours, or the reflection of light on the cornea, known as the corneal reflection (CR) or glints. Gaze direction is often inferred through geometric relationships between these features, most commonly using the pupil-center-corneal-reflection vector (PC-CR) [6], [7], to reduce the influence of head movements during gaze estimation. These landmarks are typically detected using infrared illumination to suppress ambient light interference and enhance contrast, enabling more accurate detection of pupil centers and glints. Classic algorithms such as Starburst and Świrski's model-based tracker estimate pupil contours using intensity gradients and ellipse fitting [8]. Other approaches employ polynomial regression or 3D geometric models of the eyeball-camera system to map image-space features to screen coordinates [6]. While these methods are highly interpretable and data-efficient, their performance deteriorates under real-world conditions involving variable illumination, occlusion, or non-frontal head poses. To improve robustness, many systems introduced user-specific calibration procedures that estimate personalized geometric parameters or use multiple infrared light sources to obtain more stable glint configurations [9]. Despite their accuracy under controlled conditions, these systems are typically limited to fixed setups with constrained head movements, which restricts their applicability in unconstrained or mobile scenarios. Nevertheless, the precision and interpretability of feature-based approaches continue to make them valuable for medical, automotive, and human-computer interaction research, where controlled environments remain common. This transition from handcrafted geometric models to data-driven representations laid the foundation for the appearance-based methods discussed below and inspired recent hybrid models that combine both paradigms.

Several studies continue to use MPIIGaze as a benchmark for appearance-based gaze estimation [10]. It contains over 213,000 RGB images from 15 participants collected over three months of natural laptop use. During data collection, participants were periodically prompted to fixate on on-screen targets, allowing the capture of natural variations in gaze direction, head pose, and lighting conditions. GazeCapture [11] expanded data diversity through large-scale crowdsourcing using an iOS application. The resulting dataset-comprising over 2.4 million frames from 1,474 users-introduced unprecedented variability in devices, lighting conditions, and head poses. It enabled the development of iTracker, a CNN-based model that combines facial images, eye crops, and a face grid to predict gaze in real time on mobile devices without calibration. RT-GENE [12] further advanced gaze estimation by capturing unstructured, natural viewing behavior using eye tracking glasses paired with a Kinect v2 camera. Motion capture markers ensured accurate gaze annotations, and a GAN-based approach was employed to reconstruct occluded facial regions. Hybrid and multimodal datasets have since gained attention for their ability to leverage both geometric and appearance cues. ETH-XGaze [13] and Gaze360 [14] extended gaze estimation to large-scale, 3D, and in-the-wild conditions, enabling models to generalize across wide head-pose ranges. These efforts underscore a growing shift toward datasets that combine controlled calibration with naturalistic scenarios-an approach also adopted in our work.

More recently, CrossGaze [15] improved 3D gaze estimation using a dual-encoder architecture that processes face and eye features separately before fusing them via cross-attention. Trained solely on Gaze360, it achieved a mean angular error of 9.94° on the challenging Front 180° subset, highlighting strong cross-domain generalization. AGE-Net [16] explores differences between left and right eyes using a dual-branch network that leverages asymmetric features at multiple levels. Zhao et al. [17] focus on domain generalization, using auxiliary training branches and loss functions to improve performance without requiring target domain data during training. Privacy concerns in gaze tracking have also received attention. PrivatEyes [18] combines federated learning with secure multiparty computation to protect user privacy during model training. They show that privacy-preserving methods can match traditional approaches while preventing information leakage. The same authors proposed DualView, a GAN-based technique for measuring privacy risks in gaze data. Additionally, Adebayo et al. [19] investigated self-supervised pretraining for gaze estimation. Their approach leverages representations learned on AFFECTNet and fine-tunes them on MPIIFaceGaze and Gaze360, demonstrating improved person-independent generalization under leave-one-person-out evaluation.

Several comprehensive surveys have helped organize this field. Lei et al. [20] review gaze estimation specifically for mobile devices, covering the entire pipeline from camera input to user interaction. They highlight mobile-specific challenges like lighting variability and device movement, advocating for lightweight and adaptive solutions. Bozkir et al. [21] provide a large-scale review of eye tracking in VR and AR, covering over 1,300 papers. Their survey creates a taxonomy of methods and discusses privacy concerns, hardware limitations, and the trade-offs between accuracy and real-time performance. They emphasize the need for privacy-aware frameworks, especially in immersive environments.

