*Review*

# Recent Progress on Eye-Tracking and Gaze Estimation for AR/VR Applications: A Review

Liwan Lin [ID], Zongyu Wu, Yijun Lu, Zhong Chen [ID] and Weijie Guo *[ID]

National Innovation Platform for the Fusion of Industry and Education in Integrated Circuits, Department of Electronic Science, School of Electronic Science and Engineering, Xiamen University, Xiamen 361005, China; 33320231150377@stu.xmu.edu.cn (L.L.); 33320221150317@stu.xmu.edu.cn (Z.W.); yjlu@xmu.edu.cn (Y.L.); chenz@xmu.edu.cn (Z.C.)
* Correspondence: wjguo@xmu.edu.cn; Tel.: +86-134-0066-2983

**Abstract**

Visual information is crucial in human life, not only providing critical support for communication, learning, and decision-making, but also playing a key role in psychology, medicine, and science. Eye-tracking and gaze estimation have promoted the development of foveated rendering in wearable virtual reality and augmented reality glasses. This review summarizes the recent development on gaze estimation and discusses the impacts of head posture, illumination, occlusion, blur, and individual bias on the accuracy of eye-tracking. The prospective development on eye-tracking employing unsupervised learning, self-supervised learning, and meta-learning have also been discussed.

**Keywords:** gaze estimation; eye-tracking; head pose; machine learning

## 1. Introduction

The development of eye-tracking technology provides unprecedented opportunities to comprehend the complexity and uniqueness of the human visual system, profoundly affecting various fields, including psychology [1], computer science [2], behavioral science [3], education [4], augmented reality (AR), and virtual reality (VR) glasses [5], among others [6]. In the context of AR and VR, eye-tracking offers unique advantages that are essential for enhancing user experience. It enables more natural and intuitive interactions by allowing users to control virtual objects and navigate interfaces through gaze alone. Additionally, eye-tracking in AR/VR can provide real-time insights into user attention and intent, facilitating adaptive content that responds dynamically to the user's focus, thus improving immersion and engagement in these environments. The process of gaze estimation for AR/VR applications typically includes some key steps in sequence. Firstly, the region of interest in the visual scene is identified. Then, geometric or appearance features are extracted from this area. Finally, a regression function is employed to determine the relationship between these features and the gaze direction. Numerous factors could affect the accuracy of gaze estimation, such as head pose variations, individual biases, blinking, occlusion, and image blur [7–9].

Initially, as shown in Figure 1, the gaze estimation systems relied on bulky mechanical apparatus to estimate the point of gaze (POG) [10]. These early designs, characterized by complex calibration procedures and limited accuracy due to mechanical drift, struggled to meet practical application requirements. The introduction of electrical sensors enabled the detection of electrical signals related to the eyeballs, such as voltage differences between

electrodes on the cornea and retina [11]. However, the hardware remained rudimentary, and the systems still possessed limited accuracy and reliability. The advent of computer vision revolutionized the field by enabling POG estimation through pupil position or eyeball contour, serving as a significant turning point for gaze estimation. After refining algorithms and improving adaptability to various conditions, some systems integrated active infrared illumination and high-speed cameras to track eye movements [12], achieving sub-centimeter accuracy in controlled settings.
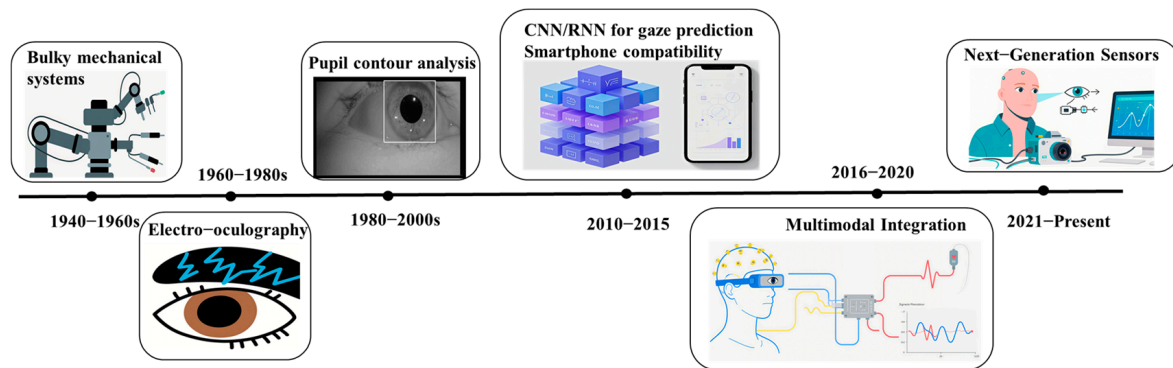


**Figure 1.** Timeline of Eye-Tracking Technology Evolution.

The advent of deep learning has revolutionized gaze estimation methodologies, transitioning from rule-based systems to data-driven architectures. Deep learning models, such as convolutional neural networks (CNNs) [13] and recurrent neural networks (RNNs) [14], have been employed to learn complex eye movement patterns. These methods have significantly improved accuracy and robustness of gaze estimation systems and enabled operation under less constrained environments, such as real-time gaze estimation through smartphone cameras [15].

This review summarizes the developments and prospects of eye-tracking technologies and gaze estimation algorithms, by elucidating their operational principles, performance benchmarks across diverse scenarios, and persistent technical challenges. In Section 2, the classification and principles of gaze estimation have been discussed. Section 3 introduces the primary sources of errors in gaze estimation algorithms and outlines commonly used datasets in the field. In Section 4, a comprehensive review and evaluation of gaze estimation algorithms have been provided. Section 5 discusses the challenges in complex scenes. Lastly, the potential future development has also been discussed.

## 2. Classification of Gaze Estimation

Gaze estimation is typically divided into two subtasks: gaze target estimation and gaze point estimation. Gaze estimation methods are generally classified into appearance-based and model-based approaches. Furthermore, model-based methods can be further categorized into corneal-reflection-based and feature-based approaches.

### 2.1. Gaze Target Estimation

Recasens et al. proposed an approach of gaze target estimation by employing a neural network and the dataset GazeFollow [16]. This method utilized the quantified spatial position of the head and a close-up image of the head to parameterize individuals for gaze prediction. A CNN had also been employed for saliency detection, generating a heatmap for gaze prediction. Notably, this method estimated gaze direction from a third-person perspective in an image, but it is effective only when both the observer and the target object appear within the same frame.

In response to this limitation, Recasens et al. [17] proposed another network that integrates both semantic and geometric understandings of frames, allowing for individual gaze tracking across video frames. To further enhance intra-frame gaze target estimation, body pose [18] and sight direction [19,20] have been utilized. By combining the output of a deep learning-based object detection with frame-by-frame gaze coordinates, Deane et al. [21] developed an automated method for detecting and annotating the content being viewed by users in each frame, thereby eliminating the need for manual intervention.

To address the challenge of out-of-frame gaze targets, Tonini et al. [22] introduced a Transformer-based architecture that automatically detects objects (including heads) within a scene and associates each head with the corresponding gaze target, enabling comprehensive and interpretable gaze analysis. Similarly, Tu et al. [23] proposed a Transformer-based method capable of simultaneously detecting the gaze targets of multiple observers. This approach overcomes the limitations of using only head images as input and significantly improves both accuracy and efficiency.

*2.2. Gaze Point Estimation*

Gaze point estimation refers to the computational process of determining the focal point of an observer. Typically, it estimates the POG on a two-dimensional plane. The mapping $(X_e, Y_e) \rightarrow (X_s, Y_s)$ links these coordinates to the gaze target, where the relationship of mapping function is described by [24].

$$X_S = a_0 + \sum_{p=1}^{n} \sum_{i=0}^{p} a_{(i,p)} X_e^{p-i} Y_e^i, \tag{1}$$

$$Y_S = b_0 + \sum_{p=1}^{n} \sum_{i=0}^{p} b_{(i,p)} X_e^{p-i} Y_e^i, \tag{2}$$

where $(X_e, Y_e)$ are coordinates derived from eye features, $(X_s, Y_s)$ are the corresponding gaze target coordinates [25], and $a_i$ and $b_i$ are coefficients. To optimize this polynomial, users are asked to focus on fixed points. This procedure enables the measurement of the discrepancy between the estimated and actual gaze positions. The primary source of discrepancy is the angular difference between the pupil axis and the visual axis of the eye, commonly referred to as the kappa angle, as illustrated in Figure 2. The smaller anterior segment represents the cornea, which is transparent and contains the pupil. The optical axis is the line connecting the center of the pupil ($P$) to the center of the cornea ($C$).
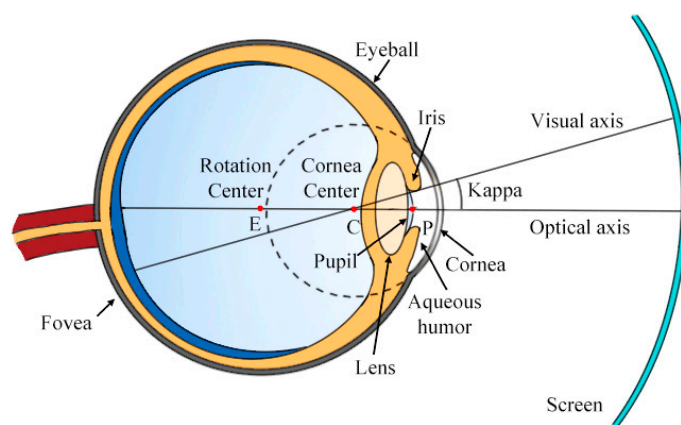


**Figure 2.** Schematic of the human eyeball, showing key anatomical parameters and reference points for gaze tracking.

When the gaze directs to a particular position, the eyes rotate to ensure that light falls precisely on the fovea—an area on the retina with a high concentration of cones [26]. The line connecting the rotation center of the eye to the fovea is known as the visual axis or line of sight. This visual axis intersects with the gaze target, defining the actual gaze vector. However, the visual axis rarely aligns perfectly with the optical axis, leading to an angular discrepancy. This angle varies among individuals, and neglecting it in model design can introduce prediction errors. Therefore, the mapping function must account for these errors, which can be minimized through calibration.

Falch et al. [27] proposed a novel webcam-based approach for gaze estimation on a computer screen. Zhou et al. [28] introduced an optimized deep neural network for 2D gaze estimation on mobile devices, including effective attention modules and metric learning. Building on the work of Krafka et al. [29], He et al. [30] proposed a few-shot personalization method for 2D gaze estimation on devices, along with an unsupervised personalization approach.

## 3. Evaluation Metrics and Datasets

### 3.1. Performance Metrics and Error Sources

The performance of gaze estimation systems is generally evaluated by angular accuracy, mean squared error, and robustness across varying input conditions, which are affected by head posture variations, blinking and occlusion, illumination changes, and inter-subject anatomical differences.

#### 3.1.1. Head Posture

The direction of sight is determined by the rotation of the eyeball, the degree of eyelid movement, and the spatial orientation of the head. In real-world visual interaction scenarios, head posture varies dynamically, directly affecting both the geometric representation of the eye region and the appearance of ocular features within the image frame. Specifically, head rotation and tilt may cause spatial displacement or morphological deformation of the ocular features, thereby reducing the accuracy of gaze estimation [31–36]. Therefore, contemporary eye-tracking systems must not only extract stable ocular features but also integrate mechanisms for real-time head pose compensation. The importance of head posture analysis becomes even more pronounced in AR and VR glasses, in which user interfaces demand uninterrupted and accurate gaze feedback across a wide range of orientations.

Methods proposed in the literature to cope with head posture variation can be broadly categorized into appearance-based, geometry-based, clustering-based, and autoencoder-based approaches.

Appearance-based methods estimate head posture directly from facial images using machine-learned templates or detection arrays. Such methods typically rely on matching an observed face to predefined templates or learned representations of facial configurations. However, visual ambiguity remains a challenge since even subtle differences in head orientation or subject identity can lead to significant prediction errors. To reduce such variance, methods incorporating Laplacian of Gaussian (LoG) filters or Gabor wavelets have been introduced to enhance feature localization and structural stability [31]. For example, Gao et al. [32] proposed a hybrid framework combining CNNs with geometric projection. This framework first classified images into head pose categories and then applied geometric transformations to refine face orientation, effectively bridging classification and estimation under pose variance.

More recently, Liu et al. [33] introduced a Transformer-based architecture, named TokenHPE, which models contextual dependencies among facial key points via tokenized representations. This method learns orientation tokens through attention-weighted aggregation of facial components, enabling robust estimation even under occlusion, dim illumination, or extreme orientations.

