Automatic Gaze Analysis: A Survey of Deep Learning Based Approaches

Shreya Ghosh Ø, Abhinav Dhall Ø, Member, IEEE, Munawar Hayat Ø, and Jarrod Knibbe Ø, Fellow, IEEE

(Survey Paper)

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).

Index Terms—Automated gaze estimation, gaze analysis, gaze tracking, human computer interaction, unsupervised and self-supervised gaze analysis.

I. 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].

Manuscript received 19 September 2021; revised 4 August 2023; accepted 14 September 2023. Date of current version 5 December 2023. Recommended for acceptance by D. Samaras. (Corresponding author: Abhinav Dhall.)

Shreya Ghosh is with Curtin University, Bentley, WA 6102, Australia (e-mail: shreya.ghosh@curtin.edu.au).

Abhinav Dhall is with Monash University, Clayton 3800, Australia, and also with the Indian Institute of Technology Ropar, Bara Phool, Punjab 140001, India (e-mail: abhinav.dhall@monash.edu).

Munawar Hayat is with Monash University, Clayton 3800, Australia (e-mail: munawar.hayat@monash.edu).

Jarrod Knibbe is with the University of Melbourne, Parkville, VIC 3052, Australia (e-mail: jarrod.knibbe@unimelb.edu.au).

Qiang Ji is with the Rensselaer Polytechnic Institute, Troy, NY 12180 USA (e-mail: jqj@rpi.edu).

Project Page: [https://github.com/i-am-shreya/EyeGazeSurvey]

Digital Object Identifier 10.1109/TPAMI.2023.3321337

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 70 s, 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], [10], [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], [25], [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], [28], [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 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 multi-person 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.

II. 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 (Section II-A), Eye movements (Section II-B), Problem Settings in Automated Gaze Analysis (Section II-C) and the associated Challenges (Section II-D).

A. 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 regions 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 12-13 mm. 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. 2 right).
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 head pose [21]. Most of the time, the combined direction of LoS and head pose provide information about where the person is looking.

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

**Smooth Pursuits.** Smooth pursuit occurs 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.

### 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 head-pose 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.

**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 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 (∼ 60 cm), mobile phone (∼ 20 cm), car (∼ 50 cm) etc. An overview is presented in Table 1.

**D. 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, ColumbiaGaze [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 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.

TABLE I

| Platform | Dist. | VA | UC | Papers |
|----------|-------|----|----|--------|
| Desktop, TV Panels | 30-50, 200-500 | ~40°-60° | Static, Sitting, Upright | [24], [26], [49]–[57] |
| 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] |

Here, Dist.: distance (in cm), VA: viewing Angle (in degrees), HMD: Head Mounted Devices, FV: Free Viewing, UC: User Condition, ET: External Target.

III. 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 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], [101], [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], [109], [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 sub-tasks in this direction are gaze following [62], [113], [114], [115], [116], gaze communication [111], human attention in goal-driven environments [117] and categorical visual search [118], visual scan-path 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.

IV. 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 (Section IV-A), Network Architecture (Section IV-B) and Level of Supervision (Section IV-C).

A. 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], MT-CNN [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.

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], [132], [133], [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]. Among 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.

B. Representative Deep Network Architectures

In this section, we provide a generic formulation and representation of gaze analysis. Given an RGB image \( I \in \mathbb{R}^{W \times H \times 3} \), 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.

1) CNN Based: Most of the recent solutions adopt a CNN based architecture [24], [25], [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], 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) [23], [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 \( I \in \mathbb{R}^{W \times H} \) as input and maps it to angular gaze vector \( g \in \mathbb{R}^{2} \). 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 \( \ell_2 \) losses between the predicted \( \hat{g} \) and actual gaze angle vectors \( g \) 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 \( I \in \mathbb{R}^{W \times H \times 3} \) 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 \( 1 \times 1 \) 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: \( W \odot U_{ce} \), 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 \( \ell_1 \) distance between the predicted and ground-truth gaze positions in the target screen coordinate system is utilized. Similarly, the \( \ell_1 \) 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:

\[
v(x, y) = \sum_{k=1}^{K} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} u(x + mr_1, y + nr_2, k)w_{nmk} + b
\]

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 \( \hat{g} \) (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:

\[
u_i = \frac{m}{2} - r'i \sin \phi \cos \theta, v_i = \frac{n}{2} - r'i \sin \theta
\]

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 [67]. This network is robust for the low resolution, image blur, glint, illumination, off-angles, off-axis, reflection, glasses and different color of iris region challenges.

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 L2 losses between the predicted and ground truth gaze vectors to train the ensemble network.

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 [146]. 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], [147], [148] 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 [149], 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 [24], [50] 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 (~30fps) 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 \( g = -\arctan \frac{\hat{y}}{\hat{x}} \) and \( \hat{\phi} = \arcsin \hat{y}_g \). 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 \( L_F \) for the quantile \( \tau \) and the angle \( \alpha \in [\theta, \phi] \) can be written as:

\[
L_F(\alpha, \sigma, \alpha_{gt}) = \max(\tau q_{\tau}^- - (1 - \tau) q_{\tau}^+, \alpha_{gt} - (\alpha - \sigma) \text{ for } \tau \leq 0.5 \text{ and } \alpha_{gt} - (\alpha + \sigma) \text{ otherwise.}
\]

Static + LSTM. In an interesting study, Palmero et al. [147] 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], [147], 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) Transformer Based: Transformer models have recently garnered 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) [150] and the other one is...
hybrid transformer in gaze estimation (GazeTR-Hybrid) [150]. GazeTR-Pure [150] 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 [150] 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.

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], [101], [102], [151].