Recent work has also emphasized multimodal learning strategies that combine eye, face, and contextual cues to improve robustness under unconstrained conditions. However, few datasets provide synchronized multimodal signals collected with both dedicated eye tracking hardware and standard RGB cameras. HybridGaze aims to address this gap by enabling the study of gaze estimation models that bridge controlled and naturalistic scenarios.

III. Methodology

This chapter presents the complete experimental methodology used to develop and evaluate the proposed gaze estimation

framework. It describes the processes of feature extraction, neural architecture design, model optimization, and performance evaluation. The approach integrates both appearance-based and geometric modalities to exploit complementary visual and spatial information. Image features provide fine-grained details of the eye region, while geometric landmarks preserve global facial structure and gaze geometry. Together, these representations form a hybrid learning pipeline designed to improve accuracy, robustness, and person-independence in gaze estimation across diverse conditions.

A. Feature Selection and Data Representation

To overcome the limitations of single-modality approaches, the proposed method integrates both visual and geometric features. Purely image-based models often degrade under large head-pose variations or partial occlusions, whereas geometry alone lacks the fine-grained information needed to capture subtle eye movements. By combining the two modalities, we exploit their complementary strengths to achieve more stable and accurate gaze estimation. We use MediaPipe's Face Mesh [22] to extract facial landmarks(Figure 1). From these, we select a subset of 40 key landmarks from the eyes region specifically optimized for gaze estimation. This selective approach reduces computational complexity while retaining the most gaze-relevant geometric information.

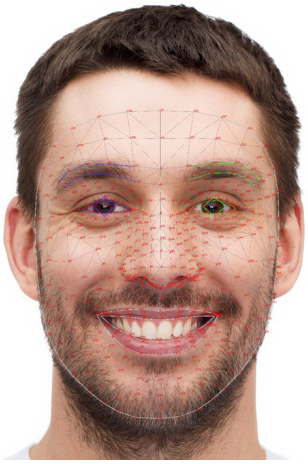

Fig. 1. MediaPipe Face Mesh showing 478 facial landmarks with emphasis on periocular regions and iris tracking points.

B. Neural Architecture Design

At the core of the proposed framework lies a hybrid deep learning model, **GazeModalNet**, specifically designed to combine appearance-based and geometry-based cues for gaze estimation. It integrates visual features extracted from eye images with spatial information derived from facial landmarks, enabling accurate gaze prediction even under challenging head poses and illumination conditions. The network follows a modular, multi-stream design in which each stream specializes in a distinct data modality, later fused into a unified representation for final gaze regression.

The architecture of the model comprises three main branches (Figure 2). The first branch processes the left and right eye images through two convolutional neural network (CNN) streams that extract appearance-based features. The second branch encodes geometric information from selected facial landmarks using a fully connected network. Finally, the outputs of all streams are merged in a feature fusion head that integrates appearance and geometry into a unified gaze representation. The following sections describe each component in detail.

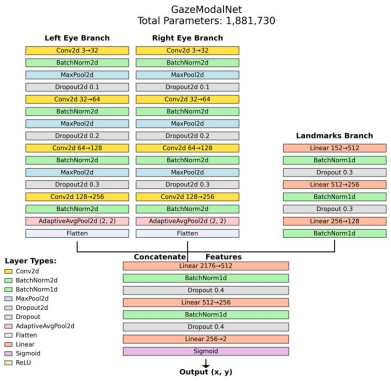

Fig. 2. GazeModalNet architecture: binocular CNN streams for eye images, a fully connected stream for geometric features, and a fusion network for final gaze prediction.

a) Eye Processing Streams: Two independent CNN branches process the left and right RGB eye images of size $(3 \times 224 \times 224)$. Both branches share weights and follow an identical configuration consisting of four convolutional blocks summarized in Table I. Each block includes a convolution layer, batch normalization, ReLU activation, max pooling, and dropout. The final block applies adaptive average pooling followed by flattening to produce a compact feature representation for each eye.

