Attention for Vision-Based Assistive and Automated Driving: A Review of Algorithms and Datasets

Iuliia Kotseruba and John K. Tsotsos

*Abstract***— Driving safety has been a concern since the first cars appeared on the streets. Driver inattention has been singled out as a major cause of accidents early on. This is hardly surprising, as drivers routinely perform other tasks in addition to controlling the vehicle. Decades of research into what causes lapses or misdirection of drivers' attention resulted in improvements in road safety through better design of infrastructure, driver training programs, in-vehicle interfaces, and, more recently, the development of driving assistance systems (ADAS) and driving automation. This review focuses on the methods for modeling and detecting spatio-temporal aspects of drivers' attention, *i.e.* where and when they look, for the two latter categories of applications.**

We start with a brief theoretical background on human visual attention, methods for recording and measuring attention in the driving context, types of driver inattention, and factors causing it. We then discuss machine learning approaches for 1) modeling gaze for assistive and self-driving applications and 2) detecting gaze for driver monitoring. Following the overview of state-of-the-art models, we provide an extensive list of publicly available datasets that feature recordings of drivers' gaze and other attention-related annotations. We conclude with a general overview of the remaining challenges, such as data availability and quality, evaluation methods, and the limited scope of attention modeling, and outline steps toward rectifying some of these issues. Categorized and annotated lists of the reviewed models and datasets are available at https://github.com/ykotseruba/attention_and_driving

*Index Terms***— Visual attention, driving, gaze prediction, driver assistance, drowsiness, distraction, self-driving, review.**

I. Introduction

DRIVING, despite being commonplace, is a demanding activity that involves multiple concurrent tasks. Besides keeping the vehicle within the road boundaries, drivers observe other road users, anticipate potential hazards, and deal with distractions from both inside and outside the vehicle. Drivers rely primarily on vision to make decisions [1], thus understanding how drivers observe the scene, how it affects their reasoning, and what causes lapses in attention is crucial for ensuring road safety, especially given the existing evidence that temporary distractions and sub-optimal visual scanning skills increase risk of accidents [2], [3].

Technology for assistive and automated driving aims to reduce traffic accidents caused by human error, and significant progress has been made towards this goal in recent years. For example, advanced driver assistance systems (ADAS) are gradually becoming standard even in low- and mid-priced commercial vehicles. More than 30% of vehicles sold in the USA in 2016 were equipped with passive sensors, such as rearview cameras, parking proximity sensors, and blind-spot detection [4]. Active assistance features, *e.g.* lane departure detection, emergency braking, and adaptive cruise control, have become standard in more than 200 car models produced by major manufacturers in the past five years [5]. According to recent estimates, ADAS can potentially eliminate up to one-third of accidents caused by light vehicles on highways [6].

Although existing ADAS can detect specific hazards and automatically take measures to avoid imminent collisions, ultimately, they act independently of the drivers' state or intentions. Driver monitoring systems (DMS) offer a complementary approach to safety by estimating drivers' inattention to alert them or safely stop the vehicle if the driver is not responsive. Currently, most commercial DMS rely on vehicle measures such as steering or lateral control to assess drivers' state [7], however, the next generation monitoring systems will use in-vehicle cameras to observe drivers, analyze where they are looking, and issue warnings to direct their attention back to the road or towards critical objects/events.

Widespread deployment of vision-based DMS is necessary for partially- or highly-automated driving systems corresponding to SAE Levels 2-4 [8]. Past research shows that drivers who are not actively controlling the vehicle (*e.g.* when using full or partial automation) and perform a supervisory role are more prone to distractions [9], [10]. The safety of switching to manual control depends on whether the driver is distracted or fatigued [11], [12]; therefore monitoring drivers' state and providing feedback is necessary. Together, ADAS and DMS are expected to offer significant improvements in road safety. For example, DMS have been included in Euro NCAP 2025 roadmap to zero road fatalities by 2050 [13], and similar initiatives are likely to be proposed in other countries.

Finally, autonomous vehicles (AVs) are seen by many as the ultimate solution to eliminating some [14] and potentially all crashes [15] caused by driver error (as defined in [16]). Given recent successes of biologically-inspired attention mechanisms in various perceptual tasks [17], many self-driving approaches now incorporate attention to improve perceptual and decision-making abilities as well as their explainability.

Manuscript received 5 April 2021; revised 10 November 2021, 13 April 2022, and 15 June 2022; accepted 22 June 2022. Date of publication 12 July 2022; date of current version 7 November 2022. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), in part by the Air Force Office for Scientific Research (USA), and in part by the Canada Research Chairs Program through Grants to JKT. The Associate Editor for this article was Z. Zheng. *(Corresponding author: Iuliia Kotseruba.)*

The authors are with the Department of Electrical Engineering and Computer Science, York University, Toronto, ON M3J 1P3, Canada (e-mail: yulia_k@eecs.yorku.ca; tsotsos@eecs.yorku.ca).

Digital Object Identifier 10.1109/TITS.2022.3186613

In sum, vision-based assistive and autonomous driving solutions rely on sophisticated algorithms that observe and analyze drivers' behavior and relate it to the events unfolding in the traffic scenes. This review summarizes past works and current state-of-the-art in estimating and modeling drivers' attention, surveys publicly available datasets and discusses open problems. To limit the scope of the review, we focus on the algorithms that use machine learning techniques to model drivers' spatial and temporal attention allocation to objects and areas inside and outside the vehicle.

The paper is structured as follows. Section III provides a brief theoretical background on drivers' attention and inattention. In Section IV we discuss approaches for driver monitoring that rely on drivers' gaze or appearance for in-vehicle gaze estimation, inattention detection, action anticipation, and awareness estimation. Section V covers algorithms for modeling drivers' attention allocation in the traffic scene for assistive and autonomous driving applications. Section VI provides an extensive list of publicly available datasets that contain recordings of driver gaze and other attention-related annotations that enable the design, development, and evaluation of the models discussed in the previous sections. Finally, in Section VII we conclude the review with the general discussion of open problems and limitations of current research and suggest steps toward rectifying some of the issues.

II. Literature Search

To gather a representative set of papers for review, we conducted a thorough search using Google Scholar with the following query words: *eye*, *gaze*, *fixation*, *glance*, *eye-tracker*, *attention*, *drowsiness*, *fatigue*, *inattention*, *distraction*, and *driver*. We limited the search to papers published from 2010 to 2021 (inclusive) in premier intelligent transportation, robotics, and computer vision venues, including but not limited to Transactions on Intelligent Transportation Systems, Intelligent Vehicles Symposium (IV), International Conference on Intelligent Transportation Systems (ITSC), International Journal of Robotics Research (IJRR), International Conference on Intelligent Robots and Systems (IROS), International Conference on Robotics and Automation (ICRA), International Conference on Computer Vision and Pattern Recognition (CVPR), International Conference on Computer Vision (ICCV), European Conference on Computer Vision (ECCV). The choice of the past decade is motivated by growing interest in developing driving assistance and self-driving systems during this time period and recent breakthroughs in machine learning that promise to make such systems viable for broad deployment.

Since search terms include commonly used words, a large portion of the 3011 papers initially returned by the search engine was excluded as not relevant upon examining their titles and abstracts. We also excluded the following: 1) studies using modes of transportation other than cars (*e.g.* bicycles, motorcycles, trucks, buses, trains), 2) studies that rely *only* on indirect methods to assess drivers' attention (*e.g.* ego-vehicle sensor information), 3) studies that focused on drivers with medical issues or under the influence of alcohol or drugs, and 4) uncited papers over 5 years old. As a result, 204 papers were selected for this review.

III. Theoretical Background

Due to space constraints, we cannot discuss all aspects of human visual attention. However, in the following section, we will provide a brief theoretical background helpful for understanding how attention is defined, recorded, measured, and operationalized for applications in the driving domain.

A. Drivers' Attention

1) What Is Attention, and Why Is It Needed?: Vision is a primary source of information for driving [1]. However, drivers do not process the entire scene at once and instead sequentially focus on its various elements. This is caused by the biological properties of human vision, where acuity (resolution) is highest in the center of the visual field (fovea and parafovea) and drops off towards the periphery due to the non-uniform distribution of receptors in the retina [18], [19]. Eye movements help bring portions of the scene into the central field for closer examination [20].

2) Types of Eye Movements: In the driving domain, gaze movements are commonly used as a proxy for attention. The literature we reviewed is dedicated to analyzing episodes when the gaze is held steady (*fixations* and *glances*) and transitions between them (*saccades*), while stabilizing eye movements and vergence were not considered. *Fixations* indicate gaze held at a single point and last from a fraction of a second to several seconds. *Glance* (termed *dwell* in psychology [21]) refers to gaze maintained within some area of interest (AOI). As defined in [22], glance starts from the moment the gaze moves inside the AOI until it moves out. Duration of fixations and glances measure what areas or objects the driver attended to [23], [24] inside the vehicle or in the traffic scene, whereas saccades are indicative of their intentions and decisions [25].

3) Types of Attention Mechanisms: Eye movements are determined by attentional control mechanisms subdivided into two groups: *bottom-up* and *top-down* [26]. The former is guided by the saliency of the objects or areas in the scene that attracts gaze [27], [28]. Top-down attention is driven by the task [29], *i.e.* it focuses on the objects or events relevant for the task, whereas salient but task-irrelevant stimuli have a lesser effect.

Both are likely involved in driving, but their relative contribution and interaction are still not fully understood. Experimental evidence points to the dominant role of task-based attention in driving [30]–[32]. At the same time, salient stimuli such as bright digital billboards also tend to attract drivers' involuntary attention even though they are irrelevant to driving [33], suggesting presence of bottom-up influences.

4) Gaze Recording Equipment: Eye trackers provide the most accurate recordings of foveal vision. Tower-mounted models offer the highest precision and sampling rates [34] at the expense of significantly limiting subjects' head movements. Remote and head-mounted eye trackers have lower precision but allow normal head movement and are thus more suitable for experiments involving active control of the vehicle. At the same time, eye trackers remain expensive and susceptible to data loss due to calibration issues [35], [36]. Video cameras offer a cost-effective and nearly maintenance-free alternative to eye trackers but require labor-intensive manual coding to extract gaze information. This process involves annotating each video frame with a text label specifying the approximate drivers' gaze direction, typically subdivided into coarse areas of interest (AOIs) [37], *e.g.* rearview mirror, windshield, or speedometer. Multiple annotators are often employed to reduce errors caused by drivers' individual characteristics and subtle eye movements [2], [38]. Another source of error is low sampling rate of the cameras which may bias the data towards prolonged glances since short fixations and saccades may not be captured [37].

5) Effect of Recording Conditions: The choice of on-road versus in-lab conditions is a trade-off between realism and replicability. Recording gaze in an actual vehicle in traffic offers the most ecologically valid conditions, but driving simulators provide a more cost-effective solution that can be more reliably replicated across multiple subjects [39]. Therefore, results obtained in a driving simulator need to be verified against conclusions made using on-road data (*validity*). *Absolute validity*, *i.e.* the exact numerical match between measures obtained in simulation and on-road is preferred to *relative validity* indicating similar trends, but both are acceptable [40].

Validity depends not only on the fidelity of the simulator (how accurately it reproduces the environment and vehicle controls) but also on the measures being considered. According to [41], most of the research focus thus far has been on validating driving performance measures (*e.g.* lane and speed maintenance, crash rate, *etc.*) and few studies examined attention-related measures. While measures such as hazard anticipation and fixation durations have been validated across different types of simulators [42], [43], comparisons between on-road vehicles and simulators have not been conclusive. For example, a study in [44] reports differences in road fixations, and [45] showed greater gaze dispersion in the simulator than on-road. In a recent experiment by Robbins *et al.* [46], mean fixation durations recorded in a high-fidelity driving simulation were similar to an on-road experiment but only for medium- to high-demand situations (such as turning at intersections).

The last result points to another factor affecting the validity of the results, the in-lab environment itself [47]. Numerous studies confirm that in-lab settings affect the transfer of findings to on-road conditions due to short session durations [48], [49], overexposure to rare events [50], low risk [51], small subject groups, and lack of diversity within them [52].

B. Drivers' Inattention

Due to associated safety risks, most of driving literature is dedicated to inattention rather than attention. According to the commonly accepted definition by Regan *et al.* [53], *inattention* during driving is operationalized as "insufficient, or no attention, to activities critical for safe driving".

1) Taxonomy of Inattention Types: Besides the definition of inattention, Regan *et al.* [53] provide a taxonomy of inattention types (Figure 1) that distinguishes between five subtypes of inattention: 1) restricted attention (due to physical obstructions or blinks), 2) misprioritized attention, 3) neglected attention (*e.g.* not checking the blind spot while changing lane), 4) cursory attention (looking in the right direction but failing to process the information), and 5) diverted attention (distraction by driving-related or non-driving-related tasks and events). Restricted attention and attention diverted towards non-driving-related tasks are two types that have been investigated theoretically and modeled in practice (*e.g.* drowsiness [54]–[56] and distraction [7], [57]–[60]). Other types of inattention, such as misprioritized, cursory, or neglected attention, can only be identified in hindsight after a safety-critical situation has occurred and are less studied [61].

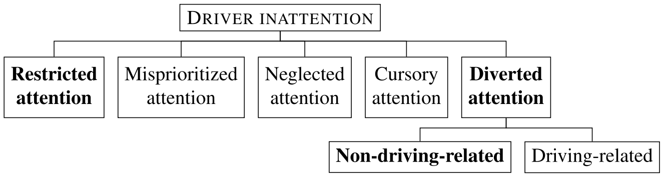


Fig. 1. Driver inattention taxonomy by Regan *et al.* [53]. Inattention types shown in bold are the focus of this review.

2) Types of Non-Driving-Related Tasks (NDRT): Two ways of grouping NDRTs have been proposed: by type (*e.g.* cell phone use, radio tuning, smoking), to determine which activities are more prevalent and pose more risk, and by demand, which includes primary modality (visual/auditory), interaction (active vs. passive), interruptibility (easy/difficult), and coding of information (verbal/spatial) [62]. Demand-based categorization is more common and better reflects what cognitive functions are affected. Based on *modality*, most tasks can be represented as a combination of one or more of the following [63]: 1) *visual* - requires averting gaze off the road (*e.g.* checking the speedometer); 2) *cognitive* - requires thinking (*e.g.* talking to the passenger or recalling information); 3) *manual* - requires taking hands off the wheel (*e.g.* smoking, drinking).[1] Demand-based categorization agrees with evidence of limited attentional resources, wherein performance in multiple tasks is reduced when those tasks compete for the same resources [66]. For example, driving as a visuo-manual activity, is affected by concurrent visual, manual, or visuo-manual tasks, although cognitive distractions can have a negative impact as well [67].

IV. Attention Estimation for Driver Monitoring

Vision-based DMS require an accurate estimate of drivers' attention towards areas inside the vehicle (to determine drivers' state and actions) and elements of the traffic scene (to identify what the driver is aware of). In this section, we review methods for in-vehicle gaze estimation (Section IV-A), inattention detection (Sections IV-B and IV-C), and action anticipation (Sections IV-D) framed as classification problems. Methods for driver awareness estimation are discussed in Section IV-E.

A. In-Vehicle Gaze Estimation

The problem of in-vehicle gaze estimation is commonly framed as multi-class classification, *i.e.* categorizing features

[1] Other proposed modality types include *biomechanical* (manual) and *auditory* [64], and *emotional* [65]. Here, we consider distractions that do not involve manual actions or visual input as *cognitive*.

related to drivers' gaze or head position with respect to predefined areas of interest (AOIs) within the car interior. It is also possible to solve this problem analytically by determining the driver's 3D gaze direction and its intersection with the 3D model of the vehicle interior, but only a few approaches do so [68], [69]. Thus most research in this field focuses on finding the best combination of features (*e.g.* head pose, gaze) and classifiers. Figure 2 shows reviewed algorithms for in-vehicle gaze estimation grouped by features they use.

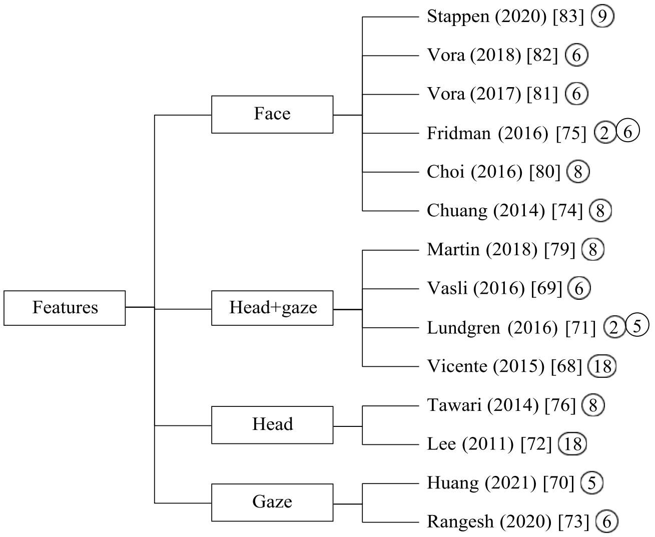


Fig. 2. In-vehicle gaze estimation models organized by the visual features: *Gaze* - gaze coordinates or eye crop, *face* - face image/landmarks, *head* - 3D head pose. Number of AOIs is shown in circles.

1) Input Sources and Feature Extraction: Specialized hardware such as eye trackers is useful for obtaining high-precision drivers' gaze direction and head pose (as in [70], [71]). Video cameras can extract similar data from images of drivers' faces but with lower precision and limited to predefined areas. At the same time, low-cost and no need for maintenance make cameras more suitable for assistive and monitoring technology. Near-infrared (NIR) imaging cameras are also used in some works, alone [72] or combined with a visible light camera [73], to make the system suitable for night driving conditions.