**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 and maps it to the latent space \( z \) via an encoder \( E \) (i.e. \( E(x) : x \rightarrow z \)). 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: \( z = \{z_g; z_h; z_a\} \). 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. \( D(E(x)) : z \rightarrow \hat{x} \)). The gaze direction is estimated from the \( x_g \) 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 \( x_r \) as input, disentangles the subject’s personal non-varying embeddings (\( z_{sd} \) and \( z_{sd}^r \)), considers pseudo-label conditions (\( c_i \) and \( c_r \)) and embedding representations (\( z_i \) and \( z_r \)). 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 \( I_e \) as input and predict the redirected eye patch as output \( \hat{I}_e \) while preserving the difference in rotation \( \Delta r = r_i - r_o = G_o(I_e) - G_o(I_e) \). 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 [152] 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 [153], [154] 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 II. 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 [155] etc. Sclera segmentation challenge was organized since 2015 to promote development in this area [133], [153], [156]. 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 used deep learning techniques [136], [144], [157].

**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], [149] 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], [25], [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 [173] 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 [173] 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. [165] 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 [176].

| Ref. | Reg. | Represent. | Level of Sup. | Model | Prediction | Validation | Plat. | Publication Venue (Publ.) and Year |
|-----|------|------------|--------------|-------|------------|------------|------|-------------------------------|
| [23] | Face [58] | Appear. | Fully-Sup. | SVM | Gaze locking | [23] | Scr. | UST 2013 |
| [159] | 3-D MM | Appear. | Fully-Sup. | ConvNet Hall | 3-D GV | [159] | RT | FFRA 2014 |
| [20] | Face, Eye [20] | Appear. | Fully-Sup. | RRF | 3-D GV | [20] | Any | CVPR 2014 |
| [160] | 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. | tTracker [25] | 2-D Scr. | [25], [87] | HH | CVPR 2016 |
| [161] | Eye | Appear. | Fully-Sup. | CNN [25] | GR Img. | [161] | Any | ECCV 2016 |
| [87] | Eye [87] | Appear. | Fully-Sup. | PINet [26] | 3-D Scr. | [87] | HH | MVA 2017 |
| [54] | Eye [163] | Appear. | Fully-Sup. | VGG-16+IP [54] | 3-D GV | [24], [54] | Scr. | ECCV 2018 |
| [26] | Eyes | Appear. | Fully-Sup. | CNN | 3-D GV | [24], [159] | Scr. | ECCV 2018 |
| [57] | Face [104] | Geo. + Appear. | Fully-Sup. | CNN [57] | 3-D GV | [23], [87] | Desk. | IJCV 2018 |
| [92] | Eye | Geo. + Appear. | Fully-Sup. | HGS+BigGAN | Eye, GV | [24], [159] | 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 | DTED+ML | 3-D GV | [24], [25] | Scr. | ICCV 2019 |
| [90] | Face | Appear. | Fully-Sup. | PixelLSTM | 3-D GV | [23], [24], [87] | BT | ICCV 2019 |
| [32] | Eye | Appear. | Fully-Sup. | SegNet [164] | Seg, Map | [32] | HMD | ICCVW 2019 |
| [66] | Face, L/R Eye | Appear. | Fully-Sup. | DGTN | GV | [66] | Desk. | CVPR 2019 |
| [165] | Face | Appear. | Fully-Sup. | MeNet | 3-D GV | [20], [24], [25] | Scr. | CVPR 2019 |
| [145] | Face, Eye | Appear. | Semi/Unsup. | RCNN | 3-D GV | [24], [159] | Desk. | CVPR 2019 |
| [136] | Eyes | Appear. | Fully-Sup. | Hybrid U-net | Seg, Map | [32] | HMD | ICCVW 2019 |
| [166] | Eyes | Appear. | Fully-Sup. | Modified ResNet | Seg, Map | [32] | HMD | ICCVW 2019 |
| [157] | 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] | Eye | Appear. | Unsup. | EyeSeg | 2-D GV | [23], [24] | Any | CVPR 2019 |
| [27] | Face, Eye | Appear. | Unsup. | InNet | 3-D GV | [24], [89] | PV | IJCVN 2019 |
| [167] | Eye | Appear. + Seg. | Fully-Sup. | Seg2Eye | Eye Img | [32] | HMD | ICCVW 2019 |
| [168] | Eye Seq. | Appear. | Unsup. | Hier. HMM | Eye Move. | [109] | Any | ECCVW 2019 |
| [170] | Eye | Appear. | Semi/Unsup. | mSegNet+Discrete. | Seg, Map | [32] | HMD | ECCVW 2019 |
| [171] | Eye | Appear. | Few-Shot | EyeSeg | Seg, Map | [32] | HMD | ECCVW 2019 |
| [101] | Face | Appear. | Fully-Sup. | STED | 3-D GV | [23], [25], [159] | Scr. | NeurIPS 2020 |
| [34] | Eye | Appear. | Fully-Sup. | Modified ResNet | GR Img | [34] | HMD | ICCVW 2020 |
| [149] | Eyes | Appear. | Fully-Sup. | ResNet+GRU | 3-D GV | [149] | Scr. | ECCV 2020 |
| [94] | Face | Appear. | Fully-Sup. | ResNet-50 | 3-D GV | [24], [25], [159] | Scr. | ECCV 2020 |
| [172] | Face | Appear. | Semi-Sup. | GRN | GV | [113] | PV | WACV 2020 |
| [60] | Face, Eye | Appear. | Fully-Sup. | RSNet+GazeNet | GV | [24], [25], [159] | Scr. | BMVC 2020 |
| [59] | Face, Eye | Appear. | Fully-Sup. | CA-Net | GV | [24], [159] | Scr. | AAAI 2020 |
| [173] | Face, Eye | Appear. | Fully-Sup. | PAF-Knet | GV | [24], [25], [35] | Scr. | TP 2020 |
| [102] | Eye | Appear.+AE | Fully-Sup. | Mf-cGAN | Eye Img | [20], [23], [24] | Scr. | WACV 2021 |
| [89] | Eye | Appear. | Fully-Sup. | APE-Net | Scr., GV | [25], [51] | Scr. | Arxiv 2021 |
| [174] | Face | Appear. | Unsup. | PureGaze | Face, GV | [20], [24], [90], [94] | Scr. | ICPR 2021 |
| [110] | Face | Appear. | Weakly-Sup. | ResNet+LSTM | GV | [25], [90], [94], [110] | Any | CVPR 2021 |
| [109] | Face | Appear. | Fully-Sup. | LAO-Net++ | LABO | Any | Any | TPAMI 2021 |
| [175] | Face, Eye | Appear. | Limited-Sup. | ResNet-50 | GV | [23], [24], [40], [90] | Any | WACV 2022 |
In order to train the deep learning based models, \( \ell_2 \) [24], [50], [51] and cosine similarity based losses [94], [149] 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], [153], [166], [177]. The most popular network architectures for eye segmentation are U-net [178], modified version of SegNet [32], RTNet [136], Eye-Net [166]. These VAE based architectures have high time and space complexity. However, recent methods [136], [166] do consider these factors without compromising the performance.