Geometry-based approaches formulate gaze estimation as a spatial relationship problem, leveraging the 6-degree-of-freedom (DOF) kinematics of head and eyeball motion to isolate and correct pose-induced distortions through coordinate transformations. Typically, these approaches normalize image appearance by projecting the eye region onto a canonical 2D view using a perspective distortion matrix. Ruzzi et al. [34] proposed a two-stream architecture utilizing conditional neural radiance fields (NeRF) to separately learn volumetric features of the face and eye regions. By applying a rigid 3D rotation matrix to the extracted features and subsequently composing them through a differentiable volume renderer, their model can precisely control gaze angle redirection under various head poses.

Yang et al. [35] developed a monocular vision-based estimation system for unconstrained human gaze, jointly extracting pose features and eye appearance descriptors and fusing them in both spatial and temporal domains to maintain gaze stability across frame sequences. However, degradation in image quality potentially results in prediction deviation.

Clustering-based strategies address pose variance by discretizing the head orientation space and training separate models for each cluster. While this reduces within-cluster variance and improves prediction accuracy, it may introduce discontinuities across cluster boundaries and require significant computational resources. Early implementations used random forest-based models [36], while more recent work has explored graph-based structures. Xin et al. [37] proposed a graph convolutional network (GCN) with an edge-vertex joint attention mechanism (EVA) to enhance intra-cluster relationships and mitigate facial feature detection instability. Their approach was evaluated on multiple benchmark datasets and showed reduced angular error in cross-pose inference. Similarly, Tian et al. [38] employed hierarchical clustering to categorize holding postures, enabling mixed-pose positioning and robust step-length estimation without complex classification procedures.

Lastly, autoencoder-based models advance gaze estimation by generating compressed latent representation of facial and ocular appearance, from which head pose can be implicitly inferred. These models learn compact yet discriminative representations that encapsulate critical pose-relevant information while suppressing noise. Hu et al. [39] proposed a multi-feature fusion gaze estimation model employing group convolution and channel–spatial attention mechanisms (GCCSAM). This framework adaptively selects and enhances relevant features from face and eye inputs, mitigating the impact of asymmetry and misalignment on gaze estimation. Their evaluation on MPIIGaze and EyeDiap datasets yielded average angular errors of 4.1° and 5.2°, respectively, highlighting the potential of attention-augmented encoders for robust gaze modeling. Ren et al. [40] introduced a feature fusion method incorporating multi-level information elements to improve the overall performance of appearance-based gaze estimation models.

These methods significantly advance gaze estimation by compensating for head pose variability, as shown in Table 1. Appearance-based and clustering-based methods are computationally efficient, whereas geometry-based and autoencoder-based models achieve higher robustness in complex motion scenarios at the cost of increased computational load.

**Table 1.** Comparison of methods for head pose estimation: input, dataset, and accuracy metrics.

| | Method | References | Input | | Dataset | Accuracy |
|---|---|---|---|---|---|---|
| | | | Eye | Face | | |
| Head posture | Appearance-based approach | [33] | | √ | BiWi | 4.66° |
| | | [32] | | √ | AFLW | 5.77° |
| | | [41] | | √ | CMU-PIE | 7.36° |
| | | | | √ | MIT-CBCL | 1.31° |
| | | | | √ | YaleB | 7.584° |
| | Geometry-based method | [42] | √ | | unknown | 2.63° |
| | | | √ | | unknown | 3.26° |
| | | [43] | √ | | MPIIGaze | 4.8° |
| | | [35] | | √ | unknown | 7.65° |
| | | [34] | | √ | ColumbiaGaze | 9.464° |
| | | | | √ | MPIIFaceGaze | 14.933° |
| | | | | √ | GazeCapture | 10.463° |
| | Clustering-based method | [36] | | √ | Biwi | 4.9° |
| | | [38] | | √ | unknown | 1.57° |
| | | [37] | | √ | AFLW2000 | 3.48° |
| | | | | √ | 300 W-LP | 3.92° |
| | | | | √ | BIWI | 2.24° |

### 3.1.2. Blinking, Occlusion, and Illumination

Blinking and occlusion constitute critical technical barriers in eye tracking systems, introducing distinct data integrity challenges. Blinking often causes a loss of eye position data, resulting in gaps and inaccuracies in gaze duration tracking. Similarly, occlusion impedes precise capture of eye movements, leading to data loss, deviations in gaze trajectories, and reduced accuracy in gaze point estimation.

To compensate for blinking, occlusion, and illumination effects, researchers have proposed enhanced network architectures capable of learning discriminative and invariant representations even from incomplete or degraded inputs. Among these, attention-guided multi-branch architectures have demonstrated considerable effectiveness. Luo et al. [44] developed CI-Net, a composite network comprising two parallel submodules: the consistency estimation network (C-Net), which captures coarse gaze direction using shared face and eye features; and the inconsistency estimation network (I-Net), which explicitly models residual estimation errors caused by partial occlusion or left–right eye asymmetry. By integrating spatial attention mechanisms across both submodules, CI-Net selectively emphasizes informative regions and suppresses noise from occluded or irrelevant features.

When unilateral eyelid occlusion occurs during blinking, state-of-the-art gaze estimation systems address prediction instability through adaptive input weighting and geometric constraint modeling. For example, CI-Net introduces a balance coefficient to weigh the contribution of each eye according to its occlusion confidence, allowing dynamic adaptation under unilateral eyelid occlusion [44]. Furthermore, residual vector fields have been employed to refine coarse predictions using the asymmetric geometric information between the two eyes.

Beyond local occlusion, structural modeling approaches have been proposed to mitigate uncertainty arising from global occlusion or self-shadowing. Nonaka et al. [45] introduced a cascaded multi-network system that first predicts head and body orientation from full-body pose, then incorporates this information into a gaze direction regressor. By capturing the spatial consistency between body motion and attention behavior, this method establishes a probabilistic prior that reduces ambiguity in gaze direction when facial features are partially unobservable. Their framework significantly improves robustness under global occlusion, especially in surveillance and group activity analysis.

To mitigate illumination-induced variability, several domain-adaptive strategies have been proposed. Cheng et al. [46] employed domain generalization to project gaze features into a shared subspace, minimizing distributional differences across lighting conditions. This approach improves cross-illumination generalization without requiring supervision from the target domain. Liu et al. [47] proposed a differential CNN (Diff-CNN), learning pairwise difference representations between samples from different lighting domains. The network emphasizes illumination-invariant components through subtraction-based feature encoding, effectively filtering out illumination-induced distortions.

Empirical results demonstrate the effectiveness of these approaches. For instance, CI-Net achieves angular errors of 3.8° (MPIIGaze), 5.4° (EYEDIAP), and 7.9° (RT-GENE), outperforming standard CNN regressors under partial occlusion [44]. Probabilistic orientation modeling introduced by Nonaka et al. reduced the mean angular error to 18.9° under full-body occlusion scenarios [45].

In summary, blinking, occlusion, and illumination remain formidable obstacles to reliable gaze estimation, particularly in unconstrained real-world settings. To address these issues, strategies such as attention-enhanced multi-path networks and domain-adaptive convolutional models have significantly enhanced the robustness of modern gaze estimation systems. Nonetheless, the challenge persists unresolved in specific scenarios, particularly real-time multi-user interactions and uncontrolled real-world environments, where frequent and unpredictable occlusion and illumination variations continue to pose fundamental limitations.

### 3.1.3. Inter-Subject Variability

The human visual system varies significantly across individuals, particularly in ocular geometry, corneal curvature, scleral reflectance, and the spatial displacement between optical and visual axes. When deep learning models trained on homogeneous populations encounter unseen individuals, systematic prediction errors emerge. These errors manifest as directionally persistent biases rather than random noise, exhibiting temporal autocorrelation and context-dependent amplification.

To address inter-subject variation, polynomial regression commonly maps image-space gaze coordinates to screen-space targets through learned nonlinear transformations [28]. While polynomial regression effectively personalizes gaze mappings, its performance depends on sufficient per-subject calibration data. Calibration-based approaches generally achieve high individual accuracy by requiring users to perform explicit calibration procedures (e.g., fixating on predefined targets), but this process can be time-consuming and may reduce the overall user-friendliness and scalability of the system, especially in real-time or consumer-grade applications.

To overcome these limitations, recent research has explored calibration-free methods, which aim to eliminate or minimize the need for explicit user calibration. Such approaches often utilize large-scale population data, robust deep learning architectures, or user-adaptive transformation layers to generalize across subjects without subject-specific calibration sessions. Bao et al. [48] proposed a personalized gaze estimation model, insert-

ing the subject-wise feature modulation layers into the backbone network. These layers adaptively transform the shared gaze representation based on the embedded identity features, allowing the model to account for inter-subject differences without re-training. Li et al. [49] introduced an event-based calibration-free gaze tracking system for wearable platforms. Their method bypasses the need for explicit per-user calibration by leveraging a high-frequency dynamic vision sensor (DVS) to capture transient eye movements at 950 Hz. Through mapping from temporal pupil contour trajectories to gaze angles, their system achieved a sub-degree estimation error without requiring individualized parameter fitting. Specifically, the reported angular error reached $0.46°$. Few-shot learning frameworks have emerged as a paradigm for personalized adaptation with minimal calibration overhead. Zhang et al. [50] proposed a meta-learning-based framework that requires fewer than five calibration samples to adapt lightweight parameter layers embedded within a shared backbone architecture. The fine-tuning process is constrained via meta-learning objectives to ensure rapid convergence and robustness to overfitting. Experimental validation demonstrated that with merely two calibration samples, the framework achieved a $1.5°$ reduction in mean angular error during cross-subject gaze estimation tasks, outperforming conventional calibration-dependent methods. Recent advancements in adaptive kappa angle estimation have enabled robust inference of angular disparities between visual and optical axes directly from ocular biometrics, circumventing reliance on explicit visual fixation labels. Zhang et al. [51] proposed a flexible, calibration-free gaze estimation method that jointly constructs the optical axis projection (OAP) and visual axis projection (VAP) planes. By using the OAP as an eye feature to predict the VAP, the method achieves linearity with natural gaze patterns, resulting in consistent 3D gaze estimation with significantly improved accuracy. While calibration-free methods significantly improve usability and adaptability, they may sacrifice a certain degree of individual accuracy compared to well-calibrated systems.

Overall, the profound impact of inter-subject variation on gaze estimation accuracy continues to drive algorithmic innovation. While calibration-based methods offer precise compensation, their operational cost limits scalability. In contrast, calibration-free, domain-invariant, and personalized learning approaches have emerged as promising alternatives, enabling robust estimation across diverse populations and unconstrained use conditions. The comparative evaluation of these methods on benchmark datasets—such as MPIIGaze, XGaze, and EVE—has demonstrated their effectiveness with personalized models reducing angular error from $4.14°$ to $2.88°$ [48,49,51].

As shown in Table 2, gaze estimation performance was systematically evaluated under multiple factors, including blinking, occlusion, illumination, and inter-individual variability.

**Table 2.** Evaluation of gaze estimation under various factors: blinking, occlusion, illumination, and individual differences. AP: average precision.

| | References | Input | | | Dataset | Accuracy |
|---|---|---|---|---|---|---|
| | | Eye | Face | Other | | |
| Blinking | [52] | √ | | | RT-BENE | AP:0.757 |
| | | | | | Eyeblink8 | AP:0.997 |
| | [53] | | √ | | RT-BENE | AP:0.653 |
| Occlusion | [45] | | | Head position | GAFA | 20.4° (3D) |
| | | | | Body image | MoDiPro | 25.6° (2D) |
| | [44] | √ | √ | | MPIIGaze | 3.8° |
| | | | | | EYEDIAP | 5.4° |
| | | | | | RT-Gene | 7.9° |

**Table 2.** *Cont.*

| References | Input | | | Dataset | Accuracy |
| --- | --- | --- | --- | --- | --- |
| | **Eye** | **Face** | **Other** | | |
| [46] | | √ | | MPIIGaze | 5.20° |
| | | | | EYEDIAP | 7.36° |
| [47] | √ | | | MPIIGaze | 4.67° |
| | | | | EYEDIAP | 3.36° |
| | | | | UT-Multiview | 4.33° |
| [48] | √ | | | EVE | 1.89° |
| | | | | MPIIGaze | 4.14°/3.02° |
| | | | | Xgaze | 2.88° |
| [54] | √ | | | - | 0.9° |
| [55] | √ | | | - | 4° |

*Illumination* spans the [46] and [47] reference groups; *Individual Differences* spans the [48], [54], and [55] reference groups.

