Investigating and Modeling the Effects of Task and Context on Drivers' Gaze Allocation

Iuliia Kotseruba

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM
ELECTRICAL ENGINEERING AND COMPUTER SCIENCE
YORK UNIVERSITY
TORONTO, ONTARIO

Abstract

Driving, despite its widespread nature, is a demanding and inherently risky activity. Any lapse in focus, such as failing to look at the traffic signals or not noticing the actions of other road users, can lead to severe consequences. Technology for driver monitoring and assistance aims to mitigate these issues, but requires a deeper understanding of how drivers observe their surroundings to make decisions.

In this dissertation, we investigate the link between where drivers look, tasks they perform, and the surrounding context. To do so, we first conduct a meta-study of the behavioral literature that documents an overwhelming importance of the top-down (task-driven) effects on gaze. Next, we survey applied research to show that most models do not necessarily make this connection and instead establish correlations between where the drivers looked and images of the scene, without explicitly considering drivers' actions and environment.

Next, we annotate and analyze the four largest publicly available datasets that contain driving footage and eye-tracking data. The new annotations for task and context show that data is dominated by trivial scenarios (e.g., driving straight, standing) and help uncover problems with the typical data recording and processing pipelines that result in noisy, missing, or inaccurate data, particularly during safety-critical scenarios (e.g., intersections). For the only dataset with the raw data available, we create a new ground truth which alleviates some of the discovered issues. We also provide recommendations for future data collection.

Using the new annotations and ground truth, we benchmark a representative set of bottom-up models for gaze prediction (i.e., those that do not represent the task explicitly). We conclude that while corrected ground truth boosts performance, the implicit representation is not sufficient to capture the effects of task and context on where drivers look.

Lastly, motivated by these findings, we propose a task- and context-aware model for drivers' gaze prediction with explicit representation of the drivers' actions and context. The first version of the model, SCOUT, improves state-of-the-art performance by over 80% overall and 30% on the most challenging scenarios. We then propose SCOUT+, which relies on the more readily available route and map information similar to what the driver might see on the in-car navigation screen. SCOUT+ achieves comparable results as the version that uses more precise numeric and text labels.

Acknowledgments

I want to thank my supervisor John Tsotsos for many years of mentorship. His support, kindness, and generosity helped me grow as a researcher and as a person. I thank my supervisory committee members Michael Brown and Marcus Brubaker for encouragement and helpful advice. I also thank the examining committee members Danica Kragic, Regina Lee, and Michael Jenkin for their kind words and thoughtful feedback.

I am grateful for the love and support that my parents, Svitlana and Viktor, always have given me.

Lastly, but most importantly, I thank my dearest friend, long-time collaborator, and life partner Amir Rasouli for his continuing friendship, care, patience, and love. I consider myself truly lucky to have Amir in my life and my present achievements would not have been possible without him by my side.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Millions of people around the world drive cars daily for a variety of everyday reasons. Because driving is so commonplace, it may seem trivial. But in reality, it is a complex and demanding visuomotor task comprised of multiple concurrent activities, including but not limited to vehicle control, navigation, and hazard anticipation and response. Driving is also both time-sensitive and safety-critical; a driver's failure to notice and respond in time to the traffic signals and actions of other road users can have dire consequences [1]. At the same time, visual skills for effectively observing the traffic and safely diverting gaze from the road (e.g., to check the mirrors) take some time to develop, from several months to several years after finishing driving school [2].

1.1 Why study drivers' attention?

Driver inattention has been identified as a significant cause of accidents since the late 1930s [3]. To this day, traffic safety continues to be the main motivation for research on attention and driving. Since drivers rely primarily on vision to make decisions [4], the central questions of interest are *where*, *when*, and *why* drivers look in different environments and situations, how it affects their reasoning, and what causes lapses in attention. Finding answers is crucial for improving road safety, especially given the existing evidence that temporary distractions and sub-optimal visual scanning skills increase risk of accidents [5, 6].

One example of how knowledge of what affects drivers' visual behaviors can help is improved societal, educational, and legal policies. For instance, after numerous studies found that bright colors and moving billboard advertisements tend to distract drivers, regulations were put in place to control the number of digital billboards along roads and frequency of ad changes [7]. Another example concerns new drivers who are over-represented in crash statistics. In the 1960s, early works by Zell [8] and Mourant & Rockwell [9] established

that even after successfully finishing driving school novice drivers lack visual skills for safe driving. In the following decades, significant efforts were made towards improving driver education, including development of better driver training methods for acquisition of the necessary visual scanning strategies [10].

Driving also provides a useful testbed for studying human visual attention as it is an activity familiar to many people of different ages and backgrounds. A traffic scene itself is a prime example of a noisy, dynamic, and complex environment that humans can actively navigate and explore by selectively sampling the information. Consequently, driving makes it possible to investigate decision-making processes and study the link between vision and motor actions in a naturalistic setting.

The main motivation for this dissertation comes from the recent advances in intelligent transportation, computer vision, and AI that aim to reduce accidents by monitoring where the driver is looking, warning them against distractions, and directing their attention to imminent danger. To achieve these objectives, it is crucial to understand how drivers behave, how they interact with other road users, and how they will transition from being in control to overseeing increasingly automated vehicles.

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 [11]. 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 [12]. According to recent estimates, ADAS can potentially eliminate up to one-third of accidents caused by light vehicles on highways [13].