C. 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.

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], [155], [156], [157], [166]. 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 Section V-A. 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 [149], [179], 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.

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)’ [110] and pseudo labelling [175]. For weak supervision, Kothari et al. [110] leverage strong gaze-related geometric constraints from two people interaction scenario. On the other hand, MTGLS [175] 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 [175]. 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 \( \leq 9 \) 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 different gaze viewpoints to multiply the data from a quantitative and a qualitative perspectives rather than manual labelling. Few other studies [20], [91], [92], [161], [181], [182], [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.

V. VALIDATION

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

A. Datasets for Gaze Analysis

With the rapid progress in the gaze analysis domain, several datasets have been proposed for different gaze analysis tasks (see Section III). 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 III. Compared with early datasets [23], [159], recently released datasets [90], [149] 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:

**ColumbiaGaze** [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.

**UT-Multiview** [20] dataset have dense gaze labels of 50 participants having 3D reconstructions of eye regions with arbitrary head poses.

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 $7 \times 5$ grid format.

**MPIIGaze** [50] dataset contains 213,659 images collected from 15 subjects during natural everyday events in front of a laptop over a three-month duration. MPIIGaze dataset is collected by showing random points on the laptop screen to the participants. Please note that in this dataset only eye region is considered as input.

**MPIIFaceGaze** [51] dataset further curated from MPIIGaze 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** [149] 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], [186], [187]. A more challenging publicly available dataset was released in sclera recognition challenges [133], [155], [156]. 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,
| Dataset                  | # Sub | Label                  | Modality | Head-Pose | Gaze | Env.       | Baseline       | # Data | Year |
|--------------------------|-------|------------------------|----------|-----------|------|------------|----------------|--------|------|
| ColumbiaGaze [23]        | 56    | 3-D                    | Image Dim.: 5184 × 3456 | 0°, ±30° | ±15°, ±10° | In | SVM Eval.: Cross-val | Total: 5880 | 2013 |
| EYEDIAP [159]            | 16    | 3-D                    | Image Dim.: HD and VGA | ±15°, ±30° | ±25°, 20° | In | Convex Hull Eval.: Hold out | Total: 237 min | 2014 |
| UT-Multiview [20]        | 50    | 3-D                    | Image Dim.: 1280 × 1024 | ±36°, ±36° | ±50°, ±36° | In | Random Reg. Forests Eval.: Hold out | Total: 64,000 | 2014 |
| OMEG [189]               | 50    | 3-D                    | Image Dim.: 1280 × 1024 | 0°, ±30° | −38° to +36°, −10° to +29° | In | SVR Eval.: LOSO | Total: 64,827 | 2015 |
| MPIIGaze [24]            | 15    | 3-D                    | Image Dim.: 1280 × 720 | ±15°, 30° | ±20°, ±20° | In | CNN variant [24] Eval.: LOSO | Total: 213,659 | 2015 |
| GazeFollow [190]         | 130,339 | 3-D                    | Image Dim.: Variable | Variable | Both | CNN variant [190] Eval.: Hold out | Total: 122,143 | 2015 |
| SynthEye [160]           | NA    | 3-D                    | Image Dim.: 120 × 80 | ±50°, ±50° | ±50°, ±50° | Syn | CNN [160] Eval.: Hold out | Total: 11,400 | 2015 |
| GazeCapture [25]         | 1450  | 2-D                    | Image Dim.: 640 × 480 | ±30°, 40° | ±20°, ±20° | Both | CNN [25] Eval.: Hold out | Total: 2,445,504 | 2016 |
| UnityEyes [191]          | NA    | 3-D                    | Image Dim.: 400 × 300 | Variable | Both | Syn | KNN Eval.: NA | Total: 1,000,000 | 2016 |
| TabletGaze [87]          | 51    | 2-D Sc.                | Image Dim.: 1280 × 720 | ±50°, ±50° | ±20°, ±20° | In | SVM Eval.: Cross-val | Total: 816 Seq. ~ 300,000 img. | 2017 |
| MPIIFaceGaze [51]        | 15    | 3-D                    | Image Dim.: 1280 × 720 | ±15°, 30° | ±20°, ±20° | 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 | CNN [70] Eval.: Hold out | Total: 280,000 | 2017 |
| RT-GENE [54]             | 15    | 3-D                    | Image Dim.: 1920 × 1080 | ±40°, ±40° | ±40°, ±40° | In | CNN [54] Eval.: Cross-val | Total: 122,531 | 2018 |
| Gaze 360 [90]            | 238   | 3-D                    | Image Dim.: 4096 × 3382 | ±90°, u/R | ±140°, ±50° | Both | Pinball LSTM Eval.: Hold out | Total: 172,000 | 2019 |
| RT-BENE [192]            | 17    | EB                     | Image Dim.: 1920 × 1080 | ±40°, ±40° | ±40°, ±40° | In | CNNs Eval.: Cross-val | Total: 243,714 | 2019 |
| NV Gaze [68]             | 30    | 3-D, Seg.              | Image Dim.: (Synthetic) 1280 × 960, 640 × 480 | Unknown | 30°×40° VF | Both | CNN [193] Eval.: Hold out | Total: 2,500,000 | 2019 |
| HUST-LBSW [145]          | 172   | EB                     | Video Dim.: 1280 × 720 | Variable | Both | MS-LSTM Eval.: Hold out | Total: 673 | 2019 |
| VACATION [111]           | 206,774 | GC                    | Video Dim.: 640 × 360 | Variable | Both | CNN Eval.: Hold out | Total: 96,993 | 2019 |
| OpenEIDS-19 [32] Track 1: Semantic Segmentation | 152  | Seg.                  | Image Dim.: 640 × 400 | Unknown | Unknown | In | SegNet [164] Eval.: Hold out | Total: 12,759 (in # SegSeq [32]) | 2019 |
| OpenEIDS-19 [32] Track 2: Synthetic Eye Generation | 152  | Gen.                  | Image Dim.: 640 × 400 | Unknown | Unknown | In | CNN Eval.: Hold out | Total: 252,690 | 2019 |
| OpenEIDS-20 [34] Track 1: Gaze Prediction | 90  | 3-D                   | Image Dim.: 640 × 400 | Unknown | ±20°, ±20° | In | Modified RestNet Eval.: Hold out | Total: 8,960 Seq., 550,400 img. | 2020 |
| OpenEIDS-20 [34] Track 2: Sparse Temporal Semantic Segmentation | 90  | Seg.                  | Image Dim.: 640 × 400 | Unknown | ±20°, ±20° | In | SegNet [164] (Power Efficient version) Eval.: Hold out | Total: 200 Seq., 29,500 img. | 2020 |
| mEBAL [142]              | 38    | EB                     | Image Dim.: 1280 × 720 | Variable | Both | VGG-16 Variant Eval.: Hold out | Total: 756,000 | 2020 |
| ETH-XGaze [94]           | 110   | 3-D                    | Image Dim.: 6000 × 4000 | ±80°, ±80° | ±120°, ±70° | In | ResNet-50 Eval.: Hold out | Total: 1,083,492 | 2020 |
| EVE [149]                | 54    | 3-D                    | Image Dim.: 6000 × 4000 | ±80°, ±80° | ±80°, ±80° | In | ResNet-18 Eval.: Hold out | Total: 12,308,334 | 2020 |
| GW [188]                 | 19    | GE                     | Image Dim.: 1920 × 1080 | Variable | Both | CNN Eval.: Hold out | Total: 3,394 | 2020 |
| LABO [110]               | 485   | 3-D                    | Image Dim.: Variable | Variable | Both | CNN Eval.: Hold out | Total: 800,000 | 2021 |
| GON [194]                | 100   | 3-D                    | Image Dim.: Variable | Variable | Both | CNN Eval.: Hold out | Total: 201,552 | 2021 |
| OpenNEDS [195]           | 44    | 3-D                    | Image Dim.: 128 × 71 | Variable | Both | GBRT Eval.: Hold out | Total: 2,086,507 | 2021 |