Feature extraction pipeline should ideally satisfy the following criteria: real-time runtime, short processing chain to avoid accumulation of error, and informativeness of features for classification of gaze zone. When using camera images, the following series of steps can be followed to extract drivers' 3D gaze direction [68]: 1) detection and tracking the driver's face; 2) detection of facial landmarks and eye features (cropped image of eyes, iris, and/or pupil location); 3) estimation of head pose (roll, yaw, and pitch angles) from facial landmarks; 4) estimation of 3D gaze vector using eye and head pose models; 5) finding the intersection point between 3D gaze direction and 3D model of the vehicle interior. Some of these steps can be omitted or simplified by use of machine learning techniques discussed below.

2) Classifiers: In the literature, the minimal set of features used for this problem consists of facial landmarks [74], [75] or head poses extracted from the landmarks [72], [76]. These features can then be fed into a classifier to determine a gaze zone. The experimental results indicate that although their computation can be performed in real-time, facial landmarks and 3D head position features cannot reliably differentiate between neighboring zones, such as a speedometer and windshield [76] or side mirror and the window next to it [72]. In both cases, the drivers either made small head movements or did not move their heads and performed eye movements instead. Temporal filtering [74], aggregating features over time [76], and feature normalization [75] led to improved performance but ultimately did not resolve the issue, leading to the conclusion that eye features are also necessary. Fridman *et al.* [77] in a thorough study estimated that eye information contributes a 5.4% increase in average accuracy, and an even larger boost of 20% was reported in [78].

Deep learning models have the advantage of combining feature extraction and classification steps. Instead of the explicit processing pipeline described above, a single convolutional neural network (CNN) pre-trained on the image classification task can be used to classify cropped frames from driver-facing videos [80]–[83]. These CNN-based models reach high accuracy and can discriminate adjacent areas better than previous methods that relied on hand-crafted features. In one study, Vora *et al.* [82] experimented with multiple face cropping methods and CNNs and determined that the upper half of the face provided optimal information. Another advantage of using CNNs is that they perform well even with uncropped images [81], thus saving computational costs associated with face detection and tracking.

3) Evaluation and Limitations: Since in-vehicle gaze estimation is a classification problem, metrics, such as accuracy, F1-score, and confusion matrix, are commonly used to evaluate the models. Accuracy and F1-score provide global performance assessment, while the confusion matrix shows accuracy per area of interest (AOI) and which areas are misclassified.

In the literature, there are large differences in the number of AOIs defined inside the vehicle: from 2 zones (driving- and non-driving related as in [71], [77]) to 18 [68], [72], with 6-8 being the most common (although the justification for the particular choice is rarely given). Naturally, more fine-grained zoning is challenging and leads to worse performance since it is more difficult to localize gaze within a smaller area [75]. As an alternative to fixed AOIs, Huang *et al.* [70] propose to cluster drivers' gaze into zones customized for individual drivers. The downside of this method is that some potentially important areas such as rear-view and side mirrors may be excluded if some drivers ignore them.

Although most models achieve high classification accuracy, some challenges remain. For example, some drivers attend to different zones by moving their heads while others move only their eyes, as noted in [72], [74] and more thoroughly investigated in [77]. Such individual differences can be captured by training the models on user-specific data rather than data aggregated across many drivers [70], [74], [75], but individual data may not be readily available. A more generic solution combining head pose and gaze direction information has been shown to mitigate this issue [76], [78]. Nevertheless,

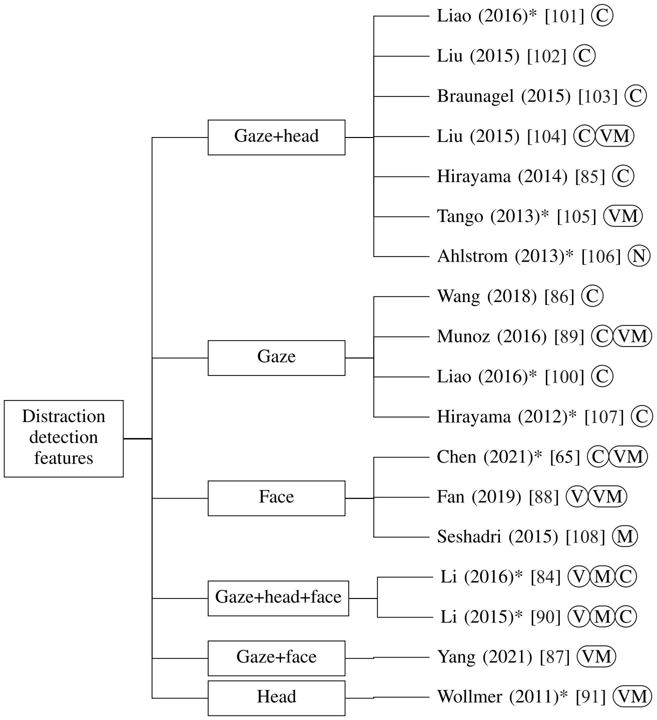


Fig. 3. Distraction detection algorithms grouped by the visual features and types of distractions they can detect. Features: *gaze* - gaze coordinates or eye crop, *face* - face crop/landmarks, *head* - 3D head pose. Distraction types: Ⓒ - cognitive, Ⓥ - visual, Ⓜ - manual, (VM) - visuo-manual, Ⓝ - non-specific. Algorithms marked with * use additional features, *e.g.* vehicle or context.

some adjacent zones remain difficult to distinguish, particularly windshield and speedometer, due to their proximity and subtle eye or head movement required to switch between them [68], [69], [71], [77], [78]. More recent CNN-based models appear to suffer less from misclassifications of these kinds [82].The presence of eyewear causes another challenge. Glasses introduce glare and occlusions, making it difficult to estimate gaze direction [81]. In [73], a pre-processing step is added to remove eyewear via a gaze-preserving generative network, however, it cannot handle thick glass frames and glare.

B. Distration Detection

Timely and reliable driver inattention detection is a prerequisite for driver monitoring systems (DMS) that aim to improve road safety by alerting drivers to dangerous behaviors. Distraction detection algorithms summarized in Figure 3 exploit changes in drivers' gaze patterns caused by secondary task involvement, *e.g.* taking eyes off the road to look at their phone or cognitive distractions. Similar to in-vehicle gaze detection, distraction detection can be solved as a multi-class classification problem. Because of differences in behavioral changes depending on the kind of distraction, the majority of the algorithms focus either on detecting a specific distraction type or distinguishing between different distractions.

1) Types of Distractions: Cognitive and visuo-manual distractions (see Section III-B) are more commonly investigated NDRT types among the reviewed papers. Unlike visual and manual distractions, cognitive tasks are difficult to induce and verify. Furthermore, tasks used to test cognitive distractions vary significantly from study to study. Some imitate natural activities, *e.g.* conversations [84] and voice-based playlist retrieval [85], and some include artificial tasks such as math quizzes [65] and n-back tasks [86]. Visual-manual tasks considered in the studies are everyday activities, such as using a cell phone for reading [87], [88] or sending messages [65], [88], radio tuning [89]–[91], and selecting a song from the playlist [91]. Since NDRTs differ in how they affect the driver and how they manifest themselves visually, algorithms must be tested on a variety of tasks to ensure robust distraction detection [87], [88], [91].

2) Input Sources and Feature Extraction: Features such as gaze coordinates or coarse AOIs (see Section IV-A) are naturally indicative of visual distractions. According to the evidence from psychological studies, longitudinal [92]–[94] and lateral [95]–[99] vehicle control measures are also sensitive to various types of distractions. Therefore, ego-vehicle information, *e.g.* speed and steering wheel rotations, is often used in addition to visual features [65], [84], [90], [91], [100], [101]. Given that most secondary activities are not instantaneous and affect the temporal distribution of gaze differently, nearly all distraction detection algorithms use temporal data, with observation lengths ranging from 2 to 10s.

Many of the reviewed algorithms use gaze and head pose features obtained via an eye-tracker [86], [100], [103], [104], [106], [108] or manually annotated [85], [89], [105], [109]. Fewer approaches incorporate explicit feature extraction pipelines (such as those discussed in Section IV-A). For instance, [87], [88], [108] rely on facial landmarks and [91] uses head pose extracted from driver-facing camera images.

Approaches that rely on gaze data alone process raw gaze coordinates to compute various statistical functionals (*e.g.* mean, standard deviation, percentiles) and apply feature selection to find the optimal set of features for a specific distraction type. For instance, Wollmer *et al.* [91] showed that head rotation angle and its derivatives are sensitive to visual-manual tasks, whereas Liao *et al.* [101] determined that gaze locations are useful for detecting cognitive distractions.

In addition to gaze and visual features, information such as vehicle data [84], [100], [101], [110], physiological signals [65], audio (to detect conversations) [84], detected objects in the scene [87], [108], and mirror-checking actions [84] have been shown to improve the accuracy of distraction detection.

3) Classifiers: Support Vector Machine (SVM) is the most commonly used classifier for this problem, used in nearly half of the reviewed algorithms [84], [100]–[106], [108]. Other approaches, such as boosting [84], extreme learning machines [102], [104], K-means [90], and Hidden Markov Models (HMM) [89] have demonstrated high performance on the distraction detection task. Recent works that use deep learning methods, such as recurrent [86], [88], [91] and convolutional neural networks [65], [87], [88], demonstrate their superior performance across most metrics and ability to extract information directly from raw images or multi-modal data.

However, existing datasets provide only a limited set of non-driving related tasks (NDRTs), therefore, features and classification approaches tend to be optimized for a narrow range of distractions. A solution proposed in [111] is more versatile as it does not focus on specific activities, but rather relies on off-road glances to detect inattention. Motivated by evidence that long off-road glances increase crash risk [112], this model uses a 2s buffer to track distractions and issue sound alarms. The buffer shortens whenever the driver looks away from the road and lengthens when they look at the road again. Driving-related glances away from the road (*e.g.* towards mirrors) are treated with a latency of 1s to prevent issuing unnecessary warnings.

4) Evaluation and Limitations: Standard classification metrics, such as accuracy, precision, recall, and F1-score, are commonly used to evaluate distraction detection models. Most achieve high accuracy and F1-scores, often over 90%, some even close to 100% [88]. However, the results of different algorithms are not directly comparable due to the prevalent use of private unpublished data and the lack of public benchmarks for this problem. Although most authors specify the number of subjects, their age, gender, and driving experience, the volume and properties of the data used for the evaluation are often defined imprecisely and in different units, *e.g.* duration [101], [103], number of video frames [108], or number of events [107]. Direct comparisons are also affected by inconsistent recording conditions across studies, *e.g.* in-lab driving simulators [101], parked vehicles [87], or on-road settings [88].

Overall, despite encouraging results, the problem of distraction detection is far from being solved since drivers' gaze distribution depends on the context and is subject to individual differences. For example, different sets of features are needed for urban and highway roads [100], therefore thorough testing in various environments is desirable [91], [105]. Driving tasks, such as vehicle following and passing, also affect gaze distribution patterns and have to be taken into account when modeling distraction [110]. Although there are commonalities in how distractions manifest themselves in different drivers [89], the system should consider individual user characteristics for the best results [84], [105].

Although the purpose of designing distraction detection algorithms is for use in vision-based ADAS, only few of the reviewed systems have been verified in practice. For example, a monitoring algorithm proposed in [88] was tested in a driving simulator to validate the effect of sound alarms on engagement in non-driving related tasks (*e.g.* texting, reaching for objects, and eating). The subjects played a truck driving game while performing NDRTs at 30s intervals. Sound alarms reduced the number of accidents and traffic tickets, however, experimental conditions were far from realistic. A more extensive field study was conducted for AttenD algorithm [106]. It involved 7 subjects who drove a test vehicle equipped with the system for one month. Despite the small subject pool, the overall changes to subjects' visual behavior were positive and pointed towards increased attention to the road ahead. Some issues were also exposed, such as data loss due to large head movements and excessive alarms caused by not taking into account drivers' intent to change lanes or brake (especially at lower speeds). Given that warning system acceptance by users is reduced significantly by false alarms [113], [114], it is important to consider HCI issues when developing monitoring algorithms and conduct user studies besides evaluation on datasets. Refer to Section VII for further discussion.

C. Drowsiness Detection

Drowsiness detection methods rely on the driver's appearance to detect signs of fatigue such as frequent blinking, closed eyes, yawning, and nodding. Similar to distraction detection, detecting drowsiness is framed as a classification problem, either binary (drowsy/alert) or multi-class for more fine-grained alertness states.

1) Input and Feature Extraction: Most algorithms rely on driver-facing video cameras to detect drowsiness. Near-infrared imaging (NIR) cameras are often used (along [118], [120], [128] or combined with visible imaging color cameras [130]) due to their versatility for day and night conditions and robustness to changes in illumination and low light conditions.

Eye, mouth, and head features are considered to be the most effective for estimating drowsiness [143], [146]. These features are detected using methods described in Section IV-A, such as face detection [121], [146] and tracking [117], locating facial landmarks [123], [146], detecting blinks [124] and eye closures [123], recognizing mouth drooping and yawning [123], and measuring head pose and head movement [124]. Eye features (blinks and closures) are further processed into standard drowsiness measures such as percentage of eye closure (PERCLOS [149]) [140], [143], [145] and blink frequency [143], [147], [148].

Detection of many drowsiness symptoms, such as slow blinking and yawning, is further improved by aggregating features across time [116]. Longer time intervals up to several minutes for the blink and eye closure features typically work best [149] but also increase the risk of missing microsleeps and "blank stares" [150], thus alerting the driver too late. Another approach is to use multiple measures, however, the design of the algorithm should account for different measures of drowsiness having different tendencies [135]. Inclusion of global context features such as continuous driving time, temperature, current time, and sleep duration has also been shown to further improve drowsiness detection [131], [137].

2) Classifiers: Rule-based approaches and thresholding are computationally the simplest methods for detecting drowsiness [132], [140], [141], [148], however, they are susceptible to differences between drivers and variations of signal due to changes in illumination and vibration of the vehicle. Learning methods such as SVM [135], [137], [145], boosting [128], or HMMs [124], are more robust in practice.

Recently, deep learning methods have been applied to drowsiness detection. For instance, Zhao *et al.* [146] use a deep belief network to classify drowsy facial expressions using a concatenation of facial landmarks and features from cropped images of drivers' eyes and mouths. Weng *et al.* [123] instead of combining the features used three DBNs to encode mouth, head, and eyes features, and HMMs to learn relationships between them for alert and drowsy states. In order to capture temporal dimension, Shih *et al.* [121] aggregate per-frame features extracted via CNN over 50 frames and feed them into a recurrent network, followed by additional temporal smoothing. Yu *et al.* [118] utilize 3D CNNs to extract generic spatio-temporal features in a single feed-forward pass. These features are then processed to extract specific information, such as the presence of eyewear, and condition of head, mouth, and eyes. Specific and generic features are fused via a feed-forward network for final drowsiness classification.

3) Evaluation and Limitations: A significant limitation of research in this field is the prevalence of private datasets. The only widely used public dataset is Driver Drowsiness Detection (DDD) [123], however, as discussed later in Section VI, it is recorded in lab conditions while subjects act drowsy. Many private datasets are also recorded in artificial conditions, *e.g.* lab [119], [127], [133] or parked vehicles [124], [148], and only a few are captured in on-road conditions, typically highways and rural roads with little traffic [132], [137], [141]. One study used recordings of drowsy passengers instead of drivers due to safety concerns [146].

The lack of benchmarks and publicly available naturalistic data make it difficult to establish state-of-the-art performance for drowsiness detection and their suitability for practical use, respectively. Despite reporting excellent results, many algorithms still struggle with certain aspects of the problem, notably extreme head angles [144], [146] and glare and occlusion from eyewear [117], [118], [143]. Individual differences across drivers can also diminish the accuracy of drowsiness detection. For example, specific signals like blink patterns are subject to considerable individual variations [141] and difficult to detect when participants have smaller eyes [148]. Vehicle vibration and variability of driver positions with respect to cameras further exacerbate these problems [144]. Furthermore, some drivers do not show visible signs of drowsiness even when fatigued [130]. More diverse publicly available multi-modal datasets collected in naturalistic conditions are part of the solution to the problems listed above [124], [130]. On the algorithmic side, including more features and personalization can potentially improve drowsiness detection results and the overall reliability of the proposed solutions but requires additional computation resources and calibration [130], [148].

Other issues are methodological. For instance, there is little agreement in the literature on inducing drowsiness. A common approach is to ask the drivers to yawn and act drowsy [119], [124], [126], [127], [129], however, there is a risk that such data is not representative of more realistic conditions. Alternatively, drowsiness and fatigue can be induced by extended [132], [137] or monotonous [131] driving sessions, driving late at night [130], or reducing sleeping hours prior to experiment [145]. Some studies also employ night shift workers for collecting naturalistic data [144], [147].

Another challenge is providing the ground truth for the recorded data since both defining types of drowsiness and recognizing them in human subjects are not trivial. Concerning the former, the Karolinska Sleepiness Scale (KSS) [151] is a standard scale defining drowsiness levels on a scale from 1 (fully alert) to 9 (fully asleep). For recognition tasks, all 9 levels are rarely used, and instead, coarser scales are formed by combining the intermediate levels. However, there is no consistency in the literature on how it is done. For example, [130] collapses the KSS scale into 2 levels, [144] uses 4, and [137] 3 levels. Likewise, there is no agreement on which KSS levels should be combined together. Depending on the study, the alert state can extend to KSS levels 3 [129], 4 [137], or 6 [143]. Intermediate levels of drowsiness may include KSS levels 6-7 [129], [143], 5-7 [137], or only 7 [147]. Only extreme drowsiness is consistently associated with KSS levels 8-9 across studies [129], [137], [143], [147]. Sleepiness scales other than KSS have also been used, *e.g.* custom 3-level scale [146] or Zilberg 4-level scale [152] in [135], although simple binary drowsy/alert assessment is the most common [116], [117], [125], [126], [134], [140], [142], [148]. The latter approach may be justified since intermediate drowsiness levels are easily misclassified as drowsy [129].