TABLE I
OVERVIEW OF THE CNN EYE STREAM ARCHITECTURE SHOWING THE LAYER CONFIGURATION AND OPERATIONS OF EACH CONVOLUTIONAL BLOCK USED TO EXTRACT APPEARANCE-BASED FEATURES FROM EYE IMAGES

Step	Block 1	Block 2	Block 3	Block 4
Conv	Conv(3 → 32)	Conv(32 → 64)	Conv(64 → 128)	Conv(128 → 256)
BatchNorm	BatchNorm2d(32)	BatchNorm2d(64)	BatchNorm2d(128)	BatchNorm2d(256)
Activation	ReLU	ReLU	ReLU	ReLU
Pooling	MaxPool2d(2)	MaxPool2d(2)	MaxPool2d(2)	AdaptiveAvgPool2d((2,2))
Dropout	Dropout2d(0.1)	Dropout2d(0.2)	Dropout2d(0.3)	Flatten

b) Landmark Processing Stream: While MediaPipe provides 478 facial landmarks, only 40 key points from the periocular region are selected for gaze estimation. The selected landmarks are flattened into a 1-D vector of size 80 (40 landmarks × 2 coordinates) and processed through a fully connected network composed of three dense layers, as shown in Table II. This stream captures the geometric configuration of the eyes and surrounding features while maintaining low computational cost.

c) Feature Fusion Network: The outputs from both CNN branches (each producing a 1024-dimensional feature vector) and the 128-dimensional landmark embedding are concatenated into a single 2176-dimensional feature representation (1024 + 1024 + 128). This combined vector is passed through

TABLE II
STRUCTURE OF THE LANDMARK PROCESSING STREAM. THE FULLY CONNECTED LAYERS TRANSFORM 2D LANDMARK COORDINATES INTO A COMPACT EMBEDDING REPRESENTING THE GEOMETRIC CONFIGURATION OF THE EYE REGION

Layer	Linear	BatchNorm1d	Activation	Dropout
Dense 1	Dense(80 → 512)	BatchNorm1d	ReLU	Dropout(0.3)
Dense 2	Dense(512 → 256)	BatchNorm1d	ReLU	Dropout(0.3)
Dense 3	Dense(256 → 128)	BatchNorm1d	ReLU	–

the fusion head detailed in Table III, which performs high-level integration and final gaze regression. Dropout regularization and batch normalization are applied at each layer to prevent overfitting and stabilize convergence.

TABLE III
STRUCTURE OF THE FEATURE FUSION NETWORK. THE CONCATENATED APPEARANCE AND GEOMETRIC EMBEDDINGS ARE INTEGRATED THROUGH FULLY CONNECTED LAYERS TO PRODUCE THE FINAL GAZE PREDICTION

Layer	Linear	BatchNorm1d	Activation	Dropout / Output
Dense 1	Dense(2176 → 512)	BatchNorm1d	ReLU	Dropout(0.4)
Dense 2	Dense(512 → 256)	BatchNorm1d	ReLU	Dropout(0.4)
Output	Dense(256 → 2)	–	Sigmoid	–

C. Training Protocol

The model is optimized using the AdamW optimizer with an initial learning rate of $\eta_0 = 1 \times 10^{-3}$. The learning rate is reduced by half whenever validation performance plateaus, subject to a lower bound:

$$\eta_t = \max(\eta_{\min}, 0.5 \cdot \eta_{t-1}), \quad \eta_{\min} = 10^{-6}.$$

To stabilize convergence, gradient clipping with a threshold of $\tau = 1.0$ is applied:

$$\nabla' = \min\left(1, \frac{\tau}{\|\nabla\|}\right) \cdot \nabla.$$

A hierarchical dropout strategy with rates ranging from 0.1 to 0.4 is employed across network components to improve generalization. Early stopping is triggered if the validation loss does not improve for ten consecutive epochs. The evolution of training loss and learning rate is shown in Figure 3, illustrating the stable convergence of the proposed model.

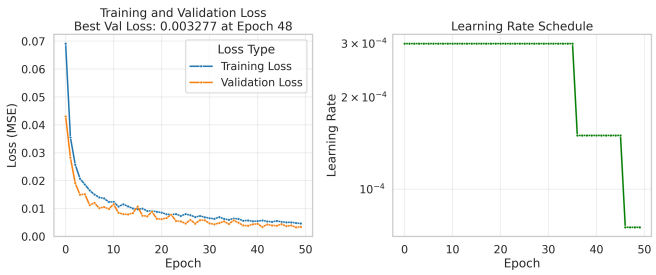

Fig. 3. Visualization of train loss and learning rate during model optimization.

D. Data Processing and Evaluation

All RGB eye images are resized to a fixed resolution of 224×224 pixels and retained in color format during training. Facial landmark coordinates are normalized relative to the screen resolution, while the eye images are normalized only by scaling pixel values to the [0, 1] range. The dataset is divided into training, validation, and test subsets.

Fig. 4. Step-by-step visualization of the preprocessing pipeline, including landmark extraction and eye image cropping.