The abbreviations used are: In: Indoor, Out: Outdoor, Both: Indoor + Outdoor, Gen.: Generation, u/R: unknown, Seq.: Sequence, VF: Visual Field, EB: Eye Blink, GE: Gaze Event [188], GBRT: Gradient Boosting Regression Trees, GC: Gaze Communication, GNN: Graph Neural Network and Seg.: Segmentation.
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], [161], [181], [182], [183] are quite popular. An early work on gaze redirection via a flow field to move eye pupil and relevant pixels from the input image to the output image. The gaze redirection task. 

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 degrees) [26], [29], [94], [149] and gaze location (in pixels or cm/mm(s)) [87], [149]. The angular error is measured between the actual gaze direction ($g \in \mathbb{R}^3$) and predicted gaze direction ($\hat{g} \in \mathbb{R}^3$) defined as $\|g - \hat{g}\|$. 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], [101], [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).

### C. 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-XGaze [94]. For this purpose, the training is performed on one dataset while the testing is conducted on the other dataset. (Refer Table IV). Further, we explore the SOTA method’s performance on different datasets to show the robustness of the model (Refer Table V). Angular error (in degrees) 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 ColumbiaGaze 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 ColumbiaGaze data has a high error in Gaze360 ($\sim 57.9^\circ$) and RT-GENE ($\sim 32.8^\circ$) datasets. A similar pattern is observed in the case of the MPIIGaze dataset as well for Pinball LSTM model. Pinball LSTM model trained on MPIIGaze dataset gives high error in adapting RT-GENE ($\sim 26.5^\circ$) and Gaze360 ($\sim 57.8^\circ$).

**Cross Dataset Generalization.** The most important factors impacting cross dataset evaluation are the gaze direction and head pose ranges. If the gaze and headpose ranges of two datasets are similar, the cross-dataset performance will enhance (e.g. see performance comparison between MPIIFaceGaze and Gaze-Capture at Table IV). From Table IV, we observe that Gaze360, Gaze-Capture, and ETH-XGaze datasets are the most challenging datasets. Training models on these three datasets would be a good choice as it has better generalization performance across different datasets. From Table IV, it is also observed that the RT-GENE and GazeCapture do not contain significant biases for training a model. By observing the cross dataset generalization performance, we can determine how diverse the training dataset is from a generalization perspective.

**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-XGaze 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.

D. Where We Stand Now?

In Table V, 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 V.