*3.2. Dataset Limitations and Domain Generalization*

The development and evaluation of gaze estimation algorithms depend fundamentally on the availability of high-quality, large-scale, and demographically diverse datasets. Unlike image recognition tasks where labels are easily obtained, gaze dataset construction requires precise spatial alignment between gaze targets and eye images, typically acquired under uncontrolled environmental conditions characterized by extreme head pose variations, heterogeneous illumination, and individual differences. These operational constraints increase the costs of data collection and annotation, resulting in limited dataset sizes and in structural inconsistencies across benchmarks [56]. One representative strategy for collecting large-scale in-the-wild data is exemplified by MPIIGaze, which utilizes an experience sampling mechanism on participants' personal laptops to prompt gaze targets across various daily contexts. Participants are periodically asked to fixate on predefined spots in-screen while their eye images are captured. Although this method enables long-term data accumulation, it introduces label uncertainty due to user distraction and changes in environmental conditions [57]. The inherent noise in such data necessitates post-processing and filtering to improve annotation reliability, which in turn limits dataset scalability.

In response to the high cost and variability of real-world gaze data collection, synthetic datasets have emerged as a complementary solution. UnityEyes, for instance, renders photo-realistic images of parameterized 3D eye models under variable illumination, head pose, and gaze direction, thereby offering precise annotations at large scale [58]. Additionally, SP-EyeGAN generates temporally plausible gaze sequences by modeling fixation and saccadic dynamics through two GAN modules—FixGAN and SacGAN—where each sub-network is tailored to distinct types of gaze transitions [59]. These synthetic datasets provide rich supervision signals, reduce the burden of manual annotation, and support pre-training for downstream tasks such as gaze zone estimation and visual saliency modeling. Despite their controllability and scalability, synthetic gaze datasets remain fundamentally constrained by systemic domain divergence from real-world deployment scenarios, where variations in ocular biometrics, skin texture, sensor noise, optical aberrations, and subject behavior introduce pronounced distributional divergence. As a result, models trained on synthetic data often suffer from degraded generalization performance when evaluated on natural images. This phenomenon, referred to as domain shift, also affects models transferred across different real-world datasets, such as MPIIGaze, GazeCapture, and EYEDIAP, due

to differences in hardware, environmental lighting, scene composition, and individual differences [60].

To address the domain shift in gaze estimation, recent methodologies have focused on domain adaptation and generalization. Bao et al. [61] proposed a rotation-enhanced unsupervised domain adaptation (RUDA) approach, in which rotated source samples are projected into a shared latent space while enforcing gaze direction consistency. A novel loss function penalizes angular deviations between original and rotated feature representations, encouraging the model to learn rotation-invariant gaze encodings robust to head poses and datasets. This rotation-based constraint enables the model to align gaze semantics without requiring explicit labels from the target domain. Beyond direct domain alignment, self-supervised techniques have also been integrated to leverage unlabeled data from the target domain. Cai et al. [62] introduced an uncertainty-aware passive adaptation framework that iteratively refines gaze predictions through epistemic uncertainty minimization. Unlike supervised approaches, their method removes the dependency on labeled target data by propagating confidence estimations across unlabeled sequences, enabling fully unsupervised and passive domain adaptation.

Despite recent advancements in cross-dataset gaze estimation frameworks, significant challenges remain. The absence of standardized evaluation protocols, the lack of unified metrics for quantifying domain similarity, and the reliance on target domain samples—even if unlabeled—limit their deployment in truly unconstrained real-world systems. Furthermore, methods such as adversarial learning [60] and ensemble modeling, while effective in certain scenarios, often incur substantial computational expenses and significant memory overheads, limiting the practical deployment on resource-constrained platforms.

### 3.3. Public Gaze Datasets

The availability of diverse, well-annotated gaze datasets is fundamental for training, evaluating, and benchmarking gaze estimation models. Publicly available datasets exhibit distinct characteristics in collection protocols, participants, acquisition environments, imaging modalities, and annotation density. These differences affect both the learning dynamics of data-driven models and their generalizability across domains. This section reviews representative gaze datasets commonly used in the literature, highlighting their structural properties and collection methods.

The MPIIGaze dataset [57] consists of 213,659 binocular eye images collected from 15 participants, with each monocular image sized at a resolution of $60 \times 35$ pixels. Data acquisition was conducted using laptop-integrated webcams during daily activities, where subjects were tasked with fixating on moving dots displayed on a screen. The number of images per participant varied significantly, ranging from 1498 to 34,745. By incorporating diverse lighting conditions and natural variations in head pose, this dataset effectively supports gaze estimation in real-world (in-the-wild) scenarios. To expand the annotated feature set beyond ocular regions, the MPIIFaceGaze dataset [63] was introduced as an extension of MPIIGaze dataset. Comprising 37,667 facial images collected from the same cohort of 15 participants, this dataset is enriched with facial landmarks and pupil centers.

Synthetic datasets have been developed to overcome the limitations of data scarcity and labor-intensive annotation. UnityEyes [58] synthesizes highly realistic eye images through a rendering pipeline that integrates high-resolution 3D facial scans with physically based modeling of eyeball geometry and material reflectance. The dataset provides pixel-level annotations, including iris center, eyelid contour, and pupil location, under user-defined gaze angles and lighting configurations. ColumbiaGaze [64] contains 5880 high-resolution images of 56 participants aged between 18 and 36. Participants were photographed under five horizontal head poses ($0°, \pm15°, \pm30°$) and seven gaze directions

($0°$, $\pm 5°$, $\pm 10°$, $\pm 15°$). The dataset includes annotations for gaze vector, head pose, and eyeglass usage. The controlled acquisition environment, combined with high-resolution imaging, enables accurate evaluation of gaze estimation algorithms, particularly in cross-pose settings. Nonetheless, the limited number of gaze angles and constrained viewing conditions reduce ecological validity. Gaze360 [65] was designed to capture unconstrained gaze behaviors across wide head pose ranges and environments. It comprises over 172,000 RGB images collected from 238 participants under indoor and outdoor scenes, with 3D gaze annotations covering a wide angular range. Captured across nine sessions, it offers diversity in lighting, backgrounds, and head positions. As one of the few large-scale datasets with continuous 3D gaze annotations, Gaze360 serves as a benchmark for evaluating the robustness of gaze estimation under real-world head and gaze dynamics. ETH-XGaze [66] addresses the need for dense sampling of gaze targets and head poses. It includes over 1 million high-resolution ($6000 \times 4000$ pixels) images captured using 18 synchronized DSLR cameras. A total of 110 participants viewed stimuli while their heads were stabilized by a chin rest. The dataset offers full 3D gaze vectors, facial landmarks, and head pose annotations, enabling fine-grained analysis of gaze across extreme head orientations. While ETH-XGaze provides excellent geometric variation, it is important to note that the implementation of head fixation restricts natural behavior and gaze spontaneity. UT-Multiview [67] contains 1.2 million eye images collected from 50 participants under controlled indoor settings. Participants were asked to fixate on a red cross embedded within a white circle, which appeared sequentially on a $16 \times 10$ grid across a monitor. Head positions were stabilized using chin rests, and images were captured from varying viewpoints. The systematic structure and grid-based target layout make UT-Multiview suitable for geometric modeling and gaze mapping studies, although the rigid head fixation may limit generalization to dynamic environments. EYEDIAP [68] provides RGB-D data collected from 16 participants recorded in laboratory conditions. It includes 94 video sequences with varying head movements and gaze behaviors, collected using a Kinect sensor and HD camera. Ground-truth gaze is obtained via LED markers and 3D calibration. EYEDIAP supports both 2D and 3D gaze estimation and is often used to validate models under depth-aware and motion-rich conditions. RT-GENE [69] combines RGB and depth modalities for real-time gaze estimation. It consists of 277,286 annotated eye images collected from 15 participants, with ground truth labels derived from mobile eye-tracking glasses. The dataset was collected using a Kinect v2 RGB-D sensor and offers synchronized eye gaze, depth maps, and full-face images. The multimodal nature allows for hybrid approaches that fuse appearance and geometric features. GazeCapture [29] is currently one of the largest publicly available datasets, comprising over 2.5 million eye images collected from 1474 participants using mobile devices. During data collection, participants tracked a moving on-screen target while the front-facing camera recorded eye images. The dataset encompasses extensive variations in gaze angle, device pose, and illumination, rendering it highly suitable for training deep appearance-based estimation models. However, label noise and head uncertainties in pose estimation complicate precise performance evaluation under real-world conditions. TabletGaze [70] comprises 816 videos from 51 participants in four different postures. Participants were instructed to fixate on dynamic targets while holding a tablet at various positions, including standing, sitting, and lying. Each video captures time-varying gaze behavior and eye-region appearance, supporting temporal modeling and real-time gaze tracking applications. In surveillance and egocentric scenarios, datasets such as GAFA [45] and GFIE [56] capture gaze behaviors from freely moving individuals. GAFA comprises 882,000 annotated video frames from indoor and outdoor recordings, providing 3D gaze labels, head orientation, and body movement cues. GFIE includes 71,799 frames from 61 participants and annotates both 2D and 3D gaze points alongside head bounding

boxes. These datasets emphasize unconstrained gaze patterns and support studies on long-range attention estimation and multimodal fusion.

These datasets differ significantly in sample size, imaging modality, head pose variability, and gaze annotation format. Their complementary properties highlight the necessity of cross-dataset training and evaluation to assess the robustness and adaptability of gaze estimation algorithms. Tables 3 and 4 provide a comprehensive comparative overview of these datasets, facilitating the selection of appropriate benchmarks based on experimental requirements. Figure 3 depicts ecological authenticity vs. annotation complexity. Figure 4 presents a normalized performance comparison of five representative datasets (MPIIGaze, MPIIFaceGaze, Gaze360, ETH-XGaze, GazeCapture) in terms of seven features, including the number of participants, data volume, yaw angle range, pitch angle range, distance, lighting conditions, and full-face inclusion status.
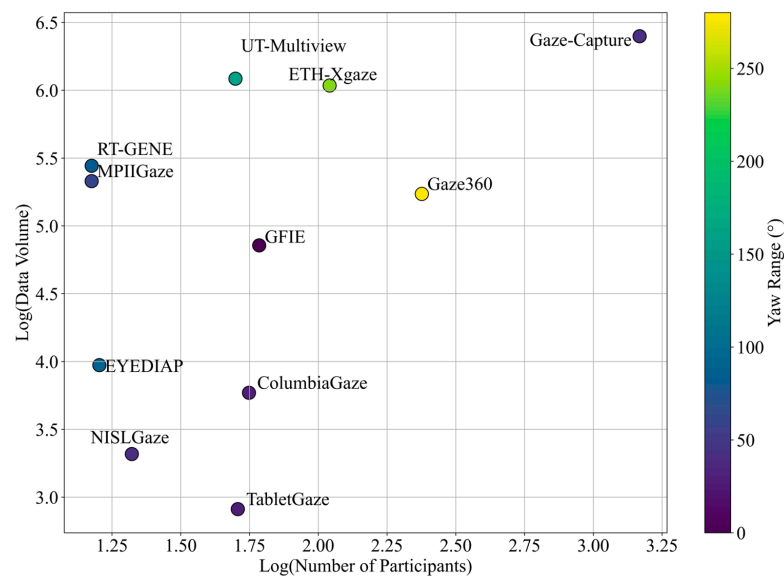


**Figure 3.** Ecological and Annotation Complexity Distribution of Gaze Datasets. The 2D scatter plot positions datasets by participant count (x-axis) and data volume (y-axis), with color intensity representing the yaw angle range. Symbol size denotes pitch angle variability, collectively capturing dataset scale (axes) and gaze diversity (visual encodings).
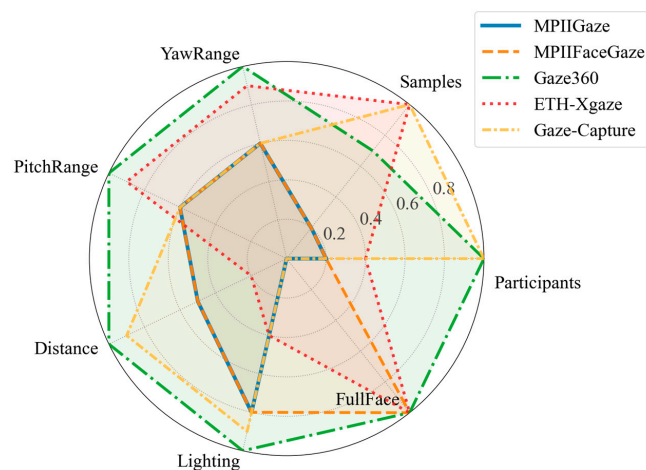


**Figure 4.** Normalized Multidimensional Comparison of Gaze Datasets. The radar chart compares five datasets (MPIIGaze, MPIIFaceGaze, Gaze360, ETH-XGaze, GazeCapture) across seven normalized features: participant count, data volume, yaw range ($\pm°$), pitch range ($\pm°$), capture distance (m), lighting conditions (controlled/uncontrolled), and full-face inclusion (binary: 0/1). Axes are min-max normalized to the [0, 1] range to enable cross-feature comparison.