Unlike conventional eye tracking evaluations that report angular accuracy and precision in degrees of visual angle, such metrics cannot be reliably computed in our setup due to variable participant distance from the laptop screen and the absence of precise camera calibration parameters. Since intrinsic and extrinsic parameters of the webcam relative to the display were not fixed or measured for each session, converting screen-space predictions into angular measures would introduce uncontrolled geometric error. Moreover, the recordings were conducted under heterogeneous real-world conditions, with variations in lighting, head pose, and device positioning, further limiting the consistency of angular metrics across sessions. Therefore, we evaluate performance directly in the normalized screen coordinate space using distance- and error-based measures, which provide a stable and reproducible basis for cross-subject comparison.

Model performance is evaluated using three error-based metrics that jointly quantify the magnitude and robustness of prediction errors: mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|.$$

Additionally, the Euclidean distance between predicted and true gaze coordinates is computed as

$$d_i = \sqrt{(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2},$$

where (x_i, y_i) and $(\hat{x}_i, \hat{y}_i)$ denote the true and predicted normalized screen positions, respectively.

To evaluate spatial precision, we report the percentage of predictions satisfying $d_i < \tau$ for thresholds τ of 1%, 5%, 10%, 15%, and 20% within the normalized screen coordinate system. In this work, *Accuracy@*τ represents the proportion of gaze predictions falling within a circular tolerance region of radius τ centered at the ground-truth gaze location. For example, *Accuracy@10%* denotes the percentage of predictions within a circle whose radius equals 10% of the normalized screen dimension. This metric provides an interpretable measure of spatial precision and directly relates to practical application requirements.

Together, these preprocessing and evaluation procedures ensure consistent normalization across samples and provide a robust quantitative basis for performance comparison.

IV. Dataset Acquisition

This chapter describes the process of data acquisition, recording setup, and statistical properties of the introduced **HybridGaze** dataset. The proposed dataset was designed to provide high-quality, synchronized gaze data that bridges laboratory precision with real-world applicability. It combines structured calibration tasks with naturalistic viewing behaviors, enabling comprehensive training and evaluation of appearance-based gaze estimation models. The synchronized acquisition of gaze coordinates and facial video ensures a direct correspondence between visual features and gaze targets.

A. Collection Procedure

Fifteen participants (9 male, 6 female, aged 21-52) took part in structured recording sessions, each lasting approximately 15 minutes. Participants were seated at a viewing distance of 55-65 cm from the laptop screen, consistent with standard gaze-tracking setups.

Each recording session consisted of five consecutive phases organized into three distinct types of tasks: two calibration sequences, one smooth-pursuit task, and two natural viewing segments. This combination of structured and free-viewing conditions was designed to capture a wide spectrum of gaze behaviors, ranging from precise fixations to spontaneous, context-driven eye movements.

Phase 1: *Calibration (3x3 grid).* Participants fixated sequentially on nine equally spaced targets presented on a 3×3 grid to establish an initial mapping between gaze coordinates and screen positions. Each target was displayed for a fixed duration while corresponding gaze samples and facial video frames were recorded.

Phase 2: *Free viewing (Movie 1).* Participants watched a five-minute segment of the open-source animated film *Big Buck Bunny* (2008) [23], eliciting natural gaze dynamics in response to motion and scene changes.

Phase 3: *Smooth pursuit (Lissajous trajectory).* A circular marker followed a Lissajous trajectory defined by sine components:

$$x(t) = 0.8 \sin(at) + 0.1,$$

$$y(t) = 0.8 \sin(bt) + 0.1,$$

where $a = 3.0$ and $b = 2.0$. The motion was remapped from normalized screen-space to pixel coordinates, resulting in a continuous and smooth tracking pattern.

Phase 4: *Free viewing (Movie 2).* A five-minute excerpt from *Sintel* (2010) [24] introduced greater visual and emotional diversity.

Phase 5: *Calibration (5x5 grid).* The session concluded with 25 fixation targets on a 5×5 grid to refine spatial calibration and assess session consistency.

All phases were presented in fullscreen mode with synchronized acquisition from a Tobii eye tracker and an RGB webcam (1280 x 720 px, 30 FPS). Frame timestamps ensured precise alignment between gaze coordinates and video frames. The Tobii tracker recorded 2D screen-space gaze points, while facial data were captured simultaneously and processed with MediaPipe Face Mesh to extract 478 landmarks per frame (468 facial and 10 iris). Individual eye regions were cropped around the outermost periocular landmarks with 10% padding to preserve eyelid and skin context.

