Automatic Gaze Analysis: A Survey of Deep Learning based Approaches

Shreya Ghosh, Abhinav Dhall, Munawar Hayat, Jarrod Knibbe, Qiang Ji

Abstract—Eye gaze analysis is an important research problem in the field of Computer Vision and Human-Computer Interaction. Even with notable progress in the last 10 years, automatic gaze analysis still remains challenging due to the uniqueness of eye appearance, eye-head interplay, occlusion, image quality, and illumination conditions. There are several open questions, including what are the important cues to interpret gaze direction in an unconstrained environment without prior knowledge and how to encode them in real-time. We review the progress across a range of gaze analysis tasks and applications to elucidate these fundamental questions, identify effective methods in gaze analysis, and provide possible future directions. We analyze recent gaze estimation and segmentation methods, especially in the unsupervised and weakly supervised domain, based on their advantages and reported evaluation metrics. Our analysis shows that the development of a robust and generic gaze analysis method still needs to address real-world challenges such as unconstrained setup and learning with less supervision. We conclude by discussing future research directions for designing a real-world gaze analysis system that can propagate to other domains including Computer Vision, Augmented Reality (AR), Virtual Reality (VR), and Human Computer Interaction (HCI). Project Page: https://github.com/i-am-shreya/EyeGazeSurvey

Index Terms—Gaze Analysis, Automated Gaze Estimation, Eye Segmentation, Gaze Tracking, Unsupervised and Self-supervised Gaze Analysis, Human Computer Interaction.

1 Introduction

Humans perceive their environment through voluntary or involuntary eye movement to receive, fixate and track visual stimuli, or in response to an auditory, or cognitive stimulus. The eye movements therefore can provide insights into our visual attention [1] and cognition (emotions, beliefs and desires) [2]. Furthermore, we rely on these insights extensively in day-to-day communication and social interaction.

Automatic gaze analysis develops techniques to estimate the position of target objects by observing the eyes' movement. However, accurate gaze analysis is a complex problem. An accurate method should be able to disentangle gaze, while being resilient to a broad array of challenges, including: eye-head interplay, illumination variations, eye registration errors, occlusions, and identity bias. Furthermore, research [3] has shown how human gaze follows an arbitrary trajectory during eye movements which poses further challenge in gaze estimation.

Research in gaze analysis mainly involves coarse or fine-grained gaze estimation. There are three aspects of gaze analysis: registration, representation, and recognition. The first step, *registration*, involves the detection of the eyes (or eye-related key points or sometimes even just the face). In the second step, *representation*, the detected eye is projected to a meaningful feature space. In the final stage, *recognition*, the corresponding gaze direction or gaze location is predicted based on the features from stage 2. Research interest in automatic gaze analysis spans in several disciplines. One of the earliest explorations of gaze analysis was conducted in 1879, when Javal et al. [4] first studied, and coined the term, *saccades*. The broader interest in gaze analysis, however, developed with the advent of eye tracking technologies (initially in 1908, before gaining momentum in the late 70s, with systems such as 'Purkinje Image' [5], 'Bright Pupil' [6]). Automated gaze analysis then gained traction in computer vision-related assistive technology [7], [8], which then propagated through HCI [9]–[11], consumer behavior analysis [12], AR and VR [13], [14], egocentric vision [15], biometric systems [16] and other domains [17], [18]. A brief chronology of the seminal gaze analysis methods with important milestones is presented in Fig. 1. The increased reliance on gaze tracking technologies, however, came with its own challenges, namely, the cost of such devices and the requirement for specific controlled settings. To overcome their limitations and handle generic unconstrained settings [19], [20], most traditional gaze analysis models rely on handcrafted low-level features (e.g., color [21], shape [21], [22] and appearance [23]) and certain geometrical heuristics [20]. Since 2015, the approach to gaze analysis has changed, turning to the deep learning [24]–[26], similar to other computer vision tasks. With the deep learning based models and the availability of the large training datasets, the challenges associated with the variation in lighting, camera setup, eye-head interplay, etc, are reduced greatly over the past few years. Although, these performance enhancements have come with the requirement of large scale annotated data, which is expensive to acquire. As such, more recently, deep learning with limited annotation has gained increasing popularity [27]–[29].

This paper surveys different gaze analysis methods by isolating their fundamental components, and discusses how each component addresses the aforementioned challenges in gaze analysis. The paper discusses new trends and developments in the field of computer vision and the AR/VR domain, from the perspective of gaze analysis. We cover recent gaze analysis techniques in the

- *S. Ghosh and M. Hayat are with Monash University. (E-mail: {shreya.ghosh, munawar.hayat}@monash.edu).*
- *A. Dhall is with Monash University and Indian Institute of Technology Ropar, India. (E-mail: abhinav.dhall@monash.edu).*
- *J. Knibbe is with the University of Melbourne. (E-mail: jarrod.knibbe@unimelb.edu.au).*
- *Q. Ji is with the Rensselaer Polytechnic Institute (E-mail: jiq@rpi.edu).*

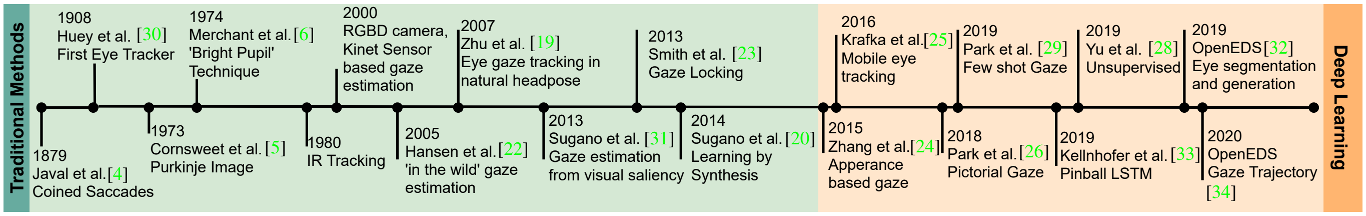


Figure 1: A brief chronology of seminal gaze analysis works. The very first gaze pattern modelling dates back to the work of Javal et al. in 1879 [4]. One of the first deep learning driven appearance based gaze estimation models was proposed in 2015 [24].

un-, self-, and weakly-supervised domain, along with validation protocols with evaluation metrics tailored for gaze analysis. We also discuss various data capturing devices, including: RGB/IR camera, tablet/laptop's camera, ladybug camera and other gaze trackers (including video-oculography [21]) are also discussed.

Due to the rapid progress in the computer vision field (Refer Fig. 1), it is increasingly useful to get thorough guidance via exhaustive survey/review articles. In 2010 and 2013, Hansen et al. [21] and Chennamma et al. [35] reviewed the state-of-the-art eye detection and gaze tracking techniques. These reviews provide a holistic view of hardware, user interface, eye detection, and gaze mapping techniques. Since these reviews were before the deep learning era, they contain the relevant features leveraged from handcrafted techniques. Afterwards in 2016, Jing et al. [36] reviewed methods for 2-D and 3-D gaze estimation methods. In 2017, Kar et al. [37] provided insights into the issues related to algorithms, system configurations, and user conditions. In 2020, a more comprehensive and detailed study of deep learning based gaze estimation methods is presented by Cazzato et al. [38]. To date, however, no comprehensive review has examined the recent trends in learning from less supervision. Moreover, all of the existing reviews focus only on gaze estimation and ignore significant works in eye segmentation, gaze zone estimation, gaze trajectory prediction, gaze redirection, and unconstrained gaze estimation in single and multiperson setting. The contributions of the paper are summarized below:

1) **A comprehensive review of automated gaze analysis.** We categorize and summarize existing methods by considering data capturing sensors, platforms, popular gaze estimation tasks in computer vision, level of supervision and learning paradigm. The proposed taxonomies aim to help researchers to get a deeper understanding of the key components in gaze analysis.
2) **Different popular tasks under one framework.** To the best of our knowledge, we are the first to put different eye and gaze related popular tasks under one framework. Apart from gaze estimation, we consider gaze trajectory, gaze zone estimation and gaze redirection tasks.
3) **Applications.** We explore major applications of gaze analysis using computer vision i.e. Augmented and Virtual Reality [13], [39], Driver Engagement [40], [41] and Healthcare [42], [43].
4) **Privacy Concerns.** We also provide a brief review of the privacy concerns of the gaze data and its possible implications.
5) **Overview of open questions and potential research directions.** We review several issues associated with the current gaze analysis frameworks (i.e. model design, dataset collection, etc.) and discuss possible future research directions.

2 Preliminaries

The human visual system is a complex cognitive process. As such, understanding and modelling human gaze has become a fundamental research problem in psychology, neurology, cognitive science, and computer vision. To lay the foundations for this review, below, we provide brief descriptions of the *Human Visual System and Eye Modelling* (Sec. 2.1), *Eye movements* (Sec. 2.2), *Problem Settings in Automated Gaze Analysis* (Sec. 2.3) and the associated *Challenges* (Sec. 2.4).

2.1 Human Visual System and Eye Modelling

Computer vision based human visual perception methods estimate gaze quantitatively from image or video data. These methods analyze the visible region of the eyes (the iris and sclera, see Fig. 2), and attempt to approximate the unobservable features of the eyes (which are integral to determining gaze direction). These approximations can be based on models of the human eyes, derived from movement patterns over time, or learned via representation learning over large scale data. For gaze estimation, we typically approximate the eye as a sphere with a radius of 12-13mm. Subsequently, we model gaze direction with respect to the optical axis, also called the *line of gaze (LoG)*, or the visual axis, the *line of sight (LoS)* (see Fig. 2right). The line of gaze (LoG) connects the pupil, cornea and eyeball center. Conversely, the line of sight (LoS) is the line connecting the fovea and center of the cornea. Generally, the LoS is considered as the *true direction of gaze*. The intersection point of the visual and optical axis is called the *nodal point of eye* (anatomically the cornea center), which typically encodes a subject dependent angular offset. This individual offset is the main motivation behind having subject dependent calibration for gaze tracking devices. According to prior studies [45], [46], the fovea is located around 4-5° horizontally and 1.5° vertically below the optical axis. Across a broader population, this can vary up to 3° between subjects [46]. Additionally, head-pose also plays an important role in gaze analysis. The coarse gaze direction of a subject can be determined by the position (in 3-D coordinate) and orientation (Euler angles) of the headpose [21]. Most of the time, the combined direction of LoS and head pose provide information about where the person is looking.

2.2 Eye Movements and Foveal Vision

We perceive our environment through eye movements. These movements can be voluntary or involuntary, and help us to acquire, fixate, and track visual stimuli (see Fig. 2). Eye movements are divided into three primary categories: saccades, smooth pursuits, and fixations.

Saccades. Saccades are rapid and reflexive eye movements, primarily used for adjustment to a new location in the visual environment. It can be executed voluntarily or involuntarily, as

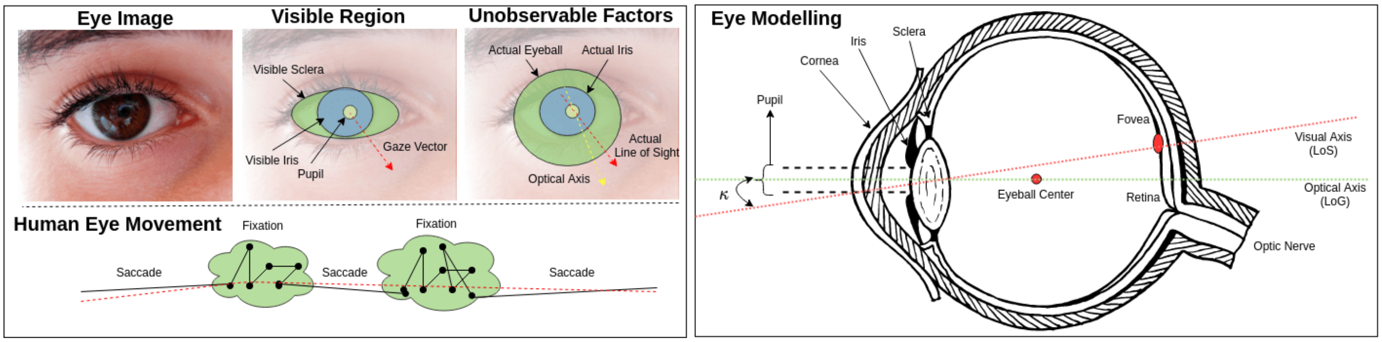


Figure 2: **Top Left:** Overview of the human visual system, eye modelling and eye movement. For computer vision based automated gaze analysis, we consider an image containing eyes (left) as input. Thus, such methods analyze the visible eye regions (middle) and predict the 2-D/3-D gaze vector as output. However, there are unobservable factors by which we can predict the true gaze direction (right) which requires person-specific information and other factors [44]. **Bottom Left:** Apart from static image based gaze estimation, the dynamic eye movement is another line of research in computer vision that provides cues regarding human behavioural traits. **Right:** The actual modelling of gaze with respect to eye anatomy. We only highlight the relevant parts i.e. pupil, cornea, iris, sclera, fovea, LOS and LOG. The angle between LOG and LOS is called angle of kappa (κ).

a part of optokinetic measure [47]. Saccades typically last for between 10 and 100 ms.

Smooth Pursuits. Smooth pursuit occur while tracking a moving target. This involuntary action depends on the range of the target's motion as astonishingly, human eyes can follow the velocity of a moving target to some extend.

Fixations (Microsaccades, Drifts, and Tremors). Fixations are eye movements in which the focus of attention stabilizes over a stationary object of interest. Fixations are characterized by three types of miniature eye movements: *tremor, drift* and *microsaccades* [47]. During Fixations, the miniature eye movements occur due to the noise present in the control system to hold gaze steadily. This noise occurs in the area of fixation, around 5° visual angle. For simplification of the underlying natural process, this noise is ignored during fixation.

Foveal Vision. The *fovea centralis* region of human eye is responsible for the perception of sharp and high resolution human vision. In order to perceive the environment, it is necessary to direct the foveal vision to select region of interest (the process is termed as 'foveation'). This sharp foveal vision decays rapidly within the range of 1-5°. Beyond this limit, human vision is blurred, and low resolution. This is termed as *peripheral vision*. Our peripheral vision plays an important role in our overall visual experience, especially for motion detection. On an abstract level, our visual perception is the result of our brains merging our foveal and peripheral vision.

2.3 Gaze Estimation: Problem Setting

The main task of gaze estimation is to determine the line of sight of the pupil. Fig. 4 depicts a typical visual sensor based real-time gaze estimation setup consisting of user, data capturing sensor(s) and visual plane. The main calibration factors in this setting are:

- Estimation of *camera calibration* parameters, which include both intrinsic and extrinsic camera parameters.
- Estimation of *geometric calibration* parameters, which include the relative position of the camera, light source and screen.
- Estimation of *personal calibration*, which include headpose and eye-specific parameters such as cornea curvature, the nodal point of the eye, etc.

In some of the applications, the calibration parameters are estimated in task-specific settings. For example, users are requested to fixate their gaze to some pre-defined points for calibration. Similarly, user-specific information is registered once in the devices for subject-specific calibrations. With the advances in computer vision and deep learning, nowadays gaze estimation techniques are developed with appearance based features and do not require an explicit calibration step. For example, *CalibMe* [48] is a fast and unsupervised calibration technique for gaze trackers, designed to overcome the burden of repeated calibration.

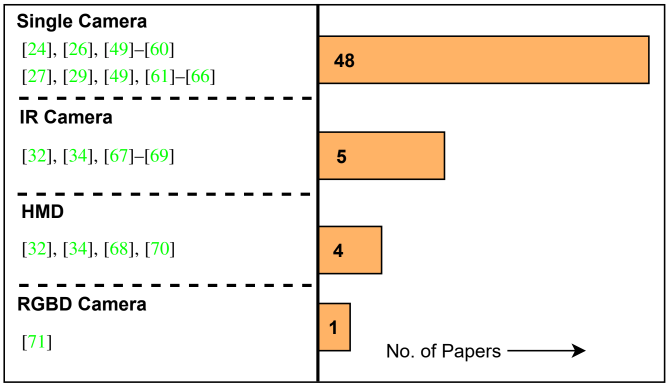


Figure 3: The plot shows the popularity of different data capturing devices across research articles over the past 10 years. Here, HMD: Head Mounted Device, RGBD: RGB Depth camera.

Role of Data Capturing Sensors. Visual stimuli provide valuable information for computer vision based gaze estimation techniques. A trade-off of widely used sensors is mentioned in Fig. 3. These data capturing sensors are mainly divided into two categories: *Intrusive* and *Non-intrusive*. An array of methods used specialized hardware which requires physical contact with human skin or eyes are termed as *intrusive sensors*. The widely used intrusive sensors are head-mounted devices (HMD), electrodes, or scleral coils [69], [72]. These devices may cause unpleasant user experience, and the accuracy of these systems depends on the tolerance, accurate subject-specific calibration and other factors of the devices. On the other hand, data capturing devices that do not require physical contact [73] are termed *non-intrusive* sensors. Mainly, RGB, RGBD and IR cameras fall under this category. These methods face several challenges, which include partial occlusion of the iris by the eyelid, varying illumination condition, head pose, specular

Table 1: Attributes of different platforms widely used in gaze analysis. Here, Dist.: distance (in cm), VA: viewing Angle (in °), HMD: Head Mounted Devices, FV: Free Viewing, UC: User Condition, ET: External Target.

Platform	Dist.	VA	UC	Papers
Desktop, TV Panels	30-50, 200-500	$\sim$ 40 °, 40°-60°	Static, Sitting, Upright	[24], [26], [49]–[57] [29], [49], [61]–[66] [58]–[60], [71], [76]
HMD	2-5	55°-75°	Independent (Leanback, Sitting, Upright)	[32], [34], [77], [78]
Automotive	50	40°-60°	Mobile, Sitting, Upright	[40], [79]–[85]
Handheld	20-40	5°-12°	Leanfwd, Sitting, Standing, Mobile	[25], [63], [86]–[89]
ET/ FV	–	–	Leanfwd, Sitting, Standing, Upright	[27], [90]

reflection in case the user wears glasses, the inability to use standard shape fitting for iris boundary detection, and other effects including motion blur and over saturation of images [73]. To deal with these challenges, most of the existing gaze estimation methods have been performed under constrained environments like constrained head pose, controlled illumination conditions, and camera angle. Among all of the aforementioned factors, pupil visibility plays an important role as robust gaze estimation needs accurate pupil-center localization. Fast and accurate pupil-center localization is still a challenging task [74], particularly for images with low resolution. A trade-off of widely used sensors are mentioned in Fig. 3.

Role of Headpose. Gaze estimation is a challenging task due to eye-head interplay. Head-pose plays the most important role in gaze estimation. The gaze direction of a subject is determined by the combined effect of position and orientation of head pose and eyeball. One can change gaze direction via eyeball and pupil movement by maintaining stationary or dynamic head-pose or by moving both. Usually, this process is subject dependent. People adjust their head-pose and gaze to maintain a comfortable posture. Thus, the gaze estimation task needs to consider both gaze and head-pose at the same time for inference. As a result of this, it is more common to consider head-pose information in the gaze estimation methods implicitly or explicitly [24], [50], [75].