Existing ADAS can identify hazards and autonomously prevent collisions but operate without considering intentions or the condition of the driver. To address this issue, driver monitoring systems (DMS) aim to enhance safety by detecting driver inattention and, if necessary, alerting them or stopping the vehicle when the driver is unresponsive. Currently, most commercial DMS rely on vehicle measures, such as steering or lateral control, to assess drivers' state [14]. 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 hazardous objects and events.

The need for deployment of vision-based DMS is also dictated by the proliferation of partially- or highly-automated driving systems corresponding to SAE Levels 2-4 [15]. Past

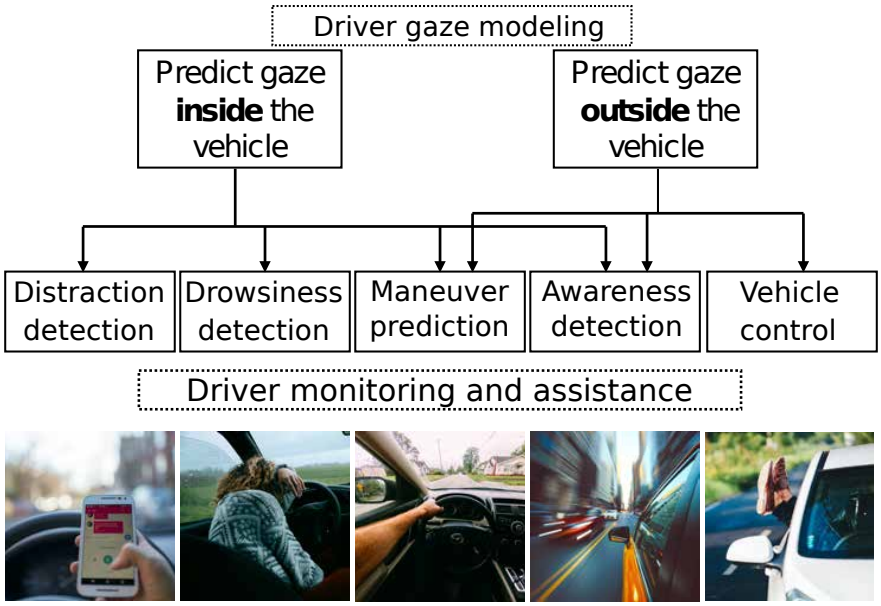


Figure 1.1: Driving applications that use human gaze. Methods for estimating drivers' gaze, whether looking into or out of the vehicle, are needed for vision-based driver monitoring, assistive, and automated driving systems. Distraction and drowsiness detection algorithms rely primarily on identification of gaze zones within the vehicle. Maneuver prediction and estimating drivers' awareness of the surroundings benefit from estimation of drivers' gaze inside and outside the vehicle. Automated driving systems simulate drivers' attention for better planning and explainability.

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 [16, 17]. Since these systems rely on a human driver being in the loop and may hand over control at any moment, the safety of switching to manual control depends on whether the driver is distracted or fatigued [18, 19]; thus monitoring the driver's 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 European New Car Assessment Programme (Euro NCAP) 2025 roadmap to zero road fatalities by 2050 [20], and similar initiatives are likely to be introduced in other countries.

In sum, vision-based assistive and autonomous driving solutions rely on sophisticated algorithms that observe and analyze drivers' attention allocation and actions and relate them to the events unfolding in the traffic scenes. Currently, the problem of driver assistance and monitoring is subdivided into several related topics: distraction/drowsiness detection, maneuver prediction, awareness detection, and planning for automated vehicle control. All of these make use of drivers' gaze information (see Figure 1.1). Monitoring inattention relies on detection of driver's gaze inside the vehicle (e.g., prolonged glances away from the windshield)

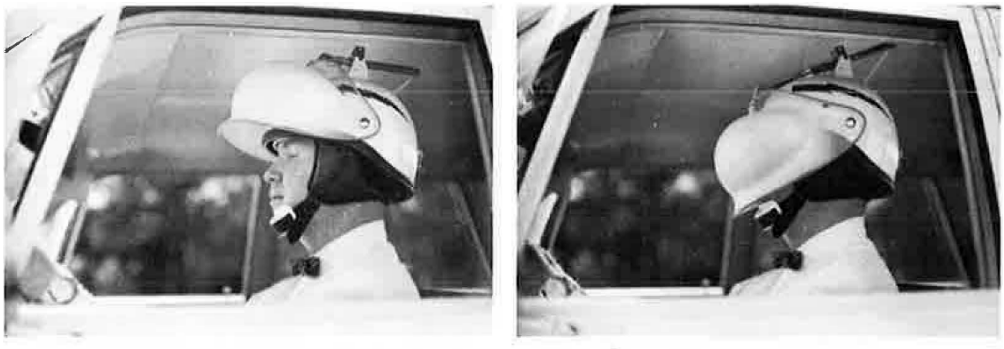
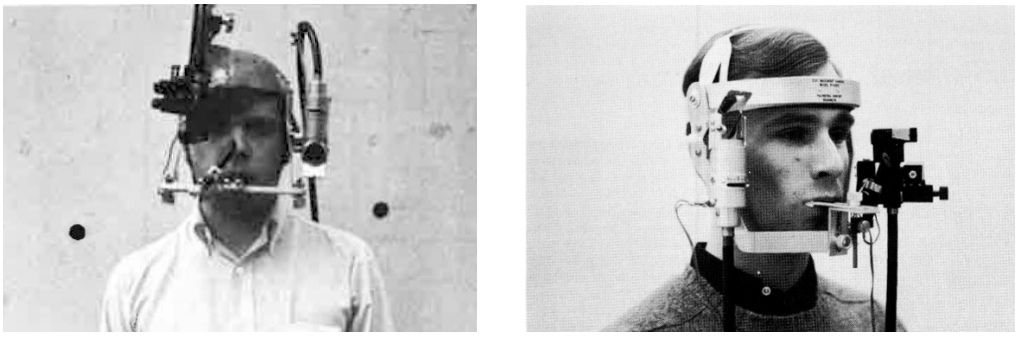