Recordings were performed on a 17.3-inch laptop (2560 x 1440 px display) equipped with a Tobii eye tracker and 720p webcam. Although the framework is hardware-agnostic, all sessions used the same device to ensure consistent geometry and viewing conditions. The setup was relocated between participants while preserving identical hardware and software configurations. Figure 5 shows an example of the recording setup used during data acquisition.

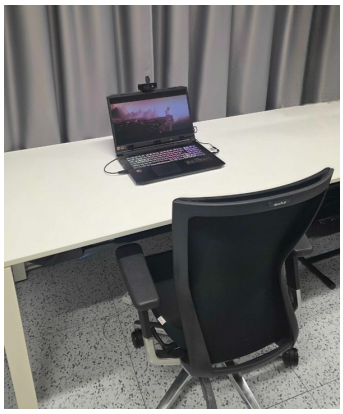

Fig. 5. Spatial distribution of collected gaze data (left) and target marker positions (right).

B. Dataset Characteristics

The acquired dataset contains 151,128 samples collected from 15 participants across five session types. After filtering, 140,424 samples (92.9%) were retained as valid, while 10,704 samples (7.1%) were discarded due to negative coordinates. An additional 0.8% of samples fall outside screen boundaries but are not marked as invalid.

The spatial distribution of gaze samples and corresponding target positions is shown in Figure 6. Participants contributed comparable amounts of data (Figure 7), confirming balanced participation across subjects.

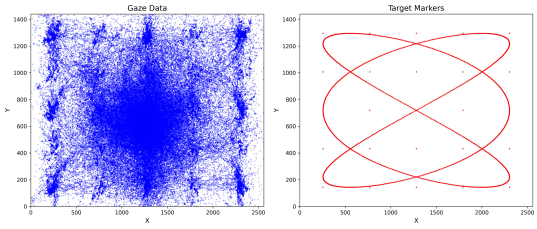

Fig. 6. Spatial distribution of collected gaze data (left) and target marker positions (right).

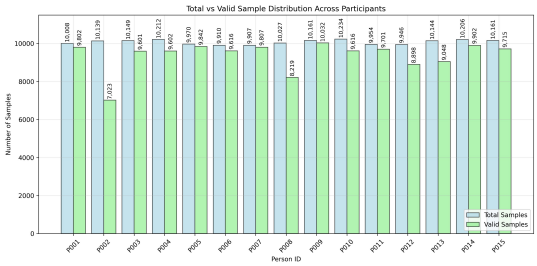

Fig. 7. Distribution of collected samples across participants before (blue) and after (green) filtering.

The varying complexity and duration of the five experimental phases are reflected in the session-wise statistics shown in Figure 8. The free-viewing movie sessions (*video1* and *video2*) contain the largest number of samples due to their longer duration and the continuous gaze dynamics they elicit. Calibration phases (3×3 and 5×5) provide precisely annotated gaze-target correspondences that serve as strong supervised learning signals, while the smooth-pursuit phase captures continuous temporal dynamics essential for modeling natural eye movement trajectories.

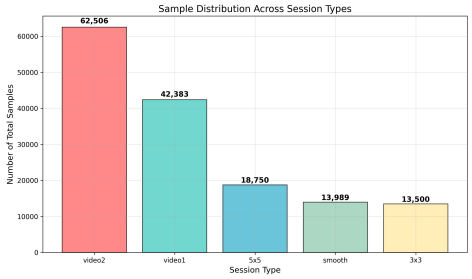

Fig. 8. Sample distribution across different session types.

The **HybridGaze** dataset will be made publicly available to promote reproducible research and enable benchmarking across varied viewing conditions. By combining structured calibration, dynamic tracking, and naturalistic free-viewing data, it provides a comprehensive foundation for developing and evaluating hybrid appearance-based gaze estimation models.

V. Implementation

A. Data Collection and Eye Tracking

The eye tracking framework, implemented in Python, uses a lightweight client-server architecture for synchronized capture of gaze coordinates and facial imagery. Its core component, the `EyeTracker` class, runs a TCP/IP socket server that continuously receives gaze data in screen-space from the hardware, ensuring low latency and easy device integration. In parallel, the software captures RGB video streams from a standard camera and synchronizes them with the gaze data, enabling frame-accurate alignment between visual content and eye movement. MediaPipe Face Mesh was used for real-time landmark extraction, ensuring compatibility between training and inference. Eye crops and landmark coordinates were automatically normalized across screen resolutions and camera configurations.

Calibration procedures are defined in external configuration files specifying marker layouts, timing, and collection parameters, providing flexibility while maintaining consistency across sessions. The recorded gaze data serve both as ground-truth labels for supervised training and as a reference for comparing hardware- and model-based gaze estimation.