Role of Visual Plane. The visual plane is the plane containing the gaze target point i.e. where the subject is looking, which is often termed as Point of Gaze (PoG). The distance between the user and the visual plane varies a lot in a real-world setting. Thus, recent deep learning based methods do not rely on the distance or placement of the visual plane. The most common gaze analysis setup uses a RGB camera placed at 20 - 70 cm from the user in unconstrained setting i.e. without any invasive sensors or fixed setup. In different real-world settings, the visual plane could be desktop ($\sim$ 60cm), mobile phone ($\sim$ 20cm), car ($\sim$ 50cm) etc. An overview is presented in Table 1.

2.4 Challenges

Data Annotation. Generally, deep learning based methods require large amount of annotated data for generalized representation learning. Curating large scale annotated gaze datasets is non-trivial [24], [87], [90], time consuming and requires expensive equipment setup. Current dataset recording paradigms via wearable sensors may lead to uncomfortable user experience and it require expert knowledge. Another common aspect of the current datasets is the constrained environment in which they are recorded (For example, CAVE [23] dataset is recorded on indoor environment with headpose restricted). Recently, a few datasets [87], [90] have been proposed to address this gap by recording in unconstrained indoor and outdoor environments. Another challenge associated with data annotation is participant's cooperation. However, it is assumed that participants fixate their gaze as per the given instructions [24], [87], [90]. Despite these attempts [24], [87], [90], data annotation still remains complex, noisy and time-consuming. Self/weakly/un-supervised learning paradigms [27], [28] could be helpful to address the dataset creation and annotation challenges.

Subjective Bias. Another challenge for the automatic gaze analysis method is subjective bias. Individual differences in the nodal points of human eyes makes automatic and generic gaze analysis way more difficult. In an ideal scenario, any gaze analysis method should encode rich features corresponding to eye region appearance, which provides relevant information for gaze analysis. To address this challenge, few-shot learning based approach has been widely adapted [29], [91], where the motivation is to adapt to new subject with minimum subject specific information. Another way to deal with the subjective bias is combining classical eye model based approaches with geometrical constraints [92] as this approach has the potential to generalize well across subjects.

Eye Blink. Blinks are an involuntary and periodic motion of the eyelids. They pose a challenge for gaze analysis, as blinks result in missed frames of data. A few recent works [40], [90] assume that the head pose information is a suitable replacement for gaze during blinking based on a common line of sight between a subject's headpose and gaze. However, it is noted that a large shift in the gaze is possible after subject re-opens their eyes. To simplify the situation, some gaze analysis methods ignore eye-blink data (e.g., [57], [87]) and some treat blinks as a separate class of data (.e.g, [41], [93]). A possibility for real world deployment of such a system is generating gaze labels by interpolating from neighbouring frames' labels when blinks are detected [40].

Data Attributes. Several factors, such as eye-head interplay, occlusion, blurred image, and illumination can influence the performance of a gaze analysis model. The presence of any subset of these attributes can degrade the performance of a system [87], [90]. Many methods use face alignment [24], [25] and 3-D head pose estimation [24] as a pre-processing step. However, face alignment on images captured in an unconstrained environment based images may introduce noise in a system. To overcome this, recent approaches [27], [57], [90], [94] avoid these pre-processing steps and show increase in gaze prediction performance.

Another critical challenge in gaze estimation is eye-head interplay. Prior studies generally address this issue via implicit training [25], [95] or provide the head pose information separately as a feature [24]. Similarly, it is challenging to estimate gaze under partial occlusion. When the head's yaw rotation is greater than 90°, one side of the face becomes occluded w.r.t. the camera. A few prior works [24], [25] avoid these scenarios by disregarding these frames. Kellnhofer et al. [90], however, argue that when the head yaw angle is in the range 90°- 135°, the partial visibility still provides relevant information about the gaze direction. This study

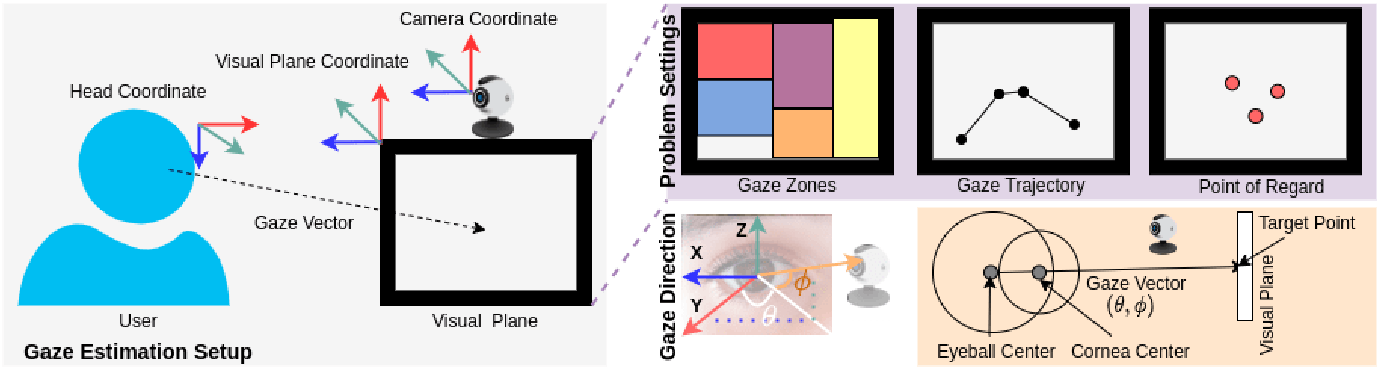


Figure 4: Overview of gaze estimation setups (See Sec. 2.3 for more details). A traditional gaze analysis setup considers the effect of head, visual plane and camera coordinates. The gaze analysis tasks include gaze zone, point of regard, gaze trajectory estimation etc. (See Sec. 3) The gaze vector is defined by the angles (θ, ϕ) in polar co-ordinate systems as shown in the gaze direction part.

also proposes quantile regression via pinball loss to mitigate the effect of partial occlusion in training data in terms of uncertainty. Despite all of these attempts, gaze estimation still remains challenging in presence of these attributes. There is still have scope to eliminate the effects of these attributes and make the gaze analysis model more robust for the real-world deployment.

Application Specific Challenges. Gaze analysis also has application-specific requirements, for example, coarse or fine gaze estimation in AR, VR, Robotics, egocentric vision and HCI. Thus, a working algorithm behind any eye-tracking devices needs to fit in the application environment.

3 Gaze Analysis in Computer Vision

We provide a breakdown of different gaze analysis tasks for vision based applications. Any statistical gaze modeling mainly estimates the relation between the input visual data and the point of regard/gaze direction.

2-D/3-D Gaze Estimation. Most of the existing studies consider gaze estimation as either the gaze direction in 3-D space or as the point of regard in 2-D/3-D coordinates (see Fig. 4). We can divide the gaze estimation methods into the following types:

1) Geometric Methods: These geometric methods compute a gaze direction from the geometric model of the eye (see Fig. 2), where the anatomical structure of the eye is considered to get the 3-D gaze direction or gaze vector. These methods were widely used in prior to more deep learning approaches [21]. These recent deep learning based approaches implicitly model these geometric parameters during the learning process, and, as such, do not explicitly require the often noisy subject specific parameters, such as cornea radii, cornea center, angles of kappa (i.e. Refer κ in Fig. 2), iris radius, the distance between the pupil center and cornea center, etc.

2) Regression Methods: Regression based methods [24], [26], [28], [96] map visual stimuli (image or image-related features) to gaze coordinates or gaze angles in 2-D/3-D. The output mapping is application-specific. For example, such techniques are often used to map 2-D/3-D gaze coordinate mainly maps people's focus of attention to the screen coordinates (for human-computer interaction based applications such as engagement or attention monitoring). Regression based methods can be divided into two types: the *parametric* and *non-parametric* approaches. Parametric approaches (e.g., [28], [96]) assume gaze trajectories as a polynomial, where the task is to estimate the parameters of the polynomial equation. *Non-parametric* approaches directly work on the mappings in spite of calculating the intersection between the gaze direction and gazed object explicitly [24], [26], [57]. The recent deep learning based approaches are non-parametric [26], [28], [54], [90], [97].

Trajectory Prediction. Gaze estimation has potential applications in AR/VR especially in Foveated Rending (FR) and Attention Tunneling (AT), where the future eye trajectory prediction is highly desirable. To meet this requirement, a new research direction (i.e. future gaze trajectory prediction) has been recently introduced [34]. Here, possible future gaze locations can be estimated based on the prior gaze points, content of the visual plane or their combination. Thus, the problem statement can be formulated as follows: given n number of prior gaze points, the algorithm will predict the m future frames' gaze direction in a user-specific setting.

Gaze Zone. In many gaze estimation based applications such as driver gaze [27], [40], [41], [93], gaming platforms [98], website designing [99], etc., the exact position or angle of the line of sight of the pupil is not required. Thus, a gaze zone approach is utilized in these cases for estimation. Here, the gaze zone refers to an area in 2-D or 3-D space. For example, in a simplistic driver gaze zone estimation, the driver could be looking straight ahead, at the steering wheel, at the radio, or at the mirrors. Similarly, another example is detecting more visually salient zone/region during website designing [99].

Gaze Redirection. Due to the challenges in different posed gaze conditions, generation on the go is gaining popularity [34], [100]. It aims to capture subject-specific signals from a few eye images of an individual and generate realistic eye images for the same individual under different eye states (gaze direction, camera position, eye openness etc.). The gaze redirection can be performed in both controlled and uncontrolled way [100]–[102]. Apart from these, eye rendering is another research direction to generate realistic eye given the appearance, gaze direction of a person. It has potential applications in virtual agents, social robotics, behaviour generation and in the animation industry [103].

Unconstrained Gaze Estimation. Gaze estimation in an unconstrained setting can be divided into two types:

1) Single Person Setting: In webcam or RGB camera based gaze estimation approaches, geometric model based eye tracking [104], [105] is typically used, as it is fast and does not require training data. At the same time, however, it relies on accurate eye location and key points detection, which is hard to achieve in real-world environments. Deep learning based methods [97], [106] have eliminated this issue to some extent, however, it still remains a challenge as it does not generalize well in different settings.

2) Multi-Person Setting: In unconstrained multi-person settings, it is very difficult to track the eyes. For example, in a social interaction scenario, understanding the gaze behaviour of each person provides important cues to interpret social dynamics [107]. To this end, a new research direction is introduced where the problem is defined as whether the people are Looking At Each Other (LAEO) in a given video sequence [108]–[110]. Similarly, gaze communication [111] and GazeOnce [112] are another line of research aligned with this field.

Visual Attention Estimation. Human visual attention estimation is another line of research which mainly focuses on *where the person is looking* irrespective of eyes visibility. The popular subtasks in this direction are gaze following [62], [113]–[116], gaze communication [111], human attention in goal-driven environments [117] and categorical visual search [118], visual scanpath analysis in visual question answering [119], and naturalistic environment [120], [121]. These methods are mostly driven by saliency in the scene, head orientation, or any other task at hand. Visual attention based approaches have the potential to localize the gaze target directly from scene information which in turn enhances the scalability of naturalistic gaze behaviour patterns.

4 Gaze Analysis Framework

We break down a gaze analysis framework into its fundamental components (Fig. 5) and discuss their role in terms of *eye detection and segmentation* (Sec. 4.1), *Network Architecture* (Sec. 4.2) and *Level of Supervision* (Sec. 4.3).

4.1 Eye Detection and Segmentation

Eye registration is the first stage of gaze analysis and requires detection of the eye and the relevant regions of interest.

Eye Detection Methods. The main aim of the eye detection algorithms is to accurately identify the eye region from an input image. Eye detection algorithms need to operate in challenging conditions such as occlusion, eye openness, variability in eye size, head pose, illumination, and viewing angle, while balancing the trade-off in appearance, dynamic variation and computational complexity. Prior works on eye detection can be divided into three categories: shape based [22], appearance based [49], [105], [122], [123] and hybrid method [123]. The most popular libraries for eye and facial point detection are Dlib [106] OpenFace [104], [105], MTCNN [124], Duel Shot Face Detector [125], FaceX-Zoo [126].

The pupil and iris region of the eye is usually darker than the sclera which provides an important cue to differentiate or localize the pupil. The pupil center localization use dedicated and costly devices [69], [72], which requires person-specific pre-calibration. To overcome this limitation, the deep learning based pupil localization methods use ensembles of randomized trees [127], local self similarity matching [73], adaptive gradient boosting [128], hough regression forests [129], deep learning based landmark localization models [97], [106], heterogeneous CNN models [130], etc. In prior literature, the choice of eye registration process is influenced by the correlation between the input image and the learning objective of the proposed method. Apart from this, the trade-off between the accuracy of eye localization and running time complexity of the algorithm is optimized in a task specific way. In this context, *OpenFace* and *Dlib* are the most popular. Moreover, the choice of eye/face registration process may also depend on their ability to detect eye components in different challenging real world conditions.

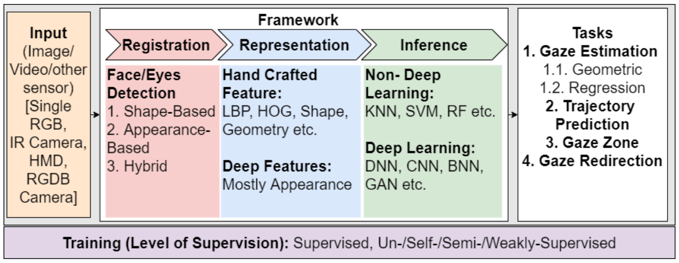


Figure 5: A generic gaze analysis framework has different components including registration, gaze representations and inference. Although in the deep learning based approaches, there is a high overlap between the representation and inference module. Refer Sec. 4 for more details.

Eye Segmentation. The main task of eye segmentation is pixel-wise or region-wise differentiation of the visible eye parts. In general, the eye region is divided into three parts: sclera (the white region of the eyes), iris (the colour ring of tissue around the pupil) and pupil (the dark iris region). Prior studies [131]–[134] on eye segmentation mainly explore to segment the iris and sclera region. Few studies [32], [34] include the pupil region in the segmentation task as well. Eye segmentation is widely used in the biometric systems [135] and prior for synthetic eye generation [136].

Eye Blink Detection. Eye blinks are the involuntary and periodic activity that can help to judge the cognitive activity of a person (e.g. driver's fatigue [137], lie detection [138]). KLT trackers and various sensors are also widely used to get the eye motion information to track eye blink [139]. The existing eye blink detection approaches aim to solve a binary classification problem (blink/no blink) either in a *heuristic based* or *data-driven* way. The *heuristic based* approaches mainly include motion localization [139] and template matching [140]. As these methods are highly reliable on pre-defined thresholds, they could be sensitive to subjective bias, illumination and head pose. To overcome this limitation, the *data-driven* approaches infer on the basis of appearance based temporal motion features [139], [141] or spatial features [142]. In *hybrid approach* [143], multi-scale LSTM based framework is used to detect eye blink using both spatial and temporal information.

4.2 Representative Deep Network Architectures

In this section, we provide a generic formulation and representation of gaze analysis. Given an RGB image $\mathbf{I} \in \mathbb{R}^{W \times H \times 3}$, a deep learning based model mapped it to task-specific label space. The input RGB image is usually the face or eye regions. Based on the primary network architectures adopted in the literature, we classify the models into the following categories: *CNN based, Multi-Branch network based, Temporal based, Transformer Based* and *VAE/GAN based.* An overview is shown in Fig. 6.

4.2.1 CNN based

Most of the recent solutions adopt a CNN based architecture [24]–[26], [50], [51], [92], which aims to learn end-to-end spatial representation followed by gaze prediction. The adopted model is often a modified version of the popular CNNs in vision (e.g. AlexNet [27], VGG [54], ResNet-18 [67], [90], ResNet-50 [94],

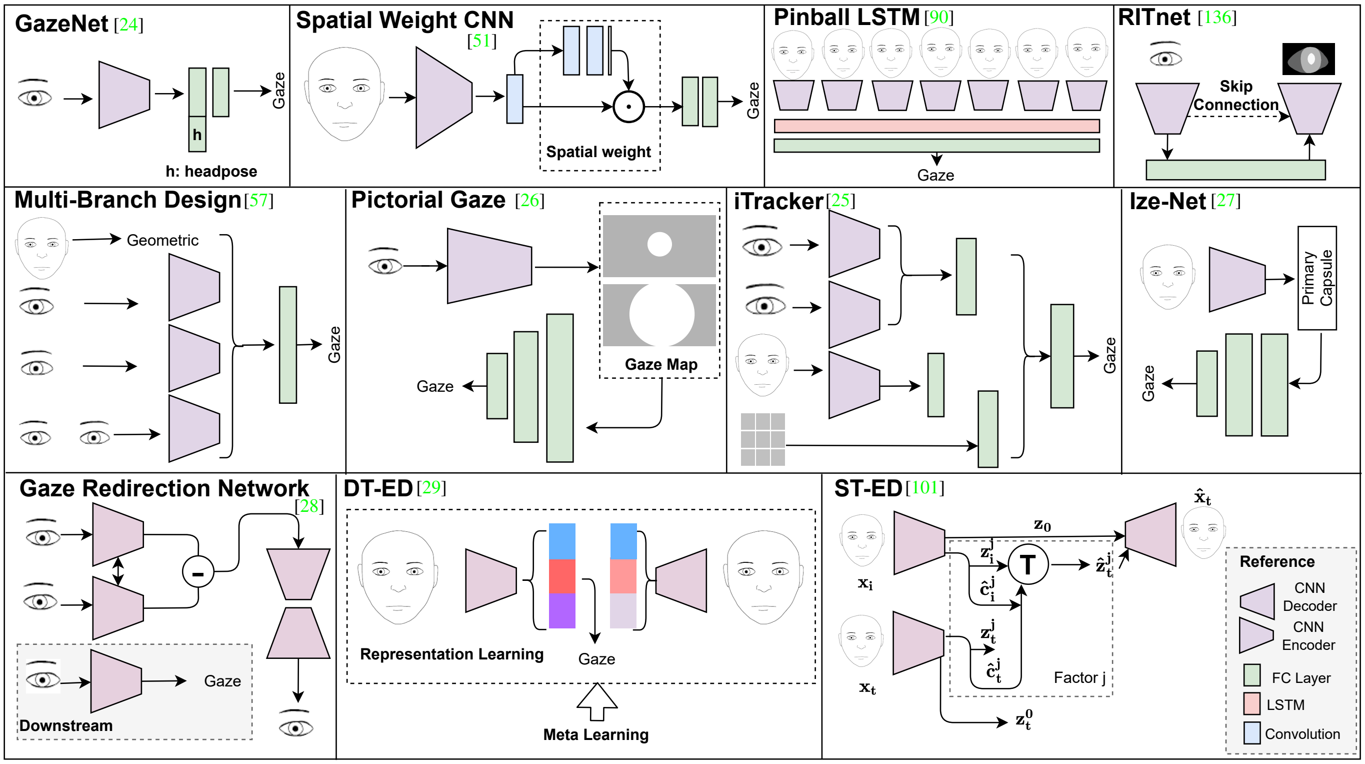


Figure 6: A brief overview of different pipelines used for gaze analysis tasks. Refer Sec. 4.2 for more details of the networks.