Figure 1.2: A helmet for studying drivers' attention. A semi-transparent visor can be closed and opened during driving to temporarily cover the driver's field of view. Source: [21].

or characteristics of gaze in general (e.g., eyes closed). Prediction of upcoming maneuvers makes use of both driver's gaze towards elements of the vehicle car, such as mirrors and dashboard, and outside, towards signs and intersections. Similarly, understanding of what the driver is aware in the traffic scene can be extracted from gaze information combined with vehicle telemetry. Because human attention is highly optimized and robust, understanding how drivers effectively analyze the scene can inform design of self-driving vehicles. On the other hand, knowing limitations of human drivers is also necessary to ensure coordination between autonomous and manually driven cars in traffic.

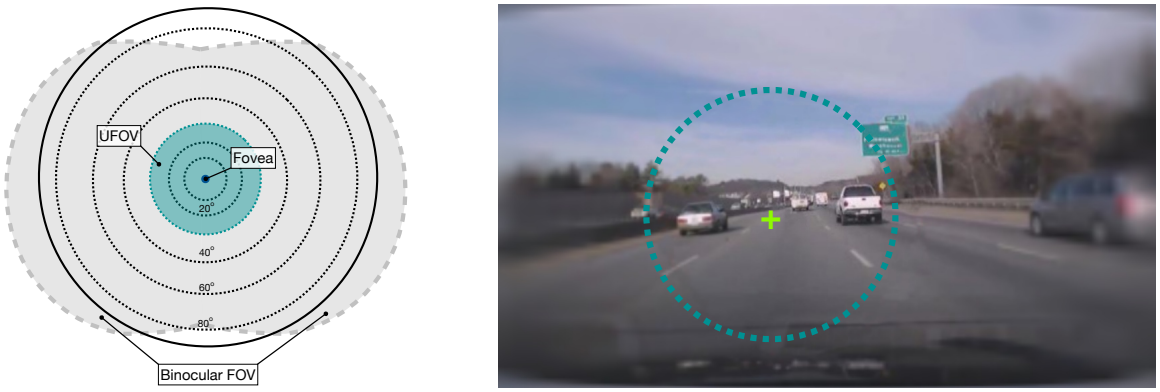
1.2 Gaze and attention

A natural first step towards understanding drivers' attention allocation is gathering objective and precise data. The most widely used method for this purpose is eye-tracking technology, which can directly and accurately determine where the person focused their gaze. Although early mechanical eye trackers had been invented in the late 1800s, commercial models were not widely available until the 1970s [22]. Until then, many studies relied on vehicle control measurement instruments, custom-built eye trackers, or devices for indirect measurement of drivers' visual behaviors. The latter included a contraption that reduced the driver's peripheral vision [23], a helmet with a shield that lowered to control the driver's visual input [21] (shown in Figure 1.2), and wooden planks placed in front of the windshield to limit the view of the road [24]. Early mobile eye trackers (Figure 1.3), despite being unwieldy, intrusive, and imprecise, afforded valuable insights into drivers' gaze behavior. However, before we discuss what has been achieved with the much more precise and high-resolution recordings offered by modern eye-tracking technology, we need to establish why eye movements are needed in the first place and whether studying them gives a full picture of what the driver perceived.



(a) (b)

Figure 1.3: Early models of mobile eye trackers available in 1970s. Source: a) [25], b) [24].



(a) (b)

Figure 1.4: Visualization of the human visual field: a) a schematic illustration of the human binocular visual field with fovea in the center and 30° useful field of view (UFOV) from which information can be extracted in one fixation (shaded blue area); b) a foveated image of the traffic scene (using algorithms proposed in [26, 27]) with fixation point (green cross) and UFOV (blue circle). Images modified from [28].

1.2.1 Foveal vision as a proxy for attention

What is fovea? Human vision is foveated, i.e., visual acuity is highest in the center of the visual field and degrades towards the periphery due to the non-uniform distribution of receptors in the retina [29, 30] and neural tissue dedicated to visual processing in the cortex [31]. Within the visual field, the fovea (2° in diameter) and parafovea (up to 5°) form *central vision*, and the *periphery* covers the remainder (Figure 1.4a). Since only a portion of the scene is captured in the high resolution, eye movements are needed to bring areas of interest into the fovea for more detailed examination. Such explicit change of foveal focus is also known as *overt attention*.