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 fixed distance between the user and the visual screen. This impacts the generalization capability of model’s performance. Apart from that, there are person specific constraints as well. As the nodal point of the human eye has subject-specific offset (which varies around 2-3°) [45], [46], 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 ColumbiaGaze, MPlIGaze, and ETH-XGaze. Thus, datasets containing more realistic situation is highly desirable in this field.

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 [203], unsupervised gaze target discovery [95], few-shot learning [29], pseudo-labelling [175] 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.

### TABLE IV

**CROSS-DATASET STUDY. CROSS DATASET GENERALIZATION STUDY ON DIFFERENT GAZE ESTIMATION DATASETS IN TERMS OF ANGULAR ERROR (IN DEGREES)**

| Model                | Train          | Test            |
|----------------------|----------------|-----------------|
| Pinball-LSTM [90]    | ColumbiaGaze   | ColumbiaGaze    |
|                      | MPlIGaze       | MPlIGaze        |
|                      | RT-GENE        | RT-GENE         |
|                      | Gaze360        | Gaze360         |
|                      |               | MPlIGaze        |
|                      |               | EYEDIAP         |
|                      |               | Gaze-Capture    |
|                      |               | RT-GENE         |
|                      |               | Gaze360         |
|                      |               | ETH-XGaze       |
|                      | 12.3°          | 32.8°           |
|                      | 26.5°          | 57.8°           |
|                      | 18.9°          | 56.6°           |
|                      | 12.1°          | 13.4°           |

Please note that the results are adopted from respective papers [90], [94].

### TABLE V

**WHERE WE STAND NOW. COMPARISON OF THE PERFORMANCE OF DIFFERENT MODELS FOR APPEARANCE BASED GAZE ESTIMATION ON BENCHMARK DATASETS**

| Task               | Datasets          | Method: AE | Method: AE | Method: AE | Method: AE | Method: AE | Method: AE | Method: AE |
|--------------------|-------------------|------------|------------|------------|------------|------------|------------|------------|
|                    |                    | [97]: 6.2° | [24]: 6.3° | [97]: 7.5° | [97]: 11.5°| [51]: 4.9° | [51]: 14.9°| [84]: 4.5° |
|                    |                    | [51]: 10.0°|            |            |            |            |            |            |
|                    |                    | [199]: 4.2°| [64]: 5.21°| [51]: 6.53°| [54]: 10.0°| [54]: 4.66°| [64]: 13.73°| [54]: 8.0° |
|                    |                    | [51]: 10.0°|            |            |            |            |            |            |
|                    |                    | [200]: 3.9°| [94]: 4.8° | [94]: 6.5° | [24]: 5.9° | [64]: 4.42°| [175]: 12.83°| [64]: 8.38°|
|                    |                    | [51]: 10.0°|            |            |            |            |            |            |
|                    |                    | [26]: 3.8° | [97]: 4.6° | [64]: 6.19°| [28]: 5.7° | [59]: 4.27°| [54]: 12.26°| [90]: 7.0° |
|                    |                    |            |            |            |            |            |            |            |
|                    |                    | [90]: 6.2° | [26]: 4.5° | [54]: 6.02°| [165]: 5.5°| [150]: 4.18°| [59]: 11.2° | [59]: 8.27°|
|                    |                    |            |            |            |            |            |            |            |
|                    |                    | [28]: 3.42°| [54]: 4.30°| [90]: 5.36°| [145]: 5.4°| [90]: 4.06°| [90]: 11.0° | [150]: 7.12°|
|                    |                    |            |            |            |            |            |            |            |
|                    |                    | [175]: 3.00°| [150]: 4.10°| [150]: 5.33°| –          | [20]: 4.06°| [150]: 11.0°| [20]: 7.0° |
|                    |                    | [57]: 2.22°| [175]: 4.07°| [20]: 5.10°| –          | [20]: 4.06°| [20]: 10.04°| [20]: 6.92°|

Please note that we adopted the results from the respective references [24], [54], [95], [90], [94], [97]. For ETH-XGaze, other than baseline [94] three methods got top-3 ranking in the GAZE challenge and their respective GitHub codes are also available online. Here, AE: Angular Error (in degree) and * methods are supervised methods.
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 [146], [204]. 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 \( \sim 100^\circ \) and a vertically oriented axis spanning \( \sim 70^\circ \) [204]. Due to the lack of labeled temporal gaze trajectory data, there are only a few studies [34], [147] that focus on tasks related to the gaze trajectory.

Discussion. Despite the steady progress of the field, there is a domain gap in terms of benchmark datasets and real-time deployment of the computational gaze models. Thus, data curation in more realistic situations is highly desirable. As the data curation process is tiresome and error prone, computational models trained with limited supervision paradigm have immense potential to eliminate the concern. Apart from these two aspects, gaze trajectory modelling and human visual attention estimation are two important fields which need much attention due to their potential application in real-world environments.

VI. APPLICATIONS

A. Gaze in Augmented Reality and Virtual Reality

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], [205], [206], [207]. 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], [206]. 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 point of 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 [209].

B. 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], [210]. In the literature, the problem is treated as a gaze zone estimation problem. A summary of the gaze estimation benchmarks is shown in Table VI. 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 [211], head-mounted devices [77], [212] and other systems [213], [214].
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 future of gaze estimation and its applications.

VIII. 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 future of gaze estimation and its applications.
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 Section II-D. 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., ColumbiaGaze, TabletGaze, MPIIGaze, EYEDIAP, ETH-XGaze etc). It would be an interesting direction to explore gaze estimation in a more flexible setting.