Capsule network [27]). These CNNs learn from a single-stream of RGB images (e.g. face, left or right eye patch) [24], [51], or multiple streams of information (e.g. face, and eye patches) [25], [57], and prior knowledge based on eye anatomy or geometrical constraints [26].

GazeNet. It is the extended version of the first deep learning based gaze estimation method [24] which aims to capture low level and high level appearance feature by using convolution operation. GazeNet takes a grayscale eye patch image $\mathbf{I} \in \mathbb{R}^{W \times H}$ as input and maps it to angular gaze vector $\mathbf{g} \in \mathbb{R}^2$. As headpose provides relevant features for gaze direction, the headpose vector is also added in the FC layer for better inference (Refer top left image in Fig. 6). The Extended version [50] is adapted from the VGG network which further boosts the performance. To train these models, the sum of the individual ℓ_2 losses between the predicted $\hat{\mathbf{g}}$ and actual gaze angle vectors $\mathbf{g}$ is considered.

Spatial Weight CNN. It is a full face appearance based gaze estimation method [51] which uses a spatial weighting mechanism for encoding the important locations of the facial image $\mathbf{I} \in \mathbb{R}^{W \times H \times 3}$ via the standard CNN architecture (Refer top row second column image in Fig. 6). This weighting mechanism (aka attention) automatically assign more weight to the regions contributing more towards gaze estimation. It includes three additional 1×1 convolutional layers followed by a ReLU activation. Given a $N \times H \times W$ dimensional activation map (U) as input (where N, H and W are the number of feature channels, height and width of the output), the spatial weights module learns the weight matrix W from element-wise multiplication of W with the original activation U via the following function: $W \bigodot U_c$, across the channel dimensions. Thus, the model learns to assign more weight to the specific regions, which in turn eliminates unwanted noise in the input. For 2-D gaze estimation, the ℓ_1 distance between the predicted and ground-truth gaze positions in the target screen coordinate system is utilized. Similarly, the ℓ_1 distance between the predicted and ground-truth gaze angle vectors in the normalized space is used for 3-D gaze estimation.

Dilated Convolution. Another interesting architecture for gaze estimation is dilated-convolutional layers which preserve spatial resolution while increasing the size of the receptive field without compromising the number of parameters [64]. It aims to capture the slight change in pixels due to eye movement. Given an input feature map U of kernel size $N \times M \times K$ (N: height, M:width, K:channel with weights W and bias b) and dilation rates (r_1, r_2), the output feature map v can be defined as follows:

$$v(x, y) = \sum_{k=1}^{K} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} u(x + nr_1, y + mr_2, k) w_{nmk} + b$$

The dilated convolution is applied in facial and left/right eye patches before inferring the gaze. For training the network, cross entropy loss is used in the label space. Representation learning via MinENet [144] also relies on dilated and asymmetric convolutions to provide context to the segmented regions of the eye by increasing the receptive field capacity of the model to learn contextual information.

Bayesian CNN. Another variant of CNN is Bayesian CNN which is used for robust and generalizable eye-tracking under different conditions [145]. Instead of predicting eye gaze using a single trained eye model, it performs eye tracking using an ensemble of models, hence alleviating the over-fitting problem, is more robust under insufficient data, and can generalize better across datasets. Compared to the point based eye landmark estimation methods, the BNN model can generalize better and it is also more robust under challenging real-world conditions. Additionally, the extended version of the BCNN (i.e. the single-stage model to multi-stage, yielding the cascade BCNN) allows feeding the uncertainty information from the current stage to the next stage to progressively improve the gaze estimation accuracy. This could be an interesting area for further study.

Pictorial Gaze. Pictorial gaze [26] aims to model the relative position of eyeball and iris to get the gaze direction. The network [26] consists of two parts: 1) regression from eye patch image to intermediate gazemap followed by 2) regression from gazemap to gaze direction vector g (Refer second row second column image in Fig. 6). The gazemap is an intermediate representation of a simple model of the human eyeball and iris in terms of $m \times n$ dimensional image, where, the projected eyeball diameter is $2r = 1.2n$ and the iris centre coordinates (u_i, v_i) are as follows: $u_i = \frac{m}{2} - r' sin\ \phi\ cos\ \theta, v_i = \frac{n}{2} - r' sin\ \theta$ where, $r' = r\ cos\ (sin^{-1}\frac{1}{2})$ and gaze direction $g = (\theta, \phi)$. Basically, the iris is an ellipse with major-axis diameter of r and minor-axis diameter of $r|cos\ \theta cos\ \phi|$. The first part is implemented via a stacked hourglass architecture which assumed to encode complex spatial relations including the locations of occluded key points. Consequently, for the second part, a DenseNet architecture is used which maps the intermediate gazemap to the gaze vector $\hat{g}$. It is trained via gaze direction regression loss defined as: $||g - \hat{g}||_2$ as well as cross-entropy loss between predicted and ground-truth gazemaps for all pixels.

Ize-Net. This framework is used for coarse to fine gaze representation learning. Here, the main idea is to learn coarse gaze representation by dividing the gaze locations to gaze zones. Further, the gaze zone is mapped to the finer gaze vector. The proposed network [27] (Refer second row right image in Fig. 6) is a combination of convolution and primary capsule layer. After the convolution layers, the primary capsule layer is appended whose job is to take the features learned by convolution layers and produce combinations of the features to consider face symmetry into account. This network is trained for coarse gaze zone which is fine-tuned for downstream 2-D/3-D gaze estimation.

EyeNet. consists of modified residual units as the backbone, attention blocks and multi-scale supervision architecture. This network is robust for the low resolution, image blur, glint, illumination, off-angles, off-axis, reflection, glasses and different colour of iris region challenges.

4.2.2 Multi-Branch network based

There are several works [25], [54], [57] which utilize multiple inputs for better inference.

iTracker. The iTracker framework [25] takes the left eye, right eye, detected facial region and face location in the original frame as a binary mask (all of the size 224 × 224) and predicts the distance from the camera (in cm). The model is jointly trained with Euclidean loss on the x and y gaze position. The overview of the framework is shown in the second row third column image in Fig. 6.

Multi-Branch Design. Similar to iTracker, Jyoti et al. [57] propose a framework which takes the full face, left and right eye, both eye patch as input for inferring gaze (Refer Fig. 6 second-row left image). To train this network, mean squared error between the true and predicted gaze point/direction is used.

Two-Stream VGG Network. In [54], a two-stream VGG network is used for gaze inference while taking left and right eye patch as input. Similar to prior works, it utilizes the sum of the individual ℓ_2 losses between the predicted and ground truth gaze vectors to train the ensemble network.

4.2.3 Temporal Gaze Modelling

The human gaze is a continuous and dynamic process. While scanning the environment, the concerned subject performs eye movements in terms of fixations, saccades, smooth pursuit, vergence, and vestibulo-ocular movements. Moreover, a certain image frame in time has a high correlation with the gaze direction of previous time steps. Based on this line of reasoning, several works [56], [65], [66], [90], [146], [147] have leveraged temporal information and eye movement dynamics to enhance gaze estimation performance as compared to image based static methods. Given a sequence of frames, here the task is to estimate the gaze direction of the concerned person. For this modelling, popular recurrent neural network structures have been explored (e.g. GRU [148], LSTM/bi-LSTM [90]).

Multi-Modal Recurrent CNN. Palmero et al. [56] have proposed a multimodal recurrent CNN framework in which the learned static features of all the input frames of a given sequence are fed to a many-to-one recurrent module for predicting the 3D gaze direction of the last frame in the sequence. Their approach improves the state-of-the-art gaze estimation performance significantly (i.e. by 4% on EYEDIAP dataset).

DGTN. Wang et al. [66] have proposed Dynamic Gaze Transition Network (DGTN) based on semi-Markov approach which models human eye movement dynamics. DGTN first computes per-frame gaze using a CNN which is further refined using the learned dynamic information.

Improved iTracker + bi-LSTM. Bidirectional recurrent module based temporal modeling methods have been introduced in [65] which rely on both past and future frames. It is quite beneficial for low-to-mid resolution images and videos having a low frame rate ($\sim$ 30 fps) despite its reduced applicability for real-time application as future frames are not usually available.

Pinball LSTM. Similarly to encode contextual information along temporal domain, pinball LSTM [90] is proposed. This video based gaze estimation model using bidirectional LSTM considers 7-frame sequence to estimate the gaze direction of the central frame. Fig. 6 first row third column image illustrates the architecture of the model. The facial region from each frame is provided as input to the backbone CNN having ResNet-18 architecture. It maps input image to 256-dimensional feature space. Further, a two layer bidirectional LSTMs map these features to label space via a FC layer with an error quantile estimation i.e. (θ, ϕ, σ), where (θ, ϕ) is the predicted gaze direction in spherical coordinates corresponding to the ground-truth gaze vector in the eye coordinate system g as $\theta = -arctan\ \frac{g_x}{g_z}$ and $\phi = arcsin\ g_y$. On the other hand, σ corresponds to the offset from the predicted gaze i.e. $\theta + \sigma$ and $\phi + \sigma$ in the 90% quantiles of its distribution and $\theta - \sigma$ and $\phi - \sigma$ are in the 10% quantiles. The pinball loss is computed as follows: given the ground truth label $y = (\theta_{gt}, \phi_{gt})$, the loss L_τ for the quantile τ and the angle $\alpha \in \{\theta, \phi\}$ can be written as:

$$L_\tau(\alpha, \sigma, \alpha_{gt}) = max(\tau \hat{q_\tau}, -(1 - \tau)\hat{q_\tau})$$

where, $\hat{q_\tau} = \alpha_{gt} - (\alpha - \sigma)$, for $\tau \leq 0.5$ and $\alpha_{gt} - (\alpha + \sigma)$ otherwise. This loss enforces θ and ϕ to converge to their ground truth values.

Static + LSTM. In an interesting study, Palmero et al. [146] analyze the effect of sequential information for appearance-based gaze estimation using a static CNN network followed by a recurrent module to capture eye movement dynamics. The model is developed based on high-resolution eye-image sequences performing a stimulus-elicited fixation and saccade task in a VR scenario. The proposed model learns eye movement dynamics with accurate localization of gaze movement transition.

Discussion. Despite the initial efforts that confirm the benefits of leveraging temporal information [56], [65], [66], [90], [146], there still have scope to explore eye dynamics in a task-driven real environment. Sometimes it is difficult to capture eye movement dynamics accurately using videos having low-resolution image frames with poor frame rates. Thus, it is still challenging how and why temporal information enhances gaze estimation performance for eye movement dynamics. Moreover, a deep understanding of eye movement patterns is required as the existing datasets are only based on task-based elicitation. It is also important to integrate the existing bio-mechanic eye models with data to achieve robust and data-efficient eye tracking. There are several open problems in this line of research in terms of eye movement dynamics (i.e. gaze directions, velocities, and gaze trajectories) in task-dependent as well as natural behaviors.

4.2.4 *Transformer based*

Transformer models have recently gotten attention for their notable performance on a broad range of vision tasks. Similarly, in the gaze estimation domain, there are two types of transformers used to date which are designed on top of the ViT framework. The first one is the *pure transformer in gaze estimation (GazeTR-Pure)* [149] and the other one is *hybrid transformer in gaze estimation (GazeTR-Hybrid)* [149]. *GazeTR-Pure* [149] takes the cropped face as input along with an extra token. The extra token is a learnable embedding which aggregates the image features together. On the other hand, *GazeTR-Hybrid* [149] is comprised of CNN and transformer. It is based on the fact that gaze estimation is a regression task and it is quite difficult to get the perception of gaze with only local patch based correlation. These models take advantage of the transformer's attention mechanism to improve gaze estimation performance. These are initial explorations using the transformer backbone. There is an immense possibility to explore this architecture.

4.2.5 *VAE/GAN based*

Variational autoencoders and GANs have been used for unsupervised or self-supervised representation learning (Refer Fig. 6). Here, the latent space feature of the autoencoder model is used for gaze estimation inference [28], [29], [101]. Apart from representation learning, VAE and GAN based models are widely used for gaze redirection tasks [100]–[102], [150].

DT-ED. For representation learning via gaze redirection in a person independent manner, variational autoencoders are often utilized [29], [101]. Disentangling Transforming Encoder-Decoder (DT-ED) framework [29] takes an input image x and maps it to the latent space z via an encoder E (i.e. $E(x) : x \to z$). In the latent space, DT-ED disentangles three important factors relevant to gaze, i.e. gaze direction (z_g), head orientation (z_h), and the appearance of the eye region (z_a). Thus, z can be expressed as: $z = \{z_a; z_g; z_h\}$. The framework disentangles these factors by explicitly applying constraints related to gaze and headpose rotations. Further, a decoder D maps z back to the redirected image (i.e. $D(E(x)) : z \to \hat{x}$). The gaze direction is estimated from the z_g part of the latent embedding. The overall illustration is shown in Fig. 6 (bottom row middle image).

ST-ED. Similarly, the Self-Transforming Encoder-Decoder (ST-ED) architecture [101] (Refer Fig. 6 bottom row right image) takes a pair of images x_i and x_t as input, disentangles the subject's personal non-varying embeddings (z_i^0 and z_t^0), considers pseudo-label conditions ($\hat{c}_i$ and $\hat{c}_t$) and embedding representations (z_i and z_t). The learning objective for transformation depends on the pseudo condition labels which consider extraneous factors in absence of ground truth annotation.

Gaze Redirection Network. The main motivation behind the unsupervised gaze redirection network [28] is capturing generic eye representation via gaze redirection (Refer Fig. 6 bottom row left image). The framework takes eye patch I_i as input and predict the redirected eye patch as output I_o while preserving the difference in rotation $\Delta_r = r_i - r_o = G_\phi(I_i) - G_\phi(I_o)$. In this work, gaze redirection is used as a pretext task for representation learning.

RITnet. RITnet [136] (Refer Fig. 6 top row right image) is a hybrid version of U-Net and DenseNet based upon Fully Convolutional Networks (FCN). To balance the trade-off between the performance and computational complexity, it consists of 5 Down-Blocks in the encoder and 4 Up-Blocks in the decoder where the last layer of the encoder block is termed as the bottleneck layer. Each Down-Block has 5 convolution layers with LeakyReLU activation and the layers share connections with previous layers similar to DenseNet architecture. Similarly, each Up-Block has 4 convolution layers with LeakyReLU activation. All Up-Blocks have skip connection with their corresponding Down-Block which is an effective strategy to learn representation. To train the model, the following loss functions are used: *1) Standard cross-entropy loss (CEL)* is applied pixel-wise to categorize each pixel into four categories (i.e. background, iris, sclera, and pupil). *2) Generalized Dice Loss (GDL)* penalize the pixels on the basis of the overlap between the ground truth pixel and corresponding prediction. *3) Boundary Aware Loss (BAL)* weights each pixel in terms of distance with its two nearest neighbour. This loss helps to avoid CEL confusion in boundary region. *4) Surface Loss (SL)* helps to recover small regions and contours via distance based scaling. The overall loss is defined as follows:

$$\ell = \ell_{CEL}(\lambda_1 + \lambda_2 \ell_{BAL}) + \lambda_3 \ell_{GDL} + \lambda_4 \ell_{SL}.$$

Similarly, another lightweight model [151] uses MobileNet with depth-wise separable convolution for efficiency. It also utilize a squeeze and excitation (SE) module for performance enhancement by modelling channel independence. Moreover, the heuristic filtering of the connected component is utilized to enforce biological coherence in the network. Few works [152], [153] also use multi-class classification strategy for rich representation learning.

Other Statistical Modelling. The statistical inference based mapping is performed based on k-NN [87], support vector regression [23], [87] and random forest [20], [87]. A brief overview of these methods is summarised in Table 2. Prior deep learning works on semantic eye segmentation is mainly focused on iris or sclera segmentation via Fuzzy C Means clustering, Otsu's binarization, k-NN [154] etc. Sclera segmentation challenge was organized since 2015 to promote development in this area [133], [154], [155]. Recently, the OpenEDS challenge was organized in 2019 by Facebook Research in which eye segmentation was one of the sub-challenges. Most of the methods in this challenge use deep learning techniques [136], [144], [156].

Table 2: A comparison of gaze analysis methods with respect to registration (Reg.), representation (Represent.), Level of Supervision, Model, Prediction, validation on benchmark datasets (validation), Platforms (Plat.), Publication venue (Publ.) and year. Here, GV: Gaze Vector, Scr.: Screen, LOSO: Leave One Subject Out, LPIPS: Learned Perceptual Image Patch Similarity, MM: Morphable Model, RRF: Random Regression Forest, AEM: Anatomic Eye Model, GRN: Gaze Regression Network, ET: External Target, FV: Free Viewing, HH: HandHeld Device, HMD: Head Mounted Device, Seg.: Segmentation and GR: Gaze Redirection, LAEO: Looking At Each Other.