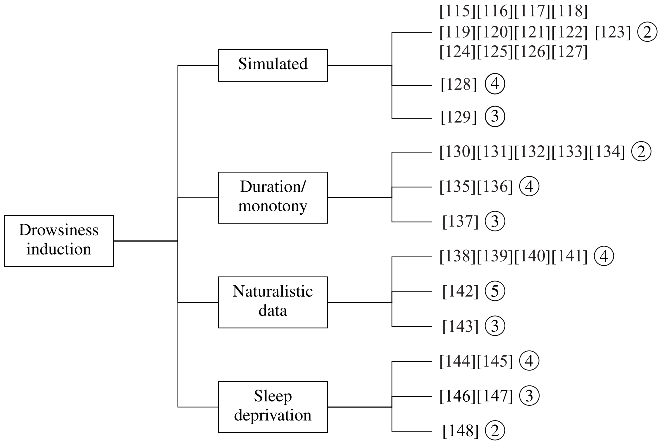


Fig. 4. Drowsiness detection algorithms grouped by the type of drowsiness induction method and number of drowsiness levels (specified in circles).

Given a specific drowsiness scale, the next step of assigning labels to the data is not straightforward. Some studies rely on self-reported drowsiness scores [130], [143], [147], observer-rated sleepiness [135], [144], or both [137]. According to evidence from psychological experiments, neither is bias-free: observer ratings may not be reliable [153], [154] and self-rated sleepiness does not always correlate with driving performance [155]. Simulated data where drivers acted drowsily is devoid of issues with assigning ground truth and is a preferred choice for most studies (Figure 4). A question remains whether such data is realistic enough. For successful application in practice, user studies and validation experiments of various sleepiness assessment methods in different contexts are needed.

D. Driver Maneuver Recognition and Prediction

Recognizing and predicting drivers' actions is another valuable feature for driver monitoring and assistive technology. Knowing what the driver is doing or intends to do next can help direct their attention to the right objects and reduce unnecessary warnings. Since drivers' gaze is linked to the goal and actions being performed [156], it can be exploited to recognize and anticipate drivers' maneuvers.

1) Feature Extraction and Classification: Similar to distraction and drowsiness detection, action recognition and prediction are often framed as a multi-class classification problem: features aggregated across observation time and classifiers are used to predict the upcoming maneuver.

For this task, the approximate direction of drivers' gaze is often used unless eye-tracking data is available as in [157]. The processing pipeline for obtaining gaze features usually includes face detection and tracking, followed by facial landmark detection, extraction of gaze zones [79], [84], [158], gaze duration, frequency, and blinks [79], [84]. Alternatively, an implicit representation of gaze can be used, such as tracked facial landmarks aggregated over time as proposed in [159], [160] or mirror-checking actions [84].

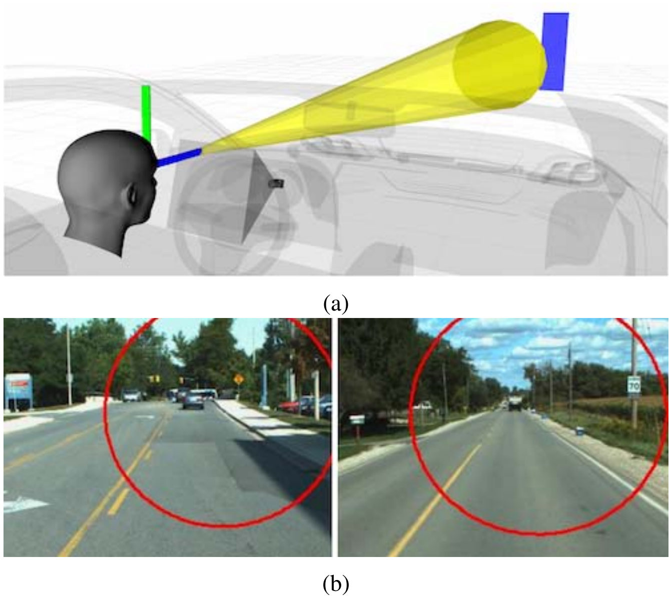
A variety of methods have been proposed to classify actions based on the temporal features above. For example, Li *et al.* [84] use boosting [161] with a combination of mirror-checking actions and vehicle dynamics features. Martin *et al.* [79], [158] model maneuvers using a multivariate normal distribution (MVN) of spatio-temporal descriptors that capture gaze duration towards relevant AOIs. Besides discriminative models, temporal modeling that fits the data more naturally has also been applied. Jain *et al.* [159] and Akai *et al.* [157] propose auto-regressive input-output Hidden Markov Models (HMMs) to classify driver's actions given driver gaze and vehicle dynamics. Recurrent networks are also effective for multi-modal data [160] but lack the explainability of HMMs.

2) Evaluation and Limitations: Since action prediction is typically framed as classification, common metrics such as precision, recall, and F1 are used to evaluate the results. As expected, it is more difficult to predict maneuvers several seconds in advance, thus precision and recall improve as time-to-maneuver (TTM) decreases [158]. Besides issues with face detection and tracking due to illumination changes [159], some maneuvers are generally more difficult to predict because of the overlap in behaviors (*e.g.* mirror checking is not always a precursor to lane changing [158]) and lack of visual cues from the driver (*e.g.* when they are familiar with the route or make a turn from the dedicated lane [159]). Inclusion of scene and route information may help handle such cases.

E. Driver Awareness Estimation

Inattention detection systems discussed so far do not consider the environment and what the driver is aware of. However, a better understanding of driver behavior, current task, and context is desirable for more effective driver assistance systems that could, for example, verify whether the driver attended to relevant objects and provide situation-specific warnings. This section reviews systems that make steps in this direction by associating driver attention to objects of interest in the traffic scene.

1) Visual Features and Processing: Algorithms discussed in this section determine drivers' awareness of vulnerable road users, signs, and traffic signals. Most of them follow a similar procedure: 1) detect drivers' 3D gaze direction, 2) convert gaze to vehicle's frame of reference, 3) detect objects in the scene and their properties, and 4) match drivers' gaze with objects to identify whether they were fixated.



(a)

(b)

Fig. 5. Visualization of the attention cone: a) view from inside the vehicle, b) projection of the cone onto the traffic scene. Sources: a) [162], b) [163].

Drivers' gaze direction estimation uses approaches discussed in Section IV-A, whereas detecting objects in the scene relies on off-the-shelf algorithms [164], classical vision pipelines [163], [165], [166], or manually annotated bounding boxes [167]. Distances to the detected objects and their relative velocities may be inferred from a stereo camera [166], provided by range sensors [166], [168]–[171], or determined using simple heuristics [164].

Different strategies have been proposed for detecting what objects the driver observed. The simplest solution is to check whether the driver's gaze falls within the object's bounding box [164], [167], [170]. Since inaccurate measurement of 3D gaze direction can result in large errors, especially for objects far away, the authors of [163], [166] propose to treat attention as a cone projected from the drivers' eyes towards the windshield (Figure 5a). Some algorithms also take into account that drivers retain information about objects for some time after looking at them [164], [171], [172], as well as other properties of the scene, such as weather and proximity of other road users [172]. For example, Schwehr *et al.* [168], [173] model the joint probability distribution of the object states in the 2D vehicle coordinate system, object coordinates, and the driver's gaze direction in 2D to estimate which objects have been fixated or tracked. Ahlstrom *et al.* [172] modify the AttenD algorithm (described earlier in Section IV-B) to include elements of context via additional buffers for targets of relevance which, besides traffic ahead and behind, include intersections. Properties, such as proximity to other road users and weather adaptations, are also accounted for in the model. Zhu *et al.* [174] use SAGAT [175], a method for measuring situation awareness (SA), to associate it with various gaze-, memory- and object-related features. A combination of these three types of features achieved over 70% agreement with human SA results.

2) Evaluation and Limitations: Evaluating driver awareness models is not trivial, however indvidual modules can be evaluated quantitatively. For example, measuring gaze estimation error [176], [177], road user trajectory prediction accuracy [162], object detection [166], [178], *etc*.. In contrast, there are no unified approaches for evaluating the performance of the entire system. Cross-model comparisons are virtually impossible since algorithms do not share the same definition of outputs, objects of interest, or application scenarios. Typically, a qualitative evaluation of the individual models is provided based on several illustrative scenarios [162], [165], [166], [168], [169], however it gives little idea of how robust, effective, and usable this system might be. Although some algorithms were tested with human subjects, many were not done in realistic driving conditions: some involved staged pedestrian crossing [179], routes on a university campus [164], and tests in driving simulators [172], [180].

In order to make a viable monitoring system that takes into account driver awareness, several fundamental issues that stem from the properties of the human visual system and limitations of recording equipment must be resolved. For instance, establishing the exact point of gaze (PoG) is difficult even with precise eye trackers, as shown in a series of experiments by Schwehr *et al.* [177]. The authors conclude that regardless of the choice of model for projecting the gaze into the scene, the point of gaze is always off the target by tens of pixels and that the error is primarily caused by the imprecise measurement of gaze by eye tracker. Given that eye trackers provide the most accurate data, it is likely that vision-based systems will suffer from the same issue. A cone of gaze used in many studies instead of PoG alleviates some of the imprecision but does not localize the targets well (Figure 5b). An assumption that the drivers detect all objects within the intersection area [163], [166] may not be accurate. According to Kim *et al.* [181], gaze is generally correlated with lower levels of situation awareness (as defined in [182]), but gaze alone is not sufficient to predict SA. A regression model that takes into account proximity of the gaze to the target and awareness score explained only 50% of the variance in the data, therefore other factors should be considered. Additional indicators that may be useful are vehicle control [180] and braking intention [183], as well as detectability of signs [184]–[186] and pedestrians [187] in traffic scenes depending on the visual properties of the objects and the scene, *e.g*. illumination, visual clutter, visibility, *etc*.

V. Models of Attention for Assistive and Autonomous Driving

Algorithms discussed in this section do not detect drivers' gaze direction and state but rather model what objects and events need to be attended to for safe driving. Models intended for use in driver assistance systems rely on human data to predict where safe drivers should look in specific conditions. Algorithms for autonomous driving utilize mechanisms inspired by human attention to focus on what is important to improve decision-making and make it more transparent.

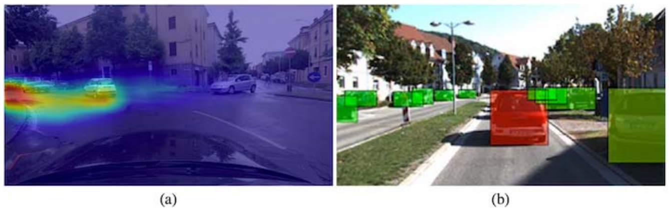


Fig. 6. Sample outputs of driver gaze estimation algorithms. a) Pixel-level saliency map for vehicle performing a left turn. b) Visualization of object-level importance scores for vehicle following. Red color indicates higher relevance/importance in both images. Sources: a) [198], b) [201].

A. Modeling Visual Attention in the Traffic Scene

In the past decade, a number of methods have been proposed for modeling the spatial distribution of drivers' gaze in traffic scenes. Given a single image or a sequence of images of the scene, these algorithms output saliency maps (or heatmaps) where higher pixel values (usually within [0, 1] range) indicate areas of interest, risk, or importance (Figure 6a). Fewer algorithms assign importance scores to specific objects. In this case, higher scores are associated with objects relevant to the ego-vehicle (*e.g*. lead vehicle shown in red in Figure 6b).

1) Bottom-Up and Top-Down Influences: As discussed in Section III, bottom-up and top-down influences affect the distribution of drivers' gaze. Bottom-up (or data-driven) features are associated with unexpected or unusual elements in the scene and are often computed using existing saliency algorithms [188]–[192]. Top-down features are task-specific and hence more varied. For example, [188], [189] use vanishing point because drivers often focus on it to get an optimal view of the road ahead [25], [193]. Other options include drivers' actions [190]–[192], current driving task [194], [195], previous fixation locations [190]–[192], and vehicle telemetry [190]–[192], [196]. Other common features include semantic segmentation maps [197]–[199] and detected objects [196], [200]–[202].

Various spatio-temporal features can be helpful for capturing the dynamic nature of the traffic environment and drivers' gaze changes. For instance, optical flow is useful for identifying the direction and magnitude of motion in the scene [198], [203]–[205]. More recently, following successful applications in video action recognition problems [206], 3D convolutional networks have become a popular choice for encoding spatiotemporal data [194], [199], [202], [207]. Some approaches use recurrent networks, combined with individually encoded frames [208], [209] or with a set of frames processed via 3D convolutional layers [197].

While it is possible to learn associations between individual scene images and human saliency maps without explicit task representation using features extracted via convolutional neural networks [198], [203], [204], [210], [211], the results are difficult to interpret and analyze. Approaches that use bottom-up and top-down features are more transparent as they directly control their influence on the resulting predictions.

For example, a weighted sum of bottom-up saliency maps and high-level features (vanishing point and center bias) was proposed by Deng *et al.* [188], [189]. They later extended this method by learning the weights for individual features using a random forest [189]. Tavakoli *et al.* [203] in experiments with regression models, demonstrate that bottom-up features are weakly correlated with task-driven gaze, however, an ensemble model that combines bottom-up and top-down influences leads to improved results. Borji *et al.* in several works [190]–[192] examined the contributions of different types of features using a Hidden Markov Model. In these experiments, top-down features (*e.g.* actions and previous gaze location) correlated with the human gaze better than bottom-up ones (*e.g.* saliency maps), however combination of both types of features performed best.

Two recent models, HammerDrive [194] and MEDIRL [195], demonstrate that explicitly modeling the underlying driving task is beneficial for gaze prediction performance. In HammerDrive, a separate module recognizes maneuvers (lane change and lane-keeping) from the vehicle telemetry. The result is used to reweight the output of the ensemble of bottom-up saliency predictors. MEDIRL applies inverse reinforcement learning to learn a policy for visual attention given the present agent state, which includes local and global context, and driving task (braking, lane-keeping, and merging).

2) Evaluation and Limitations: Evaluation of driver gaze models follows the procedure and metrics established in free-viewing saliency research (see [212] for a review). These metrics assess how similar are predicted saliency maps to those of human drivers in terms of saliency value magnitudes, statistical distribution properties, and salient locations [213].

To further verify the quality and human-likeness of the generated saliency maps, the following human experiments were proposed: subjects viewed videos with superimposed drivers' gaze or saliency maps produced by the model and were asked to choose which one came from a human driver [198] or was more consistent with good driving practices [208]. Since participants were not able to reliably discern natural and artificial gaze maps, it was interpreted by the authors as evidence in favor of the models. Whether such predicted patterns lead to safer driving remains unclear, particularly when data for training and evaluation is recorded in lab (see discussion in Section III).

A particular challenge related to driving gaze data is that it is comprised of common driving scenarios (*e.g.* vehicle following, driving on a straight road) with few surprising events or interactions with other road users. For example, in DR(eye)VE [198], the drivers encounter relatively few other road users and do not perform maneuvers often, resulting in center-biased gaze distributions. This has consequences for models since they learn the dominant gaze behaviors and thus fail to predict gaze for scenarios that occur rarely. Xia *et al.* [208] propose to mitigate the prevalence of common driving scenarios using two strategies. First, they curate the training data by focusing on abnormal events (*e.g.* braking). Second, they implement a weighted sampling strategy that selects frames with abnormal gaze distribution more frequently during training. To measure how well the models learn the underlying driving task, some authors compute metrics over segments of data where drivers' gaze distribution significantly differs from the mean due to maneuvers or actions of other road users [207], [208].

B. Attention for Self-Driving Vehicles

Self-driving technology aims to improve safety by eliminating human error. But to match or exceed human driver performance, AI-driven systems require solving multiple problems in many areas of computer vision and robotics. Perception alone involves overcoming significant challenges in object detection, tracking, scene segmentation, depth, and optical flow estimation (see [214]). Decision-making for motion planning, behavior selection, and vehicle control rely on precise mapping and localization [215] as well as understanding the behaviors of vulnerable road users [216].

Current self-driving systems that tackle these issues can be broadly subdivided into modular and end-to-end [217]. The former use dedicated modules for various processing stages, whereas the latter are unified systems that convert input from sensors directly to control commands. Due to the overwhelming number of self-driving approaches, we limit the review to end-to-end driving models that use attention for perception and reasoning to improve models' performance and transparency. The role of attention is to identify objects or areas in the scene that are most relevant for the current driving task and safety. The general principle of many attentional mechanisms is reweighting of the features according to a query that could be a literal question or a different set of features (*e.g.* hidden state of the recurrent model representing the current context) [218]. The weights themselves can be used to analyze what parts of the input had a larger influence on the output and investigate intermediate processing for better explainability of the model's decisions.

1) Types and Uses of Attention: Spatial attention is widespread in self-driving models because it retains the spatial arrangement of features and computes attention weights that can be traced back to the locations in the environment. Weights visualized as a heatmap can be interpreted as objects or areas in the scene that were important for the current output of the model. Spatial attention is usually applied to intermediate and final layers of the feature extraction step. For example, in [219], multiple attention modules are inserted after intermediate and last layers of CNN to gradually refine features. Kim *et al.* [220] insert the attentional module after the convolutional feature extraction and also condition spatial attention weights on the hidden state from the previous timestep.

