Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 19, No. 3, 2025

Deep Learning-Based Appearance-Based aze Estimation

Taowei Ge

School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

Abstract: Gaze estimation is a technology that primarily predicts gaze position or direction through eyes, facial images, or videos. It is widely used in fields such as gaming, healthcare, intelligent driving, and offline retail. In recent years, the rapid development of deep learning, with its end-to-end capability and robustness, has transformed many computer vision tasks and has been correspondingly applied in the field of gaze estimation. For recent deep learning-based gaze estimation methods, this article first introduces the fundamentals of gaze estimation, then summarizes methods based on CNN, Transformer, hybrid CNN-Transformer designs, and large models. It also presents commonly used mainstream gaze estimation datasets and applications. Finally, it provides an outlook on future trends and challenges in the field of gaze estimation.

Keywords: Gaze estimation; Gaze point; Convolutional neural network; Gaze estimation dataset; Deep learning; CNN; Transformer; LLM.

1. Introduction

The human eye not only acquires information from the external world but also transmits information outward. Human gaze is a primary channel for conveying attention, intention, and cognitive state; therefore, reliable gaze estimation is of fundamental importance. Gaze estimation technology has been widely applied in domains such as brick-and-mortar retail, medicine, and autonomous driving.

Conventional gaze estimation methods can be broadly divided into two categories. The first is model-based approaches, which rely on specialized hardware (e.g., infrared cameras) to capture ocular features such as the pupil center and corneal reflections, and then estimate gaze direction using geometric and optical models of the eyeball. Although these methods can achieve high accuracy, they are often cumbersome to deploy and exhibit limited generalization ability. The second category is traditional appearance-based approaches, which typically use hand-crafted eye features combined with regression models such as random forests or support vector regression (SVR) to directly predict gaze direction. These methods avoid complex physiological modeling and do not require specialized equipment, but they still suffer from poor robustness and generalization in complex real-world environments. With the rapid development of deep learning, appearance-based deep gaze estimation has emerged as a powerful alternative. Such methods can automatically extract gaze-related features from eye images or videos and perform end-to-end regression of the gaze direction, without relying on dedicated hardware or explicit physiological models. Moreover, they generally exhibit much stronger generalization across diverse and challenging conditions than traditional methods, which has led appearance-based deep gaze estimation to become the mainstream research direction. Researchers have conducted extensive explorations using a variety of network architectures, including CNN-based models, Transformer-based models, hybrid CNN–Transformer designs, and approaches built upon large-scale deep models.

This paper aims to provide a comprehensive and systematic review of appearance-based gaze estimation methods driven by deep learning. First, we introduce the fundamental concepts and problem formulations of gaze estimation. Then, we categorize and analyze existing methods according to their architectural paradigms, including CNN-based, Transformer-based, hybrid CNN–Transformer, and large-model-based approaches, with an emphasis on their model designs and respective advantages and limitations. Next, we summarize commonly used public datasets and evaluation metrics, and compare representative gaze estimation methods across different benchmarks. We further review practical application scenarios of gaze estimation. Finally, we discuss the current challenges in this field and outline promising directions for future research.

2. Background

2.1. 2D and 3D Gaze Estimation

The core task of gaze estimation is to predict a user’s point of regard or gaze direction from captured eye or facial images. According to the form of the output, gaze estimation can be broadly categorized into 2D and 3D gaze estimation [1].

In 2D gaze estimation, the user’s point of regard is represented directly as a two-dimensional coordinate on the display screen. The objective is to learn a mapping from the input image space to the screen coordinate system. This task is typically formulated as a regression problem, and the most commonly used evaluation metric is the Euclidean distance between the predicted and ground-truth gaze points.

In 3D gaze estimation, the goal is to predict a three-dimensional unit direction vector that originates from the eyeball center and points toward the user’s target of fixation. This provides a more intrinsic representation, as it explicitly describes the orientation of the eyeball in 3D space. In practical applications, the corresponding 2D point of regard on a screen can be obtained by intersecting this 3D gaze vector with a known display plane, thus establishing a link between 3D and 2D gaze estimation. A standard evaluation metric for 3D gaze estimation is the angular error, i.e., the angle between the predicted vector and the ground-truth vector.

2.2. Fundamental Concepts of Eye Tracking: Physiological Basis and Imaging Principles

Gaze estimation techniques are grounded in the physiological structure of the human eye and the principles of optical imaging. A clear understanding of these basic concepts is a prerequisite for understanding different technical approaches.

(1) Key physiological structures and feature points

The pupil is the circular opening at the center of the iris, responsible for regulating the amount of light entering the eye. In many traditional gaze estimation methods, the pupil center is used as a primary geometric reference point.

The iris is the colored part of the eye, commonly referred to as the "eye color." It lies between the cornea and the lens. Compared with the pupil, which undergoes significant dilation and constriction under varying illumination, the iris center is much less affected by lighting changes and can therefore serve as a more stable geometric reference.

Corneal reflection: When the eye is illuminated by an infrared (IR) light source, one or more small, very bright highlights appear on the corneal surface, known as corneal reflection points or glints. Traditional gaze estimation methods compute the gaze vector by exploiting the geometric relationship between the pupil center and the corneal reflection points (glints).

(2) Optical axis and visual axis

The optical axis is an idealized straight line connecting the corneal center and the pupil center. It represents the physical symmetry axis of the eyeball.

The visual axis is the actual line of sight that connects the corneal center and the fovea in the macular region, which is the area of highest visual acuity on the retina. The intersection of the visual axis with the display plane is referred to as the point of gaze (PoG) [2].

The kappa angle is the small angle between the optical axis and the visual axis. Due to the presence of this kappa angle, a subject-specific calibration procedure is required to compensate for the difference when converting from an optical-axis-based direction to the true visual axis.

2.3. Taxonomy of Gaze Estimation Methods

2.3.1. Model-based methods

Model-based methods [3] infer the eye gaze direction and point of regard from the geometric relationships among facial landmarks, the pupil center, and the corneal reflection points. The main advantage of such approaches lies in their clear physical interpretation and strong explainability. However, they also have notable drawbacks: they generally require specialized hardware, and typically demand time-consuming personal calibration for each user. Their performance is highly sensitive to the precision of the hardware setup, and often degrades significantly in complex scenarios, such as extreme head poses or challenging environmental conditions.

2.3.2. Deep learning-based appearance methods

In deep learning-based appearance methods, deep neural networks learn an end-to-end mapping from raw eye or face images to the gaze direction. The key breakthrough of this approach is the complete abandonment of hand-crafted feature design; instead, the network automatically learns which features are most informative for gaze estimation directly from data. As a result, deep learning-based methods generally achieve better performance than traditional gaze estimation techniques in complex environments and exhibit stronger generalization across varying conditions. Owing to their powerful feature learning capability and robustness in real-world scenarios, deep learning-based appearance methods have become the dominant research paradigm in contemporary gaze estimation.

2.4. Fundamentals of CNNs and Transformers

2.4.1. Convolutional Neural Networks (CNNs)

Convolutional neural networks (CNNs) are inspired by the human visual cortex. The earliest CNN model, LeNet [4], was proposed by Yann LeCun. A typical CNN architecture consists of convolutional layers, pooling layers, nonlinear activation layers, and fully connected layers. Among these, the convolutional layers are the most critical components, as they extract local features from input images via convolution operations.

Each convolutional layer applies multiple learnable kernels that slide over the input data with a specified stride, producing corresponding feature maps through convolution. Pooling layers are usually placed after convolutional layers to perform downsampling on the feature maps, thereby reducing the dimensionality of the data and mitigating overfitting. Nonlinear activation layers introduce nonlinearity into the network, enabling the model to approximate more complex functions and substantially enhancing its representational capacity. Fully connected layers are typically located at the end of a CNN; they learn the mapping between the features extracted by preceding layers and the final task objective, ultimately outputting the prediction results.