Ref.	Reg.	Represent.	Level of Sup.	Model	Prediction	Validation	Plat.	Publ.	Year
[23]	Face [157]	Appear.	Fully-Sup.	SVM	Gaze locking	[23]	Scr.	UIST	2013
[158]	3-D MM	Appear.	Fully-Sup.	Convex Hull	3-D GV	[158]	ET.	ETRA	2014
[20]	Face, Eye [20]	Appear.	Fully-Sup.	RRF	3-D GV	[20]	Any	CVPR	2014
[159]	Eye	Appear.	Fully-Sup.	CNN+CLNF	3-D GV	[24]	Any	ICCV	2015
[24]	Face, L/R Eye	Appear.	Fully-Sup.	CNN [24]	3-D GV	[24]	Scr.	CVPR	2015
[25]	Face, L/R Eye	Appear.	Fully-Sup.	iTracker [25]	2-D Scr.	[25], [87]	HH	CVPR	2016
[160]	Eye	Appear.	Fully-Sup.	CNN [25]	GR Img.	[160]	Any	ECCV	2016
[87]	Eye [161]	Appear.	Fully-Sup.	SVR	2-D Scr.	[87]	HH	MVA	2017
[54]	Eye [162]	Appear.	Fully-Sup.	VGG-16+FC [54]	3-D GV	[24], [54]	Scr.	ECCV	2018
[26]	Eyes	Appear.	Fully-Sup.	CNN	3-D GV	[24], [158]	Scr.	ECCV	2018
[57]	Face [104]	Geo.+Appear.	Fully-Sup.	CNN [57]	3-D GV	[23], [87]	Desk.	ICPR	2018
[92]	Eye	Geo.+Appear.	Fully-Sup.	HGSM+c-BiGAN	Eye, GV	[24], [158]	Any	CVPR	2018
[64]	Face, L/R Eye	Appear.	Fully-Sup.	Dilated CNN	3-D GV	[23]–[25]	Scr.	ACCV	2018
[29]	Face	Appear.	Few-Shot	DT-ED+ML	3-D GV	[24], [25]	Scr.	ICCV	2019
[90]	Face	Appear.	Fully-Sup.	Pinball LSTM	3-D GV	[23], [24], [87]	ET	ICCV	2019
[32]	Eye	Appear.	Fully-Sup.	SegNet [163]	Seg. Map	[32]	HMD	ICCVW	2019
[66]	Face, L/R Eye	Appear.	Fully-Sup.	DGTN	GV	[66]	Desk.	CVPR	2019
[164]	Face	Appear.	Fully-Sup.	MeNet	3-D GV	[20], [24], [25]	Scr.	CVPR	2019
[145]	Face, Eye	Appear.	Semi/Unsup.	BCNN	3-D GV	[24], [158]	Desk.	CVPR	2019
[136]	Eyes	Appear.	Fully-Sup.	Hybrid U-net	Seg. Map	[32]	HMD	ICCVW	2019
[165]	Eyes	Appear.	Fully-Sup.	Modified Resnet	Seg. Map	[32]	HMD	ICCVW	2019
[156]	Eyes	Appear.	Fully-Sup.	Eye-MMS	Seg. Map	[32]	HMD	ICCVW	2019
[144]	Eyes	Appear.	Fully-Sup.	Dilated CNN	Seg. Map	[32]	HMD	ICCVW	2019
[91]	Eyes	Appear.+Seg.	Few-shot	GR	2-D GV	[23], [24]	Any	CVPR	2019
[28]	Eyes	Appear.	Unsup.	GR	2-D GV	[23], [24]	Any	CVPR	2019
[27]	Face, Eye	Appear.	Unsup.	IzeNet	3-D GV	[23], [87]	FV	IJCNN	2019
[166]	Eye	Appear.+Seg.	Fully-Sup.	Seg2Eye	Eye Img.	[32]	HMD	ICCVW	2019
[167]	Eye Seq.	Appear.	Unsup.	Hier. HMM	Eye Move.	[168]	Any	ECCVW	2019
[169]	Eye	Appear.	Semi/Unsup.	mSegNet+Discre.	Seg. Map	[32]	HMD	ECCVW	2019
[170]	Eye	Appear.	Few-Shot	EyeSeg	Seg. Map	[32]	HMD	ECCVW	2019
[101]	Face	Appear.	Fully-Sup.	ST-ED	GR	[23], [25], [158]	Scr.	NeurIPS	2020
[34]	Eye	Appear.	Fully-Sup.	Modified ResNet	GR Img.	[34]	HMD	ECCVW	2020
[148]	Eyes	Appear.	Fully-Sup.	ResNet-18+GRU	PoG,3-D GV	[148]	Scr.	ECCV	2020
[94]	Face	Appear.	Fully-Sup.	ResNet-50	3-D GV	[24], [25], [90], [158]	Scr.	ECCV	2020
[171]	Face	Appear.	Semi-Sup.	GRN	GV	[113]	FV	WACV	2020
[60]	Face, Eye	Appear.	Fully-Sup.	RSN+GazeNet	GV	[24], [25], [158]	Scr.	BMVC	2020
[59]	Face, Eye	Appear.	Fully-Sup.	CA-Net	GV	[24], [158]	Scr.	AAAI	2020
[172]	Face, Eye	Appear.	Fully-Sup.	FAR-Net	GV	[24], [54], [158]	Scr.	TIP	2020
[102]	Eye	Appear.+AEM	Fully-Sup.	MT c-GAN	Eye Img.	[20], [23], [24]	Scr.	WACV	2021
[89]	Face, Eye	Appear.	Fully-Sup.	AFF-Net	Scr., GV	[25], [51]	Scr.	Arxiv	2021
[173]	Face	Appear.	Unsup.	PureGaze	Face, GV	[20], [24], [90], [94]	Scr.	Arxiv	2021
[110]	Face	Appear.	Weakly-Sup.	ResNet-18+LSTM	GV	[25], [90], [94], [110]	Any	CVPR	2021
[109]	Face	Appear.	Fully-Sup.	LAEO-Net++	LAEO	[108]	Any	TPAMI	2021
[174]	Face, Eye	Appear.	Limited-Sup.	ResNet-50	GV	[23], [24], [40], [90]	Any	WACV	2022

4.2.6 Discussion

In an attempt to summarize the recent deep network based gaze analysis methods, we present some main take away points as follows:

- The overall gaze estimation methods are divided into two broad categories: *1) 2-D Gaze Estimation:* In this context the proposed methods map the input image to 2-D Point of Regard (PoR) in the visual plane. The visual planes could either be the observable object or screen. Non deep learning methods or early deep learning methods [21], [24], [50], [51] perform these mappings. *2) 3-D Gaze Estimation:* The 3-D gaze estimation basically considers the gaze vector instead of 2-D PoR. The gaze vector is the line joining the pupil center point with the point of regard. Recent works [26], [29], [94], [110], [148] mainly relies on 3-D gaze estimation methods. The choice of gaze estimation methods rely on the application and requirement.
- Single branch CNN based architectures [24]–[26], [50], [51], [92] are widely used over the past few years for progressive improvements on benchmark datasets. The input to these networks are restricted to single eye, eye patch or face. Thus, to further boost the performance, multi branch networks are proposed which utilize eyes, face, geometric constraints, visual plane grid as input.
- Both single or multi branch networks depend on spatial information. However, eye movement is dynamic in nature. Thus, few recent proposed architectures [90], [92] use temporal information for inference.
- For representation learning, VAE and GAN based architectures [28], [29], [101] are explored. However, it is observed

that these architectures could have high time complexity as compared to single or multi branch CNN.

- *Prior based appearance encoding* is another line of approaches for encoding rich feature representation. Few works have defined priors based on eye anatomy [26], geometrical constraint [172] as biases for better generalization. Despite direct appearance encoding, Park et al. [26] proposed an intermediate pictorial representation, termed a 'gazemap' (refer Fig. 6) of the eye to simplify the gaze estimation task. Similarly, the 'two eye asymmetry' property is utilized for gaze estimation [172] where the underlying hypothesis is that despite the difference in appearances of two eyes due to environmental factors, the gaze directions remains approximately the same. The CNN based regression model is assumed to be independent of identity distribution, however, due to the subject-specific offset of the nodal point of the eyes, gaze datasets have identity specific bias. Xiong et al. [164] inject this bias as a prior by mixing different models. Similarly, to handle this offset, the gaze is decomposed into the subject independent and dependent bias for performance enhancement and better generalization [175].
- In order to train the deep learning based models, ℓ_2 [24], [50], [51] and cosine similarity based losses [94], [148] are used. However, a novel pinball loss [90] is proposed to model the uncertainty in gaze estimation, especially in unconstrained settings.
- Similarly for deep learning based eye segmentation approaches, the eye image to segmentation mapping is performed in a non-parametric way which implicitly encodes shape, geometry, appearance and other factors [32], [67], [136], [152], [165], [176]. The most popular network architectures for eye segmentation are U-net [177], modified version of SegNet [32], RITnet [136], EyeNet [165]. These VAE based architechtures have high time and space complexity. However, recent methods [136], [165] do consider these factors without compromising the performance.

4.3 Level of Supervision

Based on the type of supervision, the training procedure can be classified into the following categories: *fully-supervised, Semi-/Self-/weakly-/unsupervised.*

4.3.1 Fully-Supervised.

Supervised learning paradigm is the most commonly used training framework in gaze estimation literature [20], [23], [24], [26], [87], [97] and eye segmentation literature [11], [133], [135], [136], [144], [154]–[156], [165]. As the fully-supervised methods require a lot of accurately annotated data. Accurate annotation of gaze data is a complex, noise-prone, and time-consuming task and sometimes it requires expensive data acquisition setups. Moreover, there is a high possibility of noisy or wrong annotation due to distraction in participation during data collection, eye blink activity and inherent measurement errors in data curation settings. Variation in data curation setup limits merging multiple datasets for supervision. Dataset specific data acquisition processes are discussed in Sec. 5.1. Thus, the research community is moving towards learning with less supervision.

Multi-Task Learning. Multi-task learning incorporates different tasks which provide auxiliary information as a bias to improve model performance. The auxiliary information can be Gaze+Landmark [61], PoG+Screen saliency [148], [178], Gaze+Depth [71], Gaze+Headpose [52], Segmentation+Gaze [67] and Gaze-direction+Gaze-uncertainty [90]. These gaze aligned tasks facilitate strong representation learning with additional task based supervision.

4.3.2 Semi-/Self-/Weakly-/Unsupervised.

To a large extent, the supervised deep learning based methods' performance depends on the quality and quantity of annotated data. However, manual labeling of gaze data is a complex, time consuming and labor extensive process. On this front, Semi-/Self-/Weakly-/Unsupervised Learning paradigms provide a promising alternative to enable learning from a vast amount of readily available non-annotated data. For learning paradigms with less supervision, the important methods are described in detail below:

Weakly-supervised and Learning from Pseudo Labels. Weakly supervised learning aims to bridge the gap between the fully-supervised and fully-unsupervised techniques. Till date in gaze estimation domain, the weak supervision has been performed via 'Looking At Each Other (LAEO)' [179] and pseudo labelling [174]. For weak supervision, Kothari et al. [179] leverage strong gaze-related geometric constraints from two people interaction scenario. On the other hand, MTGLS [174] framework leverages from non-annotated facial image data by three complementary signals i.e. (1) the line of sight of the pupil, (2) the head-pose and (3) the eye dexterity.

Unsupervised and Self-supervised Representation Learning. Self supervised learning has emerged as a popular technique for learning meaningful representations from vast amount of non-annotated data. It requires pseudo labels for any pre-designed task which is often termed as *Auxiliary* or *Pretext* Task. The pre-designed task is mostly aligned with the gaze estimation. Dubey et al. [27] propose a pretext task where the visual regions of the gaze are divided into zones by geometric constraints. These pseudo labels are utilized for representation learning. Yu et al. [28] use subject specific gaze redirection as a pretext task. Swapping Affine Transformations (SwAT) [180] is the extended version of Swapping Assignments Between Views (SwAV), a popular self supervised learning framework used for gaze representation learning using different augmentation techniques. The self-supervised representation learning has the potential to eliminate the major drawback of gaze data annotation which is quite difficult and error prone. Future directions in this area may include designing better pre-text tasks and combining multiple pre-text tasks to jointly pre-train the models [174]. In addition, combining data-driven eye tracking with model-based eye tracking can be another future direction as model-based eye tracking can provide pseudo-labels or pre-train the models.

Few Shot Learning. Few-shot learning aims to adapt to a new task with very few examples [29], [91]. The main challenge in few shot paradigm is over-fitting issue since highly over-parameterized deep networks are involved to learn from only a few training samples. To this front, mainly gaze redirection strategy [91] and Few-shot Adaptive GaZE Estimation (FAZE) [29] frameworks are proposed. Among them, FAZE is shown a two stage adaptation strategy. In the first stage, a rotation aware latent space embedding is learned based on encoder-decoder framework. Further, adaptation is performed on top of the features using MAML which is a popular meta-learning paradigm. FAZE is able to adapt to a new subject with ≤ 9 sample which is quite promising.

Learning-by-synthesis. The term 'learning-by-synthesis' is coined by Sugano et al. [20]. The main objective is to synthesize

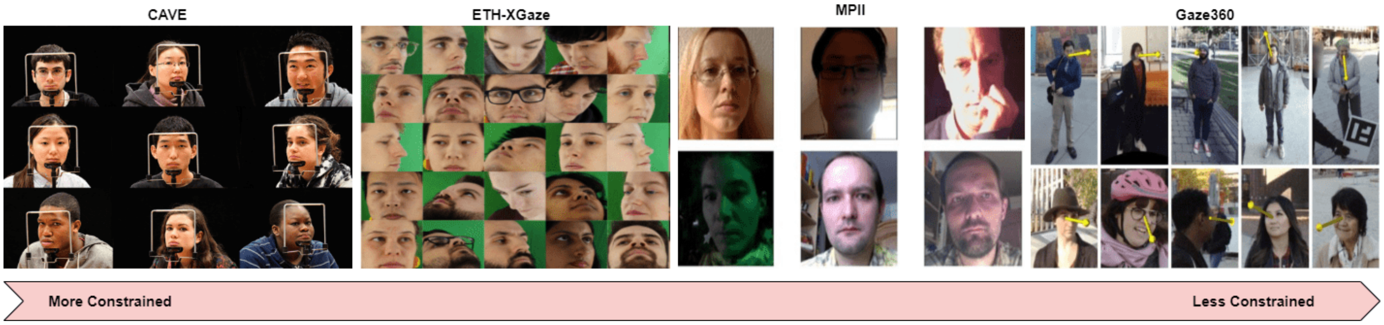


Figure 7: Data collection procedure in different settings for benchmark datasets. From left to right the examples are from CAVE [23], Eth-XGaze [94], MPII [50] and Gaze360 [90] datasets. The leftmost one is more constrained and the rightmost one is less constrained. Images are taken from respective datasets [23], [50], [90], [94]. Refer Table 3 for more details.

different gaze viewpoints to multiply the data from a quantitative and a qualitative perspectives rather than manual labelling. Few other studies [20], [91], [92], [160], [181]–[183] also adopt data generation methods which can address the diversity in terms of headpose and eye rotation. However, these generative models have high computational complexity and are constrained by the quality of generated images.

Discussion. In gaze estimation domain, there are still very few works in semi-/self-/weakly-/unsupervised learning paradigms. Among these works, learning from pseudo labels have their limitations as the label space contains noise. On the other hand, gaze redirection synthesis is based on the availability of same or different person's data with eye rotation or the prior knowledge of rotation angle. Thus, learning robust and generalizable gaze representations using minimal or no supervision still remains an open-ended research question. One possible direction to alleviate this problem is to combine data-driven eye tracking with model-based eye tracking to produce physically plausible eye tracking models that are data efficient during training and generalize better during testing.

5 Validation

In this section, we review the commonly followed evaluation procedures on various datasets along with the metrics adopted in the literature.

5.1 Datasets for Gaze Analysis

With the rapid progress in the gaze analysis domain, several datasets have been proposed for different gaze analysis tasks (see Sec. 3). The dataset collection technique has evolved from constrained lab environments [23] to unconstrained indoor [50], [51], [54], [87] and outdoor settings [90] (Refer Fig. 7). We provide a detailed overview of the datasets in Table 3. Compared with early datasets [23], [158], recently released datasets [90], [148] are typically more advanced with less bias, improved complexity, and larger in scale. These are better suited for training and evaluation. We describe a few important datasets below:

CAVE [23] contains 5,880 images of 56 subjects with different gaze directions and head poses. There are 21 different gaze directions for each person and the data was collected in a constrained lab environment, with 7 horizontal and 3 vertical gaze locations.

The *Eyediap* dataset [184] was designed to overcome the main challenges associated with the head pose, person and 3-D target variations along with changes in ambient and sensing conditions.

TabletGaze [87] is a large unconstrained dataset of 51 subjects with 4 different postures and 35 gaze locations collected using a tablet in an indoor environment. TabletGaze dataset is also collected in a 7 × 5 grid format.

MPII [50] gaze dataset contains 213,659 images collected from 15 subjects during natural everyday events in front of a laptop over a three-month duration. MPII gaze dataset is collected by showing random points on the laptop screen to the participants. Further, Zhang et al. [51] curate *MPIIFaceGaze* dataset with the hypothesis that gaze can be more accurately predicted when the entire face is considered.

RT-GENE dataset [54] is recorded in a more naturalistic environment with varied gaze and head pose angles. The ground truth annotation was done using a motion capture system with mobile eye-tracking glasses.

Gaze360 [90] is a large-scale gaze estimation dataset collected from 238 subjects in unconstrained indoor and outdoor settings with a wide range of head pose.

ETH-XGaze [94] is a large scale dataset collected in a constraint environment with a wide range of head pose, high-resolution images. The dataset contains images from different camera positions, illumination conditions to add more challenges to the data.

EVE [148] is also collected in constraint indoor setting with different camera views to map human gaze in screen co-ordinate.

Similar to gaze estimation several benchmark datasets have been proposed over the past few years for eye and sclera segmentation. The datasets collected for sclera segmentation is in a constraint environment and with very few subjects [185]–[187]. A more challenging publicly available dataset was released in sclera recognition challenges [133], [154], [155]. Recently, a large scale dataset termed as *OpenEDS: Open Eye Dataset* [32], is released which contains eye images collected by using a VR head-mounted device. Additionally, there was two synchronized eye facing cameras having a frame rate of 200 Hz. The data was collected under controlled illumination and contains 12,759 images with eye segmentation masks collected from 152 participants.

Data Generation/Gaze Redirection. Since gaze data collection and annotation is an expensive and time-consuming process, the research community moves towards a data generation process for benchmarking with a large variation in data attributes. Prior works in this domain generate both synthetic and real images. The methods are based on Generative Adversarial Networks (GANs). To capture the possible rotational variation in images, gaze redirection techniques [91], [160], [181]–[183] are quite popular. An early work on gaze manipulation [195] uses pre-recording of several potential eye replacements during test time. Further,

Table 3: **Datasets.** A comparison of gaze datasets with respect to several attributes (i.e. number of subjects (# sub), gaze labels, modality, headpose and gaze angle in yaw and pitch axis, environment (Env.), baseline method, data statistics (# data), and year of publication.) The abbreviations used are: In: Indoor, Out: Outdoor, Both: Indoor + Outdoor, Gen.: Generation, u/k: unknown, Seq.: Sequence, VF: Visual Field, EB: Eye Blink, GE: Gaze Event [179], GBRT: Gradient Boosting Regression Trees, GC: Gaze Communication, GNN: Graph Neural Network and Seg.: Segmentation.