**Table 3.** Comparison of gaze datasets: data characteristics, gaze and head pose annotations, and environmental conditions.

| Dataset | RGB/RGB-D | Participants | Full Face | Amount of Data | Distance | Illumination Conditions |
|---|---|---|---|---|---|---|
| MPIIGaze [57] | RGB | 15 | No | 213,659 images | 40~60 cm | Daily life |
| MPIIFaceGaze [63] | RGB | 15 | Yes | 213,659 images | 40~60 cm | Daily life |
| UnityEyes [58] | RGB | - | No | user defined | user defined | User defined |
| ColumbiaGaze [64] | RGB | 56 | Yes | 5880 images | 200 cm | Lab |
| Gaze360 [65] | RGB | 238 | Yes | ~172 K images | ~200 cm | Daily life |
| ETH-XGaze [66] | RGB | 110 | Yes | 1,083,492 images | 100 cm | Lab |
| UT-Multiview [67] | RGB | 50 | No | 1,216,000 images | 60 cm | Lab |
| EYEDIAP [68] | RGB-D | 16 | Yes | 94 videos | 80~120 cm | Lab |
| RT-GENE [69] | RGB-D | 15 | Yes | 277,286 images | ~182 cm | - |
| NISLGaze [71] | RGB | 21 | Yes | 2079 videos | 90 cm | - |
| Gaze-Capture [29] | RGB | 1474 | Yes | >2.5 M images | Close | Daily life |
| TabletGaze [70] | RGB | 51 | Yes | 816 videos | 30~50 cm | Lab |
| GAFA [45] | RGB | - | Yes | 882,000 videos | 50 cm~7 m | Daily life |
| GFIE [56] | RGB-D | 61 | Yes | 71,799 videos | 1.04 m ~ 6.48 m | Daily life |
| **Dataset** | **Gaze Pitch** | **Gaze Yaw** | **Head Pose Annot.** | **Gaze Pose Annot.** | **Head Pose Orient.** | |
| MPIIGaze [57] | $-5°$~$20°$ | $-40°$~$20°$ | Yes | Yes | Frontal | |
| MPIIFaceGaze [63] | $-5°$~$20°$ | $-40°$~$20°$ | Yes | Yes | Frontal | |
| UnityEyes [58] | user defined | user defined | Yes | Yes | All | |
| ColumbiaGaze [64] | $-10°$~$10°$ | $-15°$~$15°$ | 5 orient | Yes | Frontal | |
| Gaze360 [65] | $-50°$~$50°$ | $-140°$~$140°$ | Yes | Yes | All | |
| ETH-XGaze [66] | $-70°$~$70°$ | $-120°$~$120°$ | Yes | Yes | All | |
| UT-Multiview [67] | $-55°$~$65°$ | $-80°$~$80°$ | Yes | Yes | All | |
| EYEDIAP [68] | $-45°$~$45°$ | $-45°$~$45°$ | Yes | Yes | Frontal | |
| RT-GENE [69] | $-30°$~$30°$ | $-40°$~$40°$ | Yes | Yes | All | |
| NISLGaze [71] | $-21.48°$~$20.76°$ | $-21.25°$~$21.04°$ | Yes | Yes | All | |
| Gaze-Capture [29] | $-20°$~$20°$ | $-20°$~$20°$ | - | Yes | Frontal | |
| TabletGaze [70] | $-15°$~$0°$ | $-20°$~$10°$ | - | Yes | Frontal | |
| GAFA [45] | $-75°$~$75°$ | $-150°$~$150°$ | Yes | Yes | All | |
| GFIE [56] | - | - | Yes | Yes | All | |

## 4. Algorithms

### 4.1. Model-Driven Methods

Model-driven gaze estimation methods estimate gaze direction through geometric and optical principles, rather than relying purely on learned statistical mappings derived from data. By explicitly modeling eye anatomy, light paths, and camera projection geometry, these approaches establish a direct, physics-based link between observable eye features and the target gaze direction.

#### 4.1.1. Monocular Geometry

A fundamental challenge in gaze estimation lies in accurately inferring a 3D gaze vector from monocular eye images, particularly when direct depth cues are unavailable. Geometry-based approaches address this challenge by exploiting the physical principles of optical reflection and refraction within the human eye. Specifically, these methods reconstruct the optical axis from corneal reflections and infer the visual axis through calibrated angular correction. The framework typically relies on a known arrangement of light sources, cameras, and geometric assumptions about the eyeball and corneal surface.

The estimation of the optical axis relies on leveraging infrared light reflection from the corneal surface, typically using infrared LEDs as structured illumination sources. The positions of the reflected light spots (glints) are captured, and a mathematical model incor-

porating eye structure parameters is constructed. Based on image processing algorithms, the relative spatial positions of the glint and pupil center are then analyzed to derive the optical axis direction and angle. Uniquely estimating the corneal center requires at least two cameras and two non-collinear LEDs to avoid ambiguity. Increasing the number of LEDs provides additional geometric constraints, thereby reducing estimation uncertainty and enhancing robustness against noise. As illustrated in Figure 5, accurate estimation of the gaze vector necessitates accounting for the refractive distortion introduced by the corneal surface. Given the known corneal center C, two calibrated cameras can determine the 3D pupil position through back-projection based on the corneal refraction center. Therefore, the optical axis is fully specified. The visual axis can be determined by incorporating the individual-specific angular offset between the optical axis and the visual axis.
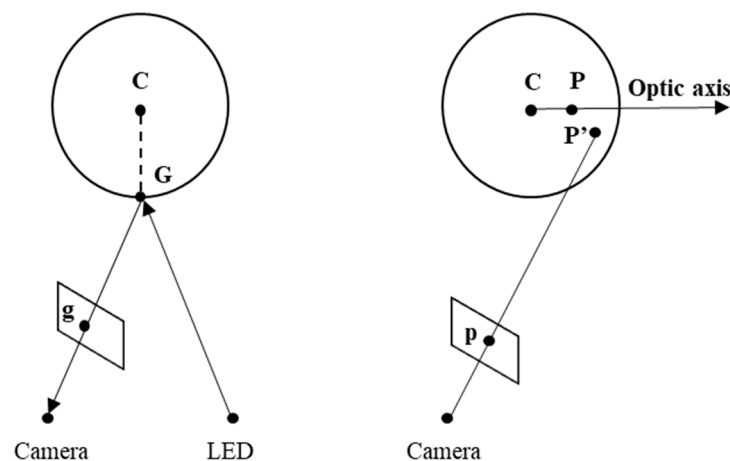


**Figure 5.** A geometric model for estimating the center of corneal curvature and the optical axis [72]. Point C denotes the center of corneal curvature, while G indicates the glint point induced by the LED light source. P represents the true 3D position of the pupil center, and P′ is the apparent pupil center after corneal refraction. And, p and g denote the 2D image-plane projections of P′ and G, respectively.

Guestrin et al. [54] systematically formalized the conditions under which geometric reconstruction is solvable. They showed that when two cameras and two LEDs are available, single-point calibration is sufficient to estimate the gaze vector. With only one camera and two LEDs, multi-point calibration is required to solve for both anatomical and device parameters. In the most constrained configuration—one camera and one LED—the model becomes underdetermined unless five anatomical parameters are assumed and the camera-to-cornea distance is fixed. Zhu et al. [73] proposed a simplified framework by treating the corneal surface as a fixed spherical mirror to reduce the complexity. Under this geometric simplification assumption, the virtual image of the illuminating LED formed via corneal reflection becomes invariant to camera position. By assuming collinearity among the optical center, LED, and the virtual reflection image, the corneal center can be estimated using observations from two cameras. Similarly, the refracted pupil is modeled as coaxial with its physical counterpart, facilitating direct derivation of the optical axis from dual-camera observations. While this model simplifies both theoretical derivation and practical implementation, it imposes strict assumptions that may not hold under natural head movements or individual anatomical variability. Consequently, this approximation introduces systematic bias in gaze estimation results.

The monocular geometric model offers a physically interpretable and mathematically rigorous framework for gaze estimation. However, it is subject to practical limitations including the need for multi-sensor configurations, high-precision calibration, and strong geometric assumptions. These constraints have motivated the development of hybrid models that integrate geometric priors with learning-based corrections.

### 4.1.2. Stereo Triangulation and Gaze Depth Estimation

Stereo triangulation constitutes a key strategy in model-driven methods, addressing the critical depth ambiguity inherent in monocular imaging systems. By employing multiple spatially separated cameras with overlapping fields of view, this approach enables precise 3D reconstruction of the gaze point through geometric triangulation principles. In multi-camera parallax triangulation frameworks, synchronized camera arrays capture corneal refraction patterns from distinct angular perspectives, leveraging binocular disparity to estimate pupil centroid depth via Perspective-n-Point (PnP) algorithm [74], which converts the parallax observed from different viewpoints into 3D spatial coordinates. It is important to note that under the spherical cornea model, the observed pupil center is formed after light passes through the cornea and is subject to refraction, resulting in a geometric offset from the actual pupil center. To mitigate this, the stereo systems incorporate corneal refractive compensation [75]. Based on the optical system of the cornea, Wan et al. [75] derived a forward transformation from the real pupil to the refracted virtual pupil and proposed a reverse transformation to recover the real pupil axis from the observed pupil contour image. Swirski and Dodgson [76] proposed a 3D eye model fitting approach that derives a unique solution by fitting a set of eye images. This approach operates under the assumption of a perfect spherical eye geometry, initializing key parameters including the eye center coordinates, ocular radius, and 3D pupil position. These parameters are then iteratively optimized by aligning the model to the original eye images. The estimated parameters are defined in the coordinate system of the eye camera, necessitating a coordinate transformation to the scene camera coordinate system. This transformation relies on a 3D translation vector and a rotation matrix that maps the gaze vector into the scene camera space.

Active stereo vision with structured illumination further enhances feature detection and depth precision. Wang et al. [77] leveraged extended displays to achieve high-density 3D measurements, combining stereo deflectometry with a single-shot cross-sine wave pattern. This method reconstructs the depth of eye surface and normal maps with low latency, enabling gaze direction estimation from the geometric properties of the eye.

Beyond triangulation, it is essential to understand the behavior of depth reconstruction error as a function of disparity and system parameters. As shown in the Figure 6, let the cameras be rectified and parallel, with baseline $B$ and focal length $f$. The disparity $d$ represents the pixel displacement (i.e., parallax) of the same 3D point between the two cameras. The depth estimate is then given by the following:

$$Z(d) \approx \frac{Bf}{d} \tag{3}$$

Therefore, in a binocular stereo vision system, increasing the baseline distance and camera focal length is the main direction to reduce depth error.
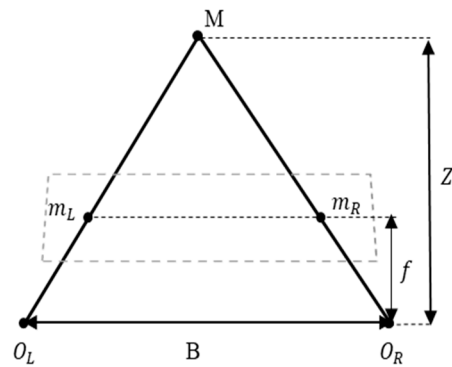
**Figure 6.** Stereo Triangulation of Binocular Errors.

Traditional geometry-based 3D gaze estimation algorithms often encounter challenges with noise in eyeball depth convergence [78]. Wang et al. [79] developed an online system that combines triangulation for initial gaze position estimation with a subsequent calibration step to refine depth accuracy. This system enhances the geometry-based 3D gaze estimation algorithm by using a screen-centered coordinate system, estimating gaze depth coordinates through the calculation of horizontal eye coordinates and the line-of-sight discrepancy. Yuan et al. [80] proposed a geometric method for high-precision, real-time head pose estimation from 2D face image, aiming to reduce the computational burden of traditional methods. However, the head posture angle is determined solely by the rotation matrix, which can be affected by external factors, leading to inaccuracies. Meyer et al. [81] optimized a geometric eye model by integrating distance and rotational velocity measurements from multiple static Light Field Imaging (LFI) sensors. This fusion strategy strengthens gaze direction estimation by compensating for individual sensor limitations, thereby improving the robustness of traditional geometric approaches in dynamic environments.