Computing attention weights for image regions and specific objects in the scene instead of pixels or individual features is also possible. Cultrera *et al.* [221] use simple spatial attention that is trained as part of the network. The model first extracts features from the input image via pre-trained CNN,

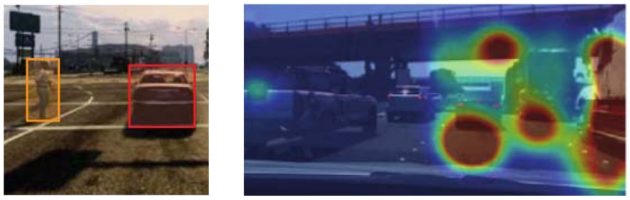


Fig. 7. Visualized spatial attention weights: a) object-level, b) pixel-wise. Sources: a) [222], b) [220].

groups them into coarse regions, and passes them through the pooling layer. Then, an attention block consisting of a fully-connected layer followed by softmax activation produces weights that are element-wise multiplied with the output of the pooling layer. In [222], features corresponding to individual objects in each frame are extracted first. An object-level attention network is then applied to a concatenation of local object feature and global image features to output a scalar score indicating the object's relevance. Top-k objects are then passed to a policy network that outputs a discretized action. He *et al.* [223] and Wei *et al.* [224] propose methods for computing sparser attention weights for input features which results in a more selective and compact focus of attention and reduced computation.

Recently, the Transformer architecture [225] has been shown effective for many vision tasks [226]. Transformers process sequential data without relying on recurrence, and instead use several identical blocks composed of multi-head attention, feed-forward neural network, residual connection, and layer normalization. Attention module is a key element that reweights input according to current task or context. Stacking several attention blocks with different initialization within a multi-head layer allows learning to focus on different parts of the input.

The flexibility of Transformers for various input modalities [226] and the inherent interpretability of learned attention scores [227], [228] lend themselves well to the self-driving domain. For example, Chitta *et al.* [229] use a Transformer to encode image patch features, ego-vehicle velocity, and positional embeddings. An additional Neural Attentional Field module then identifies parts of encoded input that are relevant for query waypoint in the bird's eye view (BEV) image. Waypoints produced by the model are then used to generate vehicle control commands. Prakash *et al.* [230] propose TransFuser, a model composed of multiple transformer blocks for gradually fusing feature maps of different modalities (image and LiDAR BEV) at multiple resolutions. The waypoints generated from these features are shown to be effective in guiding the vehicle. Li *et al.* [231] make use of Transformers for developing a model for perception and prediction from multi-modal data comprised of LiDAR sweeps, images of the scene, and high-definition maps. Features of all detected road users in the scene are passed to the Interaction Transformer module that identifies for every actor all other relevant actors. Experiments on naturalistic driving data show that focusing on the most relevant agents helps reduce the number of collisions.

Some approaches leverage human data for training attention modules. For example, in [245] human gaze is used to train a foveal visual encoder that selects informative locations in the scene, crops patches, and processes them in more detail. The peripheral visual encoder extracts convolutional features from the entire image to provide global context. The two are combined via a planner to produce vehicle speed. In [246], a gaze model trained on human eye-tracking data is used to control the amount and spread of dropout to improve the accuracy of control commands during imitation learning. Gaze-modulated dropout is lower in highly salient areas and higher in irrelevant areas and offers better performance than fixed uniform or center-biased dropout. Instead of the human gaze, Kim *et al.* in a series of works, leverage human textual annotations using visuo-linguistic techniques. In [239] they propose an explanation module for the vehicle controller that generates a textual explanation for the action and a spatial attention map that highlights relevant regions in the scene image. In [238] the authors use textual advice to generate vehicle control commands and spatial attention maps that influenced the decision. In the most recent paper [247], they use attention for simultaneous generation of control commands using natural language and textual and visual explanations.

2) Evaluation and Limitations: The driving performance of self-driving algorithms is evaluated in simulated environments [229], [230], [248], using pre-recorded datasets [220], [224], [247], [249], or both [222]. In simulations, safety metrics measure the number of collisions or infractions and the distance traveled between them [222], [229], whereas on pre-recorded data, metrics assess how well the algorithm matches vehicle data, *e.g.* in terms of acceleration, heading, or generated ego-vehicle trajectory [231], [239], [249].

Performance of attention is evaluated quantitatively via ablation studies to show improved vehicle control [224], [231], [245], [247], [248], reduction in traffic incidents [222], [224], [230], [248], and match with human data, such as gaze [245] or textual annotations [247].

Qualitative evaluations of attention modules commonly use visualizations primarily to demonstrate that the algorithm focuses on portions of the environment relevant for safe driving. Object-level attention (Fig. 7a), where importance scores are visualized for individual objects, is easier to interpret and is more common [222], [224], [229]–[231]. Pixel-wise attention scores (Fig. 7b) used in some studies [220], [245], [249] often do not match specific objects in the scene and, in some instances, do not adequately reflect the decisions made by the system [220].

VI. Datasets

High-quality publicly available data are crucially important for applied research, particularly for a complex and dynamic task such as driving. As discussed in previous sections, driving data must capture a wide range of scenarios and conditions, as well as sufficiently large and diverse pool of participants. Thus large-scale data are a must for adequate evaluation of models and benchmarking the overall progress. This section covers a number of public datasets for a range of applications and properties of drivers' attention they represent (Table I).

TABLE I

PUBLIC DATASETS FOR STUDYING DRIVERS' (IN) ATTENTION WITH LINKS TO THE CORRESPONDING PROJECT PAGES AND DATA PROPERTIES. THE DATASETS ARE SORTED BY AVAILABILITY OF EYE-TRACKING DATA AND YEAR OF PUBLICATION (IN REVERSE CHRONOLOGICAL ORDER). THE FOLLOWING ABBREVIATIONS ARE USED IN THE TABLE. VIDEO DATA: S - SCENE-FACING CAMERA, D - DRIVER-FACING CAMERA, RGB - 3-CHANNEL IMAGE, IR - INFRA-RED, DEPTH - DEPTH SENSOR, MOCAP - MOTION CAPTURE. ANNOTATIONS: TL - TEXT LABELS, BB - BOUNDING BOXES, FL - FACIAL LANDMARKS, OCCL - OCCLUSION, SEM - SEMANTIC SEGMENTATION MAPS. FRAME COUNTS MARKED WITH * ARE ESTIMATED BASED ON THE LENGTHS OF THE VIDEOS AND CAMERA FRAME RATE

Dataset/Reference	Driver inattention	Hazards	Video data	Eye-tracking	Vehicle data	Annotations	Recording conditions	Vehicle control	# subjects	# frames	# videos
MAAD [205]	+		S^{RGB}	+	+	TL	simulator	-	23	60K	8
TrafficSaliency [211]			S^{RGB}	+			simulator	-	28	77K*	16
DADA-2000 [232]		+	S^{RGB}	+		BB, TL	simulator	-	20	658K	2000
DR(eye)VE [198]	+		S^{RGB}	+	+		on-road	+	8	555K	74
BBD-A [208]		+	S^{RGB}	+	+		simulator	-	45	378K*	1232
C42CN [233]	+	+	S^{RGB}	+			simulator	+	68	-	456
TETD [188]			S^{RGB}	+			simulator	-	20	100	
3DDS [190]			S^{RGB}	+			simulator	+	10	192K	108
LISA v2 [73]	+		$D^{RGB, IR}$				simulator	+	13	3.4M	-
DGAZE [234]			S^{RGB}, D^{RGB}			BB	simulator	-	20	100K	20
DGW [235]	+		D^{RGB}			TL	on-road	-	338	-	586
DMD [236]	+		$D^{RGB, DEPTH, IR}$			TL, BB	on-road, simulator	+	37	4.4M*	-
Drive&Act [237]	+		S^{RGB}			SEM, TL	simulator	+	15	9.6M	29
HAD [238]			S^{RGB}		+	TL	simulator	-	-	3.4M*	5675
RLDD [129]	+		D^{RGB}			TL	simulator	-	60	3.2M*	180
BBD-X [239]			S^{RGB}		+	TL	simulator	-	-	8.4M	6984
HDD [240]			S^{RGB}		+	BB, TL	simulator	-	-	1.2M*	137
DDD [123]	+		$D^{RGB, IR}$			TL	simulator	+	36	486K*	360
DAD [241]		+	S^{RGB}			BB, TL	simulator	-	-	175K	620
Brain4Cars [159]	+		S^{RGB}, D^{RGB}		+	TL	on-road	+	10	2M	-
DROZY [242]	+		$D^{RGB, DEPTH}$			TL, FL	simulator	-	14	500K	-
DIPLECS [243]			S^{RGB}		+		on-road	+	1	54K*	1
YawDD [244]	+		D^{RGB}			TL, BB	simulator	-	107	-	342

A. Driving Datasets With Drivers' Eye-Tracking Data

Datasets with gaze recording accompanied by driving footage are naturally relevant for studying drivers' attention. Within this group, DADA-2000 [232] and BDD-A [208] focus on hazardous scenarios, C42CN [233] on secondary non-driving tasks, and the rest (MAAD [205], TrafficSaliency [211], DR(eye)VE [198], C42CN [233], TETD [188], and 3DDS [190]) provide gaze data for everyday driving. Hazard perception datasets contain short clips featuring abnormal events, whereas everyday driving datasets consist of longer video recordings.

Only DR(eye)VE dataset is recorded on-road, however due to difficulties in replicating the routes and traffic conditions across subjects, each video is associated with only one driver's gaze recording. 3DDS and C42CN recorded in a low-fidelity driving simulator aggregated gaze data from multiple subjects.

Eye-tracking data for the remaining datasets were recorded while subjects passively viewed driving footage on a computer monitor. As mentioned in Section III-A, such conditions lack ecological validity compared to on-road driving but are replicable across many subjects. As a result, there are measurable differences in gaze allocation between the two setups. For example, Xia *et al.* [208] reported that subjects who passively viewed videos from the DR(eye)VE dataset looked at more driving-related objects than the drivers whose gaze was recorded originally. Further analysis is needed to establish whether these changes are significant and how they affect safety.

B. Driving Datasets Without Eye-Tracking Data

Datasets in this group usually contain videos from driver-facing cameras from which gaze may be inferred or driving videos with driver attention annotations. Multiple techniques have been developed to associate recordings of drivers with the attended areas inside and outside the vehicle. For naturalistic driving data, the most common method is manual coding of gaze from driver-facing camera recordings. To avoid a labor-intensive and error-prone annotation process, drivers are instructed to look at specific areas or markers in the vehicle or the scene while seated in the parked vehicle. This approach is taken in LISA v2 [73], DGW [235], and DMD [236] datasets. In LISA v2, subjects were filmed under different lighting conditions (daytime, nighttime and harsh lighting) with and without eyeglasses. Additionally, the subjects were asked to rotate their head to capture head motions typical for actual driving. The DGW dataset followed a similar procedure

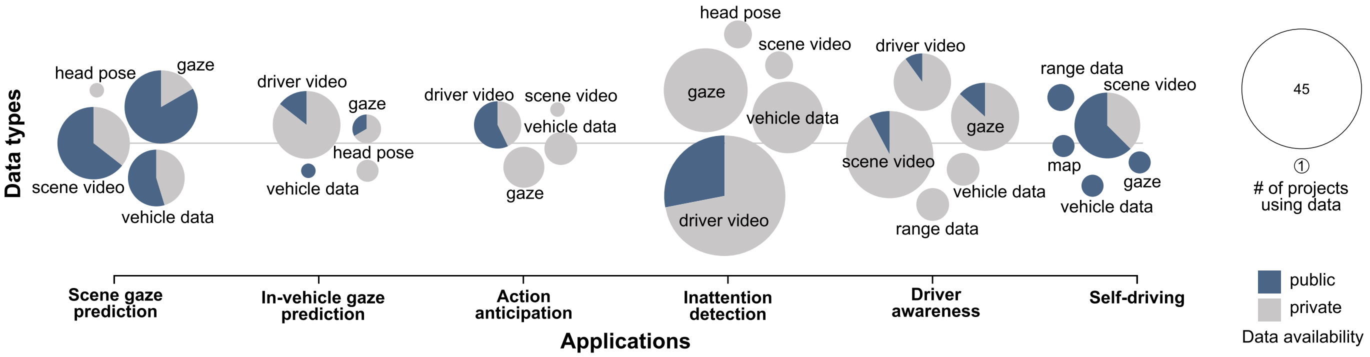


Fig. 8. Types of data, data availability, and use in applications. The size of the circles reflect the number of publications using the corresponding data type for a specific application. Within each circle, the proportion of public and private data is shown.

using a much larger and diverse pool of participants. To automate labeling, participants were asked to look at one of the 9 markers placed in the vehicle and speak the corresponding zone number, which was transcribed using a speech-to-text network. DGAZE dataset uses an in-lab setup where videos of the drivers are captured against the backdrop of the vehicle interior while they are looking at the annotated objects in the traffic scene. This makes it possible to associate drivers' appearance and the objects they attend to. However, it is yet to be determined whether such an in-lab setup will translate well to realistic on-road conditions.

Datasets without a driver video stream provide textual labels for drivers' actions and attention allocation in terms of objects and events in the scene deemed important by the annotators. For example, HDD [240], HAD [238], and BDD-X [239] provide causal explanations for drivers' actions, *e.g.* the presence of crossing pedestrians, or a vehicle ahead slowing down. Besides textual descriptions, HDD also provides bounding boxes for objects that the driver should look at when performing maneuvers. DAD [241] focuses on more extreme scenarios and contains videos of accidents recorded via dashboard cameras. The annotations include textual labels specifying the type of accident, temporal labels indicating when the dangerous situation occurs, and bounding boxes for important objects.

A number of datasets have been proposed for studying driver inattention due to drowsiness or involvement in secondary tasks. DROZY [242], YawDD [244], DDD [123], and RLDD [129] capture drowsy drivers. YawDD features recordings of a diverse set of drowsy drivers demonstrating a wide range of behaviors in varying conditions. Some drawbacks of this dataset are scripted actions and recording in a stationary vehicle. DDD is another scripted dataset where subjects were recorded laughing, talking, and looking to the sides, besides acting normal and drowsy, while playing a driving video game in a low-fidelity simulator. RLDD contains images of people captured with mobile phone cameras against neutral backgrounds. Subjects were asked to record themselves when they felt alert, low-vigilant, or drowsy, making sure that the state was authentic. Their self-assessed alertness score was used as a ground truth. Finally, DROZY is the only dataset where subjects experienced prolonged waking under the supervision and where physiological signals accompanied self-evaluated levels of drowsiness.

Large-scale naturalistic data for studying driving- and non-driving related activities are available in Brain4Cars [159] and DMD [236]. Both contain extensive footage of driver-facing cameras synchronized with traffic views, textual labels, and associated vehicle information useful for detection and anticipation of drivers' behavior.

C. Naturalistic Driving Studies (NDS)

NDS are organized efforts to collect large-scale data on the natural behaviors of drivers over extended periods of time. For one of the first such studies, 100-car NDS, drivers used their private vehicles with instrumentation installed to collect rich visual and vehicle data from 2002 to 2004 in the USA [250]. The largest NDS to date, SHRP2, conducted in 2010-2013 also in the USA, involved over 3000 drivers who generated 50M miles of travel (with 372 crashes) and 2 petabytes of data which is still being analyzed [251]. Unlike datasets listed in the Table I, NDS data is not freely accessible due to privacy concerns. For example, researchers interested in obtaining data from SHRP2[2] are required to complete training and provide a research proposal. Fees may also be charged depending on the data requested.

D. Data Availability and Properties

Overall, the datasets listed in Table I contain data recorded in multiple locations, with hundreds of subjects, and accompanied by rich annotations. However, much of this data has been released only recently and is not equally distributed across application domains. As shown in Figure 8, a large portion of the models covered in this survey are developed using private unpublished data. For instance, research on driver monitoring, such as in-vehicle gaze prediction, action anticipation, inattention detection, and driver awareness estimation, largely relies on private data sources, whereas scene gaze prediction and attention for self-driving are studied primarily on publicly available datasets.

[2] https://insight.shrp2nds.us

Another limitation of many public datasets is that attention-related data is often recorded in laboratory conditions. Particularly, for eye-tracking data, such conditions have not been validated (see in Section III-A). For datasets without eye-tracking data, manually annotated events and objects often serve as substitutes for attention. However, the procedures for obtaining and verifying the correctness of such annotations are often not discussed, making their validity for on-road conditions a concern.

Finally, there are some gaps in the data types that the open datasets provide, limiting their use in specific applications. Figure 8 shows different kinds of annotations and their respective usage for various purposes. For example, to the best of our knowledge, there are no open datasets containing driving footage synchronized with the in-vehicle view and driver gaze information for applications such as in-vehicle gaze prediction, action anticipation, inattention detection, and driver awareness.

VII. General Discussion and Conclusions

Over the past decade, significant progress has been made towards detecting and modeling properties of drivers' attention for use in assistive and automated driving. For example, detecting where the driver is looking can be used to monitor drivers' alertness and attention, anticipate their maneuvers, and estimate their awareness of the surrounding traffic situation. Research on drivers' attention allocation has also benefited self-driving. Attentional mechanisms inspired by human visual attention or trained on human gaze data help autonomous vehicles focus on important objects and can also be used to explain their decision-making. Nevertheless, many challenges and open problems remain to be solved to make attention-based driving assistive systems viable for production.

A. Data Availability and Quality

1) More Public Datasets and Models Are Needed: Overall, close to 80% of all works that we reviewed in this survey relied on unpublished private datasets, and less than 10% published relevant code for the models and statistical analysis. But, as was shown in Figure 8, data availability also depends on the application area. For instance, more than two-thirds of driver gaze estimation in the traffic scene and self-driving models are based on public data, and many provide source code, whereas this is not the case for other applications. The lack of public data severely hinders the ability of researchers to reproduce the results of others and draw comparisons between different approaches. Moreover, without established benchmarks estimating actual progress in the area and identifying future challenges is nearly impossible, especially since many unpublished datasets are not accompanied by the information on the recording conditions, characteristics of the subjects, and tasks they performed. Although benchmarks are not without issues, much of the recent progress in computer vision and natural language processing can be attributed to high-quality open large-scale data. Similar tendencies can already be observed in some research areas discussed in this survey, *e.g.* scene gaze estimation, self-driving, and drowsiness detection.