Dataset	# Sub	Label	Modality	Head-Pose	Gaze	Env.	Baseline	# Data	Year
CAVE [23]	56	3-D	Image Dim.:5184 × 3456	$0^\circ, \pm30^\circ$	$\pm15^\circ, \pm10^\circ$	In	SVM **Eval.:**Cross-val	**Total:**5880	2013
EYEDIAP [158]	16	3-D	Image Dim.: HD and VGA	$\pm15^\circ, 30^\circ$	$\pm25^\circ, 20^\circ$	In	Convex Hull **Eval.:** Hold out	**Total:**237 min	2014
UT MV [20]	50	3-D	Image Dim.:1280 × 1024	$\pm36^\circ, \pm36^\circ$	$\pm50^\circ, \pm36^\circ$	In	Random Reg. Forests **Eval.:**Hold out	**Total:**64,000	2014
OMEG [188]	50	3-D	Image Dim.: 1280 × 1024	$0^\circ, \pm30^\circ$	$-38^\circ$to $+36^\circ$, $-10^\circ$to $+29^\circ$	In	SVR **Eval.:**LOSO	**Total:**44,827	2015
MPIIGaze [24]	15	3-D	Image Dim.: 1280 × 720	$\pm15^\circ, 30^\circ$	$\pm20^\circ, \pm20^\circ$	In	CNN variant [24] **Eval.:**LOSO	**Total:**213,659	2015
GazeFollow [189]	130,339	3-D	Image Dim.: Variable	Variable	Variable	Both	CNN variant [189] **Eval.:**Hold out	**Total:**122,143	2015
SynthesEye [159]	NA	3-D	Image Dim.:120 × 80	$\pm50^\circ, \pm50^\circ$	$\pm50^\circ, \pm50^\circ$	Syn	CNN [159] **Eval.:**Hold out	**Total:**11,400	2015
GazeCapture [25]	1450	2-D	Image Dim.:640 × 480	$\pm30^\circ, 40^\circ$	$\pm20^\circ, \pm20^\circ$	Both	CNN [25] **Eval.:** Hold out	**Total:**2,445,504	2016
UnityEyes [190]	NA	3-D	Image Dim.:400 × 300	Variable	Variable	Syn	KNN **Eval.:**NA	**Total:** 1,000,000	2016
TabletGaze [87]	51	2-D Sc.	Video Dim.: 1280 × 720	$\pm50^\circ, \pm50^\circ$	$\pm20^\circ, \pm20^\circ$	In	SVR **Eval.:**Cross-val	**Total:**816 Seq. $\sim$ 300,000 img.	2017
MPIIFaceGaze [51]	15	3-D	Image Dim.: 1280 × 720	$\pm15^\circ, 30^\circ$	$\pm20^\circ, \pm20^\circ$	In	CNN variant [51] **Eval.:**LOSO	**Total:**213,659	2017
InvisibleEye [70]	17	2-D Sc	Image Dim.: 5 × 5	Unknown	2560 × 1600 pixel VF	In	ANN [70] **Eval.:** Hold out	**Total:**280,000	2017
RT-GENE [54]	15	3-D	Image Dim.:1920 × 1080	$\pm40^\circ, \pm40^\circ$	$\pm40^\circ, -40^\circ$	In	CNN [54] **Eval.:**Cross val	**Total:**122,531	2018
Gaze 360 [90]	238	3-D	Image Dim.:4096 × 3382	$\pm90^\circ$, u/k	$\pm140^\circ, -50^\circ$	Both	Pinball LSTM **Eval.:** Hold out	**Total:** 172,000	2019
RT-BENE [191]	17	EB	Image Dim.: 1920 × 1080	$\pm40^\circ, \pm40^\circ$	$\pm40^\circ, -40^\circ$	In	CNNs **Eval.:** Cross val	**Total:** 243,714	2019
NV Gaze [68]	30	3-D, Seg.	Image (Synthetic) Dim.:1280 × 960, 640 × 480	Unknown	$30^\circ \times 40^\circ$ VF	Both	CNN [192] **Eval.:** Hold out	**Total:** 2,500,000	2019
HUST-LEBW [143]	172	EB	Video Dim.: 1280 × 720	Variable	Variable	Both	MS-LSTM **Eval.:** Hold out	**Total:** 673	2019
VACATION [111]	206,774	GC	Video Dim.: 640 × 360	Variable	Variable	Both	GNN **Eval.:** Hold out	**Total:** 96,993	2019
OpenEDS-19 [32] Track 1: Semantic Segmentation	152	Seg.	Image Dim.: 640 × 400	Unknown	Unknown	In	SegNet [163] **Eval.:** Hold out	**Total:**12,759 (in # SegSeq [32])	2019
OpenEDS-19 [32] Track 2: Synthetic Eye Generation	152	Gen.	Image Dim.: 640 × 400	Unknown	Unknown	In	**Eval.:** Hold out	**Total:** 252,690	2019
OpenEDS-20 [34] Track 1: Gaze Prediction	90	3-D	Image Dim.: 640 × 400	Unknown	$\pm20^\circ, \pm20^\circ$	In	Modified ResNet **Eval.:** Hold out	**Total:** 8,960 Seq., 550,400 img.	2020
OpenEDS-20 [34] Track 2: Sparse Temporal Semantic Segmentation	90	Seg.	Image Dim.: 640 × 400	Unknown	$\pm20^\circ, \pm20^\circ$	In	SegNet [163] (Power Efficient version) Eval.: Hold out	**Total:** 200 Seq. 29,500 img.	2020
mEBAL [142]	38	EB	Image Dim.: 1280 × 720	Variable	Variable	In	VGG-16 Varient **Eval.:** Hold out	**Total:** 756,000	2020
ETH-XGaze [94]	110	3-D	Image Dim.: 6000 × 4000	$\pm80^\circ, \pm80^\circ$	$\pm120^\circ, \pm70^\circ$	In	ResNet-50 **Eval.:** Hold out	**Total:** 1,083,492	2020
EVE [148]	54	3-D	Image Dim.: 6000 × 4000	$\pm80^\circ, \pm80^\circ$	$\pm80^\circ, \pm80^\circ$	In	ResNet-18 **Eval.:** Hold out	**Total:** 12,308,334	2020
GW [179]	19	GE	Image Dim.: 1920 × 1080	Variable	Variable	In	RNN **Eval.:** Hold out	**Total:** $\sim$ 5,800,000	2020
LAEO [110]	485	3-D	Image Dim.: Variable	Variable	Variable	Both	ResNet-18+LSTM **Eval.:** Hold out	**Total:** 800,000	2021
GOO [193]	100	3-D	Image Dim.: Variable	Variable	Variable	Both	ResNet-50 **Eval.:** Hold out	**Total:** 201,552	2021
OpenNEEDS [194]	44	3-D	Image Dim.: 128 × 71	Variable	Variable	VR	GBRT **Eval.:** Hold out	**Total:** 2,086,507	2021

Table 4: **Cross-Dataset Study.** Cross dataset generalization study on different gaze estimation datasets in terms of angular error (in °).

Model	**Test→ Train ↓**	**Datasets**					
		CAVE	MPIIFace	RT-GENE	Gaze360		
Pinball-LSTM [90]	CAVE	–	12.3°	32.8°	57.9°		
	MPIIFace	12.4°	–	26.5°	57.8°		
	RT-GENE	24.2°	18.9	–	56.6°		
	Gaze360	9.0°	12.1	13.4°	–		
		MPIIGaze	EYEDIAP	Gaze-Capture	RT-GENE	Gaze360	ETH-X Gaze
ETH-X Gaze [94]	MPIIGaze	–	17.9°	6.3°	14.9°	31.7°	34.9°
	EYEDIAP	16.9°	–	14.2°	15.6°	33.7°	41.7°
	Gaze-Capture	4.5°	13.7°	–	14.7°	30.2°	29.4°
	RT-GENE	12.0°	21.2°	13.2°	–	34.7°	42.6°
	Gaze360	10.3°	11.3°	12.9°	26.6°	–	17.0°
	ETHXGaze	7.5°	11.0°	10.5°	31.2°	27.3°	–

Kononenko et al. [196] propose wrapping based gaze redirection using supervised learning, which learns the gaze redirection via a flow field to move eye pupil and relevant pixels from the input image to the output image. The gaze re-direction methods may struggle with extrapolation since it depends on the training samples and training methods. Moreover, these works suffer from low-quality generation and low redirection precision. To overcome this, Chen et al. [100] propose a MultiModal-Guided Gaze Redirection (MGGR) framework which uses gaze-map images and target angles to adjust a given eye appearance via learning. The other approaches are mainly based on random forest [196] and style transfer [197]. Random forest is used to decide the possible gaze direction and in style transfer, the appearance based feature is mainly encoded. Sela et al. [197] propose a GAN based framework to generate a large dataset of high-resolution eye images having diversity in subjects, head pose, camera settings and realism. However, the GAN based methods lack in their capability to preserve content (i.e. eye shape) for benchmarking. Buhler et al. [166] synthesize person-specific eye images with a given semantic segmentation mask by preserving the style and content of the reference images. In summary, we can say that although a lot of effort has been made to generate realistic eye images, but due to several limitations (perfect gaze direction, image quality), these images are not used for benchmarking.

5.2 Evaluation Strategy

In this section, we describe the most widely used gaze metrics in the gaze analysis domain.

Gaze Estimation. The most common practice to measure the gaze estimation accuracy/error is in terms of angular error (in °) [26], [29], [94], [148] and gaze location (in pixels or cm/mm(s)) [87], [148]. The angular error is measured between the actual gaze direction ($\mathbf{g} \in \mathbb{R}^3$) and predicted gaze direction ($\hat{\mathbf{g}} \in \mathbb{R}^3$) defined as $\frac{\mathbf{g}.\hat{\mathbf{g}}}{||\mathbf{g}||.||\hat{\mathbf{g}}||}$. On the other hand, Euclidean distance is measured between the original and predicted point of gaze (PoG).

Gaze Redirection. The gaze redirection evaluation is performed in both quantitative and qualitative manner [100]–[102]. The quantitative analysis is done in terms of angular gaze redirection error estimated between the predicted values and their intended target values. As in this task, the moment of the eye pupil is pre-defined, thus, this angular error weakly quantifies how perfectly the eye redirection occurs, although the method for measuring the angle has some inherent noise. For qualitative analysis, the Learned Perceptual Image Patch Similarity (LPIPS) metric is used which measures the paired image similarity in the gaze redirection task.

Eye Segmentation. Commonly used evaluation metric for eye segmentation methods, is average of the mean Intersection over Union (mIoU). Although, for the recent OpenEDS challenge [32], the mIoU metric is calculated for all classes and model size (S) is calculated as a function of a number of trainable parameters in megabytes (MB).

5.3 Cross Dataset Analysis.

Datasets play an important role in defining the research progress made in gaze analysis. Apart from serving as a source for training models, it helps to quantify the performance measure. In the gaze analysis domain, the aim of dataset curation is to capture the real-world scenario setting as close as possible. Thus, it is necessary to evaluate the robustness and generalizability of the models across different data acquisition setups for better adaptation.

On this front, we explore two aspects: First of all, we explore the cross-dataset generalizability of gaze estimation methods based on two SOTA models i.e. Pinball-LSTM [90] and ETH-X-Gaze [94]. For this purpose, the training is performed on one dataset while the testing is conducted on the other dataset. (Refer Table 4). Further, we explore the SOTA method's performance on different datasets to show the robustness of the model (Refer Table 5). Angular error (in °) is used as an evaluation metric. Further to generalize across datasets, we calculate the mean angular error across datasets. Below are some of the important observations inferred from our experimental results.

Data Collection Settings. Dataset collection setup plays an important role in generalizability and method's robustness. As the CAVE dataset is collected in a constrained setup and it has high-resolution images, the models trained on this data fail to adapt well to the data with low resolution and synthetic images. Thus, the pinball-LSTM trained on CAVE data has a high error in Gaze360 ($\sim$ 57.9°) and RT-GENE ($\sim$ 32.8°) datasets. A similar pattern is observed in the case of the MPII dataset as well. Model trained on this dataset gives high error in adapting RT-GENE ($\sim$ 26.5°) and Gaze360 ($\sim$ 57.8°).

Cross Dataset Generalization. By observing the cross dataset generalization performance, we can determine how diverse the training dataset is from a generalization perspective. From Table 4, we observe that Gaze360, Gaze-Capture, and ETH-X Gaze datasets are the most challenging datasets. Training models on these two datasets would be a good choice as it has better generalization performance across different datasets. In contrast, RT-GENE and Gaze-capture contain significant biases and training models on them will lead to poor cross-dataset generalization performance.

Robust Modelling. In order to study the robustness of a model trained on any dataset, we recommend evaluating the model on Gaze360, Gaze-Capture, and ETH-X Gaze datasets. These datasets exhibit multiple variations in terms of background environment, eye visibility, occlusion, low-resolution images and could serve as important indicators for real-world adaptation. We also recommend training the models on all benchmark datasets together and expect a better generalization than training on individual datasets in novel or in-the-wild settings.

5.4 Where We Stand Now?

In Table 5, we analyze dataset specific improvements made by different methods over the past few years. In the following, we discuss some of the important observations from Table 5.

Gaze Estimation on Constrained Setup. Most popular *2D-3D gaze estimation* datasets [23], [50], [94] are collected in constrained scenarios where there is a certain distance between the user and the visual screen. Moreover, as the nodal point of the human eye has subject-specific offset (which varies around 2-3°), it is difficult to reduce the angular error beyond a certain limit using visible regions of the eyes. On this front, the performance of some of the gaze estimation methods [24], [51], [94] seems to have plateaued on constrained datasets such as CAVE, MPII, and Eth-X-Gaze (Refer to Figures in Table 5).

Gaze Estimation in Unconstrained Setup. Gaze estimation in unconstrained environments still remains largely unresolved mainly due to the unavailability of large-scale annotated data. Gaze360 [90] and GazeCapture [25] are two popular public-domain datasets available for this purpose. Especially in the Gaze360 dataset, in many cases, the eyes are not visible which makes it more challenging to track where the person is looking. It is quite difficult to estimate gaze in a naturalistic environment, more exploration along this line is highly desirable.

Gaze Estimation with Limited Supervision. Gaze Estimation with limited supervision is a promising research direction. As manual annotation of gaze data is an error-prone process, there is a high possibility of noise in labeling. To this end, proposed approaches are mainly based on 'learning-by-synthesis' [20], hierarchical generative models [92], conditional random field [198], unsupervised gaze target discovery [95], few-shot learning [29], [91], pseudo-labelling [174] and self/unsupervised [27], [28]. While domain specific knowledge has been utilized for these approaches, developing robust methods from limited amount of annotated data with enhanced generalization across different real-life scenarios still largely remains unresolved.

Visual Attention Estimation. Eye visibility plays an important role in estimating the gaze direction of a person. To this end, visual attention estimation mainly focuses on where the person is looking irrespective of eye visibility. To facilitate research along this direction, GazeFollow [113] and VideoAttentionTarget [114] datasets have been proposed. Some important research directions to explore include scene saliency, visual search, and human scan path [117], [119].

Gaze Trajectory Modelling. Gaze Trajectory modeling and estimation is another line of research that requires further research attention [34]. Natural gaze dynamics consist of a continuous sequence of gaze events such as *fixations, saccades, pursuit, vestibulo-ocular reflex, optokinetic reflex, vergence,* and *blinks* [**?**]. These aforementioned dynamics can be influenced by saliency, task-relevant information, and environmental factors. Natural eye movements of humans span an elliptical region with a horizontally oriented axis greater than $\sim 100°$ and a vertically oriented axis spanning $\sim 70°$. Due to the lack of labeled temporal gaze trajectory data, there are only a few studies [34], [146] that focus on tasks related to the gaze trajectory.

6 Applications

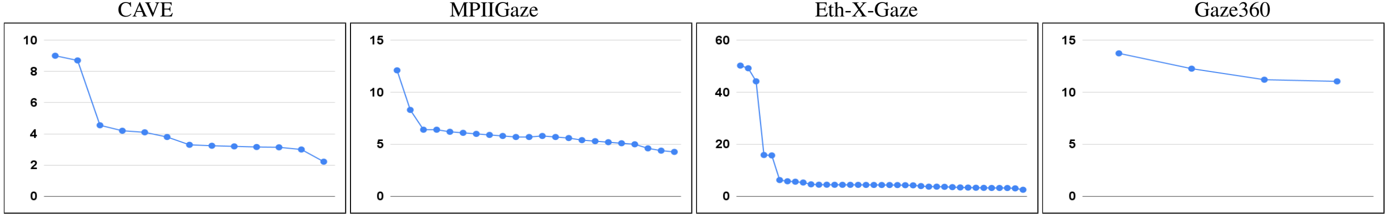
6.1 Gaze in Augmented Reality, Virtual Reality and 360° Video Streaming

We are witnessing great progress in the adaptation of VR, AR and 360° Video Streaming technology. Eye-tracking has the potential to bring revolution in the AR/VR and 360° video streaming for immersive video application space since it can enhance the device's awareness by learning about users' attention at any given point in time. Consequently, user's focus based optimization reduces power consumption by these devices [34], [199]–[201]. In this section, we will cover the importance of eye-tracking technology and how it enables better user experience in AR/VR and 360° Video Streaming devices including eye inter pupillary distance for estimating image perception quality, person identification or state estimation by their eye gaze pattern, improve interactions, etc.

Foveated Rendering (a.k.a gaze-contingent eye tracking) is a process designed to show the user only a portion of what they are looking at in full detail [32], [34], [200]. The focus region follows the user's visual field. Graphics displayed with foveated rendering better matches the way we see objects. Usually, the user watches the AR/VR environment or 360° video using head mounted display devices. The existing platforms stream the full 360° scene while the user can view only a small part of the scene which spans about 90° - 120° horizontally, 90° vertically. Quantitatively, it is less than 20% of the whole scene. Thus, a significant amount of power and network bandwidth is wasted for the display which is never utilized in viewing. In ideal condition, the display will be only in the user's visual field while blurring the periphery. Following are the three important benefits of the user's visual field based rendering process: *1. Improved Image Quality:* It can enable 4k displays on the current generation graphics processing units (GPUs) without degradation in performance. *2. Lower cost:* Similarly, the end-users can run AR/VR and 360° Video Streaming based applications on low-cost hardware without compromising the performance. *3. Increased Frame Rate per Second (FPS):* The end-user can run at a higher frame rate using the same graphical settings. There are two types of foveated rendering: *dynamic foveated rendering* and *static foveated rendering*. Dynamic foveated rendering follows the user's gaze trajectory using eye-tracking and renders a sharp image in the required region, but this eye tracking is challenging in many scenarios. On the other hand, static foveated rendering considers a fixed area of the highest resolution at the center of the viewer's device irrespective of the user's gaze. It depends on the user's head movements, thus, facing a challenge in eye-head interplay as the image quality is drastically reduced if the user looks away from the center of the field of view. The main key aspects of accurate *eye position/visual attention* estimation ahead of time is to enhance user experience via providing high image quality in the subject's visual focus area. It requires person-specific calibration as nodal

Table 5: **Where we stand now.** Chronological comparison of the performance of different models for the gaze-related tasks on related benchmark datasets.

Task	Methods	Datasets							Year
		CAVE	MPIIGaze	EYEDIAP	UT MV	MPIIFace	Gaze360	ETH-X Gaze	
Gaze Estimation	GazeNet [24]	–	5.70°	7.13°	6.44°	5.76°	–	–	2015
	Dilated-Conv. [64]	–	4.39°	6.57°	–	4.42°	13.73°	–	2018
	Landmark based [97]	8.7°	8.3°	26.6°	–	–	–	–	2018
	RT-GENE [54]	–	4.61°	6.30°	–	4.66°	12.26	–	2019
	Pinball-LSTM [90]	9.0°	12.1°	5.58°	–	12.1°	11.04°	4.46°	2020
	CA-Net [59]	–	4.27°	5.63°	–	4.27°	11.20°	–	2021
	GazeTR-Hybrid [149]								



In the diagrams, the y-axis represents the angular error (in °) and the x-axis represents the timeline.