The role of peripheral vision. The area outside of the fovea, referred to as the periphery, comprises more than 95% of the visual field [32] (see Figure 1.4a). Although acuity decreases

with eccentricity, peripheral vision is not merely a blurry version of the foveal vision. The qualitative differences between the two include perception of contrast [33], color [34], motion [35, 36], size, shape [37], and category [38] of objects (see [39] for a comprehensive review). Crowding is another well-known property of peripheral vision, defined as an inability to discern fine details and properties of the individual objects at high eccentricities [40, 41, 32].

Historically, research on attention and driving focused on foveal vision. As a result, little is known about peripheral perception for everyday tasks like driving [42], where relevant cues often appear at high eccentricities and are affected by crowding and loss of resolution (Fig 1.4b). For a long time, the role of peripheral vision has been relegated to detecting global motion cues (e.g., optical flow [43]) to aid lateral control [44, 45, 46]. More recent research shows that other tasks, such as hazard detection [47, 48, 49] and sign recognition [50], can be done with peripheral vision alone. However, it is not safe to do so as reaction times to hazards and miss rates for target detection increase with eccentricity [51, 52, 48].

Functional field of view (UFOV). A concept of useful (or functional) field of view (UFOV) describes visual field ($\approx 30^\circ$ in diameter) from which information can be extracted in a single fixation [53] and is widely used in the transportation literature (see reviews in [54, 55, 28]). An illustration of UFOV is shown in Figure 1.4a.

Based on UFOV, Ball et al. [56] developed a test to map out the functional field of view, processing speed, divided and selective attention of older drivers who may be at risk for an accident due to vision impairments. UFOV test examines processing speed, divided and selective attention using foveal identification and peripheral localization tasks. The tasks are performed separately and together, with and without distractors, while the presentation duration and eccentricity of stimuli within UFOV are manipulated. These conditions simulate peripheral detection and orientation tasks that drivers perform while focusing on the forward roadway.

The inability to perceive safety-critical information due to reduced UFOV in older drivers has been linked to increased risk of crashes in multiple independent studies [57, 58, 59, 60, 61]. Changes in cognitive workload [62, 63, 64] and fatigue [65] can also cause reduction of UFOV.

Looking $\neq$ awareness. So far we established that foveal vision can give an idea of where the person is looking. But it is important to remember that looking at something is not the same as being aware of it. Even if an object or event falls within the foveal region, it may still be missed. One such example is change blindness, occurring when one fails to notice a significant and obvious change in the discontinuous scene [66]. During driving, blinking, saccades, and physical obstructions may cause such discontinuities. But despite favorable conditions, effects of change blindness are usually mitigated by motion and illumination cues across the discontinuity [67]. When such cues are removed in controlled experiments, drivers

do not see changes in the traffic scene even when explicitly told to look for them [68, 69, 70].

Inattention blindness, i.e., failure to notice an event or object when attention is diverted elsewhere [66], is another common phenomenon. Some examples are reduced attention to road signs due to driving without awareness, especially on familiar routes [71], and failing to see hazards while talking to a passenger [72]. Inattention blindness is linked to "looked but failed to see" (LBFTS) errors that cause many accidents, particularly violations of right-of-way at T-intersections when drivers fail to see an approaching motorcycle, bicycle, or another car and pull out in front of it [73, 72, 74, 75].

Switching attention between multiple concurrent tasks also reduces performance in some or all tasks, especially when they compete for the same attentional resources [76]. For example, driving as a visuomotor activity is affected the most by visual (reading), manual (eating), or visuomanual (texting) tasks (see Section 2.3.2).

1.2.2 What determines where we look?

The question of what prompts explicit changes in gaze locations is a major topic of research in attention. Mechanisms of attentional control are commonly divided into *bottom-up* and *top-down* [77]. The former is viewed as an involuntary process guided by the saliency of objects that attract human gaze [78, 79]. In contrast, top-down attention is driven by the demands of the current task [80, 81, 82]. *Selection history* [83] (also referred to as *memory-based attention* [84, 85]), has been proposed as a third possible mechanism. Since it is based on prior experience and rewards, it is neither fully stimulus- nor task-driven. Although the effects of selection history have not been investigated as thoroughly, it is easy to think of analogies in driving. For example, when traveling on a familiar route, drivers often rely on their memory in addition to perception [86, 87].

Both bottom-up and top-down control are involved during driving, but their relative contributions and how they are integrated remain open research problems. There is evidence that top-down attention plays a large role in driving. For instance, drivers' visual strategies change when priorities of driving tasks are manipulated (e.g., drivers monitor speedometer more often when maintaining speed is prioritized over vehicle following) [88]. Likewise, change blindness depends on the relevance to the task, e.g., changes to trees and buildings are missed more often than changes to vehicles and signs [70].

Bottom-up attention also has a role in driving. For example, one can make stop-and-go judgments after viewing scenes where only salient regions are visible (e.g., traffic signals, signs, brake lights) [89]. However, since bottom-up cues are not optimal, information from non-salient areas is required for better decision-making, reinforcing the role of experience in

efficient scene processing. For example, motorcyclists are usually not the most salient (e.g., due to smaller size and dark clothing) and can be missed in a cluttered traffic scene [73, 74], but experienced drivers utilize top-down strategies to mitigate this issue [73].