2.4.2. Transformer

The Transformer [5] was first introduced in the paper "Attention Is All You Need." It completely discards traditional convolution and recurrent structures for sequence modeling, and addresses key limitations of earlier architectures (such as RNNs and CNNs) in terms of parallel computation and long-range dependency modeling.

The core framework of the Transformer is an encoder–decoder architecture. Each encoder block consists of two sub-layers: a self-attention layer and a position-wise feed-forward network. Each decoder block contains these two sub-layers plus an additional encoder–decoder attention layer. Self-attention plays a central role in the Transformer: it dynamically computes the pairwise relevance among all elements in the input sequence, thereby enabling effective modeling of long-range dependencies. This design allows the Transformer to perform highly parallel computation while capturing global contextual relationships, laying the foundation for many subsequent advances in computer vision and natural language processing.

3. Appearance-Based Gaze Estimation Methods with Deep Learning

3.1. Taxonomy of CNN-Based Gaze Estimation Methods

Driven by the rapid development of deep learning, appearance-based gaze estimation methods have made remarkable progress. To systematically review current research, we categorize CNN-based gaze estimation methods along four dimensions: network architecture type, input feature type, output task type, and application scenario. For each category, we introduce representative studies and discuss their background, core methodology, input–output design, innovations, experimental results, and limitations. In what follows, we elaborate on these methods according to the above taxonomy.

3.1.1. Network Architecture Types

3.1.1.1 Lightweight CNN Architectures

MobileNet [6] and SqueezeNet [7] are typical representatives of lightweight convolutional neural networks. Their primary design goal is to reduce the number of parameters and computational cost. This is of particular importance in gaze estimation, where models are often deployed on edge devices and thus must be both lightweight and sufficiently accurate.

In the specific scenario of driver attention estimation, the end-to-end CNN architecture proposed in [8] replaces the traditional multi-stage pipeline and shows excellent generalization across drivers and camera setups. This study clearly demonstrates that, when directly processing field-of-view images captured in the driving cabin, a SqueezeNet-based model can achieve an inference speed of 166.7 Hz. Even when face detection is added as a preprocessing step, the system can still maintain a real-time performance of approximately 16 Hz.Subsequent research has further focused on architectural lightweighting. For example, SLeNet, proposed in [9], uses monocular eye images in combination with auxiliary mouth-corner features learned in a branch network. Standard convolutions are replaced with depthwise separable convolutions, and the number of fully connected layers is reduced. On the MPIIGaze [12] dataset, SLeNet achieves lower mean squared error (MSE) than LeNet and VGG-16, while reducing the number of parameters by an order of magnitude.The LiAGE model proposed in [10] first uses Ghost convolutions to extract features from the left and right eyes, then fuses them via a squeeze-and-excitation (SE) module. The fused features are concatenated with geometric features modeled by GATv2, and finally regressed to 2D screen points through fully connected layers. The model contains only 0.62 M parameters, yet its accuracy on the GazeCapture dataset is comparable to the current state of the art, with a much smaller model size.By 2025, lightweight gaze estimation has seen further architectural innovations. GazeCapsNet, introduced in [11], is the first method to apply capsule networks to gaze estimation. Building upon this, it replaces the traditional iterative routing mechanism in capsule networks with a self-attention routing scheme, substantially improving computational efficiency. The model has only 11.7 M parameters and an inference speed of around 20 ms per frame, making it highly suitable for deployment on edge devices.

In summary, lightweight gaze estimation models have evolved from straightforward adoption of small CNNs to more advanced structure-aware designs, such as depthwise separable convolutions and improved capsule networks. These approaches effectively reduce parameter count while preserving accuracy and generalization capability.

3.1.1.2 Large-Scale Deep CNN Architectures

Large-scale deep CNNs (e.g., ResNet, VGG) possess strong feature extraction capabilities and have achieved higher accuracy in gaze estimation tasks. Similar to many other computer vision problems, deeper CNN architectures with more parameters tend to learn more robust and generalizable gaze representations from large-scale datasets.GazeNet, proposed in [12], employs VGG-16 as the backbone network. It extracts appearance features from monocular eye images, fuses them with head pose vectors, and then regresses to gaze angles through fully connected layers. GazeNet achieves state-of-the-art performance in cross-dataset evaluations on MPIIGaze and EYEDIAP, and is the first appearance-based gaze estimation model built upon a deep VGG-16 backbone.Building on this work, [13] proposes a dual-stream VGG-16 architecture that extracts features from both eyes. An additional head pose estimation network is used to obtain head pose vectors, which are concatenated with eye features. Moreover, a GAN-based module is introduced to semantically inpaint regions occluded by eyeglasses after removing the eye tracker, thereby restoring realistic eye appearance. The fused features are then used to predict gaze direction. This approach achieves state-of-the-art performance on UT Multi-View, MPIIGaze, and RT-GENE. The authors explicitly state that “deeper networks are better suited to this type of data,” empirically confirming that deep architectures exhibit superior performance and generalization under natural conditions with large head pose variations.In parallel, Gaze360, proposed in [46], uses ResNet-18 combined with an LSTM to process 7-frame image sequences. It outputs both the gaze direction and an uncertainty estimate via quantile regression, which reflects the reliability of the predicted gaze. Prior methods generally predicted gaze from a single image frame, making them vulnerable to motion blur and illumination variations.To address this, [15] introduces VGE-Net, in which CAVE is first used to generate multiple candidate gaze maps. Each candidate map is passed through a DenseNet to predict a candidate gaze direction. A GDF module then adaptively fuses these candidate directions via weighted summation to obtain the final output. VGE-Net achieves state-of-the-art performance on MPIIGaze [12], EYEDIAP, and Columbia.The AGG framework in [16] further proposes replacing the traditional fully connected regression layer with an analytic geometric module (GPM) containing only 10 parameters. This significantly suppresses overfitting and domain shift in the last layer, while retaining deep backbones such as ResNet-18 and VGG-16, thereby offering a new perspective on optimizing the final layers of large-scale deep CNNs.

In summary, deep CNNs indeed achieve higher predictive accuracy on complex, large-scale datasets. However, simply increasing network depth can exacerbate overfitting, and the limited receptive field of CNNs restricts their ability to capture global context. Combining CNNs with Transformers offers a promising way to alleviate these issues.

3.1.2. Input Feature Types

3.1.2.1 Local Appearance Input

In recent years, a line of work has focused on achieving accurate gaze estimation using only local eye-region images [18]. These methods compensate for the absence of head pose and global facial context by incorporating explicit structural priors or learning physically consistent representations. They rely solely on eye-region images and adopt different representation learning strategies to enhance geometric awareness and generalization.

A representative early work is Deep Pictorial Gaze (DPG) proposed in [18] in 2018. DPG decomposes gaze estimation into two stages: first, it regresses a semantic intermediate representation called a “gazemap” from a single-eye image. The gazemap is a Boolean map encoding the iris and eyeball geometry. Then, a lightweight network regresses gaze angles from this gazemap. The method demonstrates strong robustness and generalization across multiple datasets.In contrast to the explicit intermediate representation used in DPG, [19] proposes an unsupervised learning approach in 2020. It learns a gaze representation from paired eye images using an "align-and-redirect" task, while introducing a deformation-field regularization to enforce consistency with physical motion patterns. With only a small number of calibration samples, this method can achieve accurate gaze estimation.More recently, [20] proposes MSGazeNet (2024), which adopts a multi-stream input structure. In addition to raw eye images, segmentation masks of the iris and eyeball extracted by a segmentation network are fed into the model. Gaze is regressed by fusing multiple channels. MSGazeNet does not rely on head pose annotations and exhibits strong robustness in cross-domain scenarios, highlighting the effectiveness of segmentation priors.