Task	Datasets	Methods (Eval.AI)						Year
Trajectory Estimation	Team-name →	Random_B	caixin	EyMazing	fgp200709d	vipl_gaze	Baseline	
	OpenEDS2020 [34]	3.078°	3.248°	3.313°	3.347°	3.386°	5.368°	2020
	Team-name →	XiaodongWang	Hebut_Lyx	tetelias	TCS_Research	Baseline	AnotherShot	
	OpenNEEDS [194]	1.68°	1.75°	1.99°	2.05°	7.18°	7.94°	2021

Task	Dataset	Methods							Year
Gaze	EmotiW2020 →	DD_Vision	SituAlgorithm	Overfit	DeepBlueAI	UDECE	X-AWARE	Baseline	
Zone	DGW [40]	82.52%	81.51%	78.87%	75.88%	74.57%	71.62%	60.98%	2020

Task	Datasets	Methods						Year
Visual Attention		Human	[114]	[62]	[113]	Center	Random	
	GazeFollowing [113] (AUC ↑)	0.924	0.921	0.896	0.878	0.633	0.504	2015
		ST-GNN	CNN+LSTM	CNN+SVM	CNN+RF	CNN	Chance	
	GazeCommunication [111]	55.02%	24.65%	36.23%	37.68%	23.05%	16.44%	
	1. Atomic-Level (Top-1%)	55.90%	–	–	–	–	22.70%	2019
	2. Event-Level (Top-1%)	Behavioural Agreement	CNN	RNN	LSTM	GRU	Scanpath	
	VisualSearch [118]	0.714	0.621	0.677	0.684	0.664	Direction	
	1. Microwave (MultiMatch)	0.701	0.633	0.673	0.669	0.659	Direction	2019
	2. Clock (MultiMatch)	ST-CNN +LSTM	ST-GNN	Gaze+Saliency +LSTM	Gaze+Saliency	GazeFollow	Random	
	SharedAttention [114]	83.3%	71.4%	66.2%	59.4%	58.7%	22.70%	2020

point of human have subject specific offset. Thus, generalizing it across user poses a challenge in the gaze analysis community to address [202]. On the other hand, user's viewpoint prediction ahead of time could face a lot of challenges as human eye movement is ballistic in nature. The visual attention of the user can therefore change abruptly based on the content in the screen. Thus, the prediction algorithm needs to take care of imperfect prediction as well and it needs to integrate with bit rate control. This process will enable user-specific recommendation and other facilities to enhance user experience [203].

6.2 Driver Engagement

With the progress in autonomous and smart cars, the requirement for automatic driver monitoring has been observed and researchers have been working on this problem for a few years now [40], [79], [81], [204]. In the literature, the problem is treated as a gaze zone estimation problem. A summary of the gaze estimation benchmarks is shown in Table 6. The proposed methods can be classified into two categories:

Sensor Based Tracking. These mainly utilize dedicated sensors integrated hardware devices for monitoring the driver's gaze in real-time. These devices require accurate pre-calibration and additionally these devices are expensive. Few examples of these sensors are Infrared (IR) camera [205], head-mounted devices [77], [206] and other systems [207], [208].

Table 6: Comparison of driver gaze estimation datasets with respect to number of subjects (# Sub), number of zones (# Zones), illumination conditions and labelling procedure.

References	# Sub	# Zones	Illumination	Labelling
Choi et al. [84]	4	8	Bright & Dim	3-D Gyro.
Lee et al. [85]	12	18	Day	Manual
Fridman et al. [82]	50	6	Day	Manual
Tawari et al. [79]	6	8	Day	Manual
Vora et al. [41]	10	7	Diff. day times	Manual
Jha et al. [77]	16	18	Day	Head-band
Wang et al. [209]	3	9	Day	Motion Sensor
DGW [40]	338	9	Diff. day times	Automatic
MGM [204]	60	21	Diff. day times	Multiple Sensors

Image processing and vision based methods. These are mainly focused on two types of methods: head-pose based only [80],

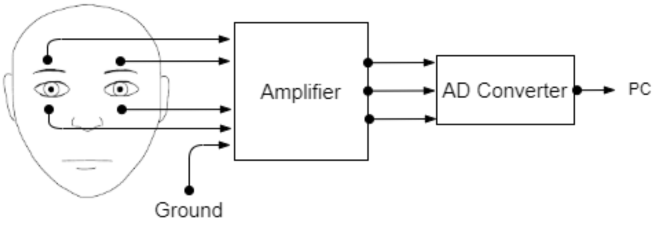


Figure 8: Electro-oculogram (EOG) based gaze estimation method [215]. This prototype opens the possibility of communication for severely disabled people. Refer Sec. 6.3 for more details.

[85], [209], [210] and both head-pose and eye-gaze based [79]–[84]. Driver's head pose provides partial information regarding his/her gaze direction as there may be an interplay between eyeball movement and head pose [83]. Hence, methods relying on head pose information may fail to disambiguate between the eye movement with fixed head-pose. Thus, the methods relying on both head pose and gaze based prediction are more robust.

6.3 Gaze in Healthcare and Wellbeing

Gaze is widely used in healthcare domain to enhance the diagnosis performance. Generally, eye movement patterns is widely used as behavioral bio-markers of various mental health problems including depression [211], post traumatic stress disorder [212] and Parkinson's disease [42]. Similarly, individuals diagnosed with Autism Spectral Disorder display gaze avoidance in social scenes [42]. Even intoxication including alcohol consumption and/or other drugs usage reflects on eye and gaze properties, especially, decreased accuracy and speed of saccades, changes in pupil size, and an impaired ability to fixate on moving objects. A recent survey [42] discusses the potential applications in healthcare including concussion [43], multiple sclerosis [213].

Physiological Signals. A gaze estimation system could be one of the communication methods for severely disabled people who cannot perform any type of gestures and speech. Sakurai et al. [214] developed an eye-tracking method using a compact and light electrooculogram (EOG) signal. Further, this prototype is improved via the usage of the EOG component which strongly correlated with the change of eye movements [215] (Refer Fig. 8). The setup can detect object scanning only by eye and face muscle movements. The experimental results open the possibility of eye-tracking via EOG signals and a Kinect sensor. Research along this direction can be extremely useful for disabled people.

7 Privacy in Gaze Estimation

Due to the rapid progress over the past few years, gaze estimation technologies have become more reliable, cheap, compact and observe increasing use in many fields, such as gaming, marketing, driver safety, and healthcare. Consequently, these expanding uses of technology raise serious privacy concerns. Gaze patterns can reveal much more information than a user wishes and expects to give away. By portraying the sensitivity of gaze tracking data, this section provides a brief overview of privacy concerns and consequent implications of gaze estimation and eye-tracking. Fig. 9 shows the overview of the privacy concerns, including common data capturing scenarios with their possible implications. A recent analysis [216] of the literature shows that eye-tracking data may implicitly contain information about a user's biometric identity [217], personal attributes (such as gender, age, ethnicity, personality traits, intoxication, emotional state, skills etc.) [218]–[220], physical and mental health [42], [211]. Few eye-tracking measures may even reveal underlying cognitive processes [17]. The widespread consideration of eye-tracking enhance the potential to improve our lives in many directions, but the technology can also pose a substantial threat to privacy. Thus, it is necessary to understand the sensitiveness of gaze data from a holistic perspective to prevent its misuse.

8 Conclusion and Future Direction

Gaze analysis is a technology in search of an application in several domains mainly in assistive technology and HCI. The wide applications of gaze related technology is growing rapidly. Thus, it opens a lot of research opportunity ahead of the community. Here, in this paper, we present an overall review of gaze analysis frameworks with different perspectives from different point of view. Beginning with the preliminaries of gaze modelling and eye movement, we further elaborate on challenges in this field, overview of gaze analysis framework and its possible applications in different domains. For eye analysis, mainly geometric and appearance properties are widely explored in prior works. Despite recent progress, the gaze analysis remains challenging due to eye head interplay, occlusion and other challenges mentioned in Sec. 2.4. Thus, there is a scope for future development in this respect. Moreover, all of the proposed datasets in this domain are collected in constraint environments. In order to overcome these limitations, the generative adversarial network based data generation approach has come into play. Due to several image quality-related issues, these datasets are not used for benchmarking. Automatic labelling of images based on accurate heuristic could be explored to reduce the data annotation burden greatly. Future directions for the eye and gaze trackers include:

Gaze Analysis in Unconstrained Setup: The most precise methods for gaze estimation is via intrusive sensors, IR camera and RGBD camera. The main drawback of these systems is that their performance degrades when used in real-world settings. In future, gaze estimation models should consider these situations. Although several current efforts in this direction employ techniques, yet further research is needed. Moreover, most of the current gaze estimation benchmark datasets require the proper geometric arrangement as well as user cooperation (e.g., CAVE, TabletGaze, MPII, Eyediap, ETH-XGaze etc). It would be an interesting direction to explore gaze estimation in a more flexible setting.

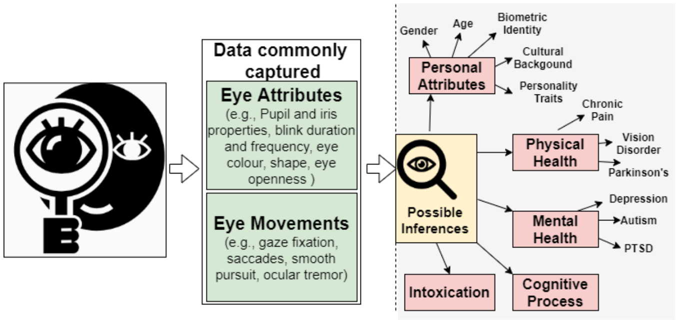


Figure 9: The possible privacy concerns related to gaze analysis framework [216]. Please refer Sec. 7 for more details.

Learning with Less Supervision: With the surge in unsupervised, self-supervised, weakly supervised techniques in this domain, more exploration in this direction is required to eliminate the dependency on ground truth gaze label which could be error-prone due to data acquisition limitations.

Gaze Inference: Apart from localizing the eye and determining gaze, the gaze patterns provides vital cues for encoding the cognitive and affective states of the concerned person. More exploration and cross-domain research could be another direction to encode visual perception.

AR/VR: Eye tracking has potential application in AR/VR including Foveated Rending (FR) and Attention Tunneling. The gaze based interaction require low latency gaze estimation. In these applications, the visual environment presents a high-quality image at the point where the user is looking while blurring the other peripheral region. The intuition is to reduce power consumption without compromising the perceptual quality as well as user experience. However, eye movements are fast and involuntary action which restrict the use of this techniques (in FR) due to the subsequent delays in the eye-tracking pipelines. In order to address this issue, a new research direction i.e. future gaze trajectory prediction has been recently introduced [34]. More exploration along this direction is highly desirable.

Eye Model and Learning Based Hybrid Approaches: Traditional geometrical eye model based and appearance guided learning based approaches have complimentary advantages. The geometrical eye model based methods does not require training data. Moreover, it has strong generalization capability but it is highly relied on relevant eye landmark localization performance. Accurate localization of eye landmarks is quite challenging in real world settings as the subject could have extreme headpose, occlusion, illumination and other environmental factors. On the other hand, the learning based approaches can encode eye appearance feature but it does not generalize well across different setups. Thus, a hybrid model which can take the advantage of both scenarios could be a possible research direction for gaze estimation and eye tracking domain.

Multi-modal/Cross-modal Gaze Estimation: Over the past decade, head gesture synthesis has become an interesting line of research. Prior works in this area have mainly used handcrafted audio features such as energy based features [221], MFCC (Mel Frequency Cepstral Coefficent) [222], LPC (Linear Predictive Coding) [222] and filter bank [222], [223] to generate realistic head gesture. The main challenge in this domain is audio data annotation for head motion synthesis which is a noisy and error prone process. Prior works approach this problem via multi-stream HMMs [221], MLP based regression modelling [222], bi-LSTM [223] and Conditional Variational Autoencoder (CVAE) [224]. In vision domain, mainly visual stimuli is utilized for gaze estimation. As the audio signal is non-trivial for gaze estimation, yet, it has the potential to coarsely define the gaze direction [225]. Research along this direction have potential to estimate gaze in challenging situation where visual stimuli fails.

The techniques surveyed in this paper focus on gaze analysis from different perspective, however, these techniques can be useful for other computer vision and HCI tasks. Gaze analysis and its widespread applications is a unique and well-defined topic, which have already influenced recent technologies. Scholarly interest in gaze estimation is established in a large number of disciplines. It primarily originates from vision-related assistive technology which further propagates through other domains and attracts a lot of future research attention across various fields.

References

[1] H. Liu and I. Heynderickx, "Visual attention in objective image quality assessment: Based on eye-tracking data," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 7, pp. 971–982, 2011.

[2] A. Frischen, A. P. Bayliss, and S. P. Tipper, "Gaze cueing of attention: visual attention, social cognition, and individual differences." *Psychological Bulletin*, vol. 133, p. 694, 2007.

[3] D. Purves, Y. Morgenstern, and W. T. Wojtach, "Perception and reality: why a wholly empirical paradigm is needed to understand vision," *Frontiers in systems neuroscience*, vol. 9, p. 156, 2015.

[4] E. Javal, "Essai sur la physiologie de la lecture," *Annales d'Ocilistique*, vol. 80, pp. 97–117, 1878.

[5] T. N. Cornsweet and H. D. Crane, "Accurate two-dimensional eye tracker using first and fourth purkinje images," *JOSA*, vol. 63, no. 8, pp. 921–928, 1973.

[6] J. Merchant, R. Morrissette, and J. L. Porterfield, "Remote measurement of eye direction allowing subject motion over one cubic foot of space," *IEEE transactions on biomedical engineering*, no. 4, pp. 309–317, 1974.

[7] M. Borgestig, J. Sandqvist, R. Parsons, T. Falkmer, and H. Hemmingsson, "Eye gaze performance for children with severe physical impairments using gaze-based assistive technology—a longitudinal study," *Assistive technology*, vol. 28, no. 2, pp. 93–102, 2016.

[8] F. Corno, L. Farinetti, and I. Signorile, "A cost-effective solution for eye-gaze assistive technology," in *IEEE International Conference on Multimedia and Expo*, vol. 2, 2002, pp. 433–436.

[9] A. W. Joseph and R. Murugesh, "Potential eye tracking metrics and indicators to measure cognitive load in human-computer interaction research," *Journal of Scientific Research*, vol. 64, no. 1, 2020.

[10] J. Pi, P. A. Koljonen, Y. Hu, and B. E. Shi, "Dynamic bayesian adjustment of dwell time for faster eye typing," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 10, pp. 2315–2324, 2020.

[11] Z. Chen, D. Deng, J. Pi, and B. E. Shi, "Unsupervised outlier detection in appearance-based gaze estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[12] M. Wedel and R. Pieters, "A review of eye-tracking research in marketing," in *Review of marketing research*. Routledge, 2017, pp. 123–147.

[13] A. Patney, J. Kim, M. Salvi, A. Kaplanyan, C. Wyman, N. Benty, A. Lefohn, and D. Luebke, "Perceptually-based foveated virtual reality," in *SIGGRAPH Emerging Technologies*. ACM, 2016, pp. 1–2.

[14] R. T. Azuma, "A survey of augmented reality," *Presence: Teleoperators & Virtual Environments*, vol. 6, no. 4, pp. 355–385, 1997.

[15] F. Ragusa, A. Furnari, S. Battiato, G. Signorello, and G. M. Farinella, "Ego-ch: Dataset and fundamental tasks for visitors behavioral understanding using egocentric vision," *Pattern Recognition Letters*, vol. 131, pp. 150–157, 2020.

[16] A. K. Jain, R. Bolle, and S. Pankanti, *Biometrics: personal identification in networked society*. Springer Science & Business Media, 2006, vol. 479.

[17] M. K. Eckstein, B. Guerra-Carrillo, A. T. M. Singley, and S. A. Bunge, "Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development?" *Developmental cognitive neuroscience*, vol. 25, pp. 69–91, 2017.

[18] M. A. Miller and M. T. Fillmore, "Persistence of attentional bias toward alcohol-related stimuli in intoxicated social drinkers," *Drug and Alcohol Dependence*, vol. 117, no. 2-3, pp. 184–189, 2011.

[19] Z. Zhu and Q. Ji, "Novel eye gaze tracking techniques under natural head movement," *IEEE Transactions on biomedical engineering*, vol. 54, no. 12, pp. 2246–2260, 2007.

[20] Y. Sugano, Y. Matsushita, and Y. Sato, "Learning-by-synthesis for appearance-based 3d gaze estimation," in *IEEE Computer Vision and Pattern Recognition*, 2014, pp. 1821–1828.

[21] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 3, pp. 478–500, 2009.

[22] D. Hansen and A. Pece, "Eye tracking in the wild," *Computer Vision and Image Understanding*, 2005.

[23] B. Smith, Q. Yin, S. Feiner, and S. Nayar, "Gaze locking: passive eye contact detection for human-object interaction," in *ACM User Interface Software & Technology*, 2013.

[24] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *IEEE Computer Vision and Pattern Recognition*, 2015, pp. 4511–4520.

[25] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in *IEEE Computer Vision and Pattern Recognition*, 2016, pp. 2176–2184.

[26] S. Park, A. Spurr, and O. Hilliges, "Deep pictorial gaze estimation," in *European Conference on Computer Vision*, 2018, pp. 721–738.

[27] N. Dubey, S. Ghosh, and A. Dhall, "Unsupervised learning of eye gaze representation from the web," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–7.

[28] Y. Yu and J. Odobez, "Unsupervised representation learning for gaze estimation," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–13, 2020.

[29] S. Park, S. D. Mello, P. Molchanov, U. Iqbal, O. Hilliges, and J. Kautz, "Few-shot adaptive gaze estimation," in *IEEE International Conference on Computer Vision*, 2019, pp. 9368–9377.

[30] E. B. Huey, "The psychology and pedagogy of reading: With a review of the history of reading and writing and of methods, texts, and hygiene in reading," 1908.

[31] Y. Sugano, Y. Matsushita, and Y. Sato, "Appearance-based gaze estimation using visual saliency," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 2, pp. 329–341, 2012.

[32] S. J. Garbin, Y. Shen, I. Schuetz, R. Cavin, G. Hughes, and S. S. Talathi, "Openeds: Open eye dataset," *arXiv preprint arXiv:1905.03702*, 2019.

[33] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, "Gaze360: Physically unconstrained gaze estimation in the wild," in *IEEE International Conference on Computer Vision*, 2019, pp. 6912–6921.

[34] C. Palmero, A. Sharma, K. Behrendt, K. Krishnakumar, O. V. Komogortsev, and S. S. Talathi, "Openeds2020: Open eyes dataset," *arXiv preprint arXiv:2005.03876*, 2020.

[35] H. Chennamma and X. Yuan, "A survey on eye-gaze tracking techniques," *arXiv preprint arXiv:1312.6410*, 2013.