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 [227], MFCC (Mel Frequency Cepstral Coefficient) [228], LPC (Linear Predictive Coding) [228] and filter bank [228], [229] 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 [227], MLP based regression modelling [228], bi-LSTM [229] and Conditional Variational Autoencoder (CVAE) [230]. 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 [231]. 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 Trans. Circuits Syst. Video Technol., vol. 21, no. 7, pp. 971–982, Jul. 2011.
[2] A. Frischen, A. P. Bayliss, and S. P. Tipper, “Gaze cueing of attention: Visual attention, social cognition, and individual differences,” Psychol. Bull., vol. 133, 2007, Art. no. 694.
[3] D. Purves, Y. Morgenstern, and W. T. Wotjak, “Perception and reality: Why a wholly empirical paradigm is needed to understand vision,” Front. Syst. Neurosci., vol. 9, 2015, Art. no. 156.
[4] E. Jalav, “Essai sur la physiologie de la lecture,” Annales d’Ocristiqute, 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,” J. Opt. Soc. Amer., vol. 63, no. 3, pp. 921–928, 1973.
[6] J. Merchant, R. Morrissite, and J. L. Porterfield, “Remote measurement of eye direction allowing subject motion over one cubic foot of space,” IEEE Trans. Biomed. Eng., vol. BME-21, no. 4, pp. 309–317, Jul. 1974.
[7] M. Borstig, J. Sandqvist, R. Parsons, T. Falker, and H. Hemmingsson, “Eye gaze performance for children with severe physical impairments using gaze-based assistive technology—a longitudinal study,” Assistive Technol., 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 Proc. IEEE Int. Conf. Multimedia Expo, 2002, pp. 433–436.
[60] X. Zhang, Y. Sugano, A. Bulling, and O. Hilliges, “Learning-based region selection for end-to-end gaze estimation,” in Proc. Brit. Mach. Vis. Conf., 2020, pp. 86–100.

[61] Y. Yu, G. Liu, and J.-M. Odobez, “Deep multitask gaze estimation with a continuous gaze distribution,” in Proc. Eur. Conf. Comput. Vis. Workshops, 2018, pp. 456–474.

[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 Proc. Eur. Conf. Comput. Vis. Workshops, 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 Proc. CHI Conf. Hum. Factors Comput. Syst., 2018, pp. 1–12.

[64] Z. Chen and B. E. Shi, “Appearance-based gaze estimation using dilated-convolutions,” in Proc. Asian Conf. Comput. Vis., 2018, pp. 309–324.

[65] X. Zhu, J. Lin, J. Jiang, and S. Chen, “Learning a 3D gaze estimator with improved itracker combined with bidirectional LSTM,” in Proc. IEEE Int. Conf. Multimedia Expo, 2019, pp. 850–855.

[66] K. Wang, H. Su, and Q. Ji, “Neuro-inspired eye tracking with eye movement dynamics,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 9831–9840.

[67] Z. Wu, S. Rajendran, T. Van As, V. Badrinarayanan, and A. Rabinovich, “EyeNet: multi-task deep network for off-axis eye gaze estimation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 3683–3691.

[68] J. Kim et al., “NVGaze: An anatomically-informed dataset for low-latency, near-eye gaze estimation,” in Proc. CHI Conf. Hum. Factors Comput. Syst., 2019, pp. 1–12.

[69] D. Xia and Z. Ruan, “IR image based gaze estimation,” in Proc. IEEE ACIS Int. Conf. Softw. Eng., Artif. Intell., Netw., Parallel/Distrib. Comput., 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,” in Proc. ACM Int. Conf. Mobile, Wearable Ubiquitous Technol., vol. 1, no. 3, pp. 21–27, 2017.

[71] D. Lian et al., “RGDB based gaze estimation via multi-task CNN,” in Proc. AAAI Conf. Artif. Intell., 2019, pp. 2488–2495.

[72] A. Tsukada, M. Shino, M. Devyver, and T. Kanade, “Illuminant-free gaze estimation method for first-person vision wearable device,” in Proc. IEEE Int. Conf. Comput. Vis. Workshops, 2011, pp. 2084–2091.

[73] M. Leo, D. Cazzato, T. De Marco, and C. Distanta, “Unsupervised eye pupil localization through differential geometry and local self-similarity,” Public Library Sci., vol. 9, no. 8, 2014, Art. no. e102829.

[74] C. Gou, Y. Wu, K. Wang, K. Wang, F. Wang, and Q. Ji, “A joint cascaded face detector for simultaneous eye detection and eye state estimation,” Pattern Recognit., vol. 67, pp. 23–31, 2017.

[75] J. Pi and B. E. Shi, “Task-embedded online eye-tracker calibration for improving robustness to head motion,” in Proc. 11th ACM Symp. Eye Tracking Res. Appl., 2019, pp. 309–324.

[76] S. Jha and C. Busso, “Probabilistic estimation of the gaze region of the driver using dense classification,” in Proc. IEEE Int. Conf. Intell. Transp. Syst., 2018, pp. 697–702.

[77] Z. Hu, S. Li, C. Zhang, K. Yi, G. Wang, and D. Manocha, “DGaze: CNN-based gaze prediction in dynamic scenes,” IEEE Trans. Visu. Comput. Graph., vol. 26, no. 5, pp. 1902–1911, May 2020.

[78] 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 Proc. IEEE Conf. Intell. Transp. Syst., 2014, pp. 988–994.

[79] A. Tawari and M. M. Trivedi, “Robust and continuous estimation of driver gaze zone by dynamic analysis of multiple face videos,” in Proc. IEEE Intellect. Veh. Symp., 2014, pp. 254–265.

[80] B. Vasi, S. Martin, and M. M. Trivedi, “On driver gaze estimation: Explanations of geometric and data driven approaches,” in Proc. IEEE 19th Int. Conf. Intell. Transp. Syst., 2016, pp. 655–660.

[81] L. Fridman, P. Langhans, J. Lee, and B. Reimer, “Driver gaze estimation without using eye movement,” IEEE Intell. Syst., vol. 31, no. 3, pp. 49–56, May/June 2015.

[82] L. Fridman, J. Lee, B. Reimer, and T. Victor, “‘Owl’ and ‘lizard’: Patterns of head pose and eye pose in driver gaze classification,” IET Comput. Vis., vol. 10, no. 4, pp. 308–314, 2016.
W. Wang, J. Shen, X. Dong, A. Borji, and R. Yang, “Inferring salient objects from human fixations,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 3756–3764.