Although these methods differ in their specific strategies—DPG's explicit intermediate representation, the unsupervised physically grounded representation in [19], and multi-stream prior injection in MSGazeNet—they share a common feature: they depend solely on eye-region images and leverage structural design or learning mechanisms to extract geometric features effectively. This demonstrates that accurate gaze estimation is feasible with eye-region input alone and provides a viable pathway for low-annotation and cross-domain generalization.

3.1.2.2 Global Appearance Input

Methods that rely on local eye-region input for gaze estimation offer several clear advantages: the input is small, the model size can be reduced, and inference is fast. However, they also have limitations, such as high sensitivity to head pose and the lack of rich contextual information. To address these issues, some researchers have proposed using full-face images as input. Full-face input provides richer context and stronger robustness to various perturbations [17].Early studies mainly aimed to demonstrate that full-face images outperform eye-only input. The work presented in [21] at CVPRW is the first to perform gaze estimation using full-face appearance. It uses entire face images as input and introduces a spatial weighting mechanism that dynamically adjusts the contributions of different facial regions. On the MPIIGaze [12] and EYEDIAP datasets, this method improves accuracy by 14.3% and 27.7%, respectively, and maintains strong performance even under extreme head poses. This highlights the advantages of full-face input in realistic scenarios.Subsequent work pushes gaze estimation towards finer-grained modeling. CA-Net, proposed in [22], adopts a "coarse-to-fine" collaborative estimation strategy. It first extracts coarse gaze features from the full face, and then refines them using residual information from eye-region patches. CA-Net achieves state-of-the-art 3D gaze estimation performance on MPIIGaze [12] and EYEDIAP.To tackle the challenge of domain shift, [23] proposes PureGaze, which is the first to introduce domain generalization into gaze estimation. By employing self-adversarial learning, it extracts gaze-relevant components from full-face features while suppressing nuisance factors such as facial expressions. Using only source-domain data for training, PureGaze significantly improves cross-dataset performance.

Overall, full-face input methods have evolved from simple holistic feature use to more advanced multimodal fusion and domain generalization designs. As a result, they can better handle key challenges such as free head motion and illumination variation in unconstrained environments.

3.1.2.3 Hybrid Input

Earlier studies predominantly relied on single-modality, single-input designs. In contrast, many recent methods adopt hybrid input strategies that integrate multiple modalities such as face, eyes, and head pose to improve accuracy.

A pioneering work in this direction is the iTracker model proposed in [14]. iTracker employs a multi-branch CNN architecture that takes full-face images, left and right eye images, and a face-grid mask as inputs. Trained on the large crowdsourced GazeCapture dataset, it is the first to achieve real-time gaze estimation (10–15 fps) on mobile devices and reaches state-of-the-art performance on both GazeCapture and MPIIFaceGaze, thereby laying the foundation for subsequent hybrid-input gaze estimation approaches.To further enhance performance in natural environments, [13] introduces RT-GENE. This model uses semantically inpainted left and right eye images (where the regions originally occluded by the eye tracker have been restored) and head pose vectors as hybrid inputs. It achieves state-of-the-art performance on MPIIGaze [12], EYEDIAP, and several other datasets, and shows strong cross-dataset generalization.In 2019, [12] proposes the GazeNet model along with the MPIIGaze [12] dataset, and systematically analyzes the benefits of hybrid input. Normalized eye images are fed into a 13-layer convolutional network; the extracted features are then concatenated with head rotation angles in the fully connected layers. The model achieves an average angular error of 10.8° on MPIIGaze [12]. However, these early hybrid approaches mostly conduct simple feature concatenation.To address this limitation, [24] introduces AFF-Net, an adaptive feature fusion network that employs a squeeze-and-excitation attention mechanism to adaptively fuse binocular features based on their similarity, while simultaneously using global facial features and eye bounding-box information to guide hybrid feature extraction.Further, [25] proposes EG-Net, which adopts a dual-branch architecture. A base CNN processes full-face images, whereas an EE-Net branch handles binocular images, applying compound scaling across depth, width, and resolution, and incorporating attention mechanisms. This is also the first application of compound model scaling strategies to gaze estimation.By 2025, [26] proposes GA3CE, which takes depth maps, RGB images, camera intrinsics, and 2D bounding boxes for the head and full body as inputs, and predicts gaze direction by reasoning over normalized spatial information.

In summary, hybrid input methods combine multimodal information and advanced fusion mechanisms to effectively cope with challenges such as illumination changes and extreme head poses in complex real-world scenarios, thereby continuously improving the generalization performance of gaze estimation models.

3.1.3. Output Task Types

3.1.3.1 Direction Regression

In CNN-based gaze estimation, a large body of work formulates the output task as direction-angle regression, directly predicting parameters such as yaw and pitch, and thereby mapping appearance features to a 3D gaze vector.

A representative early work is GazeNet [12], which uses VGGNet to extract features from eye images and combines them with head pose features to directly regress gaze direction. To improve stability under dynamic conditions, Gaze360 [46] introduces an LSTM to aggregate temporal information from video sequences and regress 3D gaze vectors, while simultaneously outputting uncertainty estimates.In 2025, [27] proposes ADGaze, which adopts a classification-then-regression framework. It uses an anisotropic Gaussian label distribution to guide gaze direction learning, thereby improving prediction accuracy. In the same year, GA3CE [26] is proposed, which uses multi-stream inputs and directly predicts gaze direction through reasoning over normalized spatial information.

Overall, direction regression methods have evolved from simple coordinate mapping to end-to-end frameworks that integrate spatiotemporal context and multimodal information. This continuous evolution has improved accuracy and robustness in complex real-world environments and provides key support for practical deployment of gaze estimation technologies.

3.1.3.2 Screen Coordinate Regression

In CNN-based gaze estimation research, screen coordinate regression methods aim to directly predict the 2D point of gaze (PoG) on the display from facial or eye images. This provides an intuitive and practical output form for human–computer interaction applications.A representative early work is the iTracker model proposed in 2016 in [14]. iTracker uses a multi-branch CNN architecture that fuses full-face images, binocular eye images, and a face-grid mask to directly regress 2D screen coordinates.For tablet usage scenarios, [28] constructs the TabletGaze dataset in 2017 and proposes a method that combines HoG features with random forests to regress 2D gaze points. The authors also systematically analyze the impact of head pose and eyeglasses occlusion on estimation accuracy.In recent years, [29] proposes a novel cross-task few-shot learning framework. It leverages a pretrained 3D gaze estimation network as a prior and integrates a differentiable projection module with learnable screen parameters. This design enables accurate PoG regression on new devices using only a small number of calibration samples.

In summary, screen coordinate regression methods have evolved from early end-to-end regression using large-scale data, to multi-branch and multi-stream architectures, and further to cross-task few-shot learning frameworks. These advances continuously improve the accuracy of PoG prediction and enhance the practicality of gaze estimation in real-world interactive systems.

3.2. Transformer-Based Gaze Estimation Methods

Within deep learning, the Transformer architecture first achieved major breakthroughs in natural language processing (NLP). Subsequently, Vision Transformer (ViT) [30] successfully extended this architecture to image classification and other computer vision tasks, demonstrating strong global modeling capabilities. It is therefore natural that researchers began to explore the application of Transformers to gaze estimation.

The first work to apply Transformers to gaze estimation is [31], which proposed two models: a pure Transformer and a hybrid Transformer. Although the pure Transformer variant underperforms the hybrid model, this study marks the first attempt to introduce Transformers into the gaze estimation domain. Building on this, TransGaze [32] employs ViT to perform end-to-end regression directly from binocular eye images and full-face images, achieving new state-of-the-art (SOTA) results on multiple datasets.Furthermore, [33] proposes a general architecture named Gaze Transformer (GaT), which can handle both image and video inputs. GaT is combined with a self-training weakly supervised framework (ST-WSGE) that leverages 2D gaze-following annotations to generate pseudo 3D gaze labels, thereby significantly improving performance across multiple datasets and achieving SOTA results on several benchmarks. In addition, [34] introduces a lightweight hybrid model that combines Transformer and graph neural network (GNN) components. Despite having only 3.72M parameters, this model attains high accuracy on multiple datasets.