Data acquisition and processing ran in real time on a laptop with an AMD Ryzen 7 5800H CPU, 32 GB RAM, and an NVIDIA GeForce RTX 3070 GPU. This setup enabled simultaneous 720p video recording, Tobii gaze capture, and online landmark extraction with real-time feedback. The same hardware was used for model inference, ensuring consistent capture and evaluation conditions. Overall, the system supports continuous acquisition and prediction at interactive frame rates, suitable for both laboratory and portable use.

B. Experiment Tracking and Workflow Management

To manage the complexity of training multiple model configurations and hyperparameter combinations, MLflow [25] was integrated for experiment tracking. MLflow automatically logs all training metrics, model parameters, and system details for each training run. This functionality proved particularly useful for monitoring gradient behavior and learning rate schedules, enabling early detection and mitigation of training instabilities. The built-in visualization tools further facilitated cross-experiment comparison, helping to identify performance trends and optimal configurations.

Fig. 9. MLflow experiment tracking interface displaying training metrics and parameter evolution.

For data preprocessing orchestration, Prefect [26] was employed to coordinate different preprocessing configurations. Rather than relying on multi-step pipelines, Prefect manages the execution of various preprocessing tasks defined in YAML configuration files. This approach streamlined experimentation with alternative data preparation strategies while ensuring reliable task scheduling and monitoring. The workflow orchestration provided by Prefect ensured consistency across experimental runs and simplified scaling to larger datasets.

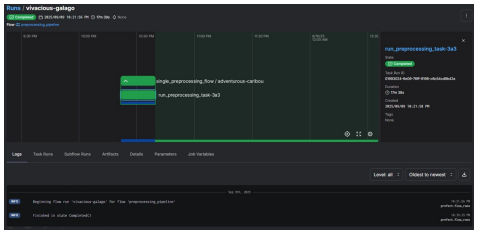

Fig. 10. Prefect dashboard visualizing preprocessing workflow execution and task monitoring.

VI. Results

All experiments presented in this section were conducted on a curated subset of the collected dataset, focusing exclusively on samples from the 3×3 and 5×5 calibration grids. Video-based data were excluded to ensure controlled experimental conditions and computational efficiency. The evaluation employs complementary metrics to assess model performance from multiple perspectives. Distance-based measures include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Distance, and Median Distance, which quantify prediction accuracy in coordinate space. Accuracy metrics report the percentage of predictions falling within circular tolerance regions at thresholds of 1%, 5%, 10%, 15%, and 20% of screen dimensions, as defined in Section III.

A. Resubstitution Performance

Resubstitution evaluation measures model performance on the same data used for training, providing an upper-bound estimate of achievable accuracy and indicating the degree of potential overfitting. While not suitable for assessing generalization, this analysis establishes a theoretical performance ceiling under ideal conditions.

TABLE IV
RESUBSTITUTION PERFORMANCE: ERROR METRICS (TRAINING = TEST SET)

Metric	MSE	RMSE	MAE	Mean Dist.	Median Dist.
Value	0.0020	0.0444	0.0265	0.0430	0.0311

TABLE V
RESUBSTITUTION PERFORMANCE: QUANTITATIVE ERROR METRICS CALCULATED ON THE TRAINING DATA. THESE VALUES REPRESENT THE UPPER PERFORMANCE BOUND OF THE PROPOSED GAZE ESTIMATION MODEL UNDER IDEAL (NON-GENERALIZED) CONDITIONS

Accuracy Threshold	1%	5%	10%	15%	20%
Value	12.09%	72.12%	94.15%	97.48%	98.58%

The resubstitution results demonstrate near-optimal performance with high accuracy across all tolerance thresholds. As shown in Table IV, the error metrics indicate precise and low-variance predictions. Accuracy values (Table V) confirm that over 94% of gaze estimates fall within 10% of screen space, highlighting the model's theoretical upper bound when evaluated on training data.

B. Group-Based K-Fold Cross-Validation

Group-based K-fold cross-validation provides robust statistical estimates of model performance through systematic data partitioning while maintaining subject independence. Our 5-fold validation divides the 15 subjects into 5 groups, ensuring that each subject's data appears in exactly one fold. This approach prevents data leakage between training and validation sets while assessing model stability across different subject groups within our dataset. To ensure computational efficiency and fair comparison across folds, each model was trained for a maximum of 20 epochs during cross-validation experiments.