1.3 Problem formulation

The main topic of this dissertation is developing a better understanding of driving as a visuomotor activity and modeling it as such to advance vision-based driver monitoring and assistive systems that rely on human gaze directly or indirectly. The first step in this process is usually estimating where the attentive driver would look. As discussed earlier in Section 1.1, this information is useful for monitoring and assistance (see Figure 1.1). Because the design of the full system is beyond the scope of this dissertation, we instead focus on obtaining the correct input, namely a good estimate of where an attentive driver should look in the environment, depending on the task (actions) and properties of the traffic scene (context). Given this estimate, the system may compare it with what the driver is currently observing and issue a warning or provide guidance if needed. Naturally, the correctness of the predicted gaze is very important. If the system has a poor understanding of how an attentive driver should view the scene, it may either a poor guidance, potentially leading to the accident, or unnecessary warning, reducing the adoption of the system.

Throughout the dissertation, we use several key terms, such as *task*, *action*, and *context*, that have broad meanings in the literature. Below we define these terms for the purposes of our problem as follows:

- *Driving task*. We consider driving task as a set of *actions* that lead to changes in the motion of the ego-vehicle. This is a common interpretation used also in the robotics, vision, and psychological literature. Typically, actions are subdivided into lateral (turns and lane changes), longitudinal (decelerating and accelerating), remaining stationary, and maintaining lane and speed. Depending on the application, there may be other actions that drivers perform in the vehicle that are not related to vehicle control. Some are driving-related, e.g., checking mirror or navigation screen, and others are not, e.g., consuming drinks or food, or using a cell phone. The latter are referred to as secondary or non-driving-related tasks (NDRT).

- *Context*. Context is a broad term that refers to any properties of the environment inside or outside the vehicle. In-vehicle context may refer to presence of automated features (e.g., cruise control) and infotainment system (e.g., navigation, head-up display). Outside the vehicle, environmental conditions (e.g., weather and visibility),

road structure (e.g., intersections), road infrastructure (e.g., traffic signals), and traffic density, may change how drivers observe and act.

The problem of drivers' gaze prediction is formulated as follows: given a set of n RGB images $I \in \mathbb{R}^{H \times W \times 3}$ recorded from the driver's perspective and (optionally) a representation of the drivers' task and environment conditions, predict where the driver is more likely to look in the last frame of the input set. Gaze is typically represented as a saliency map $S \in [0, 1]^{H \times W}$, where pixels with higher values correspond to areas in the image which are more likely to attract attention.

1.4 Outline of the dissertation

In this dissertation we link theoretical and practical aspects of research on drivers' attention. With regards to theory, we seek to identify and categorize how the effects on drivers' attention are exerted by characteristics of the drivers themselves and properties of their surroundings. On the practical side, we aim to apply these findings towards improving prediction of where drivers are more likely to look by explicitly modeling the relationship between drivers' actions and gaze. Due to the complexity of the problem, we draw on works that span multiple fields, such as psychology, human factors, human-computer interaction, and intelligent transportation, to get different perspectives and explanations for the observed phenomena. Next we briefly outline the contents of each chapter in this dissertation.

In Chapter 2, we first provide the necessary background by briefly introducing common methods for recording, processing, and aggregating human gaze data in the driving domain. We then continue with the meta-analysis of a large body of works from transportation psychology and human-computer interaction literature published over the last decade to identify a set of endogenous and exogenous influences on drivers' attention. In particular, we emphasize the significant empirical evidence that what drivers do and what is around them are the primary drivers of changes in their attention allocation. In the Chapter 3 that follows we review computational models of various aspects of drivers' attention developed within computer vision and intelligent transportation communities. These works are more application-oriented and we organize their review accordingly. The fundamental problems here are predicting drivers' gaze towards the elements of the scene and car interior. The downstream applications are distraction and drowsiness detection, driver action recognition and prediction, and driver awareness estimation. In this chapter we also identify gaps between theoretical findings in the behavioral literature and computational efforts to motivate further investigations.

The remaining chapters are dedicated to one of the two core problems mentioned earlier—predicting drivers' gaze allocation towards the traffic scene. In the Chapter 4, we review the past attempts at solving this problem, which fall into two major groups: the mainstream bottom-up approaches to learning correlations between video and human ground truth data and the less common top-down models that explicitly represent drivers' actions. While there is evidence that top-down models are more accurate, current benchmarks emphasize the overall fit to human data and lack annotations for a more detailed evaluation. Thus in this chapter, we analyze four large-scale public datasets with eye tracking data and corresponding driving footage as they form the basis for most of the research in this field. Here, we use the findings from the behavioral literature to find and correct some of the limitations of the common data recording and processing methods with respect to capturing what drivers saw and did. The chapter concludes with recommendations for future data collection efforts.

In Chapter 5, we use the annotations for drivers' actions and environmental factors from the previous chapter to analyze performance of the current state of the art models for drivers' gaze prediction. We show that the corrections introduced earlier reduce noise and lead to better model performance. Another significant finding is that bottom-up approach to learning cannot effectively capture effects of drivers' actions and surrounding context. As a result, performance of all models across all metrics drops precipitously on all non-trivial scenarios, where drivers performed maneuvers or crossed intersections.