3.3. CNN–Transformer Hybrid Methods for Gaze Estimation

In conventional computer vision tasks, hybrid architectures combining CNNs and Transformers often yield performance gains. This design paradigm has likewise been adopted in the gaze estimation community.

3.3.1. Early Fusion

CNNs and Transformers each have their own strengths and limitations. In computer vision, it is common practice to combine them, and this strategy has been extended to gaze estimation as well. The earliest work to introduce Transformers into gaze estimation is GazeTR [31], proposed in 2022. GazeTR adopts a hybrid CNN–Transformer architecture, where full-face images are first processed by a CNN backbone to extract local features, which are then passed to a Transformer for global modeling in an early fusion manner. This hybrid model outperforms the pure Transformer variant and achieves SOTA on multiple benchmarks with fewer parameters. This work lays the foundation for subsequent CNN–Transformer hybrid designs in gaze estimation.

In the same period, [35] proposes a sequential architecture that combines convolution, self-attention, and deconvolution layers. Compared with previous pure ViT-based methods, this model uses fewer parameters while maintaining competitive accuracy. Similarly, [36] introduces Res-Swin-GE and SwinT-GE. In Res-Swin-GE, image features are first extracted by ResNet-18 and then fed into a Swin-Transformer, achieving strong performance on the MPIIFaceGaze and EYEDIAP datasets.

3.3.2. Mid-Level Fusion

Mid-level fusion of CNN and Transformer features can achieve deep integration of local features and global context at different stages of representation learning.

GazeSymCAT, proposed in [37], uses a ResNet backbone augmented with dynamic channel attention (DCA) to extract features from both the face and the eyes. These features are then fused via a Transformer module. The model exhibits excellent performance on ETH-XGaze and other datasets containing extreme head poses, achieving SOTA results.In [38], the CNN–Transformer hybrid architecture is extended to multi-view gaze estimation, leading to the DV-Gaze method. DV-Gaze introduces a dual-view interactive convolution module to fuse binocular information along epipolar directions at multiple scales. The fused features are then processed by a dual-stream Transformer to aggregate global information, followed by regression of gaze direction. The model achieves leading performance on ETH-XGaze and EVE, demonstrating that multi-view methods are more advantageous than single-view approaches.For third-person gaze prediction, [39] proposes the Multi-task Mutualistic Transformer (MMTR). MMTR employs a shared encoder composed of ResNet-50 and a Transformer encoder. Mid-level fusion is performed via a mutualistic attention module

(MAM), and global–local positional encodings are introduced to enhance spatial awareness. The model jointly predicts the head bounding box and the gaze target. On GazeFollow and VideoAttentionTarget, MMTR achieves leading performance, with AUC scores of 0.936 and 0.938, respectively.

3.3.3. Late Fusion

Late fusion of CNN and Transformer branches often allows each component to fully exploit its strengths while improving computational efficiency and overall model performance.

The MRFT model proposed in [40] is the first to introduce neural architecture search (NAS) into the gaze estimation domain. NAS is used to discover a multi-branch CNN backbone that extracts feature maps at multiple resolutions. These multi-scale features are then fused by MRFT, a Transformer-based multi-resolution fusion module, to output the final gaze direction.Similarly, [41] uses ResNet-50 to extract image features, which are then partitioned into patches and augmented with positional encodings. These patch embeddings are subsequently fed into a Transformer for feature fusion, and the model jointly predicts both gaze direction and head pose.Overall, combining CNNs and Transformers in gaze estimation enables both components to fully leverage their complementary advantages. CNNs excel at capturing local details such as eye shape, while Transformers are adept at modeling global relationships, such as the coupling between head orientation and gaze direction. As a result, CNN–Transformer hybrid models tend to exhibit stronger generalization capability across diverse application scenarios.

3.4. Large-Model-Based Gaze Estimation Methods

In recent years, large-scale models have developed rapidly and have begun to be applied to gaze estimation tasks.

In 2025, [42] proposes the LightGAZE-DALK model, which combines deformable approximations of large convolutional kernels with spatial–channel self-attention, while also incorporating prior knowledge from Stable Diffusion v1.5. This design enables a relatively compact model to achieve strong generalization to diverse eye images. In deployment experiments on the iPhone 15, the LightGAZE-DALK model with 830M parameters attains an inference latency of 12.3 ms per frame. In terms of accuracy, the 830M-parameter LightGAZE-DALK achieves state-of-the-art performance on MPIIFaceGaze (3.91°) and Gaze360 (10.01°), demonstrating a favorable balance between accuracy and model complexity.Also in 2025, [43] proposes the Gaze_Lle model, which adopts Meta's DINOv2 as a generic visual feature encoder. During training, DINOv2 is kept frozen, and only a lightweight task-specific head with 2.8M parameters is trained, augmented by head-pose prompts. This approach reduces the number of trainable parameters by 95%, while achieving SOTA performance on multiple datasets (GazeFollow, VAT, ChildPlay).In the same year, [44] introduces the OmniGaz framework. OmniGaz first trains a teacher model on labeled datasets, then uses the teacher to generate pseudo-labels for a large-scale unlabeled dataset. A reward model is employed to evaluate the quality of these pseudo-labels. During scoring, the reward model integrates three signals: image features extracted by CLIP, a textual gaze description generated by InstructBLIP for the same image, and the pseudo-label itself. By fusing these three sources of information, the reward model outputs a confidence score for each pseudo-label. High-confidence pseudo-labeled samples are then combined with the original labeled data to jointly train a student model. Importantly, the large models (e.g., CLIP, InstructBLIP) only participate in the pseudo-label evaluation process and are not directly used for gaze prediction. As a result, only the student model needs to be deployed at inference time. Trained on massive web-scale image data covering a wide variety of head poses, illumination conditions, and capture environments, the student model acquires extremely strong generalization ability. It achieves outstanding performance—and SOTA results both in-domain and cross-domain—on multiple mainstream gaze datasets.In summary, large models possess rich prior knowledge and strong representational capacity. By leveraging these capabilities to guide the training of compact student models, it is possible to substantially improve generalization while keeping inference efficient. Empowering lightweight gaze estimators with large-model priors is likely to be a major trend in the future development of gaze estimation.

4. Commonly Used Datasets for Gaze Estimation

Large-scale, high-quality public datasets constitute a critical foundation for gaze estimation research. This section summarizes the main gaze estimation datasets and their characteristics. As shown in Table 1, we provide a comprehensive comparison of commonly used gaze datasets along multiple dimensions, including the number of participants, number of cameras, number of images, data modality, image type, annotation type, capture distance, head pose variation, and recording environment. These datasets span controlled laboratory settings to in-the-wild scenarios, single-user to multi-user recordings, and a variety of imaging devices and gaze annotation protocols, thereby offering diverse conditions for training and evaluating deep learning models.

(1) MPIIGaze [12]: This dataset contains a total of 213,659 images from 15 participants. The images were collected over several months during the participants' daily laptop use with a built-in webcam, capturing realistic eye appearance and natural illumination variations. The dataset provides both 2D and 3D gaze annotations.

(2) EyeDiap [45]: EyeDiap consists of 94 video sequences recorded from 16 subjects in a laboratory environment. Data collection mainly relies on an RGB-D camera and a high-definition camera. For each participant, six sequences are recorded, covering two modes: static head pose and free head movement. A limitation of this dataset is the relatively limited illumination variability.