### 4.1.3. Calibration and Parameter Estimation

In model-driven gaze estimation systems, precise calibration is critical for mitigating errors arising from inter-subject variation and camera-eye positional misalignment. Since the geometry of the human eye varies across users, a generic model without personalized adjustment often results in substantial prediction errors. Calibration serves as a compensatory mechanism to estimate both system-specific and subject-specific parameters, ensuring precise alignment between the mathematical gaze model and each user. These parameters may include the transformation matrices relating pupil or glint positions to optical axis direction, the displacement between the optical and visual axes, and the projection from gaze direction to screen coordinates. Current methodologies for gaze estimation frequently employ explicit calibration protocols that require users to fixate on predefined calibration targets displayed across the screen. By associating each image-space eye feature with its corresponding known gaze target, it becomes possible to estimate a parametric mapping function. While explicit calibration offers direct control over accuracy and provides well-defined convergence criteria, it imposes operational burdens. User cooperation, stable fixation, and strict adherence to calibration sequences are difficult to guarantee in mobile, wearable, or real-world scenarios. To reduce these dependencies, an alternative line of research explores implicit calibration methods that do not rely on labeled target data, instead inferring parameters through natural visual behavior during free-viewing tasks. By capturing eye images over time during these natural viewing processes, and assuming that the fixations of users are not uniformly random instead concentrating around perceptually meaningful regions, it becomes feasible to optimize the calibration parameters as latent variables within a probabilistic framework.

Tong et al. [82] proposed the Discretization Gaze Network (DGaze-Net), which optimizes monocular 3D gaze estimation accuracy through feature discretization and an attention mechanism. However, information loss may result from discretization, and the method could exhibit variable performance in complex environments or among different users. Multi-layer sensing was utilized by Lee et al. [23] to obtain depth gaze positions, eliminating the requirement for individual calibration. However, this method requires double Purkinje images as input, which is difficult to capture accurately. Visual saliency concepts [83] and various saliency algorithms [84] offer fresh perspectives on 3D gaze estimation calibration. Liu et al. [85] proposed a 3D gaze estimation method by using automatic calibration. By identifying 3D salient pixels in the scene as potential calibration targets through saliency detection, automatic calibration becomes achievable.

The ongoing advancement of gaze estimation systems relies heavily on calibration and parameter estimation, which are critical for optimizing their performance and usability. Whether implemented through geometric regression, probabilistic inference, or real-time adaptation, calibration is not merely a pre-processing step but an active component that bridges model predictions with real-world variability. A principled calibration strategy ensures that geometric models maintain validity across individuals and sessions, enabling gaze vectors to function as robust and interpretable indicators of visual attention.

### 4.2. Data-Driven Methods

Data-driven methods in gaze estimation have emerged as a fundamental paradigm that circumvents the need for explicit modeling of the eyeball structure or optical projection principles. These methods directly learn the mapping between visual input and gaze direction using supervised or weakly supervised learning frameworks [86,87]. This paradigm shift is driven by the inherent complexity of the human ocular system, variations in imaging conditions, and the nonlinear relationship between observable features and true gaze vectors.

In contrast, traditional geometric or appearance-based methods rely on hand-crafted parameters—such as corneal reflection points or ellipse fitting of the pupil boundary—which are highly sensitive to noise, occlusions, and pose variations. As a result, they often struggle to generalize beyond calibrated environments or specific hardware setups.

Data-driven methods are grounded in statistical learning theory and end-to-end optimization. Early studies employed classical models such as support vector regression [88] and random forests [89–91], which relied on manually extracted features to build mapping functions. However, these approaches are limited in their ability to generalize in high-dimensional feature spaces and often fail to capture complex nonlinear patterns effectively.

The advent of CNNs introduced a unified framework capable of learning spatial features directly from raw pixel data, optimized jointly with the gaze estimation objective. Wang et al. [92] proposed a unified framework that combines adversarial learning with Bayesian inference. This framework enhances a traditional CNN-based gaze estimator by incorporating an adversarial component, enhancing its sensitivity to gaze direction while remaining robustness to variations in appearance and pose. Alternatively, some studies [93,94] have adopted a hybrid model that uses a CNN to map images to eye landmarks, which are then used to estimate eye gaze.

CNNs are particularly effective at feature extraction, spatial attention, end-to-end learning, and transfer learning, enabling them to autonomously capture high-level abstractions. Although they achieve impressive performance—especially in transfer learning contexts—CNNs still face challenges such as high computational cost, heavy reliance on labeled data, and limited interpretability. Future work may focus on reducing model

complexity, designing more efficient attention mechanisms, and improving data annotation and collection processes.

The Transformer architecture, originally introduced by Vaswani et al. [95], has achieved remarkable success across wide range of computer vision tasks. Cheng and Lu et al. [96] employed Visual Transformers for gaze estimation, demonstrating significant performance improvements. However, traditional CNNs, with their limited capacity for global context modeling, have struggled to further improve prediction accuracy. To address this limitation, Li et al. [97] proposed the Swin Transformer and developed two architectures: a pure Swin Transformer model for gaze estimation (SwinT-GE), and a hybrid model (Res-Swin-GE) that combines convolutional layers with Swin Transformer modules.

Building on this, the Gaze–Swin model [98] integrates the Swin Transformer with ResNet-18 to capture both global and local facial features via a dual-branch structure. These features are concatenated and passed through a multi-layer perceptron (MLP) to predict gaze direction. In addition, the model incorporates a DA-Attention mechanism, which leverages relative position bias and scaled cosine attention to improve feature extraction accuracy. The Dropkey technique is also employed to mitigate overfitting, resulting in more accurate gaze point predictions.

A continuing challenge for data-driven methods lies in their vulnerability to degraded inputs, such as eye blinking, partial occlusion, or lens blur. To overcome this, hybrid consistency–inconsistency networks have been developed to explicitly model prediction uncertainty and recover reliable gaze estimates from compromised data. The CI-Net model, for instance, separates feature pathways for clean and corrupted inputs, applying spatial attention to reweight features based on estimated consistency. This approach has demonstrated improved robustness under adverse conditions, achieving lower angular error under occlusion compared to conventional CNN [44].

Data-driven methods depend heavily on large-scale annotated datasets, raising significant concerns about model generalization across domains and environments. In practice, models trained on a specific dataset often suffer substantial performance degradation when evaluated on unseen distributions, due to variations in illumination, camera placement, and subject. As a result, domain adaptation has become a central research focus. For instance, Bao et al. [61] proposed a rotation-enhanced unsupervised domain adaptation framework that enforces rotational consistency in the feature space, enabling the model to produce stable gaze estimates under varying orientations. Notably, this approach achieved state-of-the-art performance on cross-dataset benchmarks without requiring labeled samples from the target domain. Building on this, Cai et al. [62] introduced a passive adaptation framework that incorporates epistemic uncertainty modeling, allowing the model to self-adjust predictions dynamically during inference in novel environments.

Another major challenge in data-driven methods arises from imbalanced gaze distributions within datasets, where certain gaze directions are disproportionately represented. Attention-based methods have emerged as a promising solution, drawing inspiration from human visual attention to guide the model toward the most informative regions of the input. Unlike conventional techniques, attention mechanisms eliminate the need for separate modules for eye detection or head pose estimation, offering a more unified architecture.

Zhuang et al. [99] implemented an attention-enhanced ResNet-50 to estimate gaze points in a flight simulator. The introduction of the attention mechanism improved network performance and addressed challenges in multi-camera and multi-screen systems. Luo et al. [44] introduced a cross-attention mechanism that adaptively reweights information from facial and ocular regions. Their I-Net model selectively integrates features from a complementary network (C-Net), enhancing eye direction estimation. Attention mechanisms have also proven effective in fusing predictions. Huang et al. [100] proposed a

regression-based framework that adaptively fuses candidate gaze maps, achieving more robust estimation. Yi et al. [101] modeled the line of sight as a probabilistic distribution by sampling from random units in a deep network to construct an attention map, which then guided the aggregation of visual features for downstream tasks like action recognition.

Recent studies have further demonstrated that incorporating auxiliary tasks into the learning pipeline enhances feature robustness and generalization. Zhang et al. [102] addressed the challenge of eye blinks using cross-dataset multi-task training. Díaz et al. [103] developed the Asymmetric Multi-Task system for Gaze-driven grasping Action Forecasting (AMT-GAF) model, which jointly predicts future visual attention and grasping actions through multi-task learning. Lu et al. [104] considered the refractive effects introduced by eyeglasses and proposed a dual-objective network that simultaneously regresses the line of sight and classifies eyewear conditions.

Data-driven methods offer a scalable, adaptive, and increasingly generalizable solution to the diverse challenges of real-world gaze estimation. By leveraging advancements in network architectures, including attention mechanisms, adaptation strategies, and supervision paradigms, the capability to surpass traditional geometry-based models has been demonstrated. As a result, data-driven techniques now dominate the field and continue to drive progress in modern gaze estimation research.

In conclusion, the continuous evolution of data-driven methods, particularly through the integration of deep learning and adaptive learning techniques, significantly enhances the precision and flexibility of gaze estimation systems, enabling them to handle the complexities of dynamic, real-world environments and broadening their applicability across a wide range of domains.

## 5. Challenges and Future Directions

### 5.1. Existing Challenges and Issues

Gaze estimation remains a technically demanding task, particularly in unconstrained and complex environments. While 2D imaging methods achieve high accuracy under controlled settings, they are highly sensitive to illumination variations and require precise modeling of eye movements. These limitations reduce their reliability in real-world, dynamic scenes.

In contrast, 3D model-based methods, which incorporate depth cues and multi-view fusion, offer a more holistic understanding of gaze behavior. However, they introduce additional challenges, including increased computational cost, complexity in multi-view alignment, and the demand for real-time processing capabilities.

Traditional machine learning approaches are favored for their interpretability and effectiveness on small datasets, but they rely on manual feature engineering, resulting in that they are less suitable for complex or high-dimensional scenarios. Moreover, their generalization capability often deteriorates under diverse or unseen conditions.

Deep learning-based methods have demonstrated superior performance in learning complex patterns from large datasets. Nevertheless, many models assume a calibrated input (e.g., calibrated face images), which constrains their applicability in real-time and multi-person settings. Additional pre-processing steps, such as face cropping and calibration, further contribute to increased inference latency [105].

When selecting an appropriate gaze estimation strategy, it is essential to consider application-specific requirements, data characteristics, and computational constraints. In many practical cases, hybrid approaches that combine complementary methods may be necessary to achieve robust performance.

Data imbalance poses another persistent challenge. Two-dimensional methods exhibit relative robustness in scenarios with limited data; their effectiveness significantly

diminishes when faced with severely skewed distributions. Three-dimensional models, leveraging multi-view information, can improve estimation for underrepresented gaze directions, yet their computational demands may offset these benefits.

Traditional models generally handle moderate imbalance better than deep networks but struggle in scenarios with extreme skew or complex features. In contrast, deep learning models are prone to overfitting dominant categories, leading to suboptimal performance on underrepresented gaze targets.

In cross-dataset testing, where the training and test data come from different datasets, performance often declines significantly due to domain discrepancies. A key challenge is the unavailability of target domain labels in real-world scenarios, which prevents direct training of gaze estimators in the target domain. Additionally, as the scope of the source domain decreases, adaptive capabilities also diminish. Current methods have not yet fully resolved these problems.

To overcome these obstacles, future research must integrate weight rebalancing, data augmentation, task-specific model architectures, and regularization techniques. Yet, a unified framework that systematically addresses gaze estimation across heterogeneous scenarios is still lacking.

Privacy and ethical issues surrounding eye tracking technology are also key areas of development. Existing privacy-preserving mechanisms face challenges in balancing the dual demands of privacy and utility in the context of AR/VR applications. Privacy-preservation techniques such as plausible deniability (PD) and differential privacy (DP) mechanisms have recently been applied to eye movement data. David-John et al. [106] applied the privacy definitions of k-anonymity and PD to a dataset of eye tracking samples, introducing a privacy-utility trade-off while maintaining gaze prediction accuracy. Bozkir et al. [107] propose a novel transform-coding based differential privacy mechanism to further adapt it to the statistics of eye movement feature data. Their results provide significantly high privacy without any essential loss in classification accuracies while hiding personal identifiers. As eye tracking continues to evolve, striking a balance between technological advancement and responsible usage will be key to ensuring the broader societal acceptance and success of these systems.