TABLE VI
FIVE-FOLD GROUP-BASED CROSS-VALIDATION PERFORMANCE SUMMARY

Value	**MSE**	**RMSE**	**MAE**	**Acc@5%**	**Acc@10%**	**Acc@20%**
Mean	0.0545	0.2331	0.1674	4.72%	53.43%	76.43%
Std Dev	0.0067	0.0141	0.0179	2.50%	7.32%	9.70%

The cross-validation results show consistent model performance across folds, with low standard deviation across error and accuracy metrics. Moderate inter-fold variability reflects natural differences between participant groups but indicates overall model stability. As shown in Table VI, the model maintains robust performance at practical accuracy thresholds (10-20%), while precision decreases for stricter criteria, reflecting the inherent noise in gaze and landmark measurements.

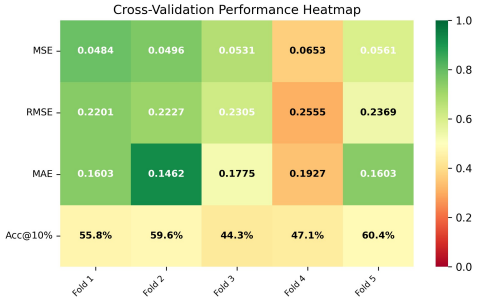

Fig. 11. Cross-validation performance heatmap showing metric variance across folds.

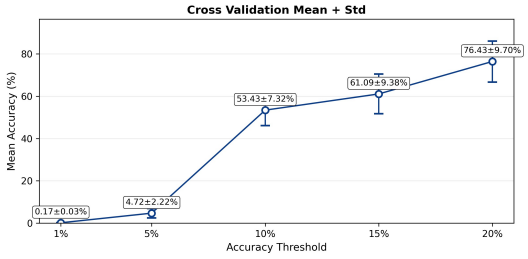

Fig. 12. Accuracy threshold analysis demonstrating model performance across tolerance levels.

C. Leave-One-Subject-Out Evaluation

The Leave-One-Subject-Out (LOSO) protocol provides a rigorous assessment of person-independent generalization. For each of 15 participants, the model was trained on data from 14

and tested on the remaining one, ensuring every subject served once as the test case. This setup closely simulates real-world conditions where the system must generalize to unseen users.

TABLE VII
LOSO EVALUATION: ERROR METRICS AVERAGED ACROSS FOLDS, REPRESENTING PERSON-INDEPENDENT MODEL PERFORMANCE

Metric	MSE	RMSE	MAE	Mean Dist.	Median Dist.
Value	0.0679	0.2606	0.0940	0.1412	0.0613

TABLE VIII
LOSO EVALUATION: ACCURACY AT MULTIPLE TOLERANCE THRESHOLDS

Threshold	1%	5%	10%	15%	20%
Accuracy	5.54%	48.12%	74.58%	82.80%	92.92%

LOSO results show a moderate increase in prediction error compared to intra-subject and cross-validation setups, as expected for person-independent testing. Still, the model achieves over 74% accuracy within a 10% tolerance, confirming strong generalization to unseen users. Overall, the proposed hybrid model demonstrates consistent performance across all validation protocols, combining high accuracy with practical generalization suitable for real-world deployment.

VII. CONCLUSION

This work introduced a new **HybridGaze** dataset for eye tracking research and a hybrid deep learning model for gaze estimation. The dataset combines structured calibration sequences with smooth pursuit and natural viewing phases, providing a comprehensive foundation for training and evaluating appearance-based and geometry-aware gaze prediction approaches. Building on this dataset, the proposed hybrid architecture **GazeModalNet** integrates image-based and geometry-based representations to predict gaze positions with high precision. Experimental results demonstrated robust and consistent performance across multiple validation protocols, including cross-subject and LOSO evaluations, confirming the model's generalization ability.

Future work will expand the dataset to include more participants and environmental variability, enabling evaluation under broader gaze behaviors. Enhancements such as head-pose estimation can improve gaze estimation in real-life scenarios, and real-time optimization through model compression will further improve applicability to interactive systems.

REFERENCES

[1] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 185–207, 2012.

[2] H. Admoni and B. Scassellati, "Social eye gaze in human-robot interaction: a review," *Journal of Human-Robot Interaction*, vol. 6, no. 1, pp. 25–63, 2017.

[3] K. A. Funes Mora, F. Monay, and J.-M. 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*, 2014, pp. 255–258.

[4] H. Griffith, D. Lohr, E. Abdulin, and O. Komogortsev, "Gazebase, a large-scale, multi-stimulus, longitudinal eye movement dataset," *Scientific Data*, vol. 8, no. 1, p. 184, 2021.