(3) GazeCapture [14]: Collected via crowdsourcing, GazeCapture comprises 2,445,504 images from 1,474 participants and is one of the largest gaze estimation datasets to date. Data are captured using front-facing cameras of mobile devices such as tablets and smartphones. Head pose is not constrained during acquisition; participants are only required to follow a moving dot on the screen. A key limitation is that the dataset provides only 2D gaze annotations, and thus it is restricted to PoG regression tasks.

(4) Gaze360 [46]: This dataset contains 172,000 images from 238 participants, captured in both indoor and outdoor environments. It covers a wide range of head poses, gaze directions, and illumination conditions, and provides 3D gaze annotations.

(5) InvisibleEye [47]: InvisibleEye combines synthetic and real data. Millimeter-scale RGB cameras embedded in eyeglass frames are used to capture eye images. Due to the extremely low resolution of the captured images (only 5×5 pixels), four cameras are used simultaneously for multi-view acquisition. In total, 280,000 near-eye images from 17 participants are collected.

Table 1. Gaze Estimation Benchmark Datasets

Dataset Name	Number of Participants	Number of Cameras	Number of Images	Data Type	Image Type	Annotation Type	Distance	Head Pose	Environment
NVGaze [48]	35	Not specified	2500000	IR RGB	Eye images	3D	Close range	Variable	Synthetic
UT-multiview [50]	50	8	64000	RGB	Eye images	2D,3D	60cm	Variable	Indoor
MPIIGaze [12]	15	1	213659	RGB	Face + eye	2D,3D	40-60cm	Frontal	Indoor
EYEDIAP [45]	16	2	62500	RGB-D	Face + eye region	3D	80-120cm	Frontal	Indoor
RT-GENE [13]	15	1	277286	RGB-D	Face + eye region	3D	80-280cm	Variable	Outdoor
UnityEye [51]	N/A	N/A	1000000	RGB	Eye images	3D	Variable	Variable	Synthetic
Gaze360 [46]	238	5	172000	RGB	Full images	3D	Variable	Variable	Indoor and outdoor
ETH-XGaze [52]	110	18	1000000	RGB	Face region	2D,3D	Not specified	Variable	Indoor
ARGaze [53]	25	N/A	1321968	RGB	Eye images	3D	Not specified	Unspecified	Indoor
GazeCapture [14]	1474	1	2445504	RGB	Face region	2D	Variable	Variable	Outdoor
Columbia [54]	56	1	5880	RGB	Full face	3D	2m	Variable	Indoor
U2Eyes [55]	N/A	N/A	5875000	RGB	Eye images	2D ,3D	Not specified	Not specified	Synthetic
MagicEyes [49]	587	N/A	800000	IR RGB	Eye images	2D,3D	Not specified	Not specified	Indoor
MPIIFACEGaze [21]	15	1	37667	RGB	Face region	3D	40-60cm	Frontal	Indoor
NISLGaze [56]	21	N/A	2079videos	RGB	Face region	3D	90cm	Variable	Indoor
TabletGaze [28]	51	1	100000	RGB	Tablet camera videos	2D	30-50cm	Frontal	Indoor
EvE [57]	54	N/A	4.2K videos	RGB	Not specified	2D,3D	Not specified	Not specified	Synthetic
InvisibleEye [47]	17	4	280000	RGB	Not specified	3D	Not specified	Not specified	Synthetic

(6) NVGaze [48]: NVGaze consists of two parts. The first is a synthetic dataset of 2 million images (1280×960) that spans variations in face shape, gaze direction, pupil and iris appearance, skin tone, and external conditions. The second is a real near-eye IR dataset with 2.5 million images (640×480) from 35 participants.

(7) MagicEyes [49]: MagicEyes is specifically designed for mixed reality (MR) devices and is collected using IR cameras integrated into MR headsets. The dataset includes 587 subjects with diverse gender, age, ethnicity, and eye color, and contains more than 800,000 images, of which 80,000 have manually annotated ground-truth labels.

For within-dataset evaluation, we categorize methods according to both the input image type and the definition of the gaze origin. Specifically, when evaluating methods that assume the eyeball center as the origin, we use eye-region images from MPIIGaze [12] and EyeDiap. For methods that define the face center as the gaze origin, we use face images from MPIIFaceGaze, EyeDiap, Gaze360, RT-GENE, and ETH-XGaze. The corresponding quantitative results are summarized in Table 2.

Table 2. Performance Evaluation of Gaze Estimation Methods

Methods	MPIIGaze	EyeDiap	MPIIFaceGaze	EyeDiap	Gaze360	RT-Gene	ETH-XGaze
GazeNet [12]	5.70°	7.13°	5.76°	6.79°	N/A	N/A	N/A
Dilated-Net [57]	4.39°	6.57°	4.42°	6.19°	13.73°	8.38°	N/A
Gaze360 [46]	4.07°	5.58°	4.06°	5.36°	11.04°	7.06°	4.46°
RT-Gene [13]	4.61°	5.30°	4.66°	6.02°	12.96°	8.60°	N/A
FullFace [21]	4.96°	6.76°	4.93°	6.53°	14.99°	10.90°	7.38°
RCNN [59]	N/A	N/A	4.39°	6.31°	11.30°	10.09°	N/A
CA-Net [22]	4.27°	5.63°	4.27°	5.27°	11.20°	8.27°	N/A
GazeTR Pure [31]	N/A	N/A	4.74°	5.72°	13.85°	8.68°	N/A

4.1. Evaluation Metrics

In gaze estimation research, performance metrics for gaze prediction vary depending on whether the target is two-dimensional or three-dimensional. The following summarizes the evaluation criteria commonly used in the literature and groups them into two main categories: metrics for 2D gaze estimation and metrics for 3D gaze estimation, with some measures applicable to both types of tasks.

4.1.1. Evaluation Metrics for 2D Gaze Estimation

2D gaze estimation primarily evaluates the prediction accuracy of the point of gaze (PoG) on the screen, and may also include the accuracy of gaze event classification and spatial attention localization.

(1) L2 distance / Average Euclidean Distance

Some studies [14] use the L2 distance, also referred to as the average Euclidean distance, to evaluate the accuracy of predicted gaze points. It is defined as the average Euclidean distance between the predicted PoG coordinates and the ground-truth PoG coordinates, and can be computed as:

$$AED = \frac{1}{n}\sum_{i=1}^{n}\sqrt{(gt\_x_i - e\_x_i)^2 + (gt\_y_i - e\_y_i)^2} \qquad (1)$$

where gt_x_i, $gt_y_i$are the ground-truth coordinates of the i-th sample, $e\_x^{(i)}, e\_y^{(i)}$ are the corresponding predicted coordinates, and $n$is the total number of samples.

(2) Pixel Distance

Pixel distance is defined as the Euclidean distance (in pixels) between a single predicted PoG and the corresponding ground-truth PoG. It is the basic unit used to compute AED and can be expressed as:

$$pix\_shift = \sqrt{(GT_x - Gaze\_X)^2 + \left(GT_y - Gaze\_Y\right)^2} \qquad (2)$$

where GT_x, GT_ydenote the ground-truth PoG coordinates for a given sample, and $Gaze_X, Gaze_Y$denote the predicted PoG coordinates.

4.1.2. Evaluation Metrics for 3D Gaze Estimation

3D gaze estimation primarily evaluates the accuracy of the gaze direction (gaze vector) or the 3D point of gaze in space.

(1) Angular Error

Angular error is defined [14] as the angle between the predicted gaze direction vector and the ground-truth gaze direction vector. It is the most commonly used metric for assessing the accuracy of gaze direction estimation.