*5.2. Prospects*

As shown in Figure 7, the integration of gaze estimation with multimodal sensor data has revolutionized cognitive state analysis. By combining gaze information with physiological signals such as electroencephalography (EEG), heart rate, and galvanic skin response (GSR), researchers can more accurately infer user attention, emotional states, and cognitive workload. Technically, multimodal fusion can be achieved through early fusion (e.g., feature concatenation), late fusion (e.g., decision-level integration), or hybrid fusion frameworks. Recent methods also employ attention mechanisms or cross-modal Transformers to align temporal and spatial features across modalities. In VR and AR, the combination of gaze estimation with head tracking and gesture recognition improves interaction quality, enabling more natural and immersive experiences [108,109]. For example, sensor synchronization and spatial–temporal encoding networks have been applied to enhance interaction fidelity. Kin et al. use eye tracking for global intuitive navigation and gesture controllers for local fine-grained navigation [110]. In the automotive field, fusing gaze data with driver monitoring systems and vehicle sensors allows for precise detection of driver fatigue, attentiveness, and emotions, contributing to safer driving and more personalized in-vehicle experiences. In clinical contexts, including medical diagnostics, psychotherapy, and rehabilitation, integrating gaze estimation with physiological signals

facilitates a deeper understanding of cognitive and emotional processes, supporting the development of individualized treatment strategies.
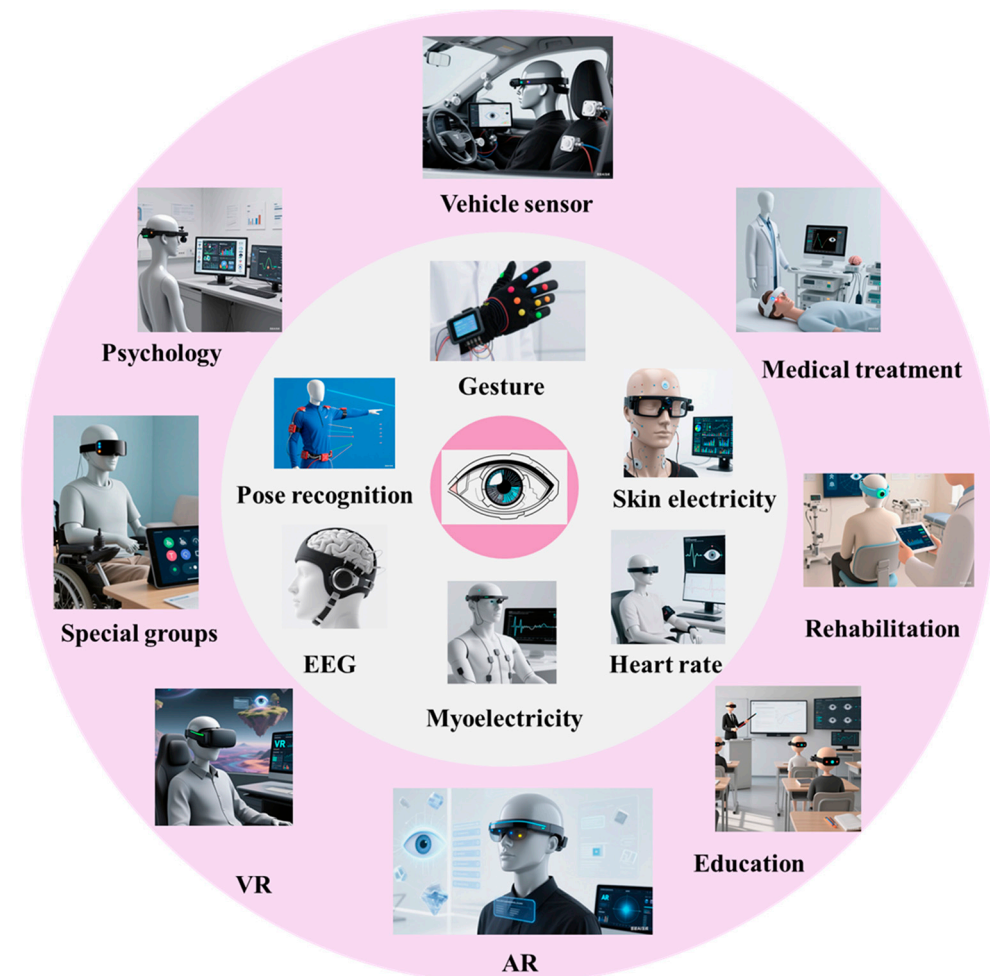


**Figure 7.** Multimodal Sensor Fusion and Application Scenarios.

In gaze estimation, unsupervised learning and self-supervised learning have emerged as effective approaches to reduce reliance on labeled data. Unsupervised learning enables models to extract meaningful visual patterns and discriminative features from large-scale unannotated gaze datasets, thereby improving the understanding of users' attention behaviors. Typical approaches include clustering methods, representation learning through autoencoders, and contrastive learning strategies. By uncovering the underlying structure of the data, it reveals correlations among different gaze behaviors, which contributes to better modeling of visual attention. In contrast, self-supervised learning leverages intrinsic properties of gaze data to generate pseudo-labels or auxiliary tasks, effectively minimizing the need for manual annotation. Common tasks include temporal order prediction, masked signal reconstruction, or gaze consistency verification, which help guide feature learning in the absence of labels. This strategy enhances the efficiency of utilizing unlabeled data and strengthens the adaptability of gaze estimation models across diverse user behaviors and environmental conditions, ultimately leading to more robust and accurate predictions.

Meta-learning focuses on enabling models to rapidly adapt to new tasks or domains using only a small number of training samples, thereby alleviating the need for extensive data collection. Its core idea lies in learning a generalizable learning strategy that allows the model to update efficiently when facing novel scenarios. In the context of eye tracking, meta-learning has achieved promising results in personalization tasks. For example, it has

reached an accuracy of 88.6% in distinguishing individuals with Autism Spectrum Disorder (ASD) from typically developing subjects. This demonstrates its potential for improving user-specific gaze estimation and adapting gaze models to complex real-world variability.

Compared to conventional cameras, event cameras offer several technical advantages, including ultra-low latency, low power consumption, high temporal resolution, wide dynamic range, asynchronous data acquisition, and sparse event-driven outputs, facilitating eye tracking and gaze estimation in dynamic or resource-constrained environments.

In recent years, HMD devices such as Meta's Quest Pro, Apple Vision Pro, and HTC Vive have exemplified different developmental approaches to eye-tracking technology across major manufacturers. Meta's latest devices incorporate compact eye-tracking modules primarily for gaze estimation and facial expression capture. While these systems are lightweight and wearable, they still fall short of high-end devices in terms of gaze accuracy and stability. The Apple Vision Pro integrates a dense array of infrared cameras combined with a sophisticated real-time calibration mechanism, significantly enhancing gaze tracking accuracy and precision—particularly in applications such as gaze-based rendering and immersive interaction—while maintaining a high level of user comfort and natural system responsiveness. However, the complexity of its hardware design results in higher cost and larger physical size. The HTC Vive series, leveraging Tobii's mature eye-tracking technology, delivers high precision and reliability, making it widely adopted in research and industrial settings. Nonetheless, its overall bulk and weight pose challenges for long-term comfort and everyday usability. With continued advancements in sensor miniaturization, low-power infrared imaging, and adaptive algorithms tailored to individual users, future eye-tracking systems are expected to achieve high precision while becoming lighter, more comfortable, and less obtrusive.

To facilitate the selection and comparison of approaches, Table 4 provides a summary of representative application areas, tasks, algorithms, and benchmark datasets commonly used in eye-tracking and gaze estimation research.

In conclusion, the combination of advanced machine learning techniques, multimodal sensor integration, and event-based imaging technologies presents prospects for the future of gaze estimation, making it more adaptive, accurate, and applicable across various real-world domains.

**Table 4.** Summary of typical application areas, representative tasks, algorithms, and datasets in eye-tracking and gaze estimation.

| Application Area | Task | Algorithms | Datasets |
|---|---|---|---|
| HCI | Contactless Interaction | CNN-based gaze regression [29] | GazeCapture [29] MPIIGaze [57] |
| | Remote Gaze Synchronization | Gaze-following CNN [17] | GazeFollow [16] |
| | Wearable Eye-based Interaction | Multi-Stage CNN + SVM [111] | TabletGaze [70] |
| AR/VR | Foveated Rendering | Real-time pupil tracking + foveated rendering [112] | GazeCapture [29] UnityEyes [58] |
| | Visual Attention Analysis | LSTM-based sequence modeling [69] | ColumbiaGaze [64] Gaze360 [65] |
| | Gaze-based Interaction Control | Multi-task CNN (appearance + geometry) [66] | MPIIGaze [57] UnityEyes [58] |
| Medical Diagnosis | Parkinson's Screening | Saccadic movement analysis (RF, SVM) [113] | Clinical datasets [113] |
| | Autism Spectrum Identification | Spatio-temporal gaze patterns (CNN + LSTM) [114] | ASD Eye-tracking datasets [115] |
| | Reading and Language Assessment | Scanpath analysis (HMM, RF) [116] | ZuCo [117] |
| Automotive | Driver Attention Monitoring | Gaze zone classification (CNN, 3D CNN) [118] | DR(eye)VE [119] |
| | Hazard Prediction | Temporal gaze prediction (LSTM) [120] | EyeTrackUAV2 [121] DR(eye)VE [119] |

## 6. Conclusions

The critical aspects of eye-tracking and gaze estimation algorithms have been reviewed. The challenges in complex scenes, as well as issues related to data imbalance and model generalization have been discussed. The future of gaze estimation algorithms in eye tracking and potential research directions or innovative approaches have also been proposed. These comprehensive discussions offer valuable insights for advancing our understanding and guiding the future pathway of gaze estimation research.

**Author Contributions:** L.L. and Z.W. contributed equally to this work. They conducted the literature review, conceptualized the structure of the manuscript, and were primarily responsible for writing the original draft. Z.C. and Y.L. provided academic guidance, critical review, and helped refine the content of the manuscript through multiple revisions. W.G. supervised the overall project, contributed to the manuscript's conceptual framework, and finalized the paper as the corresponding author. All authors have read and agreed to the published version of the manuscript.

**Data Availability Statement:** The data are available from the authors upon reasonable request.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Portugal, A.M.; Viktorsson, C.; Taylor, M.J.; Mason, L.; Tammimies, K.; Ronald, A.; Falck-Ytter, T. Infants' looking preferences for social versus non-social objects reflect genetic variation. *Nat. Hum. Behav.* **2024**, *8*, 13. [CrossRef] [PubMed]
2. Zhu, H.T.; Yang, H.; Xu, S.Q.; Ma, Y.Y.; Zhu, S.G.; Mao, Z.Y.; Chen, W.W.; Hu, Z.Z.; Pan, R.R.; Xu, Y.R.; et al. Frequency-encoded eye tracking smart contact lens for human-machine interaction. *Nat. Commun.* **2024**, *15*, 13. [CrossRef] [PubMed]
3. Poli, F.; Li, Y.L.; Naidu, P.; Mars, R.B.; Hunnius, S.; Ruggeri, A. Toddlers strategically adapt their information search. *Nat. Commun.* **2024**, *15*, 10. [CrossRef]

4. Linde-Domingo, J.; Spitzer, B. Geometry of visuospatial working memory information in miniature gaze patterns. *Nat. Hum. Behav.* **2024**, *8*, 15. [CrossRef]

5. Joo, H.-J.; Jeong, H.-Y. A study on eye-tracking-based interface for VR/AR education platform. *Multimed. Tools Appl.* **2020**, *79*, 16719–16730. [CrossRef]

6. Harezlak, K.; Kasprowski, P. Application of eye tracking in medicine: A survey, research issues and challenges. *Comput. Med. Imaging Graph.* **2018**, *65*, 176–190. [CrossRef]

7. Kaduk, T.; Goeke, C.; Finger, H.; König, P. Webcam eye tracking close to laboratory standards: Comparing a new webcam-based system and the eyeLink 1000. *Behav. Res. Methods* **2024**, *56*, 5002–5022. [CrossRef]