Y. Guan, D. Kononenko, D. Sungatullina, and V. Lemptitsky, “DeepWater: Photorealistic image resynthesis for gaze manipulation,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 311–326.

S. Yu, “Harr feature cart-tree based cascade eye detector homepage,” 2015. [Online]. Available: http://yushiqh.cn/research/eye_detection

J. Xiang and G. Zhu, “Joint face detection and facial expression recognition with MTCNN,” in Proc. 4th Int. Conf. Inf. Sci. Control Eng., 2017, pp. 424–427.

Y. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A deep convolutional encoder-decoder architecture for image segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

Y. Xiong, H. J. Kim, and Y. Singh, “Mixed effects neural networks (MeNets) with applications to gaze estimation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 7743–7752.

P. Kansal and S. Devanathan, “EyeNet: Attention based convolutional encoder-decoder network for eye region segmentation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop, 2019, pp. 3688–3693.

M. Bühler, S. Park, S. De Mello, X. Zhang, and O. Hilliges, “Content-consistent generation of realistic eyes with style,” 2019, arXiv:1911.03346.

Y. Zhu, Y. Yan, and O. Komogortsev, “Hierarchical HMM for eye movement classification,” in Proc. Eur. Conf. Comput. Vis., Springer, 2020, pp. 544–554.

O. V. Komogortsev and A. Karpov, “Automated classification and scoring of smooth pursuit eye movements in the presence of fixations and saccades,” Behav. Res. Methods, vol. 45, no. 1, pp. 203–215, 2013.

Y. Shen, O. Komogortsev, and S. S. Talathi, “Domain adaptation for eye segmentation,” in Proc. Eur. Conf. Comput. Vis., Springer, 2020, pp. 555–569.

J. Perry and A. S. Fernandez, “Eyeseg: Fast and efficient few-shot semantic segmentation,” in Proc. Eur. Conf. Comput. Vis., Springer, 2020, pp. 570–582.

P. A. Dias, D. Malafronte, H. Medeiros, and F. Odone, “Gaze estimation for assisted living environments,” in Proc. IEEE/CVF Winter Conf. Comput. Vis. Pattern Recognit., 2020, pp. 290–299.

Y. Cheng, X. Zhang, F. Lu, and Y. Sato, “Gaze estimation by exploiting two-eye asymmetry,” IEEE Trans. Image Process., vol. 29, no. 5, pp. 5259–5272, 2020.

Y. Cheng, Y. Bao, and F. Lu, “Puregaze: Purifying gaze feature for generalizable gaze estimation,” 2021, arXiv:2103.13173.

S. Ghosh, M. Hayat, A. Dhall, and J. Knibbe, “MTGLS: Multi-task gaze estimation with limited supervision,” in Proc. IEEE/CVF Winter Conf. Comput. Vis. Pattern Recognit., 2022, pp. 3223–3234.

Z. Chen and B. Shi, “Offset calibration for appearance-based gaze estimation via gaze decomposition,” in Proc. IEEE/CVF Winter Conf. Comput. Vis. Pattern Recognit., 2020, pp. 270–279.

B. Luo, J. Shen, S. Cheng, Y. Wang, and M. Pantic, “Shape constrained network for eye segmentation in the wild,” in Proc. IEEE Winter Conf. Comput. Vis. Pattern Recognit., 2020, pp. 1992–1999.

O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in Proc. Int. Conf. Med. Image Comput. Assist. Intervention, Springer, 2015, pp. 234–241.

W. Wang, J. Shen, X. Dong, A. Borji, and R. Yang, “Inferring salient objects from human fixations,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 8, pp. 1913–1927, Aug. 2020.

A. S. Fernandez, A. F. Palmeiro, S. P. Lumsden, S. Escalera, S. P. Lumsden, and A. S. Fernandez, “Towards self-supervised gaze estimation,” 2022, arXiv:2203.09749.

Z. He, A. Spurr, X. Zhang, and O. Hilliges, “Photo-realistic monocular gaze redirection using generative adversarial networks,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 6932–6941.

E. Wood, T. Bahrulsaitis, L.-P. Morency, P. Robinson, and A. Bulling, “Gazedirector: Fully articulated eye gaze redirection in videos,” Comput. Graph. Forum, vol. 37, no. 2, pp. 217–225, 2018.

K. A. Funes Mora, F. Monay, and J.-M. Odobez, “GOO: A dataset for gaze object prediction in retail environments,” in Proc. Int. Conf. Med. Image Comput. Vis. Pattern Recognit., 2022, pp. 3119–3128.

K. Cortacero, T. Fischer, and Y. Demiris, “RT-Bene: A dataset and baselines for real-time blink estimation in natural environments,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops, 2019, pp. 1190–1198.

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 Proc. ACM Symp. Eye Tracking Res. Appl., 2021, pp. 1–7.

L. Wolf, Z. Freund, and S. Avidan, “An eye for an eye: A single camera gaze-replacement method,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2010, pp. 817–824.

D. Kononenko and V. Lemptitsky, “Learning to look up: Realtime monocular gaze correction using machine learning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 4667–4675.

K. Cortacero, T. Fischer, and Y. Demiris, “RT-Bene: A dataset and baselines for real-time blink estimation in natural environments,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops, 2019, pp. 1190–1198.

K. Cortacero, T. Fischer, and Y. Demiris, “RT-Bene: A dataset and baselines for real-time blink estimation in natural environments,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops, 2019, pp. 1190–1198.

K. Cortacero, T. Fischer, and Y. Demiris, “RT-Bene: A dataset and baselines for real-time blink estimation in natural environments,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops, 2019, pp. 1190–1198.