(2) Angular Accuracy

Angular accuracy is closely related to angular error and quantifies the angular deviation between the predicted and ground-truth gaze positions. It is typically obtained by converting pixel errors into angular errors using screen parameters and the viewing distance, and can be computed as:

$$\mathrm{ang_acc} = \frac{\left(\mu \cdot \mathrm{pix_shift} \cdot \mathrm{co\,s}\left(\mathrm{mean}(\theta)\right)\right)^2}{\mathrm{Est_GP}} \qquad (3)$$

where $\mu$is the physical size of a pixel, pix_shiftis the pixel error, Est_P is the estimated distance between the eye and the PoG on the screen, and $\theta$is the horizontal angle between the gaze direction and the screen normal.

4.2. General Evaluation Metrics

The following metric can be used directly for 2D PoG estimation and, with appropriate adaptation, also for 3D gaze direction estimation.

Mean Squared Error (MSE):

MSE is defined as the average of the squared differences between the predictions and the ground-truth values. In gaze estimation [59], it can be applied directly to regression outputs such as coordinates or direction vectors, and is given by:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - gt\_y^{(i)}\right)^2 \quad (4)$$

where $n$is the number of samples, $i$is the sample index, y_i denotes the predicted value, and $gt\_y^{(i)}$ denotes the corresponding ground-truth value.

5. Application

Gaze reflects human attention and encodes rich and diverse information. It has already been applied in domains such as gaming, intelligent driving safety, healthcare, and brick-and-mortar retail, enabling gaze estimation technologies to be deployed in real-world scenarios and to create commercial value, which in turn drives continuous technical progress. This section systematically reviews application paradigms of gaze estimation across different domains and analyzes how gaze technologies are gradually entering everyday life.

5.1. Gaming

In games, gaze is not merely a simple input modality; each gaze event can be endowed with specific meaning. In 2021, [61] proposed the Eyesthetics framework, a conceptual framework for “designing and analyzing gaze-centric game interactions.” This framework characterizes gaze interaction along four dimensions: who is looking, what is controlled by the eyes, what is being looked at, and what consequences are triggered by looking at a particular location. On this basis, “where you look” is elevated from a mere coordinate to a concrete interaction mechanism, for example, narrative triggering, teammate signaling, or skill enhancement. Thus, gaze becomes not just a point in space, but an interaction resource that can create meaning and shape player experience. Beyond experience design, gaze data are also used to improve player performance and training. A 2025 study on FPS games [62] shows that experienced players tend to rely more on saccades that directly jump to targets rather than on fixations. The experiments reveal that oculomotor–manual coordination efficiency under uncertainty has a significant impact on gameplay performance, and they encourage players to develop a “zero-fixation–single-saccade” habit. Gaze has also been extended to multi-user game interaction. In [63], the authors introduce the GET paradigm for group gaze interaction. Their results show that shared gaze substantially enhances players’ sense of immersion, perceived competence, positive affect, and social presence. However, the increased demand on attentional allocation also leads to cognitive overload, which manifests as prolonged reaction times. Further research is needed to refine such systems.

5.2. Intelligent Driving and Safety

In the domain of intelligent driving and safety, gaze estimation is a key technology for assessing driver attentional state and supporting autonomous driving, and has attracted increasing attention in recent years. In 2022, [64] directly regressed gaze angles from eye images or landmarks and used synthetic data to improve annotation quality. Their system employs in-cabin cameras to capture eye images of the driver and uses deep learning models to estimate gaze direction, addressing the driver monitoring requirements of L3 autonomous driving systems. Subsequently, [65] in 2025 advanced this research direction by addressing the lack of continuous ground-truth labels for driver readiness. They introduced a novel continuous label, the Readiness Index, into the DMD dataset and proposed a bidirectional LSTM model that fuses head pose and eye movement data. On the DMD dataset, the model achieved a mean absolute error of MAE = 0.363, indicating good performance. In 2025, [66] proposed the TDGH-YOLOv7 framework, which incorporates a Transformer into YOLOv7 to enable simultaneous detection of head pose and gaze direction. The framework achieves a weighted accuracy of 95.02% on multiple datasets, with vertical and horizontal RMSEs of 2.23 and 1.68, respectively. These results significantly improve accuracy, latency, and computational efficiency in complex environments and provide strong support for real-time safety applications in advanced driver assistance and autonomous driving systems.

5.3. Healthcare

In healthcare applications, gaze behavior serves as a non-invasive biomarker and plays a crucial role in early intervention for disorders such as autism spectrum disorder (ASD) and amyotrophic lateral sclerosis (ALS). Using video and eye-tracking technology, [67] demonstrated that ASD can be preliminarily screened within just 15–20 seconds, with an accuracy exceeding 92%. This highlights the high accessibility of gaze-based screening in clinical pre-assessment and community healthcare. To promote large-scale, home-based screening, [68] introduced a gamified mobile application, Guess What, in 2022. The app uses the front-facing camera of a mobile device to collect more than 11 hours of in-home video data and analyzes gaze fixation patterns and visual scanning behaviors of children with ASD during social interactions. An LSTM model is employed to achieve preliminary ASD prediction. A key contribution is that ASD screening can be conducted at home without specialized eye-tracking hardware. In 2025, [69] applied interpretable machine learning to eye-tracking data from 93 children. Their model achieved an F1 score of 0.63 in distinguishing ASD from developmental language disorder (DLD), substantially reducing dependence on traditional clinical scales. Beyond algorithmic advances for improving diagnostic accuracy, there are also efforts to build foundational datasets for disease research. In 2024, [70] released the first multimodal EEG–eye-tracking dataset for ALS patients and healthy controls. This dataset records EEG and eye movement signals from 6 ALS patients at different disease stages and 170 healthy participants while they use a Vietnamese eye-tracking spelling system. The main contribution is the provision of a high-quality, open-source multimodal dataset that can serve as a benchmark for ALS research and brain–computer interface development, filling a critical gap in real-world multimodal data resources for this area.

5.4. Offline Retail

In brick-and-mortar retail environments, gaze estimation is mainly used to analyze consumer attention mechanisms in depth, thereby optimizing store layout and marketing strategies. In 2021, [71] conducted a study that synchronously collected eye-tracking data and analyzed Google Analytics logs. The results show that Google Analytics logs capture only about half of consumer interaction behavior, whereas eye-tracking can record all user activities on a mobile fashion retail platform. This provides a new perspective on web and app analytics, enabling companies to more accurately reconstruct customer shopping journeys and optimize mobile user experience.In 2024, [72] proposed a practical guideline for conducting eye-tracking studies in retail settings. By analyzing gaze data such as in-store fixation paths and dwell times, the guideline offers actionable insights for shelf arrangement and promotional display design.Similarly, [73] reported in 2024 that gaze patterns—such as fixation duration and count—can effectively reflect consumer goals (hedonic vs. utilitarian) and information preferences (product attributes vs. subjective experiences. Consumers with utilitarian goals tend to focus more on reviews related to product performance and spend more time reading them, whereas consumers with hedonic goals pay more attention to experiential reviews and also devote longer viewing times to them. These findings suggest that retailers can tailor their promotional wording to match the type of product and the likely consumer goal, thereby increasing consumers' willingness to purchase.

6. Summary

This paper presents a summary and investigation of gaze estimation technologies, with a particular focus on the research progress of deep learning-based gaze estimation. We review and discuss the state of the art and recent advances from seven perspectives: the fundamentals of gaze estimation, CNN-based methods, Transformer-based methods, hybrid CNN–Transformer methods, large-model-based approaches, commonly used datasets, and application domains. Although deep learning methods have made remarkable progress in recent years, they still face significant challenges in real-world deployment. Several research directions require further breakthroughs. First, there is a lack of large-scale, high-quality, real-world datasets; most existing large-scale datasets are synthetic, which limits the further development of algorithms. Second, there is a critical trade-off between computational efficiency and accuracy for models on edge devices: models must be lightweight while still achieving high accuracy and fast inference. Finally, deep learning models largely remain "black boxes"; their internal computation and decision-making processes cannot yet be clearly and reliably interpreted, which leads to a crisis of trust when deploying such models in safety-critical domains such as medicine and autonomous driving.