[5] Tobii Technology, "Tobii eye tracker: Advanced eye tracking solutions," 2024, accessed: 2024-12-22. [Online]. Available: https://www.tobii.com/

[6] E. D. Guestrin and M. Eizenman, "General theory of remote gaze estimation using the pupil center and corneal reflections," *IEEE Transactions on biomedical engineering*, vol. 53, no. 6, pp. 1124–1133, 2006.

[7] C. Zhang, J. Chi, Z. Zhang, X. Gao, T. Hu, and Z. Wang, "Gaze estimation in a gaze tracking system," *Science China Information Sciences*, vol. 54, no. 11, pp. 2295–2306, 2011.

[8] L. Świrski, A. Bulling, and N. Dodgson, "Robust real-time pupil tracking in highly off-axis images," in *Proceedings of the symposium on eye tracking research and applications*, 2012, pp. 173–176.

[9] D. Li, J. Babcock, and D. J. Parkhurst, "openeyes: a low-cost head-mounted eye-tracking solution," in *Proceedings of the 2006 symposium on Eye tracking research & applications*, 2006, pp. 95–100.

[10] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 4511–4520.

[11] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Baluja, D. Mermelstein, J. Trinkaus, A. Monroy, M. Koskela, G. Wilson *et al.*, "Eye tracking for everyone," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2176–2184.

[12] T. Fischer, H. J. Chang, and Y. Demiris, "Rt-gene: Real-time eye gaze estimation in natural environments," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[13] X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang, and O. Hilliges, "Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation," in *European conference on computer vision*. Springer, 2020, pp. 365–381.

[14] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, "Gaze360: Physically unconstrained gaze estimation in the wild," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6912–6921.

[15] A. Cătrună, A. Cosma, and E. Rădoi, "Crossgaze: A strong method for 3d gaze estimation in the wild," 2024. [Online]. Available: https://arxiv.org/abs/2402.08316

[16] Y. Shi, F. Zhang, W. Yang, G. Wang, and N. Su, "Agent-guided gaze estimation network by two-eye asymmetry exploration," in *2024 IEEE International Conference on Image Processing (ICIP)*, 2024, pp. 2320–2326. [Online]. Available: https://doi.org/10.1109/ICIP51287.2024.10648029

[17] R. Zhao, P. Tang, and S. Luo, "Improving domain generalization on gaze estimation via branch-out auxiliary regularization," 2024. [Online]. Available: https://arxiv.org/abs/2405.01439

[18] M. Elfares, P. Reisert, Z. Hu, W. Tang, R. Küsters, and A. Bulling, "Privateyes: appearance-based gaze estimation using federated secure multi-party computation," *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. ETRA, pp. 1–23, 2024.

[19] S. Adebayo, J. C. Dessing, and S. McLoone, "Slyklatent: A learning framework for gaze estimation using deep facial feature learning," 2024. [Online]. Available: https://arxiv.org/abs/2402.01555

[20] Y. Lei, S. He, M. Khamis, and J. Ye, "An end-to-end review of gaze estimation and its interactive applications on handheld mobile devices," *ACM Computing Surveys*, vol. 56, no. 2, pp. 1–38, Sep. 2023. [Online]. Available: http://dx.doi.org/10.1145/3606947

[21] E. Bozkir, S. Özdel, M. Wang, B. David-John, H. Gao, K. Butler, E. Jain, and E. Kasneci, "Eye-tracked virtual reality: A comprehensive survey on methods and privacy challenges," 2023. [Online]. Available: https://arxiv.org/abs/2305.14080

[22] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee *et al.*, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.

[23] T. Roosendaal, "Big buck bunny," in *ACM SIGGRAPH ASIA 2008 computer animation festival*, 2008, pp. 62–62.

[24] ——, "Sintel," in *ACM SIGGRAPH 2011 Computer Animation Festival*, 2011, pp. 71–71.

[25] M. A. Zaharia, A. Chen, A. Davidson, A. Ghodsi, S. A. Hong, A. Konwinski, S. Murching, T. Nykodym, P. Ogilvie, M. Parkhe, F. Xie, and C. Zumar, "Accelerating the Machine Learning Lifecycle with MLflow," *IEEE Data Eng. Bull.*, vol. 41, pp. 39–45, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:83459546

[26] Prefect Technologies, "Prefect: The modern data stack for data engineering," 2019, version 3.0+. [Online]. Available: https://www.prefect.io/