8. Zhang, T.; Shen, Y.; Zhao, G.; Wang, L.; Chen, X.; Bai, L.; Zhou, Y. Swift-Eye: Towards anti-blink pupil tracking for precise and robust high-frequency near-Eye movement analysis with event cameras. *IEEE Trans. Vis. Comput. Graph.* **2024**, *30*, 2077–2086. [CrossRef]

9. Wang, Y.; Wang, J.; Guo, P. Eye-UNet: A UNet-based network with attention mechanism for low-quality human eye image segmentation. *Signal Image Video Process.* **2023**, *17*, 1097–1103. [CrossRef]

10. Eggert, T. Eye movement recordings: Methods. *Dev. Ophthalmol.* **2007**, *40*, 15–34. [CrossRef]

11. Shi, Y.; Yang, P.; Lei, R.; Liu, Z.; Dong, X.; Tao, X.; Chu, X.; Wang, Z.L.; Chen, X. Eye tracking and eye expression decoding based on transparent, flexible and ultra-persistent electrostatic interface. *Nat. Commun.* **2023**, *14*, 3315. [CrossRef]

12. Yang, M.; Gao, Y.; Tang, L.; Hou, J.; Hu, B. Wearable eye-tracking system for synchronized multimodal data acquisition. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 5146–5159. [CrossRef]

13. Karmi, R.; Rahmany, I.; Khlifa, N. Gaze estimation using convolutional neural networks. *Signal Lmage Video Process.* **2023**, *18*, 389–398. [CrossRef]

14. Liu, B.; Lye, S.W.; Zakaria, Z.B. An integrated framework for eye tracking-assisted task capability recognition of air traffic controllers with machine learning. *Adv. Eng. Inf.* **2024**, *62*, 102784. [CrossRef]

15. Lohr, D.; Aziz, S.; Friedman, L.; Komogortsev, O.V. GazeBaseVR, a large-scale, longitudinal, binocular eye-tracking dataset collected in virtual reality. *Sci. Data* **2023**, *10*, 177. [CrossRef]

16. Recasens, A.; Khosla, A.; Vondrick, C.; Torralba, A. Where are they looking? In Proceedings of the 29th Annual Conference on Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 12 December 2015; pp. 199–207.

17. Recasens, A.; Vondrick, C.; Khosla, A.; Torralba, A. Following gaze in video. In Proceedings of the 16th IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1444–1452.

18. Backhaus, D.; Engbert, R. How body postures affect gaze control in scene viewing under specific task conditions. *Exp. Brain Res.* **2024**, *242*, 745–756. [CrossRef]

19. Chen, Y.; Zhou, J.; Gao, Q.; Gao, J.; Zhang, W. MDNN: Predicting student engagement via gaze direction and facial expression in collaborative learning. *CMES-Comput. Model. Eng. Sci.* **2023**, *136*, 381. [CrossRef]

20. Liu, S.; Huang, S.; Wang, S.; Muhammad, K.; Bellavista, P.; Del Ser, J. Visual tracking in complex scenes: A location fusion mechanism based on the combination of multiple visual cognition flows. *Inf. Fusion.* **2023**, *96*, 281–296. [CrossRef]

21. 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*, 1372–1391. [CrossRef] [PubMed]

22. Tonini, F.; Dall'Asen, N.; Beyan, C.; Ricci, E. Object-aware gaze target detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 21860–21869.

23. Tu, D.Y.; Min, X.K.; Duan, H.Y.; Guo, G.D.; Zhai, G.T.; Shen, W. End-to-end human-gaze-target detection with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 2192–2200.

24. Zhang, X.; Wang, L.; He, Y.; Mou, Z.; Cao, Y. High-speed eye tracking based on a synchronized imaging mechanism by a dual-ring infrared lighting source. *Appl. Opt.* **2024**, *63*, 4293–4302. [CrossRef]

25. Severitt, B.R.; Kübler, T.C.; Kasneci, E. Testing different function fitting methods for mobile eye-tracker calibration. *J. Eye Mov. Res.* **2023**, *16*, 10-16910. [CrossRef]

26. Curcio, C.A.; Sloan, K.R.; Kalina, R.E.; Hendrickson, A.E. Human photoreceptor topography. *J. Comp. Neurol.* **1990**, *292*, 497–523. [CrossRef]

27. Falch, L.; Lohan, K.S. Webcam-based gaze estimation for computer screen interaction. *Front. Rob. AI* **2024**, *11*, 1369566. [CrossRef]

28. Zhou, J.; Li, G.; Shi, F.; Guo, X.; Wan, P.; Wang, M. EM-Gaze: Eye context correlation and metric learning for gaze estimation. *Vis. Comput. Ind. Biomed. Art.* **2023**, *6*, 8. [CrossRef]

29. Krafka, K.; Khosla, A.; Kellnhofer, P.; Kannan, H.; Bhandarkar, S.; Matusik, W.; Torralba, A. Eye tracking for everyone. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 27–30 June 2016; pp. 2176–2184.

30. He, J.F.; Pham, K.; Valliappan, N.; Xu, P.M.; Roberts, C.; Lagun, D.; Navalpakkam, V. On-device few-shot personalization for real-time gaze estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1149–1158.

31. Algassir, A.O.M.; Manickam, S.; Anbar, M.; Nseaf, A.K. Acf-Gsvm: Cascade aggregate channel feature with gabor filters and support vector machine for enhanced face detection. *J. Theor. Appl. Inf. Technol.* **2023**, *101*, 7317–7327.

32. Gao, F.; Li, S.; Lu, S. How frontal is a face? Quantitative estimation of face pose based on CNN and geometric projection. *Neural Comput. Appl.* **2021**, *33*, 3035–3051. [CrossRef]

33. Liu, H.; Zhang, C.; Deng, Y.; Liu, T.; Zhang, Z.; Li, Y.F. Orientation cues-aware facial relationship representation for head pose estimation via transformer. *IEEE Trans. Image Process.* **2023**, *32*, 6289–6302. [CrossRef]

34. Ruzzi, A.; Shi, X.W.; Wang, X.; Li, G.Y.; De Mello, S.; Chang, H.J.; Zhang, X.C.; Hilliges, O. GazeNeRF: 3D-aware gaze redirection with neural radiance fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 9676–9685.

35. Yang, A.; Jin, Z.; Guo, S.; Wu, D.; Chen, L. Unconstrained human gaze estimation approach for medium-distance scene based on monocular vision. *Vis. Comput.* **2023**, *40*, 73–85. [CrossRef]

36. Bisogni, C.; Cascone, L.; Nappi, M.; Pero, C. IoT-enabled biometric security: Enhancing smart car safety with depth-based head pose estimation. *ACM Trans. Multimed. Comput. Commun. Appl.* **2024**, *20*, 1–24. [CrossRef]

37. Xin, M.; Mo, S.T.; Lin, Y.Z. EVA-GCN: Head pose estimation based on graph convolutional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual Conference, Nashville, TN, USA, 19–25 June 2021; pp. 1462–1471.

38. Tian, J.; Cong, L.; Qin, H. Mixed-pose positioning in smartphone-based pedestrian dead reckoning using hierarchical clustering. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 9514312. [CrossRef]

39. Hu, Z.F.; Xia, Y.L.; Luo, Y.; Wang, L. Multi-feature fusion gaze estimation based on attention mechanism. In Proceedings of the Conference on Optoelectronic Imaging and Multimedia Technology VIII, Nantong, China, 10–12 October 2021; pp. 172–182.

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*, 1080–1109. [CrossRef]

41. Zhao, H.D.; Ding, Z.M.; Fu, Y. Pose-dependent low-rank embedding for head pose estimation. In Proceedings of the 30th Association-for-the-Advancement-of-Artificial-Intelligence (AAAI) Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; pp. 1422–1428.

42. Bao, H.; Fang, W.; Guo, B.; Wang, J. Real-time wide-view eye tracking based on resolving the spatial depth. *Multimed. Tools Appl.* **2018**, *78*, 14633–14655. [CrossRef]

43. Hu, J.; Lu, Y.; Zhang, J.; Xu, J.; Yang, H. Monocular free-head gaze tracking method for driving electric sickbed. *Meas. Sci. Technol.* **2023**, *34*, 12. [CrossRef]

44. Luo, Y.; Chen, J.; Chen, J. CI-Net: Appearance-based gaze estimation via cooperative network. *IEEE Access* **2022**, *10*, 78739–78746. [CrossRef]

45. Nonaka, S.; Nobuhara, S.; Nishino, K. Dynamic 3D gaze from afar: Deep gaze estimation from temporal eye-head-body coordination. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 2182–2191.

46. Cheng, Y.H.; Bao, Y.W.; Lu, F. PureGaze: Purifying gaze feature for generalizable gaze estimation. In Proceedings of the 36th AAAI Conference on Artificial Intelligence/34th Conference on Innovative Applications of Artificial Intelligence/12th Symposium on Educational Advances in Artificial Intelligence, Virtual (Online), 22 February–1 March 2022; pp. 436–443.

47. Liu, G.; Yu, Y.; Mora, K.A.F.; Odobez, J.M. A differential approach for gaze estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 1092–1099. [CrossRef]

48. Bao, J.; Liu, B.; Yu, J. An individual-difference-aware model for cross-person gaze estimation. *IEEE Trans. Image Process* **2022**, *31*, 3322–3333. [CrossRef]

49. Li, N.; Chang, M.; Raychowdhury, A. E-Gaze: Gaze estimation with event camera. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 4796–4811. [CrossRef]

50. Zeng, Z.; Liu, S.; Cheng, H.; Liu, H.; Li, Y.; Feng, Y.; Siebert, F.W. GaVe: A webcam-based gaze vending interface using one-point calibration. *J. Eye Mov. Res.* **2023**, *16*, 1–13. [CrossRef] [PubMed]

51. Zhang, H.; Wu, S.; Chen, W.; Gao, Z.; Wan, Z. Self-calibrating gaze estimation with optical axes projection for head-mounted eye tracking. *IEEE Trans. Ind. Inf.* **2024**, *20*, 1397–1407. [CrossRef]

52. Cortacero, K.; Fischer, T.; Demiris, Y. RT-BENE: A dataset and baselines for real-time blink estimation in natural environments. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1159–1168.

53. Ding, L.; Terwilliger, J.; Parab, A.; Wang, M.; Fridman, L.; Mehler, B.; Reimer, B. CLERA: A unified model for joint cognitive load and eye region analysis in the wild. *ACM Trans. Comput.-Hum. Interact.* **2023**, *30*, 1–23. [CrossRef]

54. Guestrin, E.D.; Eizenman, M. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Trans. Biomed. Eng.* **2006**, *53*, 1124–1133. [CrossRef]

55. Hansen, D.W.; Pece, A.E.C. Eye tracking in the wild. *Comput. Vis. Image Underst.* **2005**, *98*, 155–181. [CrossRef]

56. Hu, Z.X.; Yang, Y.X.; Zhai, X.L.; Yang, D.Y.; Zhou, B.H.; Liu, J.T. GFIE: A dataset and baseline for gaze-following from 2D to 3D in indoor environments. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 8907–8916.

57. Zhang, X.; Sugano, Y.; Fritz, M.; Bulling, A. MPIIGaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 162–175. [CrossRef]

58. Wood, E.; Baltrusaitis, T.; Morency, L.P.; Robinson, P.; Bulling, A. Learning an appearance-based gaze estimator from one million synthesised images. In Proceedings of the 9th Biennial ACM Symposium on Eye Tracking Research and Applications (ETRA), Charleston, SC, USA, 14–17 March 2016; pp. 131–138.

59. Prasse, P.; Reich, D.R.; Makowski, S.; Ahn, S.; Scheffer, T.; Jäger, L.A. SP-EyeGAN: Generating synthetic eye movement data with generative adversarial networks. In Proceedings of the 2023 Symposium on Eye Tracking Research and Applications, Tubingen, Germany, 30 May–2 June 2023; p. 18.

60. Chen, D.; Wang, D.; Darrell, T.; Ebrahimi, S. Contrastive test-time adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 295–305.

61. Bao, Y.W.; Liu, Y.F.; Wang, H.F.; Lu, F. Generalizing gaze estimation with rotation consistency. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 4197–4206.

62. Cai, X.; Zeng, J.B.; Shan, S.G.; Chen, X.L. Source-free adaptive gaze estimation by uncertainty reduction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 22035–22045.

63. Zhang, X.; Sugano, Y.; Fritz, M.; Bulling, A. It's written all over your face: Full-face appearance-based gaze estimation. In Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 2299–2308.

64. Smith, B.A.; Yin, Q.; Feiner, S.K.; Nayar, S.K. Gaze locking. In Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology, St. Andrews, UK; 2013; pp. 271–280.