2) Improving Data Diversity and Fidelity Regardless of Data Accessibility: Recording conditions can have a significant effect on the data quality and model applicability in practice. Naturalistic recordings of drivers' behaviors are generally difficult to collect and analyze due to high associated costs and lack of control over the conditions and tasks that the drivers perform. As a result, large volumes of data need to be aggregated to capture specific rarely occurring events. Thus, virtually all data used for developing models is restricted in some sense, *e.g.* by using predefined routes and tasks or conducting the study in the lab or in a parked vehicle. Even though laboratory conditions may be justified for potentially dangerous experiments (*e.g.* involving drowsiness or distractions), they nevertheless affect subjects' behaviors due to low perceived risk, overexposure to rare events, and short duration of sessions [47]. Recording in the lab or in a stationary vehicle cannot capture dynamic changes in lighting, shifts of driver's position due to changes in the road angle, and data loss caused by vibrations and road bumps.

Highway and rural road scenarios, often with low traffic volume, are more common in both on-road and in-lab experiments, whereas city driving is not as well investigated. However, when it comes to drivers' attention, urban conditions are far more challenging due to the presence of intersections and vulnerable road users. Although the speed of the vehicle is lower on the city streets, drivers interact with many other agents, which requires complex attention strategies. Furthermore, bad weather, unfamiliar environment, or heavy traffic are rarely modeled. Even in large naturalistic studies, these conditions are not well represented [37], [252].

Diversity of the participant pool is also a concern. The vast majority of the works we considered record data from no more than a dozen subjects, mostly university students, and many do not provide detailed information about the characteristics of the participants. However, given the evidence of significant individual differences between drivers (as discussed in Section IV), recruiting more subjects with diverse demographic characteristics is highly desirable.

The lack of realism in datasets extends from the environment to the drivers' actions, which are often staged. For instance, it is common practice to induce distraction by asking the drivers to perform tasks at timed intervals (see Sections IV). In reality, however, secondary task engagement is voluntary and depends on many factors, including experience, environmental, and situational, as well as characteristics of the secondary task itself [64]. Forcing the subjects to engage in meaningless tasks on-demand and incentivizing high performance produces detectable changes in gaze allocation and driving performance, but such behaviors may differ from inattention occurring naturally during driving.

3) Taking Into Account the Active Nature of Driving: All available datasets consist of pre-recorded driving footage accompanied by gaze information (driver's gaze and/or gaze of passive observers) or manual annotations. As such they provide limited use for estimating the changes in drivers' gaze depending on the task. Counterfactual studies may help in testing how changes in the task or the environment may affect attention allocation [253] but it is virtually impossible to estimate the effect of the drivers' actions on other road users using pre-recorded data. Simulated environments can generate the outcomes of different actions in the same scenarios but lack realistic models of road user behavior and the environment. While the quality of rendering has been steadily improving with advances in computer graphics, the problem of modeling the actions and reactions of the surrounding pedestrians and vehicles remains far from being solved [254], [255].

B. Evaluation

1) Establishing Ground Truth: There are unresolved issues related to establishing ground truth for many applications. For example, determining specific objects or areas that the driver is observing is not trivial. A recent study by Jansen *et al.* [256] raised concerns regarding the manual annotation of gaze from driver-facing videos. Based on their analysis, the customary practice of measuring several independent annotators' agreement may not produce good quality labels as some areas of interest are easily confused (*e.g.* accuracy for the AOIs is consistently lower on the passenger side). Other factors, such as the driver's height, may also affect the results but are rarely considered. Even with precise eye-tracking data, establishing a point of gaze, especially for small or moving objects, is prone to errors, as Schwehr *et al.* [177] show in a series of experiments. Due to these limitations, models that demonstrate high performance on such ground truth may not transfer to real traffic conditions. Similarly, determining driver's cognitive state may be problematic. As discussed in Section IV-C, self-reported and observer ratings for drowsiness are often not accurate and do not correlate with driving performance. Cognitive distractions are also difficult to induce and detect (Section IV-B). Physiological indicators are more suitable for these purposes [55] but require additional sensors, making data collection more costly and the use of such systems less desirable in practice.

2) Including Safety-Focused Evaluation: Assistive and autonomous driving applications are motivated by safety concerns; however, quantitative evaluations can only assess how well they align with ground truth (which, as noted above, may not be accurate). At the same time, actual crash data is exceedingly rare. For example, in 43 thousand driving hours of driving data recorded in 100-Car NDS, 82 crashes (mostly rear-end collisions), 761 near-crashes, and 8295 incidents were recorded [2].

In the literature, two approaches are commonly taken to mitigate this issue depending on the application in question. One is collecting and annotating accident videos published online (see Section VI), and another is integrating attention into vehicle control models, and testing them in simulation to estimate crash risks (Section V-A2). Both methods have limitations. While accident datasets may provide information about various types of collisions and their timelines, annotations collected in lab conditions, whether eye-tracking data, textual labels or importance scores, are difficult to verify with regard to safety. Therefore, it cannot be guaranteed whether visual strategies learned from such data could have prevented the crash or reduced its severity. Simulated experiments provide both the active control and the ability to replicate the same scenarios, as well as accident risk estimates, but typically are not validated in on-road conditions.

There are also assessments of the risk of prolonged off-road glance durations derived from naturalistic studies [109], [257]. Currently, they are widely used in behavioral literature and as guidelines for in-vehicle infotainment system design. Although they are relevant for the design and evaluation of inattention detection algorithms, only one model within our selection of papers uses them [106]. Incidentally, it is the only model that captures the duration of the inattention, whereas the rest focus on instantaneous detection.

3) Better Coordination Between Research Areas: Research areas covered in this survey are complementary and can benefit from coordinating their efforts. For example, taking into account driver's actions helps better predict attention [194], [195] and detect inattention [109], [258]. Likewise, gaze and appearance features are useful for detecting both distraction and drowsiness (Section IV) but relatively few works investigate these problems together [259], [260].

Human-computer interaction (HCI) research is also very relevant for the design of algorithms intended for use in assistive and autonomous driving systems. To ensure the adoption of such systems, they should function seamlessly and help the driver rather than add to their cognitive load (*e.g.* by unclear or false alarms [261]). However, in the reviewed models, such considerations are rarely taken into account or verified through user studies. For example, many driver gaze prediction algorithms (Section V-A) output pixel-wise heatmaps where objects of interest or imminent hazards are highlighted. Although several studies show that target and maneuver-relevant cues can help direct drivers' gaze to those areas [262]–[264], how this guidance is realized is important. It has been shown that providing too many cues is detrimental and may obscure other important information [263], [265]. Specifically for hazards, indicating the path to avoid them [266] is more effective than pointing at the obstacle itself [267]. Another example is fatigue detection. While reliable detection is the first necessary step, it is not sufficient to provide effective countermeasures. Fatigue due to cognitive under- or overload and drowsiness caused by sleep deprivation require different interventions [268], therefore context must be modeled as well. Individual characteristics of the drivers discussed in Section IV or their driving preferences [114], [269] should also be factored in when setting thresholds for warnings.

C. Limitations of Attention Models

1) Using Gaze as a Proxy for Attention: As discussed in Section III, driving literature views attention as observable gaze changes and measures related to spatio-temporal properties in gaze, such as location and duration of fixations, and transitions between them. The assumption is often made that most driving-related information is processed in the fovea and is predominantly task-driven. Thus analyzing gaze can shed light on what the driver observed at any given time and how it affected their decision-making. However, gaze as a proxy for attention also has a number of limitations. First, gaze alone does not guarantee processing, in other words, looking at something is not equal to being aware of it (*e.g*. looked-but-failed-to-see errors [270] and change blindness [30], [271], [272] occur during driving). Second, drivers extensively use peripheral vision for vehicle control [273], [274] and hazard detection [50], [275], [276], however, gaze provides little insight into peripheral processing. Third, gaze is a result of a complex interplay between various attention control mechanisms, tasks being performed, and surrounding context. These caveats must be taken into account when analyzing eye-tracking data and designing models for various applications in the driving domain and beyond.

2) Reducing the Gap Between Behavioral Research and Implementations: Despite encouraging results, most algorithms consider only a fraction of the factors affecting drivers' attention identified in behavioral studies. For example, there is evidence that age [35], [277]–[279] and driving experience [3] affect drivers' attention allocation. Besides driver characteristics, external conditions matter. Effects of driving through intersections [280]–[282], on curved roads [283], in dense traffic [284], as well as the presence of outside distractors (*e.g*. billboards) [285], [286] have been investigated in numerous behavioral studies but are not taken into account in many implementations.

3) Incorporating Explicit Task Representation: As discussed in Section III-A, top-down factors play a large role in drivers' gaze allocation. High-level features, such as location and class of objects, vehicle telemetry, and optical flow, allow capturing only implicit dependencies between visual features, drivers' gaze data, and vehicle control signals. Such interpretation of attention poorly reflects biological properties of the human visual system and offers little control over algorithms that predict attention allocation in practice. Driving is not a uniform activity, different underlying tasks affect attention distribution differently. For example, when controlling the vehicle, the drivers focus on the road ahead and track road boundaries, periodically fixating on other road users or scanning the intersections [156]. Visual context and gaze may be ambiguous, therefore an explicit top-down signal with intended action or planned route could help better direct the model.

4) Modeling Attention Beyond Selective and Explanatory Functions: In most models of drivers' attention reviewed in Section III-A), the role of attention is reduced to highlighting and ranking objects or areas in the scenes. Other aspects of attention, such as the effect of task on attentional modulation of perception, sequential nature of processing, relation to working memory, decision-making, and allocation of cognitive resources [287], are not considered. Part of the reason is the limited scope of the proposed models. More sophisticated attention mechanisms would be necessary as models' complexity increases towards incorporating the state of the driver, and analysis of the interactions between road users, infrastructure, and drivers' actions.

In conclusion, the problem of modeling drivers' attention is of immense practical and theoretical importance. In this review, we discussed several research directions that analyze and model information on where the driver is looking for applications in driving assistance and automation. We hope that providing a broad overview of several inter-related research areas and identifying open problems will help guide future investigations and lead to improvements in road safety.

REFERENCES

[1] M. Sivak, "The information that drivers use: Is it indeed 90% visual?" *Perception*, vol. 25, no. 9, pp. 1081–1089, Sep. 1996.

[2] S. G. Klauer, T. A. Dingus, V. L. Neale, J. D. Sudweeks, and D. J. Ramsey, "The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data," NHTSA, Washington, DC, USA, Tech. Rep. DOT HS 810 594, 2006.

[3] C. Robbins and P. Chapman, "How does drivers' visual search change as a function of experience? A systematic review and meta-analysis," *Accident Anal. Prevention*, vol. 132, Nov. 2019, Art. no. 105266.

[4] (2017). *How Prevalent are Advanced Driver Assistance Systems (ADAS)?* Accessed: Nov. 1, 2021. [Online]. Available: https://rts.i-car.com/collision-repair-news/how-prevalent-are-advanced-driver-assistance-systems-adas.html

[5] (2021). *Consumer Reports. Guide to Cars With Advanced Safety Systems*. Accessed: Mar. 10, 2020. [Online]. Available: https://www.consumerreports.org/car-safety/cars-with-advanced-safety-systems/

[6] L. Yue, M. Abdel-Aty, Y. Wu, and L. Wang, "Assessment of the safety benefits of vehicles' advanced driver assistance, connectivity and low level automation systems," *Accident Anal. Prevention*, vol. 117, pp. 55–64, Aug. 2018.

[7] A. El Khatib, C. Ou, and F. Karray, "Driver inattention detection in the context of next-generation autonomous vehicles design: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 11, pp. 4483–4496, Nov. 2019.

[8] *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*, Standard J3016, SAE International, 2018.

[9] J. Gaspar and C. Carney, "The effect of partial automation on driver attention: A naturalistic driving study," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 61, no. 8, pp. 1261–1276, Dec. 2019.

[10] J. C. F. de Winter, R. Happee, M. H. Martens, and N. A. Stanton, "Effects of adaptive cruise control and highly automated driving on workload and situation awareness: A review of the empirical evidence," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 27, pp. 196–217, Nov. 2014.

[11] A. Feldhütter, C. Gold, S. Schneider, and K. Bengler, "How the duration of automated driving influences take-over performance and gaze behavior," in *Advances in Ergonomic Design of Systems, Products and Processes*. Berlin, Germany: Springer, 2017, pp. 309–318.

[12] T. Vogelpohl, M. Kühn, T. Hummel, and M. Vollrath, "Asleep at the automated wheel—Sleepiness and fatigue during highly automated driving," *Accident Anal. Prevention*, vol. 126, pp. 70–84, May 2019.

[13] (2017). *Euro NCAP 2025 Roadmap*. Accessed: Mar. 1, 2022. [Online]. Available: https://cdn.euroncap.com/media/30700/euroncap-roadmap-2025-v4.pdf

[14] A. S. Mueller, J. B. Cicchino, and D. S. Zuby, "What humanlike errors do autonomous vehicles need to avoid to maximize safety?" *J. Saf. Res.*, vol. 75, pp. 310–318, Dec. 2020.

[15] J. M. Scanlon, K. D. Kusano, T. Daniel, C. Alderson, A. Ogle, and T. Victor, "Waymo simulated driving behavior in reconstructed fatal crashes within an autonomous vehicle operating domain," Waymo, Mountain View, CA, USA, Tech. Rep., 2021. Accessed: Mar. 10, 2021. [Online]. Available: https://storage.googleapis.com/waymo-uploads/files/documents/Waymo-Simulated-Driving-Behavior-in-Reconstructed-Collisions.pdf

[16] NHTSA. (2015). *Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey*. Accessed: Mar. 10, 2021. [Online]. Available: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812115

[17] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, "An attentive survey of attention models," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 5, pp. 1–32, Oct. 2021.

[18] G. Osterberg, "Topography of the layer of the rods and cones in the human retima," *Acta Ophthalmol.*, vol. 13, no. 6, pp. 1–102, 1935.

[19] C. A. Curcio, K. R. Sloan, R. E. Kalina, and A. E. Hendrickson, "Human photoreceptor topography," *J. Comparative Neurol.*, vol. 292, no. 4, pp. 497–523, 1990.

[20] M. I. Posner, "Orienting of attention," *Quart. J. Exp. Psychol.*, vol. 32, no. 1, pp. 3–25, Feb. 1980.

[21] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer, *Eye Tracking: A Comprehensive Guide to Methods and Measures*. London, U.K.: Oxford Univ. Press, 2011.

[22] *Road Vehicles—Measurement of Driver Visual Behaviour With Respect to Transport Information and Control Systems*, Standard ISO 15007, 2020.

[23] T. W. Victor, J. L. Harbluk, and J. A. Engström, "Sensitivity of eye-movement measures to in-vehicle task difficulty," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 8, pp. 167–190, Mar. 2005.

[24] A. Borowsky, T. Oron-Gilad, A. Meir, and Y. Parmet, "Drivers' perception of vulnerable road users: A hazard perception approach," *Accident Anal. Prevention*, vol. 44, no. 1, pp. 160–166, Jan. 2012.

[25] S. Lemonnier, R. Brémond, and T. Baccino, "Discriminating cognitive processes with eye movements in a decision-making driving task," *J. Eye Movement Res.*, vol. 7, no. 4, pp. 1–14, Jul. 2014.

[26] S. Yantis, "Control of visual attention," *Attention*, vol. 1, no. 1, pp. 223–256, 1998.

[27] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognit. Psychol.*, vol. 12, no. 1, pp. 97–136, Jan. 1980.

[28] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[29] A. L. Yarbus, "Eye movements during perception of complex objects," in *Eye Movements and Vision*. Boston, MA, USA: Springer, 1967, pp. 171–211.

[30] V. Beanland, A. J. Filtness, and R. Jeans, "Change detection in urban and rural driving scenes: Effects of target type and safety relevance on change blindness," *Accident Anal. Prevention*, vol. 100, pp. 111–122, Mar. 2017.

[31] B. T. Sullivan, L. Johnson, C. A. Rothkopf, D. Ballard, and M. Hayhoe, "The role of uncertainty and reward on eye movements in a virtual driving task," *J. Vis.*, vol. 12, no. 13, p. 19, 2012.

[32] P. Konstantopoulos, P. Chapman, and D. Crundall, "Driver's visual attention as a function of driving experience and visibility. Using a driving simulator to explore drivers' eye movements in day, night and rain driving," *Accident Anal. Prevention*, vol. 42, no. 3, pp. 827–834, May 2010.

[33] M. Costa, L. Bonetti, V. Vignali, A. Bichicchi, C. Lantieri, and A. Simone, "Driver's visual attention to different categories of roadside advertising signs," *Appl. Ergonom.*, vol. 78, pp. 127–136, Jul. 2019.

[34] SR Research. *EyeLink 1000 Plus Technical Specifications*. Accessed: Nov. 28, 2020. [Online]. Available: https://www.sr-research.com/wp-content/uploads/2017/11/eyelink-1000-plus-specifications.pdf

[35] D. Stavrinos *et al.*, "Visual behavior differences in drivers across the lifespan: A digital billboard simulator study," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 41, pp. 19–28, Aug. 2016.

[36] C. J. Robbins and P. Chapman, "Drivers' visual search behavior toward vulnerable road users at junctions as a function of cycling experience," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 60, no. 7, pp. 889–901, Nov. 2018.