Lastly, taking both theoretical and practical considerations into account, in Chapter 6, we propose SCOUT, a task- and context-aware approach, that effectively models where drivers are more likely to look depending on what they do and what is around them. Here, we emphasize safety-critical scenarios, such as lateral maneuvers and crossing intersections. In these cases, the path of the vehicle may potentially overlap with the paths of other road users (either on neighboring lanes or intersecting roads) and thus drivers must be especially aware of their surroundings. We propose two variants of the model: 1) SCOUT that relies on the ground truth task information to empirically demonstrate its benefits, and 2) SCOUT+ which achieves similar results by utilizing map and route information derived from the accessible GPS data.

In Chapter 7 we summarize our findings, discuss some of the limitations of our approach, and propose directions for future research in this domain. Publications, reports, and other resources that resulted from this dissertation are listed in Appendix A.

1.5 Evaluation

Even though work presented in this dissertation is motivated by the practical need for further development of driver monitoring and assistance systems, it is difficult to evaluate our model for predicting drivers' gaze allocation in this setting. The main reason is that there are currently no clearly defined pathways of integrating these systems into existing vehicles as many related human-computer interaction issues need to be resolved first. Furthermore, even simulated human studies are currently infeasible due to the limitations of the available data described in Chapter 4.

Therefore, we evaluate our model using conventional methods accepted by the computer vision community. Specifically, we split the video data into partially overlapping observations of predefined length (16 frames or approximately 0.5 s) and compute the output for the last frame in the set as a saliency map representing where the driver is more likely to look. We then compare the output of our model to the recorded human ground truth available in public datasets. Performance is measured with respect to a set of standard metrics and results of current state-of-the-art models. In addition, we use new annotations for the common public datasets introduced in Chapter 4 to measure how well our proposed model performs in specific scenarios against other approaches.

Chapter 2

What affects where drivers look?

In the previous chapter, we discussed two main types of mechanisms, bottom-up and top-down, that guide drivers' gaze towards different elements of the traffic scene. Here, we will associate them with specific factors. To do so, we analyze a comprehensive set of 247 studies published since 2010 that explicitly measured drivers' gaze using an eye tracker[1]. For each study, we extract independent variables that were manipulated in the experiments and were shown to affect drivers' gaze distribution. We then group these variables into more general factors subdivided into two sets: *endogenous* and *exogenous* to the driver.

By *endogenous factors* we mean what the drivers are doing and their personal characteristics. The former refers to the *primary driving task* (e.g., hazard perception, vehicle following) or involvement in *secondary non-driving related tasks* (NDRT). Driver's characteristics are *demographics* (age and gender), *driving experience* (e.g., duration, frequency of driving), and *inattention* state (e.g., distracted or drowsy). *Exogenous factors* are external to the driver: *environment*, a broad category for road infrastructure, traffic conditions, weather and lighting conditions, and *automation*, a catch-all category for any system that monitors or assists the driver, including full and partial automation of vehicle control.

We will start with the short overview on recording and measuring drivers' gaze, as it will be relevant for the rest of this chapter and will provide the necessary context for dataset analysis in Chapter 4.

[1]Studies that used modes of transportation other than cars (e.g., bicycles, motorcycles, buses) and focused on drivers with medical issues were excluded. A complete list is available at https://github.com/ykotseruba/attention_and_driving/blob/2.0/behavioral.md

2.1 Recording drivers' gaze

Human data collection is the first and, perhaps, the most important step in studying drivers' attention. A traffic environment has a number of challenging characteristics, among which are its dynamic nature, diversity, and inherent risk. Given the added complexity of the driving task itself, recording an unbiased, accurate, and complete picture of what the driver perceived and did during any given trip is a non-trivial problem. As a result, there are trade-offs between ecological validity, quality, and scale of the data collection. To limit the scope of our discussion, we will focus on the three main components of data collection: recording conditions, recording equipment, and data processing.

2.1.1 Recording conditions

To categorize studies we reviewed based on the recording conditions, we considered whether they were conducted on-road in a real vehicle or in-lab using a driving simulator. We only considered studies conducted in cars since we focus on the development of ADAS for this category of vehicles. Therefore, other modes of transportation (e.g., motorcycles, trucks, buses, trains) were omitted.

On-road studies are subdivided from the least to most restrictive in the following groups: *naturalistic*, *directed*, and *closed track*.

- *Naturalistic.* Naturalistic studies represent real driving patterns because subjects use instrumented vehicles for daily activities over extended periods of time. The downsides are geographic and population biases, as well as the need to collect larger volumes of data to study specific events since drivers' behaviors are not scripted [90].
- *Directed.* During *directed* on-road studies, the drivers are given a route and may be accompanied by the researcher. Reduction in the variability of the drivers' behaviors removes the need for large-scale data collection. However, the lack of familiarity with the vehicle may affect subjects' behaviors [91, 92]. Conducting studies during the off-peak hours [93, 94, 91] reduces the risk but may bias the results. In any case, conditions are not the same across participants due to variability in the environmental conditions and vehicle control.
- *Closed track.* Closed track isolated from traffic provides somewhat realistic conditions and allows conducting experiments risky for public roads (e.g., hazard response [95]). However, reducing participants' sense of danger may potentially bias their behavior.