[36] H. Jing-Yao, X. Yong-Yue, L. Lin-Na, X.-C. ZHANG, Q. Li, and C. Jian-Nan, "Survey on key technologies of eye gaze tracking," *DEStech Transactions on Computer Science and Engineering*, no. aicencs, 2016.

[37] A. Kar and P. Corcoran, "A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms," *IEEE Access*, vol. 5, pp. 16 495–16 519, 2017.

[38] D. Cazzato, M. Leo, C. Distante, and H. Voos, "When i look into your eyes: A survey on computer vision contributions for human gaze estimation and tracking," *Sensors*, vol. 20, no. 13, p. 3739, 2020.

[39] V. Clay, P. König, and S. U. König, "Eye tracking in virtual reality," *Journal of Eye Movement Research*, vol. 12, no. 1, 2019.

[40] S. Ghosh, A. Dhall, G. Sharma, S. Gupta, and N. Sebe, "Speak2label: Using domain knowledge for creating a large scale driver gaze zone estimation dataset," *arXiv preprint arXiv:2004.05973*, 2020.

[41] S. Vora, A. Rangesh, and M. M. Trivedi, "Driver gaze zone estimation using convolutional neural networks: A general framework and ablative analysis," *IEEE Transactions on Intelligent Vehicles*, pp. 254–265, 2018.

[42] K. Harezlak and P. Kasprowski, "Application of eye tracking in medicine: A survey, research issues and challenges," *Computerized Medical Imaging and Graphics*, vol. 65, pp. 176–190, 2018.

[43] Y. Kempinski, "System and method of diagnosis using gaze and eye tracking," Apr. 21 2016, uS Patent App. 14/723,590.

[44] S. Park, "Representation learning for webcam-based gaze estimation," Ph.D. dissertation, ETH Zurich, 2020.

[45] R. H. Carpenter, *Movements of the Eyes, 2nd Rev*. Pion Limited, 1988.

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

[47] A. T. Duchowski and A. T. Duchowski, *Eye tracking methodology: Theory and practice*. Springer, 2017.

[48] T. Santini, W. Fuhl, and E. Kasneci, "Calibme: Fast and unsupervised eye tracker calibration for gaze-based pervasive human-computer interaction," in *ACM Conference on Human Factors in Computing Systems*, 2017, pp. 2594–2605.

[49] X. Zhang, Y. Sugano, and A. Bulling, "Evaluation of appearance-based methods and implications for gaze-based applications," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–13.

[50] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Mpiigaze: Real-world dataset and deep appearance-based gaze estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[51] ——, "It's written all over your face: Full-face appearance-based gaze estimation," in *IEEE Computer Vision and Pattern Recognition Workshop*, 2017.

[52] W. Zhu and H. Deng, "Monocular free-head 3d gaze tracking with deep learning and geometry constraints," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3143–3152.

[53] W. Cui, J. Cui, and H. Zha, "Specialized gaze estimation for children by convolutional neural network and domain adaptation," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 3305–3309.

[54] T. Fischer, H. J. Chang, and Y. Demiris, "RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments," in *European Conference on Computer Vision*, 2018, pp. 339–357.

[55] Y. Cheng, F. Lu, and X. Zhang, "Appearance-based gaze estimation via evaluation-guided asymmetric regression," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 100–115.

[56] C. Palmero, J. Selva, M. A. Bagheri, and S. Escalera, "Recurrent cnn for 3d gaze estimation using appearance and shape cues," *arXiv preprint arXiv:1805.03064*, 2018.

[57] S. Jyoti and A. Dhall, "Automatic eye gaze estimation using geometric & texture-based networks," in *International Conference on Pattern Recognition*. IEEE, 2018, pp. 2474–2479.

[58] G. Liu, Y. Yu, K. A. F. Mora, and J.-M. Odobez, "A differential approach for gaze estimation," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[59] Y. Cheng, S. Huang, F. Wang, C. Qian, and F. Lu, "A coarse-to-fine adaptive network for appearance-based gaze estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10 623–10 630.

[60] X. Zhang, Y. Sugano, A. Bulling, and O. Hilliges, "Learning-based region selection for end-to-end gaze estimation," in *British Machine Vision Conference (BMVC 2020)*, 2020.

[61] Y. Yu, G. Liu, and J.-M. Odobez, "Deep multitask gaze estimation with a constrained landmark-gaze model," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[62] E. Chong, N. Ruiz, Y. Wang, Y. Zhang, A. Rozga, and J. M. Rehg, "Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 383–398.

[63] X. Zhang, M. X. Huang, Y. Sugano, and A. Bulling, "Training person-specific gaze estimators from user interactions with multiple devices," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–12.

[64] Z. Chen and B. E. Shi, "Appearance-based gaze estimation using dilated-convolutions," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 309–324.

[65] X. Zhou, J. Lin, J. Jiang, and S. Chen, "Learning a 3d gaze estimator with improved itracker combined with bidirectional lstm," in *2019 IEEE international conference on Multimedia and expo (ICME)*. IEEE, 2019, pp. 850–855.

[66] K. Wang, H. Su, and Q. Ji, "Neuro-inspired eye tracking with eye movement dynamics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9831–9840.

[67] Z. Wu, S. Rajendran, T. Van As, V. Badrinarayanan, and A. Rabinovich, "Eyenet: A multi-task deep network for off-axis eye gaze estimation," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 2019, pp. 3683–3687.

[68] J. Kim, M. Stengel, A. Majercik, S. De Mello, D. Dunn, S. Laine, M. McGuire, and D. Luebke, "Nvgaze: An anatomically-informed dataset for low-latency, near-eye gaze estimation," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.

[69] D. Xia and Z. Ruan, "IR image based eye gaze estimation," in *IEEE ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, vol. 1, 2007, pp. 220–224.

[70] M. Tonsen, J. Steil, Y. Sugano, and A. Bulling, "Invisibleeye: Mobile eye tracking using multiple low-resolution cameras and learning-based gaze estimation," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, pp. 1–21, 2017.

[71] D. Lian, Z. Zhang, W. Luo, L. Hu, M. Wu, Z. Li, J. Yu, and S. Gao, "Rgbd based gaze estimation via multi-task cnn," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 2488–2495.

[72] A. Tsukada, M. Shino, M. Devyver, and T. Kanade, "Illumination-free gaze estimation method for first-person vision wearable device," in *IEEE International Conference on Computer Vision Workshop*, 2011.

[73] M. Leo, D. Cazzato, T. De Marco, and C. Distante, "Unsupervised eye pupil localization through differential geometry and local self-similarity," *Public Library of Science*, vol. 9, no. 8, 2014.

[74] C. Gou, Y. Wu, K. Wang, K. Wang, F. Wang, and Q. Ji, "A joint cascaded framework for simultaneous eye detection and eye state estimation," *Pattern Recognition*, 2017.

[75] J. Pi and B. E. Shi, "Task-embedded online eye-tracker calibration for improving robustness to head motion," in *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, 2019, pp. 1–9.

[76] D. Lian, L. Hu, W. Luo, Y. Xu, L. Duan, J. Yu, and S. Gao, "Multiview multitask gaze estimation with deep convolutional neural networks," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 10, pp. 3010–3023, 2018.

[77] S. Jha and C. Busso, "Probabilistic estimation of the gaze region of the driver using dense classification," in *IEEE International Conference on Intelligent Transportation Systems*, 2018, pp. 697–702.

[78] Z. Hu, S. Li, C. Zhang, K. Yi, G. Wang, and D. Manocha, "Dgaze: Cnn-based gaze prediction in dynamic scenes," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 5, pp. 1902–1911, 2020.

[79] A. Tawari, K. H. Chen, and M. M. Trivedi, "Where is the driver looking: Analysis of head, eye and iris for robust gaze zone estimation," in *IEEE Conference on Intelligent Transportation Systems*, 2014, pp. 988–994.

[80] A. Tawari and M. M. Trivedi, "Robust and continuous estimation of driver gaze zone by dynamic analysis of multiple face videos," in *IEEE Intelligent Vehicles Symposium*, 2014, pp. 254–265.

[81] B. Vasli, S. Martin, and M. M. Trivedi, "On driver gaze estimation: Explorations and fusion of geometric and data driven approaches," in *IEEE Intelligent Transportation Systems*, 2016, pp. 655–660.

[82] L. Fridman, P. Langhans, J. Lee, and B. Reimer, "Driver gaze estimation without using eye movement," *IEEE Intelligent Systems*, pp. 49–56, 2015.

[83] L. Fridman, J. Lee, B. Reimer, and T. Victor, "'owl'and 'lizard': patterns of head pose and eye pose in driver gaze classification," *IET Computer Vision*, vol. 10, no. 4, pp. 308–314, 2016.

[84] I. H. Choi, S. K. Hong, and Y. G. Kim, "Real-time categorization of driver's gaze zone using the deep learning techniques," in *International Conference on Big Data and Smart Computing*. IEEE, 2016, pp. 143–148.

[85] S. Lee, J. Jo, H. Jung, K. Park, and J. Kim, "Real-time gaze estimator based on driver's head orientation for forward collision warning system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 1, pp. 254–267, 2011.

[86] J. He, K. Pham, N. Valliappan, P. Xu, C. Roberts, D. Lagun, and V. Navalpakkam, "On-device few-shot personalization for real-time gaze estimation," in *IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[87] Q. Huang, A. Veeraraghavan, and A. Sabharwal, "Tabletgaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets," *Machine Vision and Applications*, vol. 28, no. 5-6, pp. 445–461, 2017.

[88] T. Guo, Y. Liu, H. Zhang, X. Liu, Y. Kwak, B. In Yoo, J.-J. Han, and C. Choi, "A generalized and robust method towards practical gaze estimation on smart phone," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[89] Y. Bao, Y. Cheng, Y. Liu, and F. Lu, "Adaptive feature fusion network for gaze tracking in mobile tablets," *arXiv preprint arXiv:2103.11119*, 2021.

[90] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, "Gaze360: Physically unconstrained gaze estimation in the wild," in *IEEE International Conference on Computer Vision*, 2019.

[91] Y. Yu, G. Liu, and J. Odobez, "Improving few-shot user-specific gaze adaptation via gaze redirection synthesis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 937–11 946.

[92] K. Wang, R. Zhao, and Q. Ji, "A hierarchical generative model for eye image synthesis and eye gaze estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 440–448.

[93] S. Vora, A. Rangesh, and M. M. Trivedi, "On generalizing driver gaze zone estimation using convolutional neural networks," in *IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2017, pp. 849–854.

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

[95] X. Zhang, Y. Sugano, and A. Bulling, "Everyday eye contact detection using unsupervised gaze target discovery," in *ACM User Interface Software and Technology*, 2017, pp. 193–203.

[96] C. H. Morimoto and M. R. Mimica, "Eye gaze tracking techniques for interactive applications," *Computer vision and image understanding*, vol. 98, no. 1, pp. 4–24, 2005.

[97] S. Park, X. Zhang, A. Bulling, and O. Hilliges, "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*, 2018, pp. 1–10.

[98] P. M. Corcoran, F. Nanu, S. Petrescu, and P. Bigioi, "Real-time eye gaze tracking for gaming design and consumer electronics systems," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 2, pp. 347–355, 2012.

[99] S. Chu, N. Paul, and L. Ruel, "Using eye tracking technology to examine the effectiveness of design elements on news websites." *Information Design Journal (IDJ)*, vol. 17, no. 1, 2009.

[100] J. Chen, J. Zhang, J. Fan, T. Chen, E. Sangineto, and N. Sebe, "Mggr: Multimodal-guided gaze redirection with coarse-to-fine learning," *arXiv preprint arXiv:2004.03064*, 2020.

[101] Y. Zheng, S. Park, X. Zhang, S. De Mello, and O. Hilliges, "Self-learning transformations for improving gaze and head redirection," *arXiv preprint arXiv:2010.12307*, 2020.

[102] J. Chen, J. Zhang, E. Sangineto, T. Chen, J. Fan, and N. Sebe, "Coarse-to-fine gaze redirection with numerical and pictorial guidance," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3665–3674.

[103] K. Ruhland, C. E. Peters, S. Andrist, J. B. Badler, N. I. Badler, M. Gleicher, B. Mutlu, and R. McDonnell, "A review of eye gaze in virtual agents, social robotics and hci: Behaviour generation, user interaction and perception," in *Computer graphics forum*, vol. 34, no. 6. Wiley Online Library, 2015, pp. 299–326.

[104] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *IEEE Winter Conference on Applications of Computer Vision*, 2016, pp. 1–10.

[105] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 59–66.

[106] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, pp. 1755–1758, 2009.

[107] P. Müller, M. X. Huang, X. Zhang, and A. Bulling, "Robust eye contact detection in natural multi-person interactions using gaze and speaking behaviour," in *Proc. ACM International Symposium on Eye Tracking Research and Applications (ETRA)*, 2018, pp. 31:1–31:10.

[108] M. J. Marin-Jimenez, V. Kalogeiton, P. Medina-Suarez, and A. Zisserman, "Laeo-net: revisiting people looking at each other in videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3477–3485.

[109] ——, "LAEO-Net++: revisiting people Looking At Each Other in videos," in *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[110] R. Kothari, S. De Mello, U. Iqbal, W. Byeon, S. Park, and J. Kautz, "Weakly-supervised physically unconstrained gaze estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 9980–9989.

[111] L. Fan, W. Wang, S. Huang, X. Tang, and S.-C. Zhu, "Understanding human gaze communication by spatio-temporal graph reasoning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5724–5733.

[112] M. Zhang, Y. Liu, and F. Lu, "Gazeonce: Real-time multi-person gaze estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 4197–4206.

[113] A. Recasens, A. Khosla, C. Vondrick, and A. Torralba, "Where are they looking?" in *Advances in Neural Information Processing Systems*, vol. 28. Curran Associates, Inc., 2015.

[114] E. Chong, Y. Wang, N. Ruiz, and J. M. Rehg, "Detecting attended visual targets in video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5396–5406.

[115] D. Tu, X. Min, H. Duan, G. Guo, G. Zhai, and W. Shen, "End-to-end human-gaze-target detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 2202–2210.

[116] B. Wang, T. Hu, B. Li, X. Chen, and Z. Zhang, "Gatector: A unified framework for gaze object prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 19 588–19 597.

[117] Z. Yang, L. Huang, Y. Chen, Z. Wei, S. Ahn, G. Zelinsky, D. Samaras, and M. Hoai, "Predicting goal-directed human attention using inverse reinforcement learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 193–202.

[118] G. Zelinsky, Z. Yang, L. Huang, Y. Chen, S. Ahn, Z. Wei, H. Adeli, D. Samaras, and M. Hoai, "Benchmarking gaze prediction for categorical visual search," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[119] X. Chen, M. Jiang, and Q. Zhao, "Predicting human scanpaths in visual question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 876–10 885.

[120] M. Kümmerer and M. Bethge, "State-of-the-art in human scanpath prediction," *arXiv preprint arXiv:2102.12239*, 2021.

[121] J. Bao, B. Liu, and J. Yu, "Escnet: Gaze target detection with the understanding of 3d scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 14 126–14 135.

[122] K. Liang, Y. Chahir, M. Molina, C. Tijus, and F. Jouen, "Appearance-based gaze tracking with spectral clustering and semi-supervised gaussian process regression," in *Proceedings of the 2013 Conference on Eye Tracking South Africa*, 2013, pp. 17–23.

[123] K. Wang and Q. Ji, "Hybrid model and appearance based eye tracking with kinect," in *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, 2016, pp. 331–332.

[124] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[125] J. Li, Y. Wang, C. Wang, Y. Tai, J. Qian, J. Yang, C. Wang, J. Li, and F. Huang, "Dsfd: Dual shot face detector," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[126] J. Wang, Y. Liu, Y. Hu, H. Shi, and T. Mei, "Facex-zoo: A pytorh toolbox for face recognition," *arXiv preprint arXiv:2101.04407*, 2021.

[127] N. Markuš, M. Frljak, I. S. Pandžić, J. Ahlberg, and R. Forchheimer, "Eye pupil localization with an ensemble of randomized trees," *Pattern Recognition*, 2014.

[128] D. Tian, G. He, J. Wu, H. Chen, and Y. Jiang, "An accurate eye pupil localization approach based on adaptive gradient boosting decision tree," in *2016 Visual Communications and Image Processing (VCIP)*. IEEE, 2016, pp. 1–4.

[129] A. Kacete, J. Royan, R. Seguier, M. Collobert, and C. Soladie, "Real-time eye pupil localization using hough regression forest," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–8.

[130] J. H. Choi, K. I. Lee, Y. C. Kim, and B. C. Song, "Accurate eye pupil localization using heterogeneous cnn models," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 2179–2183.

[131] W. Sankowski, K. Grabowski, M. Napieralska, M. Zubert, and A. Napieralski, "Reliable algorithm for iris segmentation in eye image," *Image and vision computing*, vol. 28, no. 2, pp. 231–237, 2010.

[132] P. Radu, J. Ferryman, and P. Wild, "A robust sclera segmentation algorithm," in *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2015, pp. 1–6.

[133] A. Das, U. Pal, M. A. Ferrer, M. Blumenstein, D. Štepec, P. Rot, Ž. Emeršič, P. Peer, V. Štruc, S. A. Kumar *et al.*, "Sserbc 2017: Sclera segmentation and eye recognition benchmarking competition," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2017, pp. 742–747.

[134] D. R. Lucio, R. Laroca, E. Severo, A. S. Britto, and D. Menotti, "Fully convolutional networks and generative adversarial networks applied to sclera segmentation," in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2018, pp. 1–7.

[135] A. Das, U. Pal, M. Blumenstein, and M. A. F. Ballester, "Sclera recognition-a survey," in *2013 2nd IAPR Asian Conference on Pattern Recognition*. IEEE, 2013, pp. 917–921.

[136] A. K. Chaudhary, R. Kothari, M. Acharya, S. Dangi, N. Nair, R. Bailey, C. Kanan, G. Diaz, and J. B. Pelz, "Ritnet: real-time semantic segmentation of the eye for gaze tracking," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 2019, pp. 3698–3702.

[137] N. N. Pandey and N. B. Muppalaneni, "Real-time drowsiness identification based on eye state analysis," in *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*. IEEE, 2021, pp. 1182–1187.

[138] M. Monaro, P. Capuozzo, F. Ragucci, A. Maffei, A. Curci, C. Scarpazza, A. Angrilli, and G. Sartori, "Using blink rate to detect deception: A study to validate an automatic blink detector and a new dataset of videos from liars and truth-tellers," in *International Conference on Human-Computer Interaction*. Springer, 2020, pp. 494–509.

[139] T. Drutarovsky and A. Fogelton, "Eye blink detection using variance of motion vectors," in *European Conference on Computer Vision*. Springer, 2014, pp. 436–448.

[140] A. Królak and P. Strumiłło, "Eye-blink detection system for human–computer interaction," *Universal Access in the Information Society*, vol. 11, no. 4, pp. 409–419, 2012.