[37] E. Tivesten and M. Dozza, "Driving context and visual-manual phone tasks influence glance behavior in naturalistic driving," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 26, pp. 258–272, Sep. 2014.

[38] M. Costa, A. Simone, V. Vignali, C. Lantieri, A. Bucchi, and G. Dondi, "Looking behavior for vertical road signs," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 23, pp. 147–155, Mar. 2014.

[39] J. K. Caird and W. J. Horrey, "Twelve practical and useful questions about driving simulation," in *Handbook of Driving Simulation for Engineering, Medicine, and Psychology*. Boca Raton, FL, USA: CRC Press, 2011, pp. 1–5.

[40] G. J. Blaauw, "Driving experience and task demands in simulator and instrumented car: A validation study," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 24, no. 4, pp. 473–486, Aug. 1982.

[41] R. A. Wynne, V. Beanland, and P. M. Salmon, "Systematic review of driving simulator validation studies," *Saf. Sci.*, vol. 117, pp. 138–151, Aug. 2019.

[42] G. P. Mangalore, Y. Ebadi, S. Samuel, M. A. Knodler, and D. L. Fisher, "The promise of virtual reality headsets: Can they be used to measure accurately drivers' hazard anticipation performance?" *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2673, no. 10, pp. 455–464, Oct. 2019.

[43] H. Kim, J. L. Gabbard, S. Martin, A. Tawari, and T. Misu, "Toward prediction of driver awareness of automotive hazards: Driving-video-based simulation approach," in *Proc. Hum. Factors Ergon. Soc. Annu. Meeting*, 2019, vol. 63, no. 1, pp. 2099–2103.

[44] C. Fors, C. Ahlström, and A. Anund, "Simulator validation with respect to driver sleepiness and subjective experiences: Final report of the project SleepEYE II," Swedish Nat. Road Transp. Res. Inst., Linköping, Sweden, Tech. Rep. 2013-1, 2013.

[45] J. A. Mueller, "Driving in a simulator versus on-road: The effect of increased mental effort while driving on real roads and a driving simulator," Ph.D. dissertation, College Eng., Montana State Univ., Bozeman, MT, USA, 2015.

[46] C. J. Robbins, H. A. Allen, and P. Chapman, "Comparing drivers' visual attention at junctions in real and simulated environments," *Appl. Ergonom.*, vol. 80, pp. 89–101, Oct. 2019.

[47] C. Ho, R. Gray, and C. Spence, "To what extent do the findings of laboratory-based spatial attention research apply to the real-world setting of driving?" *IEEE Trans. Human-Machine Syst.*, vol. 44, no. 4, pp. 524–530, Aug. 2014.

[48] T. A. Dingus *et al.*, "The 100-car naturalistic driving study, Phase II-results of the 100-car field experiment," NHTSA, Washington, DC, USA, Tech. Rep. DOT HS 810 593, 2006.

[49] C. D. Fitzpatrick, S. Samuel, and M. A. Knodler, "Evaluating the effect of vegetation and clear zone width on driver behavior using a driving simulator," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 42, pp. 80–89, Oct. 2016.

[50] B. Wolfe, B. Seppelt, B. Mehler, B. Reimer, and R. Rosenholtz, "Rapid holistic perception and evasion of road hazards," *J. Exp. Psychol., Gen.*, vol. 149, no. 3, p. 490, 2020.

[51] W. Chen, X. Zhuang, Z. Cui, and G. Ma, "Drivers' recognition of pedestrian road-crossing intentions: Performance and process," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 64, pp. 552–564, Jul. 2019.

[52] A. Feierle, S. Danner, S. Steininger, and K. Bengler, "Information needs and visual attention during urban, highly automated driving—An investigation of potential influencing factors," *Information*, vol. 11, no. 2, p. 62, Jan. 2020.

[53] M. A. Regan, C. Hallett, and C. P. Gordon, "Driver distraction and driver inattention: Definition, relationship and taxonomy," *Accident Anal. Prevention*, vol. 43, no. 5, pp. 1771–1781, Sep. 2011.

[54] M. Doudou, A. Bouabdallah, and V. Berge-Cherfaoui, "Driver drowsiness measurement technologies: Current research, market solutions, and challenges," *Int. J. Intell. Transp. Syst. Res.*, vol. 18, pp. 1–23, Sep. 2019.

[55] G. Sikander and S. Anwar, "Driver fatigue detection systems: A review," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 6, pp. 2339–2352, Jun. 2018.

[56] M. G. Lenné and E. E. Jacobs, "Predicting drowsiness-related driving events: A review of recent research methods and future opportunities," *Theor. Issues Ergonom. Sci.*, vol. 17, nos. 5–6, pp. 533–553, 2016.

[57] T. C. Ojsteršek and D. Topolšek, "Eye tracking use in researching driver distraction: A scientometric and qualitative literature review approach," *J. Eye Movement Res.*, vol. 12, no. 3, Sep. 2019. [Online]. Available: https://bop.unibe.ch/JEMR/article/view/JEMR.12.3.5

[58] K. J. Parnell, N. A. Stanton, and K. Plant, "Where are we on driver distraction? Methods, approaches and recommendations," *Theor. Issues Ergonom. Sci.*, vol. 19, no. 5, pp. 578–605, Sep. 2018.

[59] M. L. Cunningham and M. A. Regan, "Driver distraction and inattention in the realm of automated driving," *IET Intell. Transp. Syst.*, vol. 12, no. 6, pp. 407–413, Aug. 2018.

[60] A. S. Aghaei *et al.*, "Smart driver monitoring: When signal processing meets human factors: In the driver's seat," *IEEE Signal Process. Mag.*, vol. 33, no. 6, pp. 35–48, Nov. 2016.

[61] K. Kircher and C. Ahlstrom, "Minimum required attention: A human-centered approach to driver inattention," *Hum. Factors*, vol. 59, no. 3, pp. 471–484, 2017.

[62] W. Spiessl and H. Hussmann, "Assessing error recognition in automated driving," *IET Intell. Transport Syst.*, vol. 5, no. 2, pp. 103–111, Jun. 2011.

[63] European Commission. *Road Safety. Distraction*. Accessed: Nov. 28, 2020. [Online]. Available: https://ec.europa.eu/transport/road_safety/topics/behaviour/distraction_en

[64] T. A. Ranney, W. R. Garrott, and M. J. Goodman, "NHTSA driver distraction research: Past, present, and future," in *Proc. Int. Tech. Conf. Enhanced Saf. Vehicles*, 2001.

[65] J. Chen *et al.*, "Fine-grained detection of driver distraction based on neural architecture search," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 9, pp. 5783–5801, Sep. 2021.

[66] M. A. Recarte and L. M. Nunes, "Effects of verbal and spatial-imagery tasks on eye fixations while driving," *J. Exp. Psychol., Appl.*, vol. 6, no. 1, p. 31, 2000.

[67] F. Naujoks, D. Befelein, K. Wiedemann, and A. Neukum, "A review of non-driving-related tasks used in studies on automated driving," in *Proc. AHFE*, 2017, pp. 525–537.

[68] F. Vicente, Z. Huang, X. Xiong, F. D. L. Torre, W. Zhang, and D. Levi, "Driver gaze tracking and eyes off the road detection system," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 2014–2027, Aug. 2015.

[69] B. Vasli, S. Martin, and M. M. Trivedi, "On driver gaze estimation: Explorations and fusion of geometric and data driven approaches," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2016, pp. 655-660.

[70] J. Huang, Y. Long, and X. Zhao, "Driver glance behavior modeling based on semi-supervised clustering and piecewise aggregate representation," *IEEE Trans. Intell. Transp. Syst.*, early access, May 25, 2021, doi: 10.1109/TITS.2021.3080322.

[71] M. Lundgren, L. Hammarstrand, and T. McKelvey, "Driver-gaze zone estimation using Bayesian filtering and Gaussian processes," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 10, pp. 2739–2750, Oct. 2016.

[72] S. J. Lee, J. Jo, H. G. Jung, K. R. Park, and J. Kim, "Real-time gaze estimator based on driver's head orientation for forward collision warning system," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, pp. 254–267, Mar. 2011.

[73] A. Rangesh, B. Zhang, and M. M. Trivedi, "Driver gaze estimation in the real world: Overcoming the eyeglass challenge," in *Proc. IV*, 2020, pp. 1054–1059.

[74] M.-C. Chuang, R. Bala, E. A. Bernal, P. Paul, and A. Burry, "Estimating gaze direction of vehicle drivers using a smartphone camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 655–660.

[75] L. Fridman, P, Langhans, J, Lee, and B. Reimer, "Driver gaze region estimation without use of eye movement," *IEEE Intell. Syst.*, vol. 31, no. 3, pp. 49–56, May/Jun. 2016.

[76] A. Tawari and M. M. Trivedi, "Robust and continuous estimation of driver gaze zone by dynamic analysis of multiple face videos," in *Proc. IV*, 2014, pp. 344–349.

[77] 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, Jun. 2016.

[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. ITSC*, 2014, pp. 988–994.

[79] S. Martin, S. Vora, K. Yuen, and M. M. Trivedi, "Dynamics of driver's gaze: Explorations in behavior modeling and maneuver prediction," *IEEE Trans. Intell. Veh.*, vol. 3, no. 2, pp. 141–150, Feb. 2018.

[80] I.-H. Choi, S. K. Hong, and Y.-G. Kim, "Real-time categorization of driver's gaze zone using the deep learning techniques," in *Proc. ICBDSC*, 2016, pp. 143–148.

[81] S. Vora, A. Rangesh, and M. M. Trivedi, "On generalizing driver gaze zone estimation using convolutional neural networks," in *Proc. IV*, 2017, pp. 849–854.

[82] S. Vora, A. Rangesh, and M. M. Trivedi, "Driver gaze zone estimation using convolutional neural networks: A general framework and ablative analysis," *IEEE Trans. Intell. Veh.*, vol. 3, no. 3, pp. 254–265, Sep. 2018.

[83] L. Stappen, G. Rizos, and B. Schuller, "X-aware: Context-aware human-environment attention fusion for driver gaze prediction in the wild," in *Proc. ICMI*, 2020, pp. 858–867.

[84] N. Li and C. Busso, "Detecting drivers' mirror-checking actions and its application to maneuver and secondary task recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 4, pp. 980–992, Apr. 2016.

[85] T. Hirayama, S. Sato, K. Mase, C. Miyajima, and K. Takeda, "Analysis of peripheral vehicular behavior in driver's gaze transition: Differences between driver's neutral and cognitive distraction states," in *Proc. ITSC*, 2014, pp. 962–967.

[86] R. Wang, P. V. Amadori, and Y. Demiris, "Real-time workload classification during driving using hypernetworks," in *Proc. IROS*, 2018, pp. 3060–3065.

[87] L. Yang, K. Dong, Y. Ding, J. Brighton, Z. Zhan, and Y. Zhao, "Recognition of visual-related non-driving activities using a dual-camera monitoring system," *Pattern Recognit.*, vol. 116, Aug. 2021, Art. no. 107955.

[88] X. Fan, F. Wang, D. Song, Y. Lu, and J. Liu, "GazMon: Eye gazing enabled driving behavior monitoring and prediction," *IEEE Trans. Mobile Comput.*, vol. 20, no. 4, pp. 1420–1433, Apr. 2021.

[89] M. Muñoz, B. Reimer, J. Lee, B. Mehler, and L. Fridman, "Distinguishing patterns in drivers' visual attention allocation using hidden Markov models," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 43, pp. 90–103, Nov. 2016.

[90] N. Li and C. Busso, "Predicting perceived visual and cognitive distractions of drivers with multimodal features," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 1, pp. 51–65, Feb. 2015.

[91] M. Wollmer *et al.*, "Online driver distraction detection using long short-term memory," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 2, pp. 574–582, Mar. 2011.

[92] B. Reimer, B. Mehler, Y. Wang, and J. F. Coughlin, "The impact of systematic variation of cognitive demand on drivers' visual attention across multiple age groups," in *Proc. Hum. Factors Ergon. Soc. Annu. Meeting*, 2010, vol. 54, no. 24, pp. 2052–2055.

[93] G. F. Briggs, G. J. Hole, and M. F. Land, "Emotionally involving telephone conversations lead to driver error and visual tunnelling," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 14, no. 4, pp. 313–323, Jul. 2011.

[94] B. Reimer, B. Mehler, Y. Wang, and J. F. Coughlin, "A field study on the impact of variations in short-term memory demands on drivers' visual attention and driving performance across three age groups," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 54, no. 3, pp. 454–468, Jun. 2012.

[95] Y. Peng, L. N. Boyle, and S. L. Hallmark, "Driver's lane keeping ability with eyes off road: Insights from a naturalistic study," *Accident Anal. Prevention*, vol. 50, pp. 628–634, Jan. 2013.

[96] J. D. Lee, S. C. Roberts, J. D. Hoffman, and L. S. Angell, "Scrolling and driving: How an MP3 player and its aftermarket controller affect driving performance and visual behavior," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 54, no. 2, pp. 250–263, Apr. 2012.

[97] J. Y. Lee, M. C. Gibson, and J. D. Lee, "Error recovery in multitasking while driving," in *Proc. CHI*, 2016, pp. 5104–5113.

[98] M. Smith, J. L. Gabbard, and C. Conley, "Head-up vs. head-down displays: Examining traditional methods of display assessment while driving," in *Proc. AutomotiveUI*, 2016, pp. 185–192.

[99] K. Zeeb, A. Buchner, and M. Schrauf, "Is take-over time all that matters? The impact of visual-cognitive load on driver take-over quality after conditionally automated driving," *Accident Anal. Prevention*, vol. 92, pp. 230–239, Jul. 2016.

[100] Y. Liao, S. E. Li, W. Wang, Y. Wang, G. Li, and B. Cheng, "Detection of driver cognitive distraction: A comparison study of stop-controlled intersection and speed-limited highway," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 6, pp. 1628–1637, Jun. 2016.

[101] Y. Liao, S. E. Li, G. Li, W. Wang, B. Cheng, and F. Chen, "Detection of driver cognitive distraction: An SVM based real-time algorithm and its comparison study in typical driving scenarios," in *Proc. IV*, 2016, pp. 394–399.

[102] T. Liu, Y. Yang, G. B. Huang, Y. K. Yeo, and Z. Lin, "Driver distraction detection using semi-supervised machine learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 4, pp. 1108–1120, Apr. 2016.

[103] C. Braunagel, E. Kasneci, W. Stolzmann, and W. Rosenstiel, "Driver-activity recognition in the context of conditionally autonomous driving," in *Proc. ITSC*, 2015, pp. 1652–1657.

[104] T. Liu *et al.*, "Cluster regularized extreme learning machine for detecting mixed-type distraction in driving," in *Proc. ITSC*, 2015, pp. 1323–1326.

[105] F. Tango and M. Botta, "Real-time detection system of driver distraction using machine learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 2, pp. 894–905, Jun. 2013.

[106] C. Ahlstrom, K. Kircher, and A. Kircher, "A gaze-based driver distraction warning system and its effect on visual behavior," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 2, pp. 965–973, Mar. 2013.

[107] T. Hirayama, K. Mase, and K. Takeda, "Detection of driver distraction based on temporal relationship between eye-gaze and peripheral vehicle behavior," in *Proc. ITSC*, 2012, pp. 870–875.

[108] K. Seshadri, F. Juefei-Xu, D. K. Pal, M. Savvides, and C. P. Thor, "Driver cell phone usage detection on strategic highway research program (SHRP2) face view videos," in *Proc. CVPRW*, 2015, pp. 35–43.

[109] Y. Liang, J. D. Lee, and L. Yekhshatyan, "How dangerous is looking away from the road? Algorithms predict crash risk from glance patterns in naturalistic driving," *Hum. Factors*, vol. 54, no. 6, pp. 1104–1116, 2012.

[110] Y. Zhang and D. Kaber, "Evaluation of strategies for integrated classification of visual-manual and cognitive distractions in driving," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 58, no. 6, pp. 944–958, Sep. 2016.

[111] K. Kircher and C. Ahlström, "The driver distraction detection algorithm attend," in *Driver Distraction and Inattention*. Boca Raton, FL, USA: CRC Press, 2017.

[112] NHTSA. (2012). *Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices*. Accessed: Nov. 28, 2020. [Online]. Available: https://www.govinfo.gov/content/pkg/FR-2013-04-26/pdf/2013-09883.pdf

[113] R. D. Sorkin, "Why are people turning off our alarms?" *J. Acoust. Soc. Amer.*, vol. 84, no. 3, pp. 1107–1108, 1988.

[114] C. Wang, Q. Sun, Y. Guo, R. Fu, and W. Yuan, "Improving the user acceptability of advanced driver assistance systems based on different driving styles: A case study of lane change warning systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 1, pp. 4196–4208, Oct. 2019.

[115] M. Ahmed, S. Masood, M. Ahmad, and A. A. A. El-Latif, "Intelligent driver drowsiness detection for traffic safety based on multi CNN deep model and facial subsampling," *IEEE Trans. Intell. Transp. Syst.*, early access, Dec. 22, 2021, doi: 10.1109/TITS.2021.3134222.

[116] R. Huang, Y. Wang, Z. Li, Z. Lei, and Y. Xu, "RF-DCM: Multi-granularity deep convolutional model based on feature recalibration and fusion for driver fatigue detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 1, pp. 630–640, Jan. 2022.

[117] W. Deng and R. Wu, "Real-time driver-drowsiness detection system using facial features," *IEEE Access*, vol. 7, pp. 118727–118738, 2019.

[118] J. Yu, S. Park, S. Lee, and M. Jeon, "Driver drowsiness detection using condition-adaptive representation learning framework," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 11, pp. 4206–4218, Nov. 2019.