Overall, better ecological validity of the on-road studies imposes loss of experimental control and data interpretation challenges [96]. As a result, out of the studies we reviewed,

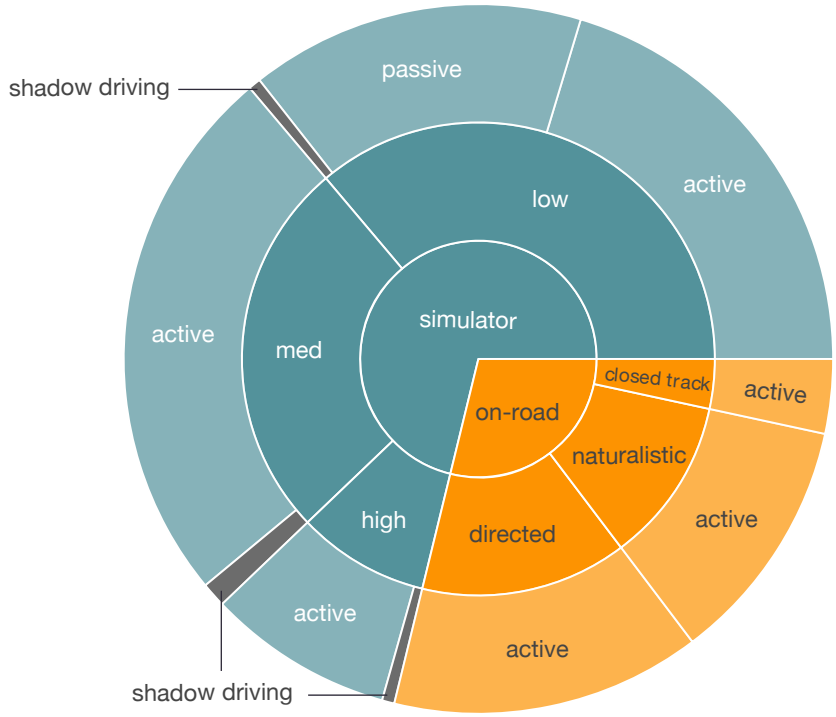


Figure 2.1: A diagram showing the distribution of the 247 reviewed studies with respect to recording conditions (innermost circle), environment fidelity (the middle ring), and types of vehicle control (the outer ring). "Shadow driving" indicates that subjects used inactive vehicle controls to match the drivers' actions in the pre-recorded videos. The segments in the diagram are proportional to the number of studies that used the corresponding data collection method.

only one-third were conducted on-road, and the rest in the driving simulators, predominantly of low and medium fidelity (see Figure 2.1). In nearly 20% of studies, subjects passively viewed the pre-recorded driving videos on the computer monitor.

Driving simulators. A good driving simulator offers several advantages over on-road conditions. Among them are the possibility to recreate the same route for all subjects, triggering events on demand, reduced behavior variability, scalability, and cost-efficiency [101]. However, the main disadvantage is that the *validity* of the simulation studies is not guaranteed. For example, absolute validity, i.e., exact match of the results obtained in the simulated and on-road conditions, is preferred, but very difficult to achieve. Thus, in most cases, relative validity is acceptable, meaning that the simulated and on-road results show similar trends but differ numerically [101].