[141] J. Cech and T. Soukupova, "Real-time eye blink detection using facial landmarks," *Cent. Mach. Perception, Dep. Cybern. Fac. Electr. Eng. Czech Tech. Univ. Prague*, pp. 1–8, 2016.

[142] R. Daza, A. Morales, J. Fierrez, and R. Tolosana, "mebal: A multimodal database for eye blink detection and attention level estimation," in *Companion Publication of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 32–36.

[143] G. Hu, Y. Xiao, Z. Cao, L. Meng, Z. Fang, J. T. Zhou, and J. Yuan, "Towards real-time eyeblink detection in the wild: Dataset, theory and practices," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2194–2208, 2019.

[144] J. Perry and A. Fernandez, "Minenet: A dilated cnn for semantic segmentation of eye features," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[145] K. Wang, R. Zhao, H. Su, and Q. Ji, "Generalizing eye tracking with bayesian adversarial learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 907–11 916.

[146] C. Palmero Cantarino, O. V. Komogortsev, and S. S. Talathi, "Benefits of temporal information for appearance-based gaze estimation," in *ACM Symposium on Eye Tracking Research and Applications*, 2020, pp. 1–5.

[147] S. Nonaka, S. Nobuhara, and K. Nishino, "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)*, June 2022, pp. 2192–2201.

[148] S. Park, E. Aksan, X. Zhang, and O. Hilliges, "Towards end-to-end video-based eye-tracking," in *European Conference on Computer Vision (ECCV)*, 2020.

[149] Y. Cheng and F. Lu, "Gaze estimation using transformer," *arXiv preprint arXiv:2105.14424*, 2021.

[150] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2107–2116.

[151] S.-H. Kim, G.-S. Lee, H.-J. Yang *et al.*, "Eye semantic segmentation with a lightweight model," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 2019, pp. 3694–3697.

[152] P. Rot, Ž. Emeršič, V. Struc, and P. Peer, "Deep multi-class eye segmentation for ocular biometrics," in *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*. IEEE, 2018, pp. 1–8.

[153] B. Luo, J. Shen, Y. Wang, and M. Pantic, "The ibug eye segmentation dataset," in *2018 Imperial College Computing Student Workshop (ICCSW 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

[154] A. Das, U. Pal, M. A. Ferrer, and M. Blumenstein, "Ssrbc 2016: Sclera segmentation and recognition benchmarking competition," in *2016 International Conference on Biometrics (ICB)*. IEEE, 2016, pp. 1–6.

[155] A. Das, U. Pal, M. Blumenstein, C. Wang, Y. He, Y. Zhu, and Z. Sun, "Sclera segmentation benchmarking competition in cross-resolution environment," in *2019 International Conference on Biometrics (ICB)*. IEEE, 2019, pp. 1–7.

[156] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, "Eye-mms: Miniature multi-scale segmentation network of key eye-regions in embedded applications," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[157] "OKAO vision, howpublished = http://www.omron.com/r_d/coretech/vision/okao.html."

[158] K. A. Funes Mora, F. Monay, and J.-M. Odobez, "Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras," in *Proceedings of the Symposium on Eye Tracking Research and Applications*, 2014, pp. 255–258.

[159] E. Wood, T. Baltrusaitis, X. Zhang, Y. Sugano, P. Robinson, and A. Bulling, "Rendering of eyes for eye-shape registration and gaze estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3756–3764.

[160] Y. Ganin, D. Kononenko, D. Sungatullina, and V. Lempitsky, "Deepwarp: Photorealistic image resynthesis for gaze manipulation," in *European conference on computer vision*. Springer, 2016, pp. 311–326.

[161] "Yu, S.:Harr feature cart-tree based cascade eye detector homepage., howpublished = http://yushiqi.cn/research/eyedetection."

[162] J. Xiang and G. Zhu, "Joint face detection and facial expression recognition with mtcnn," in *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*. IEEE, 2017, pp. 424–427.

[163] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[164] Y. Xiong, H. J. Kim, and V. Singh, "Mixed effects neural networks (menets) with applications to gaze estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7743–7752.

[165] P. Kansal and S. Devanathan, "Eyenet: Attention based convolutional encoder-decoder network for eye region segmentation," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 2019, pp. 3688–3693.

[166] M. Bühler, S. Park, S. De Mello, X. Zhang, and O. Hilliges, "Content-consistent generation of realistic eyes with style," *arXiv preprint arXiv:1911.03346*, 2019.

[167] Y. Zhu, Y. Yan, and O. Komogortsev, "Hierarchical hmm for eye movement classification," in *European Conference on Computer Vision*. Springer, 2020, pp. 544–554.

[168] O. V. Komogortsev and A. Karpov, "Automated classification and scoring of smooth pursuit eye movements in the presence of fixations and saccades," *Behavior research methods*, vol. 45, no. 1, pp. 203–215, 2013.

[169] Y. Shen, O. Komogortsev, and S. S. Talathi, "Domain adaptation for eye segmentation," in *European Conference on Computer Vision*. Springer, 2020, pp. 555–569.

[170] J. Perry and A. S. Fernandez, "Eyeseg: Fast and efficient few-shot semantic segmentation," in *European Conference on Computer Vision*. Springer, 2020, pp. 570–582.

[171] P. A. Dias, D. Malafronte, H. Medeiros, and F. Odone, "Gaze estimation for assisted living environments," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 290–299.

[172] Y. Cheng, X. Zhang, F. Lu, and Y. Sato, "Gaze estimation by exploring two-eye asymmetry," *IEEE Transactions on Image Processing*, vol. 29, pp. 5259–5272, 2020.

[173] Y. Cheng, Y. Bao, and F. Lu, "Puregaze: Purifying gaze feature for generalizable gaze estimation," *arXiv preprint arXiv:2103.13173*, 2021.

[174] S. Ghosh, M. Hayat, A. Dhall, and J. Knibbe, "Mtgls: Multi-task gaze estimation with limited supervision," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 3223–3234.

[175] Z. Chen and B. Shi, "Offset calibration for appearance-based gaze estimation via gaze decomposition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 270–279.

[176] B. Luo, J. Shen, S. Cheng, Y. Wang, and M. Pantic, "Shape constrained network for eye segmentation in the wild," in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 1952–1960.

[177] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[178] W. Wang, J. Shen, X. Dong, A. Borji, and R. Yang, "Inferring salient objects from human fixations," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 8, pp. 1913–1927, 2019.

[179] R. Kothari, Z. Yang, C. Kanan, R. Bailey, J. B. Pelz, and G. J. Diaz, "Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities," *Scientific Reports*, vol. 10, no. 1, pp. 1–18, 2020.

[180] A. Farkhondeh, C. Palmero, S. Scardapane, and S. Escalera, "Towards self-supervised gaze estimation," *arXiv preprint arXiv:2203.10974*, 2022.

[181] Z. He, A. Spurr, X. Zhang, and O. Hilliges, "Photo-realistic monocular gaze redirection using generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6932–6941.

[182] E. Wood, T. Baltrušaitis, L.-P. Morency, P. Robinson, and A. Bulling, "Gazedirector: Fully articulated eye gaze redirection in video," in *Computer Graphics Forum*, vol. 37, no. 2. Wiley Online Library, 2018, pp. 217–225.

[183] H. Kaur and R. Manduchi, "Subject guided eye image synthesis with application to gaze redirection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 11–20.

[184] K. A. Funes Mora, F. Monay, and J.-M. Odobez, "Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras," in *ACM Symposium on Eye Tracking Research and Applications*, 2014.

[185] R. Derakhshani, A. Ross, and S. Crihalmeanu, "A new biometric modality based on conjunctival vasculature," in *Proceedings of Artificial Neural Networks in Engineering*, 2006, pp. 1–8.

[186] R. Derakhshani and A. Ross, "A texture-based neural network classifier for biometric identification using ocular surface vasculature," in *2007 International Joint Conference on Neural Networks*. IEEE, 2007, pp. 2982–2987.

[187] S. Crihalmeanu, A. Ross, and R. Derakhshani, "Enhancement and registration schemes for matching conjunctival vasculature," in *International Conference on Biometrics*. Springer, 2009, pp. 1240–1249.

[188] Q. He, X. Hong, X. Chai, J. Holappa, G. Zhao, X. Chen, and M. Pietikäinen, "Omeg: Oulu multi-pose eye gaze dataset," in *Scandinavian Conference on Image Analysis*. Springer, 2015, pp. 418–427.

[189] A. Recasens*, A. Khosla*, C. Vondrick, and A. Torralba, "Where are they looking?" in *Advances in Neural Information Processing Systems (NIPS)*, 2015, * indicates equal contribution.

[190] E. Wood, T. Baltrušaitis, L.-P. Morency, P. Robinson, and A. Bulling, "Learning an appearance-based gaze estimator from one million synthesised images," in *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, 2016, pp. 131–138.

[191] K. Cortacero, T. Fischer, and Y. Demiris, "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 Workshops*, 2019, pp. 0–0.

[192] S. Laine, T. Karras, T. Aila, A. Herva, S. Saito, R. Yu, H. Li, and J. Lehtinen, "Production-level facial performance capture using deep convolutional neural networks," in *Proceedings of the ACM SIGGRAPH/Eurographics symposium on computer animation*, 2017, pp. 1–10.

[193] H. Tomas, M. Reyes, R. Dionido, M. Ty, J. Mirando, J. Casimiro, R. Atienza, and R. Guinto, "Goo: A dataset for gaze object prediction in retail environments," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3125–3133.

[194] K. J. Emery, M. Zannoli, J. Warren, L. Xiao, and S. S. Talathi, "Openneeds: A dataset of gaze, head, hand, and scene signals during exploration in open-ended vr environments," in *ACM Symposium on Eye Tracking Research and Applications*, 2021, pp. 1–7.

[195] L. Wolf, Z. Freund, and S. Avidan, "An eye for an eye: A single camera gaze-replacement method," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 817–824.

[196] D. Kononenko and V. Lempitsky, "Learning to look up: Realtime monocular gaze correction using machine learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4667–4675.

[197] M. Sela, P. Xu, J. He, V. Navalpakkam, and D. Lagun, "Gazegan-unpaired adversarial image generation for gaze estimation," *arXiv preprint arXiv:1711.09767*, 2017.

[198] B. Benfold and I. Reid, "Unsupervised learning of a scene-specific coarse gaze estimator," in *IEEE International Conference on Computer Vision*, 2011, pp. 2344–2351.

[199] T. Li, Q. Liu, and X. Zhou, "Ultra-low-power gaze tracking for virtual reality," *GetMobile: Mobile Comp. and Comm.*, vol. 22, no. 3, p. 27–31, Jan. 2019. [Online]. Available: https://doi.org/10.1145/3308755.3308765

[200] A. Patney, M. Salvi, J. Kim, A. Kaplanyan, C. Wyman, N. Benty, D. Luebke, and A. Lefohn, "Towards foveated rendering for gaze-tracked virtual reality," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, pp. 1–12, 2016.

[201] S. Park, A. Bhattacharya, Z. Yang, S. R. Das, and D. Samaras, "Mosaic: Advancing user quality of experience in 360-degree video streaming with machine learning," *IEEE Transactions on Network and Service Management*, vol. 18, no. 1, pp. 1000–1015, 2021.

[202] D. Alexandrovsky, S. Putze, M. Bonfert, S. Höffner, P. Michelmann, D. Wenig, R. Malaka, and J. D. Smeddinck, "Examining design choices of questionnaires in vr user studies," ser. CHI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–21. [Online]. Available: https://doi.org/10.1145/3313831.3376260

[203] G.-A. Koulieris, K. Akşit, C. Richardt, and R. Mantiuk, "Cutting-edge vr/ar display technologies (gaze-, accommodation-, motion-aware and hdr-enabled)," in *SIGGRAPH Asia 2018 Courses*, ser. SA '18. New York, NY, USA: Association for Computing Machinery, 2018. [Online]. Available: https://doi.org/10.1145/3277644.3277771

[204] S. Jha, M. F. Marzban, T. Hu, M. H. Mahmoud, and N. A.-D. C. Busso, "The multimodal driver monitoring database: A naturalistic corpus to study driver attention," *arXiv preprint arXiv:2101.04639*, 2020.

[205] M. W. Johns, A. Tucker, R. Chapman, K. Crowley, and N. Michael, "Monitoring eye and eyelid movements by infrared reflectance oculography to measure drowsiness in drivers," *Somnologie-Schlafforschung und Schlafmedizin*, vol. 11, no. 4, pp. 234–242, 2007.

[206] S. Jha and C. Busso, "Challenges in head pose estimation of drivers in naturalistic recordings using existing tools," in *IEEE International Conference on Intelligent Transportation Systems*, 2017, pp. 1–6.

[207] Y. Feng, G. Cheung, W.-t. Tan, P. Le Callet, and Y. Ji, "Low-cost eye gaze prediction system for interactive networked video streaming," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1865–1879, 2013.

[208] C. Zhang, Q. He, J. Liu, and Z. Wang, "Exploring viewer gazing patterns for touch-based mobile gamecasting," *IEEE Transactions on Multimedia*, vol. 19, no. 10, pp. 2333–2344, 2017.

[209] Y. Wang, G. Yuan, Z. Mi, J. Peng, X. Ding, Z. Liang, and X. Fu, "Continuous driver's gaze zone estimation using rgb-d camera," *Sensors*, p. 1287, 2019.

[210] S. S. Mukherjee and N. M. Robertson, "Deep head pose: Gaze-direction estimation in multimodal video," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2094–2107, 2015.

[211] S. Alghowinem, R. Goecke, M. Wagner, G. Parker, and M. Breakspear, "Eye movement analysis for depression detection," in *2013 IEEE International Conference on Image Processing*. IEEE, 2013, pp. 4220–4224.

[212] M. E. Milanak, M. R. Judah, H. Berenbaum, A. F. Kramer, and M. Neider, "Ptsd symptoms and overt attention to contextualized emotional faces: Evidence from eye tracking," *Psychiatry research*, vol. 269, pp. 408–413, 2018.

[213] O. Avital, "Method and system of using eye tracking to evaluate subjects," Oct. 8 2015, uS Patent App. 14/681,083.

[214] K. Sakurai, M. Yan, H. Tamura, and K. Tanno, "A study on gaze estimation system using the direction of eyes and face," in *2016 World Automation Congress (WAC)*. IEEE, 2016, pp. 1–6.

[215] K. Sakurai, M. Yan, K. Tanno, and H. Tamura, "Gaze estimation method using analysis of electrooculogram signals and kinect sensor," *Computational intelligence and neuroscience*, vol. 2017, 2017.

[216] J. L. Kröger, O. H.-M. Lutz, and F. Müller, "What does your gaze reveal about you? on the privacy implications of eye tracking," in *IFIP International Summer School on Privacy and Identity Management*. Springer, 2019, pp. 226–241.

[217] B. John, S. Koppal, and E. Jain, "Eyeveil: degrading iris authentication in eye tracking headsets," in *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, 2019, pp. 1–5.

[218] M. Erbilek, M. Fairhurst, and M. C. D. C. Abreu, "Age prediction from iris biometrics," in *5th International Conference on Imaging for Crime Detection and Prevention (ICDP 2013)*. IET, 2013, pp. 1–5.

[219] F. J. M. Moss, R. Baddeley, and N. Canagarajah, "Eye movements to natural images as a function of sex and personality," *PloS one*, vol. 7, no. 11, p. e47870, 2012.

[220] S. Hoppe, T. Loetscher, S. A. Morey, and A. Bulling, "Eye movements during everyday behavior predict personality traits," *Frontiers in human neuroscience*, vol. 12, p. 105, 2018.

[221] A. Ben Youssef, H. Shimodaira, and D. A. Braude, "Articulatory features for speech-driven head motion synthesis," *Proceedings of Interspeech, Lyon, France*, vol. 3, 2013.

[222] C. Ding, L. Xie, and P. Zhu, "Head motion synthesis from speech using deep neural networks," *Multimedia Tools and Applications*, vol. 74, no. 22, pp. 9871–9888, 2015.

[223] C. Ding, P. Zhu, and L. Xie, "Blstm neural networks for speech driven head motion synthesis," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[224] D. Greenwood, S. Laycock, and I. Matthews, "Predicting head pose from speech with a conditional variational autoencoder." ISCA, 2017.

[225] S. Ghosh, A. Dhall, M. Hayat, and J. Knibbe, "Av-gaze: A study on the effectiveness of audio guided visual attention estimation for non-profilic faces," *arXiv preprint arXiv:2207.03048*, 2022.

Shreya Ghosh is currently pursuing her PhD at Monash University, Australia. She received MS(R) degree in Computer Science and Engineering from the Indian Institute of Technology Ropar, India. She received the B.Tech. in CSE from the Govt. College of Engineering and Textile Technology (Serampore, India). Her research interests include affective computing, computer vision, Deep Learning. She is a student member of the IEEE.

Abhinav Dhall is an Assistant Professor at Indian Institute of Technology Ropar and Adjunct Senior Lecturer at Monash University . He received PhD from the Australian National University in 2014. Followed by postdocs at the University of Waterloo and the University of Canberra. He was awarded the Best Doctoral Paper Award at ACM ICMR 2013, Best Student Paper Honourable mention at IEEE AFGR 2013 and Best Paper Nomination at IEEE ICME 2012. His research interests are in computer vision for Affective computing and Assistive Technology. He is a member of the IEEE and Associate Editor of IEEE Transactions on Affective Computing.

Munawar Hayat is currently a Senior Research Fellow with Monash University, Australia. He received his PhD from The University of Western Australia (UWA). His PhD thesis received multiple awards, including the Deans List Honorable Mention Award and the Robert Street Prize. After his PhD, he joined IBM Research as a postdoc and then moved to the University of Canberra as an Assistant Professor. He was a Senior Scientist at the Inception Institute of Artificial Intelligence, UAE. He has been awarded the ARC DECRA fellowship. His research interests are in computer vision, machine learning, deep learning, and affective computing.

Jarrod Knibbe is currently with the University of Melbourne, Australia. He received his PhD from The University of Bristol in 2016. He completed a post-doc in human-centred computing at the University of Copenhagen. His research interests include interaction design and user experience, with a focus on body based user interfaces, electric muscle stimulation, and virtual reality. He has published over 25 papers at top venues in Human-Computer Interaction, including CHI, UIST, CSCW, and Ubicomp.

Qiang Ji is a Professor with the Department of Electrical, Computer, and Systems Engineering at Rensselaer Polytechnic Institute (RPI). He received his Ph.D degree in Electrical Engineering from the University of Washington. He was a program director at the National Science Foundation, where he managed NSF's computer vision and machine learning programs. He also held teaching and research positions at University of Illinois at Urbana-Champaign, Carnegie Mellon University, and University of Nevada at Reno. His research interests are in computer vision, probabilistic machine learning, and their applications. He has published over 300 papers, received multiple awards for his work, serve as an editor for multiple international journals, and organize numerous international conferences/workshops. He is a fellow of the IEEE and the IAPR.