[119] B. Reddy, Y.-H. Kim, S. Yun, C. Seo, and J. Jang, "Real-time driver drowsiness detection for embedded system using model compression of deep neural networks," in *Proc. CVPRW*, 2017, pp. 121–128.

[120] J. Yu, S. Park, S. Lee, and M. Jeon, "Representation learning, scene understanding, and feature fusion for drowsiness detection," in *Proc. ACCV*, 2016, pp. 165–177.

[121] T.-H. Shih and C.-T. Hsu, "MSTN: Multistage spatial-temporal network for driver drowsiness detection," in *Proc. ACCV*, 2016, pp. 146–153.

[122] X.-P. Huynh, S.-M. Park, and Y.-G. Kim, "Detection of driver drowsiness using 3D deep neural network and semi-supervised gradient boosting machine," in *Proc. ACCV*, 2016, pp. 134–145.

[123] C.-H. Weng, Y.-H. Lai, and S.-H. Lai, "Driver drowsiness detection via a hierarchical temporal deep belief network," in *Proc. ACCV*, 2016, pp. 117–133.

[124] I.-H. Choi, C.-H. Jeong, and Y.-G. Kim, "Tracking a driver's face against extreme head poses and inference of drowsiness using a hidden Markov model," *Appl. Sci.*, vol. 6, no. 5, p. 137, May 2016.

[125] S. Park, F. Pan, S. Kang, and C. D. Yoo, "Driver drowsiness detection system based on feature representation learning using various deep networks," in *Proc. ACCV*, 2016, pp. 154–164.

[126] E. Tadesse, W. Sheng, and M. Liu, "Driver drowsiness detection through HMM based dynamic modeling," in *Proc. ICRA*, 2014, pp. 4003–4008.

[127] B. Akrout and W. Mahdi, "A visual based approach for drowsiness detection," in *Proc. IV*, 2013, pp. 1324–1329.

[128] M. Vijay, N. N. Vinayak, M. Nunna, and S. Natarajan, "Real-time driver drowsiness detection using facial action units," in *Proc. ICPR*, 2021, pp. 10113–10119.

[129] R. Ghoddoosian, M. Galib, and V. Athitsos, "A realistic dataset and baseline temporal model for early drowsiness detection," in *Proc. CVPRW*, 2019, pp. 1–10.

[130] B. Bakker *et al.*, "A multi-stage, multi-feature machine learning approach to detect driver sleepiness in naturalistic road driving conditions," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 5, pp. 4791–4800, May 2022.

[131] K. Qian, T. Koike, T. Nakamura, B. Schuller, and Y. Yamamoto, "Learning multimodal representations for drowsiness detection," *IEEE Trans. Intell. Transp. Syst.*, early access, Aug. 26, 2021, doi: 10.1109/TITS.2021.3105326.

[132] A. Dasgupta, D. Rahman, and A. Routray, "A smartphone-based drowsiness detection and warning system for automotive drivers," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 11, pp. 4045–4054, Nov. 2019.

[133] H. Yin, Y. Su, Y. Liu, and D. Zhao, "A driver fatigue detection method based on multi-sensor signals," in *Proc. WACV*, 2016 pp. 1–7.

[134] X. Fan, Y. Sun, B. Yin, and X. Guo, "Gabor-based dynamic representation for human fatigue monitoring in facial image sequences," *Pattern Recognit. Lett.*, vol. 31, no. 3, pp. 234–243, Feb. 2010.

[135] J. Gwak, M. Shino, and A. Hirao, "Early detection of driver drowsiness utilizing machine learning based on physiological signals, behavioral measures, and driving performance," in *Proc. ITSC*, 2018, pp. 1794–1800.

[136] H. Matsuo and A. Khiat, "Prediction of drowsy driving by monitoring driver's behavior," in *Proc. ICPR*, 2012, pp. 3390–3393.

[137] W. Sun, X. Zhang, S. Peeta, X. He, and Y. Li, "A real-time fatigue driving recognition method incorporating contextual features and two fusion levels," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 12, pp. 3408–3420, Dec. 2017.

[138] T.-H. Chang and Y.-R. Chen, "Driver fatigue surveillance via eye detection," in *Proc. ITSC*, 2014, pp. 366–371.

[139] I. Garcia, S. Bronte, L. M. Bergasa, J. Almazán, and J. Yebes, "Vision-based drowsiness detector for real driving conditions," in *Proc. IV*, 2012, pp. 618–623.

[140] Z. Zhang and J. Zhang, "A new real-time eye tracking based on nonlinear unscented Kalman filter for monitoring driver fatigue," *J. Control Theory Appl.*, vol. 8, no. 2, pp. 181–188, 2010.

[141] S. Dari, N. Epple, and V. Protschky, "Unsupervised blink detection and driver drowsiness metrics on naturalistic driving data," in *Proc. ITSC*, 2020, pp. 1–6.

[142] X. Li, E. Seignez, and P. Loonis, "Vision-based estimation of driver drowsiness with ORD model using evidence theory," in *Proc. IV*, 2013, pp. 666–671.

[143] F. Friedrichs and B. Yang, "Camera-based drowsiness reference for driver state classification under real driving conditions," in *Proc. IV*, 2010, pp. 101–106.

[144] A. Joshi, S. Kyal, S. Banerjee, and T. Mishra, "In-the-wild drowsiness detection from facial expressions," in *Proc. IV*, 2020, pp. 207–212.

[145] L. Jin, Q. Niu, Y. Jiang, H. Xian, Y. Qin, and M. Xu, "Driver sleepiness detection system based on eye movements variables," *Adv. Mech. Eng.*, vol. 5, Jan. 2013, Art. no. 648431.

[146] L. Zhao, Z. Wang, X. Wang, and Q. Liu, "Driver drowsiness detection using facial dynamic fusion information and a DBN," *IET Intell. Transp. Syst.*, vol. 12, no. 2, pp. 127–133, Mar. 2018.

[147] X. Wang and C. Xu, "Driver drowsiness detection based on non-intrusive metrics considering individual specifics," *Accident Anal. Prevention*, vol. 95, pp. 350–357, Oct. 2016.

[148] C. Zhang, X. Wu, X. Zheng, and S. Yu, "Driver drowsiness detection using multi-channel second order blind identifications," *IEEE Access*, vol. 7, pp. 11829–11843, 2019.

[149] D. F. Dinges and R. Grace, "PERCLOS: A valid psychophysiological measure of alertness as assessed by psychomotor vigilance," Federal Highway Admin., Washington, DC, USA, Tech. Rep. FHWA-MCRT-98-006, 1998.

[150] L. N. Boyle, J. Tippin, A. Paul, and M. Rizzo, "Driver performance in the moments surrounding a microsleep," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 11, no. 2, pp. 126–136, 2008.

[151] T. Akerstedt and M. Gillberg, "Subjective and objective sleepiness in the active individual," *Int. J. Neurosci.*, vol. 52, nos. 1–2, pp. 29–37, Jul. 1990.

[152] E. Zilberg, Z. M. Xu, D. Burton, M. Karrar, and S. Lal, "Methodology and initial analysis results for development of non-invasive and hybrid driver drowsiness detection systems," in *Proc. AusWireless*, 2007, p. 16.

[153] A. Anund, C. Fors, D. Hallvig, T. Åkerstedt, and G. Kecklund, "Observer rated sleepiness and real road driving: An explorative study," *PLoS ONE*, vol. 8, no. 5, May 2013, Art. no. e64782.

[154] C. Ahlstrom, C. Fors, A. Anund, and D. Hallvig, "Video-based observer rated sleepiness versus self-reported subjective sleepiness in real road driving," *Eur. Transp. Res. Rev.*, vol. 7, no. 4, pp. 1–9, 2015.

[155] P. Philip *et al.*, "Fatigue, sleep restriction, and performance in automobile drivers: A controlled study in a natural environment," *Sleep*, vol. 26, no. 3, pp. 277–280, May 2003.

[156] O. Lappi, P. Rinkkala, and J. Pekkanen, "Systematic observation of an expert driver's gaze strategy—An on-road case study," *Frontiers Psychol.*, vol. 8, p. 620, Apr. 2017.

[157] N. Akai, T. Hirayama, L. Y. Morales, Y. Akagi, H. Liu, and H. Murase, "Driving behavior modeling based on hidden Markov models with driver's eye-gaze measurement and ego-vehicle localization," in *Proc. IV*, 2019, pp. 949–956.

[158] S. Martin and M. M. Trivedi, "Gaze fixations and dynamics for behavior modeling and prediction of on-road driving maneuvers," in *Proc. IV*, 2017, pp. 1541–1545.

[159] A. Jain, H. S. Koppula, B. Raghavan, S. Soh, and A. Saxena, "Car that knows before you do: Anticipating maneuvers via learning temporal driving models," in *Proc. ICCV*, 2015, pp. 3182–3190.

[160] A. Jain, A. Singh, H. S. Koppula, S. Soh, and A. Saxena, "Recurrent neural networks for driver activity anticipation via sensory-fusion architecture," in *Proc. ICRA*, 2016, pp. 3118–3125.

[161] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost: A hybrid approach to alleviating class imbalance," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 40, no. 1, pp. 185–197, Jan. 2010.

[162] M. Roth, F. Flohr, and D. M. Gavrila, "Driver and pedestrian awareness-based collision risk analysis," in *Proc. IV*, 2016, pp. 454–459.

[163] S. J. Zabihi, S. Zabihi, S. S. Beauchemin, and M. A. Bauer, "Detection and recognition of traffic signs inside the attentional visual field of drivers," in *Proc. IV*, 2017, pp. 583–588.

[164] Y. Ma, J. Wu, and C. Long, "GazeFCW: Filter collision warning triggers by detecting driver's gaze area," in *Proc. SenSys-ML*, 2019, pp. 13–18.

[165] M. Rezaei and R. Klette, "Look at the driver, look at the road: No distraction! No accident!" in *Proc. CVPR*, 2014, pp. 129–136.

[166] T. Langner, D. Seifert, B. Fischer, D. Goehring, T. Ganjineh, and R. Rojas, "Traffic awareness driver assistance based on stereovision, eye-tracking, and head-up display," in *Proc. ICRA*, 2016, pp. 3167–3173.

[167] A. Tawari, A. Møgelmose, S. Martin, T. B. Moeslund, and M. M. Trivedi, "Attention estimation by simultaneous analysis of viewer and view," in *Proc. ITSC*, 2014, pp. 1381–1387.

[168] J. Schwehr and V. Willert, "Multi-hypothesis multi-model driver's gaze target tracking," in *Proc. ITSC*, 2018, pp. 1427–1434.

[169] T. Kowsari, S. S. Beauchemin, M. A. Bauer, D. Laurendeau, and N. Teasdale, "Multi-depth cross-calibration of remote eye gaze trackers and stereoscopic scene systems," in *Proc. IV*, 2014, pp. 1245–1250.

[170] T. Bär, D. Linke, D. Nienhüser, and J. M. Zöllner, "Seen and missed traffic objects: A traffic object-specific awareness estimation," in *Proc. IV*, 2013, pp. 31–36.

[171] M. Mori *et al.*, "Measuring driver awareness based on correlation between gaze behavior and risks of surrounding vehicles," in *Proc. ITSC*, 2012, pp. 644–647.

[172] C. Ahlström, G. Georgoulas, and K. Kircher, "Towards a context-dependent multi-buffer driver distraction detection algorithm," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 5, pp. 4778–4790, May 2022.

[173] J. Schwehr and V. Willert, "Driver's gaze prediction in dynamic automotive scenes," in *Proc. ITSC*, 2017, pp. 1–8.

[174] H. Zhu, T. Misu, S. Martin, X. Wu, and K. Akash, "Improving driver situation awareness prediction using human visual sensory and memory mechanism," in *Proc. IROS*, 2021, pp. 6210–6216.

[175] M. R. Endsley, "Situation awareness global assessment technique (SAGAT)," in *Proc. NAECON*, 1988, pp. 789–795.

[176] L. Yang, K. Dong, A. J. Dmitruk, J. Brighton, and Y. Zhao, "A dual-cameras-based driver gaze mapping system with an application on non-driving activities monitoring," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 10, pp. 4318–4327, Oct. 2020.

[177] J. Schwehr, M. Knaust, and V. Willert, "How to evaluate object-of-fixation detection," in *Proc. IV*, 2019, pp. 570–577.

[178] S. Zabihi, S. S. Beauchemin, E. De Medeiros, and M. A. Bauer, "Frame-rate vehicle detection within the attentional visual area of drivers," in *Proc. IV*, 2014, pp. 146–150.

[179] K. Gillmeier, F. Diederichs, and D. Spath, "Prediction of ego vehicle trajectories based on driver intention and environmental context," in *Proc. IV*, 2019, pp. 963–968.

[180] D. Tran, J. Du, W. Sheng, D. Osipychev, Y. Sun, and H. Bai, "A human-vehicle collaborative driving framework for driver assistance," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 9, pp. 3470–3485, Sep. 2019.

[181] H. Kim, S. Martin, A. Tawari, T. Misu, and J. L. Gabbard, "Toward real-time estimation of driver situation awareness: An eye-tracking approach based on moving objects of interest," in *Proc. IV*, 2020, pp. 1035–1041.

[182] M. R. Endsley, "Design and evaluation for situation awareness enhancement," in *Proc. HFES*, 1988, vol. 32, no. 2, pp. 97–101.

[183] F. Diederichs, T. Schüttke, and D. Spath, "Driver intention algorithm for pedestrian protection and automated emergency braking systems," in *Proc. ITSC*, 2015, pp. 1049–1054.

[184] K. Doman *et al.*, "Estimation of traffic sign visibility considering local and global features in a driving environment," in *Proc. IV*, 2014, pp. 202–207.

[185] K. Doman *et al.*, "Estimation of traffic sign visibility considering temporal environmental changes for smart driver assistance," in *Proc. IV*, 2011, pp. 667–672.

[186] K. Doman *et al.*, "Estimation of traffic sign visibility toward smart driver assistance," in *Proc. IV*, 2010, pp. 45–50.

[187] D. Engel and C. Curio, "Detectability prediction in dynamic scenes for enhanced environment perception," in *Proc. IV*, 2012, pp. 178–183.

[188] T. Deng, K. Yang, Y. Li, and H. Yan, "Where does the driver look? Top-down-based saliency detection in a traffic driving environment," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 7, pp. 2051–2062, Jul. 2016.

[189] T. Deng, H. Yan, and Y.-J. Li, "Learning to boost bottom-up fixation prediction in driving environments via random forest," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 9, pp. 3059–3067, Sep. 2018.

[190] A. Borji, D. N. Sihite, and L. Itti, "Computational modeling of top-down visual attention in interactive environments," in *Proc. BMVC*, 2011, pp. 1–12.

[191] A. Borji, D. N. Sihite, and L. Itti, "Probabilistic learning of task-specific visual attention," in *Proc. CVPR*, 2012, pp. 470–477.

[192] A. Borji, D. N. Sihite, and L. Itti, "What/where to look next? Modeling top-down visual attention in complex interactive environments," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 5, pp. 523–538, May 2014.

[193] M. F. Land and D. N. Lee, "Where we look when we steer," *Nature*, vol. 369, no. 6483, pp. 742–744, 1994.

[194] P. V. Amadori, T. Fischer, and Y. Demiris, "HammerDrive: A task-aware driving visual attention model," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 5573–5585, Jun. 2022.

[195] S. Baee, E. Pakdamanian, I. Kim, L. Feng, V. Ordonez, and L. Barnes, "MEDIRL: Predicting the visual attention of drivers via maximum entropy deep inverse reinforcement learning," in *Proc. ICCV*, 2021, pp. 13178–13188.

[196] A. Pal, S. Mondal, and H. I. Christensen, "'Looking at the right stuff'—Guided semantic-gaze for autonomous driving," in *Proc. CVPR*, 2020, pp. 11883–11892.

[197] J. Fang, D. Yan, J. Qiao, J. Xue, and H. Yu, "DADA: Driver attention prediction in driving accident scenarios," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 4959–4971, Jun. 2022.

[198] A. Palazzi *et al.*, "Predicting the driver's focus of attention: The DR (eye) VE project," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1720–1733, Jul. 2019.

[199] L. Palmer, A. Bialkowski, G. J. Brostow, J. Ambeck-Madsen, and N. Lavie, "Predicting the perceptual demands of urban driving with video regression," in *Proc. WACV*, 2017, pp. 409–417.

[200] M. Gao, A. Tawari, and S. Martin, "Goal-oriented object importance estimation in on-road driving videos," in *Proc. ICRA*, 2019, pp. 5509–5515.

[201] E. Ohn-Bar and M. M. Trivedi, "Are all objects equal? Deep spatio-temporal importance prediction in driving videos," *Pattern Recognit.*, vol. 64, pp. 425–436, Apr. 2017.

[202] Z. Zhang, A. Tawari, S. Martin, and D. Crandall, "Interaction graphs for object importance estimation in on-road driving videos," in *Proc. ICRA*, 2020, pp. 8920–8927.

[203] H. R. Tavakoli, E. Rahtu, J. Kannala, and A. Borji, "Digging deeper into egocentric gaze prediction," in *Proc. WACV*, 2019, pp. 273–282.

[204] M. Ning, C. Lu, and J. Gong, "An efficient model for driving focus of attention prediction using deep learning," in *Proc. ITCS*, 2019, pp. 1192–1197.

[205] D. Gopinath *et al.*, "MAAD: A model and dataset for 'attended awareness' in driving," in *Proc. CCVW*, 2021, pp. 3426–3436.

[206] G. Yao, T. Lei, and J. Zhong, "A review of convolutional-neural-network-based action recognition," *Pattern Recognit. Lett.*, vol. 118, pp. 14–22, Feb. 2019.