65. Kellnhofer, P.; Recasens, A.; Stent, S.; Matusik, W.; Torralba, A. Gaze360: Physically unconstrained gaze estimation in the wild. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6911–6920.

66. 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 Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part V, pp. 365–381.

67. Sugano, Y.; Matsushita, Y.; Sato, Y. Learning-by-synthesis for appearance-based 3D gaze estimation. In Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 1821–1828.

68. Funes Mora, K.A.; Monay, F.; Odobez, J.M. 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, Safety Harbor, FL, USA, 26–28 March 2014; pp. 255–258.

69. Fischer, T.; Chang, H.J.; Demiris, Y. RT-GENE: Real-time eye gaze estimation in natural environments. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 339–357.

70. Huang, Q.; Veeraraghavan, A.; Sabharwal, A. TabletGaze: Dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. *Mach. Vis. Appl.* **2017**, *28*, 445–461. [CrossRef]

71. Chen, Z.; Shi, B.E. Towards high performance low complexity calibration in appearance based gaze estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 1174–1188. [CrossRef]

72. Nagamatsu, T.; Hiroe, M.; Rigoll, G. Corneal-reflection-based wide range gaze tracking for a car. In Proceedings of the Human Interface and the Management of Information. Information in Intelligent Systems, Cham, Switzerland, 26–31 July 2019; pp. 385–400.

73. Zhu, Z.; Ji, Q. Novel eye gaze tracking techniques under natural head movement. *IEEE Trans. Biomed. Eng.* **2007**, *54*, 2246–2260. [CrossRef]

74. Chen, H.; Tian, W.; Wang, P.; Wang, F.; Xiong, L.; Li, H. EPro-PnP: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 1–12. [CrossRef]

75. Wan, Z.; Xiong, C.; Chen, W.; Zhang, H.; Wu, S. Pupil-contour-based gaze estimation with real pupil axes for head-mounted eye tracking. *IEEE Trans. Ind. Inform.* **2022**, *18*, 3640–3650. [CrossRef]

76. Swirski, L.; Dodgson, N.A. A fully-automatic, temporal approach to single camera, glint-free 3D eye model fitting. In Proceedings of the Pervasive Eye Tracking Mobile Eye-Based Interact, Lund, Sweden, 13 August 2013; pp. 1–11.

77. Wang, J.; Wang, T.; Xu, B.; Cossairt, O.; Willomitzer, F. Accurate eye tracking from dense 3D surface reconstructions using single-shot deflectometry. *Nat. Commun.* **2025**, *16*, 2902. [CrossRef]

78. Duchowski, A.T.; Pelfrey, B.; House, D.H.; Wang, R. Measuring gaze depth with an eye tracker during stereoscopic display. In Proceedings of the ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization, Toulouse, France, 27–28 August 2011; pp. 15–22.

79. Wang, R.I.; Pelfrey, B.; Duchowski, A.T.; House, D.H. Online 3D gaze localization on stereoscopic displays. *ACM Trans. Appl. Percept.* **2014**, *11*, 1–21. [CrossRef]

80. Yuan, H.; Li, M.; Hou, J.; Xiao, J. Single image-based head pose estimation with spherical parametrization and 3D morphing. *Pattern Recognit.* **2019**, *103*, 107316. [CrossRef]

81. Meyer, J.; Gering, S.; Kasneci, E. Static laser feedback interferometry-based gaze estimation for wearable glasses. *IEEE Sens. J.* **2023**, *23*, 7558–7569. [CrossRef]

82. Sha, T.; Sun, J.; Pun, S.; Liu, Y. Monocular 3D gaze estimation using feature discretization and attention mechanism. *Optoelectron. Lett.* **2023**, *19*, 301–306. [CrossRef]

83. Zhang, Y.; Wu, N.; Lin, C.Z.; Wetzstein, G.; Sun, Q. GazeFusion: Saliency-guided image generation. *ACM Trans. Appl. Percept.* **2024**. [CrossRef]

84. Hu, Z.; Cai, Y.; Li, Q.; Su, K.; Lv, C. Context-aware driver attention estimation using multi-hierarchy saliency fusion with gaze tracking. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 8602–8614. [CrossRef]

85. Liu, M.; Li, Y.; Liu, H. 3D gaze estimation for head-mounted eye tracking system with auto-calibration method. *IEEE Access* **2020**, *8*, 104207–104215. [CrossRef]

86. Hu, D.; Huang, K. Semi-supervised multitask learning using gaze focus for gaze estimation. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 7935–7946. [CrossRef]

87. Ghosh, S.; Hayat, M.; Dhall, A.; Knibbe, J. MTGLS: Multi-task gaze estimation with limited supervision. In Proceedings of the 22nd IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 4–8 January 2022; pp. 1161–1172.

88. Yang, B.; Huang, J.; Chen, X.; Li, X.; Hasegawa, Y. Natural grasp intention recognition based on gaze in human–robot interaction. *IEEE J. Biomed. Health. Inf.* **2023**, *27*, 2059–2070. [CrossRef]

89. Jha, S.; Al-Dhahir, N.; Busso, C. Driver visual attention estimation using head pose and eye appearance information. *IEEE Open J. Intell. Transp. Syst.* **2023**, *4*, 216–231. [CrossRef]

90. Pan, Y.; Xu, J. Gaze-based human intention prediction in the hybrid foraging search task. *Neurocomputing* **2024**, *587*, 127648. [CrossRef]

91. Wang, S.; Niu, H.; Wei, W.; Yang, X.; Zhang, S.; Ai, M. Eye-gaze-based intention recognition for selection task by using SVM-RF. In Proceedings of the Human-Computer Interaction, Washington, WA, USA, 29 June–4 July 2024; pp. 157–168.

92. Wang, K.; Zhao, R.; Su, H.; Ji, Q. Generalizing eye tracking with bayesian adversarial learning. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 11899–11908.

93. Wang, K.; Zhao, R.; Ji, Q. A hierarchical generative model for eye image synthesis and eye gaze estimation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 440–448.

94. Park, S.; Zhang, X.; Bulling, A.; Hilliges, O. Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, Warsaw, Poland, 14–17 June 2018; pp. 1–10.

95. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.

96. Cheng, Y.; Lu, F. Gaze estimation using transformer. In Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022; pp. 3341–3347.

97. Li, Y.; Chen, J.; Ma, J.; Wang, X.; Zhang, W. Gaze Estimation based on convolutional structure and sliding window-based attention mechanism. *Sensors* **2023**, *23*, 6226. [CrossRef]

98. Zhao, R.; Wang, Y.; Luo, S.; Shou, S.; Tang, P. Gaze-swin: Enhancing gaze estimation with a hybrid CNN-transformer network and dropkey mechanism. *Electronics* **2024**, *13*, 328. [CrossRef]

99. Zhuang, J.; Wang, C. Attention mechanism based full-face gaze estimation for human-computer interaction. In Proceedings of the 2022 International Conference on Computer Network, Electronic and Automation (ICCNEA), Xi'an, China, 23–25 September 2022; pp. 6–10.

100. Huang, G.; Shi, J.; Xu, J.; Li, J.; Chen, S.; Du, Y.; Zhen, X.; Liu, H. Gaze estimation by attention-induced hierarchical variational auto-encoder. *IEEE Trans. Cybern.* **2024**, *54*, 2592–2605. [CrossRef]

101. Li, Y.; Liu, M.; Rehg, J.M. In the eye of the beholder: Gaze and actions in first person video. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 6731–6747. [CrossRef]

102. Zhang, H.; Wang, X.; Ren, W.; Noack, B.R.; Liu, H. Improving the reliability of gaze estimation through cross-dataset multi-task learning. In Proceedings of the 2022 International Conference on High Performance Big Data and Intelligent Systems (HDIS), Tianjin, China, 10–11 December 2022; pp. 202–206.

103. González-Díaz, I.; Molina-Moreno, M.; Benois-Pineau, J.; de Rugy, A. Asymmetric multi-task learning for interpretable gaze-driven grasping action forecasting. *IEEE J. Biomed. Health. Inf.* **2024**, *28*, 7517–7530. [CrossRef]

104. Lyu, J.F.; Xu, F. Towards eyeglasses refraction in appearance-based gaze estimation. In Proceedings of the 22nd IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Sydney, Australia, 16–20 October 2023; pp. 693–702.

105. Zhang, M.F.; Liu, Y.F.; Lu, F. GazeOnce: Real-time multi-person gaze estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 4187–4196.

106. David-John, B.; Butler, K.; Jain, E. Privacy-preserving datasets of eye-tracking samples with applications in XR. *IEEE Trans. Vis. Comput. Graph.* **2023**, *29*, 2774–2784. [CrossRef]

107. Bozkir, E.; Günlü, O.; Fuhl, W.; Schaefer, R.F.; Kasneci, E. Differential privacy for eye tracking with temporal correlations. *PLoS ONE* **2021**, *16*, 22. [CrossRef]

108. Jin, H.; Lin, Z.; Lai, W.; Jiang, H.; Cai, J.; Chen, H.; Hao, W.; Ye, Y.; Xu, S.; Yan, Q.; et al. Micro-LED retinal projection for augmented reality near-eye displays. *Laser Photonics Rev.* **2025**, *19*, 2402083. [CrossRef]

109. Jiang, H.; Cheng, Y.; Sun, Z.; Yuan, Z.; Jin, H.; Huo, Y.; Tseng, M.C.; Yeung, F.; Kwok, H.S.; Chen, E. Pupil-adaptive retina projection augment reality displays with switchable ultra-dense viewpoints. *Adv. Sci.* **2025**, *12*, 2416961. [CrossRef]

110. Kim, H.; Suh, K.H.; Lee, E.C. Multi-modal user interface combining eye tracking and hand gesture recognition. *J. Multimodal User Interfaces* **2017**, *11*, 241–250. [CrossRef]

111. Akinyelu, A.A.; Blignaut, P. Convolutional neural network-based technique for gaze estimation on mobile devices. *Front. Artif. Intell.* **2022**, *4*, 11. [CrossRef]

112. Patney, A.; Salvi, M.; Kim, J.; Kaplanyan, A.; Wyman, C.; Benty, N.; Luebke, D.; Lefohn, A. Towards foveated rendering for gaze-tracked virtual reality. *ACM Trans. Graph.* **2016**, *35*, 179. [CrossRef]

113. Pretegiani, E.; Optican, L.M. Eye movements in parkinson's disease and inherited parkinsonian syndromes. *Front. Neurol.* **2017**, *8*, 592. [CrossRef] [PubMed]

114. Asmetha Jeyarani, R.; Senthilkumar, R. Eye tracking biomarkers for autism spectrum disorder detection using machine learning and deep learning techniques: Review. *Res. Autism Spectr. Disord.* **2023**, *108*, 102228. [CrossRef]

115. Cilia, F.; Carette, R.; Elbattah, M.; Guérin, J.L.; Dequen, G. Eye-tracking dataset to support the research on autism spectrum disorder. In Proceedings of the Workshop on Scarce Data in Artificial Intelligence for Healthcare, Vienna, Austria, 23 July 2022; pp. 59–64.

116. Prabha, A.J.; Bhargavi, R. Predictive model for dyslexia from fixations and saccadic eye movement events. *Comput. Methods Programs Biomed.* **2020**, *195*, 13. [CrossRef] [PubMed]

117. Hollenstein, N.; Rotsztejn, J.; Troendle, M.; Pedroni, A.; Zhang, C.; Langer, N. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Sci. Data* **2018**, *5*, 180291. [CrossRef]

118. Yang, D.W.; Wang, Y.; Wei, R.; Guan, J.P.; Huang, X.H.; Cai, W.; Jiang, Z. An efficient multi-task learning CNN for driver attention monitoring. *J. Syst. Architect.* **2024**, *148*, 9. [CrossRef]

119. Alletto, S.; Palazzi, A.; Solera, F.; Calderara, S.; Cucchiara, R. DR(eye)VE: A dataset for attention-based tasks with applications to autonomous and assisted driving. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 54–60.

120. Costela, F.M.; Castro-Torres, J.J. Risk prediction model using eye movements during simulated driving with logistic regressions and neural networks. *Transp. Res. Pt. F-Traffic Psychol. Behav.* **2020**, *74*, 511–521. [CrossRef]

121. Perrin, A.-F.; Krassanakis, V.; Zhang, L.; Ricordel, V.; Perreira Da Silva, M.; Le Meur, O. EyeTrackUAV2: A large-scale binocular eye-tracking dataset for UAV videos. *Drones* **2020**, *4*, 2. [CrossRef]