K. Cortacero, T. Fischer, and Y. Demiris, “RT-Bene: A dataset and baselines for real-time blink estimation in natural environments,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops, 2019, pp. 1190–1198.

K. Cortacero, T. Fischer, and Y. Demiris, “RT-Bene: A dataset and baselines for real-time blink estimation in natural environments,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops, 2019, pp. 1190–1198.

K. Cortacero, T. Fischer, and Y. Demiris, “RT-Bene: A dataset and baselines for real-time blink estimation in natural environments,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops, 2019, pp. 1190–1198.

K. Cortacero, T. Fischer, and Y. Demiris, “RT-Bene: A dataset and baselines for real-time blink estimation in natural environments,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops, 2019, pp. 1190–1198.
[207] 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 Trans. Netw. Service Manag., vol. 18, no. 1, pp. 1000–1015, Mar. 2021.

[208] D. Alexandrovsky et al., “Examining design choices of questionnaires in VR user studies,” in Proc. CHI Conf. Hum. Factors Comput. Syst., New York, NY, USA, 2020, pp. 1–21. [Online]. Available: https://doi.org/10.1145/3313831.3376260

[209] G.-A. Koulieris, K. Akçıl, C. Richardt, and R. Mantiuk, “Cutting-edge VR/AR display technologies (gaze-, accommodation-, motion-aware and head-tracked),” in Proc. SIGGRAPH Asia Courses. New York, NY, USA, 2018, pp. 1–341. [Online]. Available: https://doi.org/10.1145/3277644.3277771

[210] 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,” 2020, arXiv:2101.04639.

[211] 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.

[212] S. Jha and C. Busso, “Challenges in head pose estimation of drivers in naturalistic recordings using existing tools,” in Proc. IEEE Int. Conf. Intell. Transp. Syst., 2017, pp. 1–6.

[213] 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 Trans. Multimedia, vol. 15, no. 8, pp. 1865–1879, Dec. 2013.

[214] C. Zhang, Q. He, J. Liu, and Z. Wang, “Exploring viewer gazing patterns for touch-based mobile gamecasting,” IEEE Trans. Multimedia, vol. 19, no. 10, pp. 2333–2344, Oct. 2017.

[215] Y. Wang et al., “Continuous driver’s gaze zone estimation using RGB-D camera,” Sensors, vol. 19, no. 6, 2019, Art. no. 1287.

[216] S. S. Mukherjee and N. M. Robertson, “Deep head pose: Gaze-direction estimation in multimodal video,” IEEE Trans. Multimedia, vol. 17, no. 11, pp. 2094–2107, Nov. 2015.

[217] S. Alghowinem, R. Goecke, M. Wagner, A. F. Kramer, and M. Breakspear, “Evidence from eye tracking,” Psychiatry Res., vol. 269, pp. 408–413, 2018.

[218] O. Avital, “Method and system of using eye tracking to evaluate subjects,” US Patent App. 14/681,083, Oct. 08, 2015.

[219] K. Sakurai, M. Yan, H. Tamura, and K. Tanno, “A study on gaze estimation system using the direction of eyes and face,” in Proc. World Automat. Congr., 2016, pp. 1–6.

[220] K. Sakurai, M. Yan, K. Tanno, and H. Tamura, “Gaze estimation method using analysis of electrooculogram signals and inertial sensor,” Comput. Intell. Neurosci., vol. 2017, 2017, Art. no. 2074752.

[221] J. L. Kröger, O.-M. Lutz, and F. Müller, “What does your gaze reveal about you? on the privacy implications of eye tracking,” in Proc. IFIP Int. Summer Sch. Privacy Identity Manage., Springer, 2019, pp. 226–241.

[222] B. John, S. Koppal, and E. Jain, “EyeVei: Degrading iris authentication in eye tracking headsets,” in Proc. 11th ACM Symp. Eye Tracking Res. Appl., 2019, pp. 1–5.

[223] 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 Res., vol. 269, pp. 408–413, 2018.

[224] 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,” 2022, arXiv:2207.03048.

Shreya Ghosh received the BTech degree in CSE from the Govt. College of Engineering and Textile Technology, Serampore, India, and the MS(R) degree in computer science and engineering from the Indian Institute of Technology Roorkee, India, and the PhD degree from the Monash University, Melbourne, Australia, in 2022. She is currently a research academic with Curtin University. Prior to this, she was a postdoctoral fellow with Monash University funded by DARPA. Her research interests include affective computing, computer vision, deep learning.

Avinav Dhall (Member, IEEE) received the PhD degree from the Australian National University, in 2014. Followed by postdocs with the the University of Waterloo and the University of Canberra. He is an assistant professor with the Indian Institute of Technology Roorkee and adjunct senior lecturer with Monash University. He was awarded the Best Doctoral Paper Award at ACM CIKM 2013, Best Student Paper Honourable mention at IEEE AFGR 2013 and Best Polytechnic Institute (RPI). His research interests are in computer vision for affective computing and assistive technology. He is a associate editor of IEEE Transactions on Affective Computing.

Munawar Hayat received the PhD degree from the University of Western Australia (UWA). He is currently a senior research fellow with Monash University, Australia. 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 with 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 received the PhD from the University of Bristol, in 2016. He is currently with the University of Melbourne, Australia. He completed a post-doc in human-centred computing with 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 more than 25 papers at top venues in Human-Computer Interaction, including CHI, UIST, CSCW, and Ubicomp.

Qiang Ji (Fellow, IEEE) received the PhD degree in electrical engineering from the University of Washington. He is a professor with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute (RPI). He was a program manager with the National Science Foundation, where he managed NSF’s computer vision and machine learning programs. He also held teaching and research positions with the University of Illinois at Urbana-Champaign, Carnegie Mellon University, and University of Nevada, Reno. His research interests are in computer vision, probabilistic machine learning, and their applications. He has published more than 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 AIPAR.