[207] A. Palazzi, F. Solera, S. Calderara, S. Alletto, and R. Cucchiara, "Learning where to attend like a human driver," in *Proc. IV*, 2017, pp. 920–925.

[208] Y. Xia, D. Zhang, J. Kim, K. Nakayama, K. Zipser, and D. Whitney, "Predicting driver attention in critical situations," in *Proc. ACCV*, 2018, pp. 658–674.

[209] A. Tawari, P. Mallela, and S. Martin, "Learning to attend to salient targets in driving videos using fully convolutional RNN," in *Proc. ITSC*, 2018, pp. 3225–3232.

[210] A. Tawari and B. Kang, "A computational framework for driver's visual attention using a fully convolutional architecture," in *Proc. IV*, 2017, pp. 887–894.

[211] T. Deng, H. Yan, L. Qin, T. Ngo, and B. S. Manjunath, "How do drivers allocate their potential attention? Driving fixation prediction via convolutional neural networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 5, pp. 2146–2154, May 2020.

[212] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 740–757, Mar. 2019.

[213] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit, "Saliency and human fixations: State-of-the-art and study of comparison metrics," in *Proc. CVPR*, 2013, pp. 1153–1160.

[214] J. Janai, F. Güney, A. Behl, and A. Geiger, "Computer vision for autonomous vehicles: Problems, datasets and state of the art," *Found. Trends Comput. Graph. Vis.*, vol. 12, nos. 1–3, pp. 1–308, 2020.

[215] C. Badue *et al.*, "Self-driving cars: A survey," *Expert Syst. Appl.*, vol. 165, Mar. 2020, Art. no. 113816.

[216] A. Rasouli and J. K. Tsotsos, "Autonomous vehicles that interact with pedestrians: A survey of theory and practice," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 900–918, Mar. 2019.

[217] A. Tampuu, M. Semikin, N. Muhammad, D. Fishman, and T. Matiisen, "A survey of end-to-end driving: Architectures and training methods," 2020, *arXiv:2003.06404*.

[218] G. Brauwers and F. Frasincar, "A general survey on attention mechanisms in deep learning," *IEEE Trans. Knowl. Data Eng.*, early access, Nov. 9, 2021, doi: 10.1109/TKDE.2021.3126456.

[219] K. Ishihara, A. Kanervisto, J. Miura, and V. Hautamaki, "Multi-task learning with attention for end-to-end autonomous driving," in *Proc. CVPRW*, 2021, pp. 2902–2911.

[220] J. Kim and J. Canny, "Interpretable learning for self-driving cars by visualizing causal attention," in *Proc. ICCV*, 2017, pp. 2942–2950.

[221] L. Cultrera, L. Seidenari, F. Becattini, P. Pala, and A. Del Bimbo, "Explaining autonomous driving by learning end-to-end visual attention," in *Proc. CVPRW*, 2020, pp. 340–341.

[222] D. Wang, C. Devin, Q.-Z. Cai, F. Yu, and T. Darrell, "Deep object-centric policies for autonomous driving," in *Proc. ICRA*, 2019, pp. 8853–8859.

[223] S. He, D. Kangin, Y. Mi, and N. Pugeault, "Aggregated sparse attention for steering angle prediction," in *Proc. ICPR*, 2018, pp. 2398–2403.

[224] B. Wei, M. Ren, W. Zeng, M. Liang, B. Yang, and R. Urtasun, "Perceive, attend, and drive: Learning spatial attention for safe self-driving," in *Proc. ICRA*, 2021, pp. 4875–4881.

[225] A. Vaswani *et al.*, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 6000–6010.

[226] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, Jan. 2022. [Online]. Available: https://dl.acm.org/doi/10.1145/3505244

[227] J. Vig, "A multiscale visualization of attention in the transformer model," 2019, *arXiv:1906.05714*.

[228] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proc. CVPR*, 2021, pp. 782–791.

[229] K. Chitta, A. Prakash, and A. Geiger, "Neat: Neural attention fields for end-to-end autonomous driving," in *Proc. ICCV*, 2021, pp. 15793–15803.

[230] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *Proc. CVPR*, 2021, pp. 7077–7087.

[231] L. L. Li *et al.*, "End-to-end contextual perception and prediction with interaction transformer," in *Proc. IROS*, 2020, pp. 5784–5791.

[232] J. Fang, D. Yan, J. Qiao, J. Xue, H. Wang, and S. Li, "DADA-2000: Can driving accident be predicted by driver attention *f* analyzed by a benchmark," in *Proc. ITSC*, 2019, pp. 4303–4309.

[233] S. Taamneh *et al.*, "A multimodal dataset for various forms of distracted driving," *Sci. Data*, vol. 4, no. 1, Dec. 2017, Art. no. 170110.

[234] I. Dua, T. A. John, R. Gupta, and C. Jawahar, "DGAZE: Driver gaze mapping on road," in *Proc. IROS*, 2020, pp. 5946–5953.

[235] S. Ghosh, A. Dhall, G. Sharma, S. Gupta, and N. Sebe, "Speak2Label: Using domain knowledge for creating a large scale driver gaze zone estimation dataset," in *Proc. ICCVW*, 2021, pp. 2896–2905.

[236] J. D. Ortega *et al.*, "DMD: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis," in *Proc. ECCV*, 2020, pp. 387–405.

[237] M. Martin *et al.*, "Drive&Act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles," in *Proc. ICCV*, 2019, pp. 2801–2810.

[238] J. Kim, T. Misu, Y.-T. Chen, A. Tawari, and J. Canny, "Grounding human-to-vehicle advice for self-driving vehicles," in *Proc. CVPR*, 2019, pp. 10591–10599.

[239] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual explanations for self-driving vehicles," in *Proc. ECCV*, 2018, pp. 563–578.

[240] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko, "Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning," in *Proc. CVPR*, 2018, pp. 7699–7707.

[241] F.-H. Chan, Y.-T. Chen, Y. Xiang, and M. Sun, "Anticipating accidents in dashcam videos," in *Proc. ACCV*, 2016, pp. 136–153.

[242] Q. Massoz, T. Langohr, C. François, and J. G. Verly, "The ULg multimodality drowsiness database (called DROZY) and examples of use," in *Proc. WACV*, 2016, pp. 1–7.

[243] N. Pugeault and R. Bowden, "How much of driving is preattentive?" *IEEE Trans. Veh. Technol.*, vol. 64, no. 12, pp. 5424–5438, Dec. 2015.

[244] S. Abtahi, M. Omidyeganeh, S. Shirmohammadi, and B. Hariri, "YawDD: A yawning detection dataset," in *Proc. MMSys*, 2014, pp. 24–28.

[245] Y. Xia, J. Kim, J. Canny, K. Zipser, T. Canas-Bajo, and D. Whitney, "Periphery-fovea multi-resolution driving model guided by human attention," in *Proc. WACV*, 2020, pp. 1767–1775.

[246] Y. Chen, C. Liu, L. Tai, M. Liu, and B. E. Shi, "Gaze training by modulated dropout improves imitation learning," in *Proc. IROS*, 2019, pp. 7756–7761.

[247] J. Kim, S. Moon, A. Rohrbach, T. Darrell, and J. Canny, "Advisable learning for self-driving vehicles by internalizing observation-to-action rules," in *Proc. CVPR*, 2020, pp. 9661–9670.

[248] K. Mori, H. Fukui, T. Murase, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Visual explanation by attention branch network for end-to-end learning-based self-driving," in *Proc. IV*, 2019, pp. 1577–1582.

[249] S. Mund, R. Frank, G. Varisteas, and R. State, "Visualizing the learning progress of self-driving cars," in *Proc. ITSC*, 2018, pp. 2358–2363.

[250] V. L. Neale, T. A. Dingus, S. G. Klauer, J. Sudweeks, and M. Goodman, "An overview of the 100-car naturalistic study and findings," NHTSA, Washington, DC, USA, Tech. Rep. DOT HS 810 846, 2005.

[251] T. A. Dingus *et al.*, "Naturalistic driving study: Technical coordination and quality control," Transp. Res. Board, SHRP 2, Tech. Rep. S2-S06-RW-1, 2015.

[252] B. E. Hammit, A. Ghasemzadeh, R. M. James, M. M. Ahmed, and R. K. Young, "Evaluation of weather-related freeway car-following behavior using the SHRP2 naturalistic driving study database," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 59, pp. 244–259, Nov. 2018.

[253] J. Bärgman, V. Lisovskaja, T. Victor, C. Flannagan, and M. Dozza, "How does glance behavior influence crash and injury risk? A 'what-if' counterfactual simulation using crashes and near-crashes from SHRP2," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 35, pp. 152–169, Nov. 2015.

[254] A. Rasouli, "Pedestrian simulation: A review," 2021, *arXiv:2102.03289*.

[255] S. W. Kim, J. Philion, A. Torralba, and S. Fidler, "DriveGAN: Towards a controllable high-quality neural simulation," in *Proc. CVPR*, 2021, pp. 5820–5829.

[256] R. J. Jansen, S. T. van der Kint, and F. Hermens, "Does agreement mean accuracy? Evaluating glance annotation in naturalistic driving data," *Behav. Res. Methods*, vol. 53, no. 1, pp. 430–446, Feb. 2021.

[257] J. Y. Lee, J. D. Lee, J. Bärgman, J. Lee, and B. Reimer, "How safe is tuning a radio: Using the radio tuning task as a benchmark for distracted driving," *Accident Anal. Prevention*, vol. 110, pp. 29–37, Jan. 2018.

[258] J. Kim, K. Kim, D. Yoon, Y. Koo, and W. Han, "Fusion of driver-information based driver status recognition for co-pilot system," in *Proc. IV*, 2016, pp. 1398–1403.

[259] D. Tran, H. M. Do, J. Lu, and W. Sheng, "Real-time detection of distracted driving using dual cameras," in *Proc. IROS*, 2020, pp. 2014–2019.

[260] C.-Y. Chiou, W.-C. Wang, S.-C. Lu, C.-R. Huang, P.-C. Chung, and Y.-Y. Lai, "Driver monitoring using sparse representation with part-based temporal face descriptors," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 1, pp. 346–361, Jan. 2020.

[261] L. Viktorová and M. Sucha, "Drivers' acceptance of advanced driver assistance systems—What to consider," *Int. J. Traffic Transp. Eng*, vol. 8, no. 3, pp. 320–333, 2018.

[262] E. Bozkir, D. Geisler, and E. Kasneci, "Assessment of driver attention during a safety critical situation in VR to generate VR-based training," in *Proc. SAP*, 2019, pp. 1–5.

[263] R. Eyraud, E. Zibetti, and T. Baccino, "Allocation of visual attention while driving with simulated augmented reality," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 32, pp. 46–55, Jul. 2015.

[264] L. Pomarjanschi, M. Dorr, and E. Barth, "Gaze guidance reduces the number of collisions with pedestrians in a driving simulator," *ACM Trans. Interact. Intell. Syst.*, vol. 1, no. 2, pp. 1–14, 2012.

[265] H. Kim and J. L. Gabbard, "Assessing distraction potential of augmented reality head-up displays for vehicle drivers," *Hum. Factors*, pp. 1–14, May 2019. [Online]. Available: https://journals.sagepub.com/doi/10.1177/0018720819844845

[266] S. S. Borojeni, L. Chuang, W. Heuten, and S. Boll, "Assisting drivers with ambient take-over requests in highly automated driving," in *Proc. AutomotiveUI*, 2016, pp. 237–244.

[267] Y. Yang, B. Karakaya, G. C. Dominioni, K. Kawabe, and K. Bengler, "An HMI concept to improve driver's visual behavior and situation awareness in automated vehicle," in *Proc. ITSC*, 2018, pp. 650–655.

[268] J. F. May and C. L. Baldwin, "Driver fatigue: The importance of identifying causal factors of fatigue when considering detection and countermeasure technologies," *Transp. Res. F, Psychol. Behav.*, vol. 12, no. 3, pp. 218–224, 2009.

[269] J. M. Fleming, C. K. Allison, X. Yan, R. Lot, and N. A. Stanton, "Adaptive driver modelling in ADAS to improve user acceptance: A study using naturalistic data," *Saf. Sci.*, vol. 119, pp. 76–83, Nov. 2019.

[270] C. B. White and J. K. Caird, "The blind date: The effects of change blindness, passenger conversation and gender on looked-but-failed-to-see (LBFTS) errors," *Accident Anal. Prevention*, vol. 42, no. 6, pp. 1822–1830, Nov. 2010.

[271] B. Metz and H.-P. Krüger, "Do supplementary signs distract the driver?" *Transp. Res. F, Traffic Psychol. Behav.*, vol. 23, pp. 1–14, Mar. 2014.

[272] I. M. Harms and K. A. Brookhuis, "Dynamic traffic management on a familiar road: Failing to detect changes in variable speed limits," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 38, pp. 37–46, Apr. 2016.

[273] D. Crundall, G. Underwood, and P. Chapman, "Driving experience and the functional field of view," *Perception*, vol. 28, no. 9, pp. 1075–1087, 1999.

[274] D. Crundall, G. Underwood, and P. Chapman, "Attending to the peripheral world while driving," *Appl. Cognit. Psychol.*, vol. 16, no. 4, pp. 459–475, 2002.

[275] A. Shahar, C. F. Alberti, D. Clarke, and D. Crundall, "Hazard perception as a function of target location and the field of view," *Accident Anal. Prevention*, vol. 42, no. 6, pp. 1577–1584, Nov. 2010.

[276] B. Wolfe, B. D. Sawyer, A. Kosovicheva, B. Reimer, and R. Rosenholtz, "Detection of brake lights while distracted: Separating peripheral vision from cognitive load," *Attention, Perception, Psychophysics*, vol. 81, no. 8, pp. 2798–2813, Nov. 2019.

[277] J. Edquist, T. Horberry, S. Hosking, and I. Johnston, "Effects of advertising billboards during simulated driving," *Appl. Ergonom.*, vol. 42, no. 4, pp. 619–626, May 2011.

[278] P. Urwyler *et al.*, "Age-dependent visual exploration during simulated day- and night driving on a motorway: A cross-sectional study," *BMC Geriatrics*, vol. 15, no. 1, p. 18, Dec. 2015.

[279] M. Zahabi *et al.*, "Effect of driver age and distance guide sign format on driver attention allocation and performance," in *Proc. Hum. Factors Ergon. Soc. Annu. Meeting*, 2018, vol. 62, no. 1, pp. 1903–1907.

[280] J. Werneke and M. Vollrath, "What does the driver look at? The influence of intersection characteristics on attention allocation and driving behavior," *Accident Anal. Prevention*, vol. 45, pp. 610–619, Mar. 2012.

[281] S. Lemonnier, R. Brémond, and T. Baccino, "Gaze behavior when approaching an intersection: Dwell time distribution and comparison with a quantitative prediction," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 35, pp. 60–74, Nov. 2015.

[282] G. Li *et al.*, "Drivers' visual scanning behavior at signalized and unsignalized intersections: A naturalistic driving study in China," *J. Saf. Res.*, vol. 71, pp. 219–229, Dec. 2019.

[283] O. Lappi, "Future path and tangent point models in the visual control of locomotion in curve driving," *J. Vis.*, vol. 14, no. 12, p. 21, 2014.

[284] J.-T. Wong and S.-H. Huang, "Attention allocation patterns in naturalistic driving," *Accident Anal. Prevention*, vol. 58, pp. 140–147, Sep. 2013.

[285] D. Topolšek, I. Areh, and T. Cvahte, "Examination of driver detection of roadside traffic signs and advertisements using eye tracking," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 43, pp. 212–224, Nov. 2016.

[286] D. Belyusar, B. Reimer, B. Mehler, and J. F. Coughlin, "A field study on the effects of digital billboards on glance behavior during highway driving," *Accident Anal. Prevention*, vol. 88, pp. 88–96, Mar. 2016.

[287] J. K. Tsotsos, *A Computational Perspective on Visual Attention*. Cambridge, MA, USA: MIT Press, 2011.

Iuliia Kotseruba received the B.Sc. degree in computer science from the University of Toronto in 2010 and the M.Sc. degree in computer science from York University in 2016. She is currently pursuing the Ph.D. degree with the Laboratory for Active and Attentive Vision supervised by Prof. John K. Tsotsos. After her M.Sc. degree, she continued to work as a Research Associate with York University. Her research focuses on using various machine learning techniques to model visual attention for cognitive systems and applications in assistive and autonomous driving.

John K. Tsotsos received the doctorate degree in computer science from the University of Toronto. He is a Distinguished Research Professor of vision science with York University. After a Post-Doctoral Fellowship in cardiology with Toronto General Hospital, he joined the Faculty of Computer Science and Faculty of Medicine, University of Toronto. In 1980, he founded the highly respected the Computer Vision Group, University of Toronto, which he led for 20 years. He was recruited to move to York University in 2000 as the Director of the Centre for Vision Research. His current research focuses on a comprehensive theory of visual attention in humans. A practical outlet for this theory forms a second focus, embodying elements of the theory into the vision systems of mobile robots. He has been a Canadian Heart Foundation Research Scholar, a fellow of the Canadian Institute for Advanced Research, and the Canada Research Chair in Computational Vision. He has received many awards and honors, including several best paper awards, the 2006 Canadian Image Processing and Pattern Recognition Society Award for Research Excellence and Service, the 1st President's Research Excellence Award by York University on the occasion of the university's 50th anniversary in 2009, and the 2011 Geoffrey J. Burton Memorial Lectureship from the U.K.'s Applied Vision Association for Significant Contribution to Vision Science. He was elected as a fellow of the Royal Society of Canada in 2010 and was awarded its 2015 Sir John William Dawson Medal for sustained excellence in multidisciplinary research and the first computer scientist to be so honored.