Overall, deep learning-based gaze estimation methods have fundamentally reshaped the research paradigm of this field and demonstrated substantial potential. With the continuous advancement of related technologies, gaze estimation is expected to make increasingly significant contributions to human society.

References

[1] Kwon Y M, Jeon J S, Ki J. 3D gaze estimation and interaction to stereo display[J]. International Journal of Virtual Reality, 2006, 5(3): 41-45.

[2] Wang K,Wang S,Ji Q. Deep eye fixation map learning for calibration-free eye gaze tracking[C]. Charleston: Ninth Biennial ACM Symposium on Eye Tracking Research & Applications (ETRA '16),2016:47-55.

[3] Kar A,Corcoran P. A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms[J]. IEEE Access, 2017, 5:16495-16519.

[4] LeCun Y,Bottou L,Bengio Y,Haffner P. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11):2278-2324.

[5] Vaswani A,Shazeer N,Parmar N, et al. Attention is all you need[C]. Long Beach: Advances in Neural Information Processing Systems 30 (NIPS 2017),2017:5998-6008.

[6] Howard A G,Zhu M,Chen B,et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications[EB/OL].(2017-04-17)[2025-10-22]https://doi.org/10.48550/arXiv.1704.04861

[7] Iandola F N,Han S,Moskewicz M W,et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size[EB/OL].(2016-02-24)[2025-10-22] https://doi.org/10.48550/arXiv.1602.07360

[8] Vora S,Rangesh A,Trivedi M M. Driver gaze zone estimation using convolutional neural networks: A general framework and ablative analysis[J]. IEEE Transactions on Intelligent Vehicles, 2018, 3(3):254-265.

[9] Zhuang Y,Zhang Y,Zhao H. Appearance-based gaze estimation using separable convolution neural networks[C]. Chongqing: 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 2021:609-612.

[10] Holla N S,Kushwaha A,Sanchi C, et al. LiAGE: Light-weight Adaptive Gaze Estimation[C]. Bengaluru: Fifteenth Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP 2024),2024:1-8.

[11] Muksimova S,Valikhujaev Y,Umirzakova S, et al. GazeCapsNet: A lightweight gaze estimation framework[J]. Sensors, 2025, 25(4):1224.

[12] Zhang Xucong, Sugano Y, Fritz M, et al. MPIIGaze: Real-world dataset and deep appearance-based gaze estimation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(1): 162–175.

[13] Fischer T,Chang H J,Demiris Y. RT-GENE: Real-time eye gaze estimation in natural environments[C]. Munich: European Conference on Computer Vision (ECCV 2018),2018:339-357.

[14] Krafka K,Khosla A,Kellnhofer P, et al. Eye tracking for everyone[C]. Las Vegas: IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2016:2176-2184.

[15] Huang G,Shi J,Xu J, et al. Gaze estimation by attention-induced hierarchical variational auto-encoder[J]. IEEE Transactions on Cybernetics, 2024, 54(4):2592-2605.

[16] Bao Y,Lu F. From feature to gaze: A generalizable replacement of linear layer for gaze estimation[C]. Seattle: CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024),2024:1409-1418.

[17] Wen Mingqi, Ren Luqian, Chen Zhenqin, et al. Survey on gaze estimation methods based on deep learning [J]. Computer Engineering and Applications.2024,60(12):18-33.

[18] Park S,Spurr A,Hilliges O. Deep pictorial gaze estimation[C]. Munich: European Conference on ComputerVision (ECCV 2018),2018:741-757.

[19] Yu Y,Odobez J-M. Unsupervised representation learning for gaze estimation[C]. Seattle: CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020),2020:7312-7322.

[20] Mahmud Z,Hungler P,Etemad A. Multistream gaze estimation with anatomical eye region isolation by synthetic to real transfer learning[J]. IEEE Transactions on Artificial Intelligence, 2024, 5(8):4232-4246.

[21] Zhang X,Sugano Y,Fritz M, et al. It's written all over your face: Full-face appearance-based gaze estimation[C]. Honolulu: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2017),2017:2299-2308.

[22] Cheng Y H,Huang S Y,Wang F, et al. A coarse-to-fine adaptive network for appearance-based gaze estimation[C]. New York: AAAI Conference on Artificial Intelligence (AAAI 2020),2020:10623-10630.

[23] Cheng Y H,Bao Y W,Lu F. PureGaze: Purifying gaze feature for generalizable gaze estimation[C]. Vancouver: AAAI Conference on Artificial Intelligence (AAAI 2022),2022:436-443.

[24] Bao Y,Cheng Y,Liu Y,Lu F. Adaptive feature fusion network for gaze tracking in mobile tablets[C]. Milan: International Conference on Pattern Recognition (ICPR 2020),2021:9936-9943.

[25] Wu Xinmei, Li Lin, Zhu Haihong, et al. EG-Net: Appearance-based eye gaze estimation using an efficient gaze network with attention mechanism[J]. Expert Systems with Applications, 2024, 238:122363.

[26] Kawana Y,Shiba S,Kong Q,Kobori N. GA3CE: Unconstrained 3D gaze estimation with gaze-aware 3D context encoding[C]. Nashville: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025),2025:3081-3090.

[27] Li Duantengchuan,Wang Shutong,Zhao Wanli, et al. ADGaze: Anisotropic Gaussian label distribution learning for fine-grained gaze estimation[J]. Pattern Recognition, 2025, 164:111536.

[28] Huang Q,Veeraraghavan A,Sabharwal A. TabletGaze: Dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets[J]. Machine Vision and Applications, 2017, 28(5-6):445-461.

[29] Cheng, Yihua, Hengfei Wang, Zhongqun Zhang, Yang Yue, Bo Eun Kim, Feng Lu and Hyung Jin Chang. "3D Prior is All You Need: Cross-Task Few-shot 2D Gaze Estimation." 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025): 23891-23900.

[30] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition atscale[EB/OL].(2020-10-22)[2025-10-22]https://arxiv.org/abs/2010.11929

[31] Cheng Y,Lu F. Gaze estimation using transformer[C]. Montreal: International Conference on Pattern Recognition (ICPR 2022),2022:3341-3347.

[32] Ye Lang, Wang Xinggang, Yao Jingfeng, et al. Transgaze: exploring plain vision transformers for gaze estimation[J]. Machine Vision and Applications, 2024, 35,128.

[33] Vuillecard P,Odobez J-M. Enhancing 3D gaze estimation in the wild using weak supervision with gaze following labels[C]. Nashville: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025),2025:13508-13518.

[34] Yahyaabadi R, Nikan S. Efficient 2D/3D Gaze Estimation Using TGGNet: A Transformer Graph Approach[J/OL]. IEEE Transactions on Cognitive and Developmental Systems. [2025-03-28]. https://doi.org/10.1109/TCDS.2025.3600102.

[35] Oh J O,Chang H J,Choi S I. Self-Attention with Convolution and Deconvolution for Efficient Eye Gaze Estimation from a Full Face Image[C]. New Orleans,LA, USA:2022 CVF Conference on Computer Visionand Pattern Recognition Workshops (CVPRW),2022:4988-4996.

[36] Li Yujie, Chen Jiahui, Ma Jiaxin, et al. Gaze Estimation Based on Convolutional Structure and Sliding Window-Based Attention Mechanism[J]. Sensors, 2023, 23(13): 6226.

[37] Zhong Y,Lee S H. GazeSymCAT: A symmetric cross-attention transformer for robust gaze estimation under extreme head poses and gaze variations[J]. Journal of Computational Design and Engineering,2025,12(3):115-129.

[38] Cheng Y,Lu F. DVGaze: Dual-View Gaze Estimation[C]. Paris, France:2023 IEEE/CVF International Conference on Computer Vision (ICCV),2023:20575-20584.

[39] Chen W,Chai Y,Wu X J,et al. Privileged Information-Guided Multitask Mutualistic Transformer for Gaze Prediction[J]. IEEE Transactions on Multimedia, 2025, 27:7353-7368.

[40] Nagpure V, Okuma K. Searching efficient neural architecture with multi-resolution fusion transformer for appearance-based gaze estimation[C]. Waikoloa, HI, USA: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023: 890–899.

[41] Karmi R, Mastouri R, Rahmany I, et al. An Appearance-based VisionTransformer Network for Enhanced Gaze Estimation[J]. Signal, Image and Video Processing, 2025, 19: 742.

[42] CHEN X, CHEN M, CHEN Y, et al. Large Generative Model Impulsed Lightweight Gaze Estimator via Deformable Approximate Large Kernel Pursuit [J]. IEEE Transactions on Image Processing, 2025, 34: 1149-62.

[43] Ryan F,Bati A,Lee S,et al. Gaze-LLE: Gaze Target Estimation via Large-Scale Learned Encoders[C]. Nashville, TN, USA:2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),2025:28874-28884.

[44] Qu H, Wei J, Shu X, et al. OmniGaze: Reward-inspired Generalizable Gaze Estimation In The Wild[EB/OL].(2025-10-15)[2025-10-23]https://arxiv.org/abs/2510.13660

[45] 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[C]. Safety Harbor, FL, USA: Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA), 2014: 255-258.

[46] Kellnhofer P, Recasens A, Stent S, et al. Gaze360: Physically Unconstrained Gaze Estimation in the Wild[C]. Seoul, Korea (South): 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019: 6911-6920.

[47] Tonsen M, Steil J, Sugano Y, et al. Invisibleeye: Mobile eye tracking using multiple low-resolution cameras and learning-based gaze estimation[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2017, 1(3): 1-21.

[48] Kim J, Stengel M, Majercik A, et al. Nvgaze: An anatomically-informed dataset for low-latency, near-eye gaze estimation[C]. Glasgow, UK: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019: 1-12.

[49] Wu Z, Rajendran S, van As T, Zimmermann J, Badrinarayanan V, Rabinovich A. MagicEyes: A Large Scale Eye Gaze Estimation Dataset for Mixed Reality [EB/OL]. arXiv:2003.08806, 2020.

[50] Sugano Y, Matsushita Y, Sato Y. Learning-by-Synthesis for Appearance-based 3D Gaze Estimation[C]. Columbus, OH, USA: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014: 1821-1828.

[51] Wood E, Baltrušaitis T, Morency L P, et al. Learning an Appearance-Based Gaze Estimator from One Million Synthesised Images[C]. Charleston, SC, USA: Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications (ETRA), 2016: 131-138.

[52] Zhang Xucong, Park S, Beeler T, et al. ETH-XGaze: A Large Scale Dataset for Gaze Estimation Under Extreme Head Pose and Gaze Variation[C]. Glasgow, UK: European Conference on Computer Vision (ECCV 2020), 2020: 365-381.

[53] Yan Zihan, Wu Yue, Shan Yifei, et al. A dataset of eye gaze images for calibration-free eye tracking augmented reality headset[J]. Scientific Data, 2022, 9: 115.

[54] Smith B A, Yin Qi, Feiner S, et al. Gaze locking: passive eye contact detection for human-object interaction[C]. St Andrews, UK: The 26th Annual ACM Symposium on User Interface Software and Technology (UIST 2013), 2013: 271-280.

[55] Porta S, Bossavit B, Cabeza R, et al. U2Eyes: A Binocular Dataset for Eye Tracking and Gaze Estimation[C]. Seoul, Korea (South): 2019 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2019: 3660-3664.

[56] Chen Z K, Bertram E S. Towards high performance low complexity calibration in appearance-based gaze estimation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 1174-1188.

[57] Park S, Aksan E, Zhang X, et al. Towards end-to-end video-based eye-tracking[C]. Glasgow, UK: European Conference on Computer Vision (ECCV 2020), 2020: 747-763.

[58] Chen Zhaokang, Shi Bertram E. Appearance-Based Gaze Estimation Using Dilated-Convolutions[C]//Computer Vision – ACCV 2018. pp. 309–324.

[59] Palmero C, Selva J, Bagheri M A, et al. Recurrent CNN for 3D gaze estimation using appearance and shape cues[C]. Newcastle upon Tyne, UK: British Machine Vision Conference (BMVC), 2018: 1–11.

[60] Fang Yi, Tang Jiapeng, Shen Wang, et al. Dual Attention Guided Gaze Target Detection in the Wild[C]. Nashville, TN, USA: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 11390–11399.

[61] Ramirez Gomez A, Lankes M. Eyesthetics: making sense of the aesthetics of playing with gaze[J]. Proceedings of the ACM on Human-Computer Interaction, 2021, 5(CHI PLAY): 1–24.

[62] Yang L, Zhang W, Li P, et al. The aiming advantages in experienced first-person shooter gamers: Evidence from eye movement patterns[J]. Computers in Human Behavior, 2025, 165: 108573.

[63] Acartürk C, Fal M, Çakır M P. User performance and engagement in multi-user gaming environments: An experimental investigation through the group eye tracking (GET) paradigm[J]. Entertainment Computing, 2024, 51: 100714.

[64] Nikan S, Upadhyay D. Appearance-based gaze estimation for driver monitoring[C]. New Orleans: NeurIPS 2022 Workshop on Gaze Meets ML, 2022: 1–13.

[65] Kazemi M, Rezaei M, Azarmi M. Evaluating driver readiness in conditionally automated vehicles from eye-tracking data and head pose[J]. IET Intelligent Transport Systems, 2025, 19(1): e70006.

[66] Shah SM, Gan Zengkang, Sun Zhaoyun, et al. AI-enabled driver assistance: monitoring head and gaze movements for enhanced safety[J]. Complex & Intelligent Systems, 2025, 11: 297.

[67] Ahuja K, Bose A, Jain M, et al. Gaze-based screening of autistic traits for adolescents and young adultsusing prosaic videos[C]. Guayaquil: ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS ’20), 2020: 324.

[68] Varma M, Washington P, Chrisman B, et al. Identification of Social Engagement Indicators Associated With Autism Spectrum Disorder Using a Game-Based Mobile App: Comparative Study of Gaze Fixation and Visual Scanning Methods[J]. Journal of Medical Internet Research, 2022, 24(2): e31830.

[69] Antolí A, Rodríguez-Lozano FJ, Juan Cañas J, et al. Using explainable machine learning and eye-tracking for diagnosing autism spectrum and developmental language disorders in social attention tasks[J]. Frontiers in Neuroscience, 2025, 19: 1558621.

[70] Ngo T D, Kieu H D, Nguyen M H, et al. An EEG & eye-tracking dataset of ALS patients & healthy people during eye-tracking-based spelling system usage [J]. Scientific Data, 2024, 11(1).

[71] Tupikovskaja-Omovie Z, Tyler D. Eye tracking technology to audit google analytics: Analysing digital consumer shopping journey in fashion m-retail [J]. International Journal of Information Management, 2021, 59: 102294.

[72] Nordfält J, Ahlbom C-P. Utilising eye-tracking data in retailing field research: A practical guide [J]. Journal of Retailing, 2024, 100(1): 148-60.

[73] Chen L, Jing K, Mei Y. The effect of consumption goals on review helpfulness: Behavioral and eye-tracking research [J]. Journal of Retailing and Consumer Services, 2024, 76.