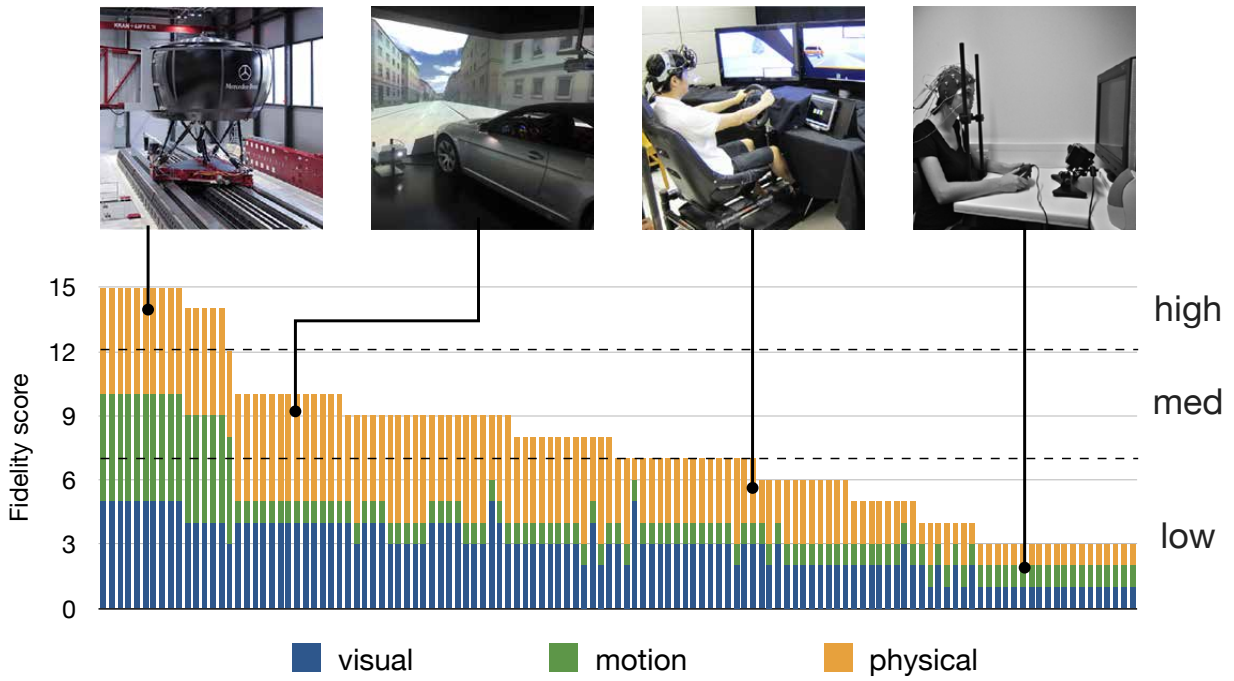


Figure 2.2: Simulator fidelity scores across 247 reviewed studies (one bar per study): visual (blue), motion (green) and physical (yellow). Fidelity levels: high (12-15), medium (7-11) and low (1-6). Samples above the bar plot (left to right): 1) high, 2) medium; 3), 4) low fidelity. Sources: 1) [97], 2) [98], 3) [99], 4) [100].

Another important factor is the fidelity of the simulator, of how faithfully it can reproduce the environment inside and outside the vehicle. Contrary to intuition, simulator fidelity does not always correlate with the validity of the results [102]. Part of the issue is that the concept of fidelity is not well defined and thus inconsistently used in the literature. Here, we adopt the scoring system based on visual, motion, and physical fidelity proposed in [102], which separates simulators into *low*, *medium*, and *high* fidelity. Figure 2.2 shows the distribution of the reviewed studies based on the simulator fidelity.

Visual fidelity is determined by the field of view, which has been related to validity [103]. Recently, virtual reality (VR) headsets capable of providing 360° view have been used to study driver's attention, but, to date, only one study validated VR in a driver hazard anticipation task [104]. The effect of visual realism of the simulation is less studied, and findings are inconclusive. For example, gaze patterns when watching driving footage vs a video game differed in one study [105] but were the same in another [106]. Robbins et al. [107] established that visual behavior at real and simulated intersections was similar as long as the task demands were at least moderate. In this study, low demand was associated with rural roads and moderate to high demand with suburban and urban environments.

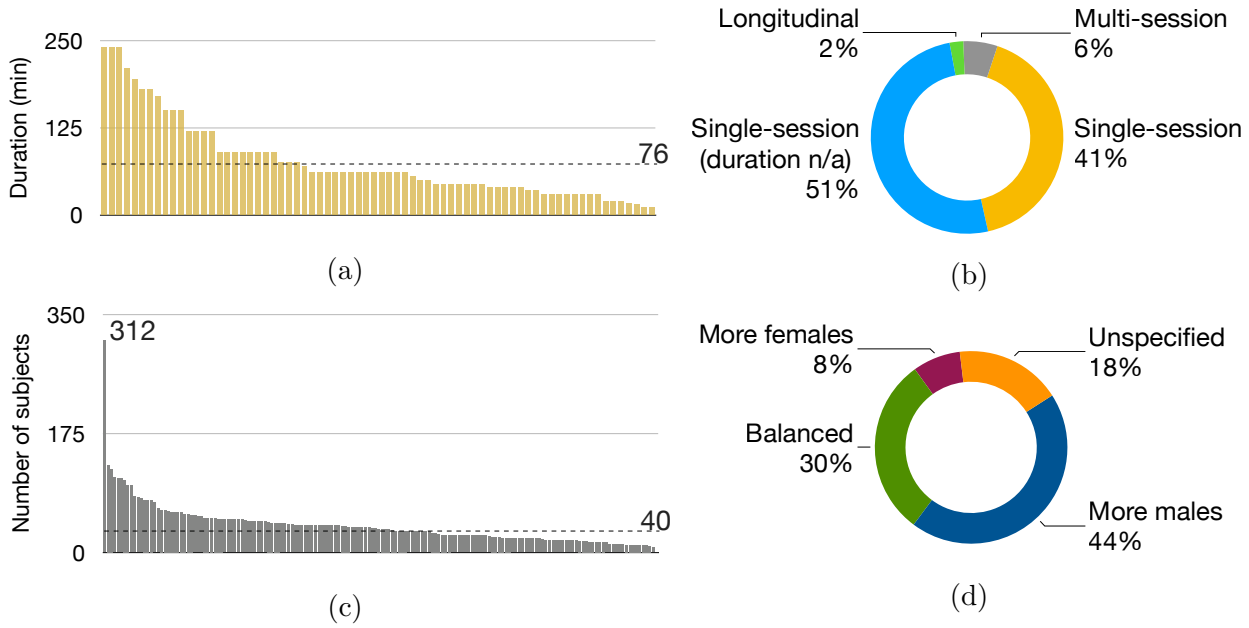


Figure 2.3: Properties of the reviewed behavioral studies: a) single session duration (one bar per study) and average duration across all studies (dashed line); b) proportions of single-session, multi-session and longitudinal studies, c) number of subjects (one bar per study) and average across all studies (dashed line); d) subject genders (balanced studies are those where the difference between the number of participants of different genders is within 5% of the total number of subjects).

Physical and *motion fidelity* of the driving simulators vary significantly depending on the appearance and authenticity of controls, from a full vehicle cab to a steering wheel mounted on the desk. Although visuomotor coordination during driving is well-studied, there are not many investigations into changes in attention distribution caused by a lack of active vehicle control. Compared to passive viewers, drivers engaged in control scanned the environment less and had slower reaction times to hazards [108]. In some studies conducted in low-fidelity simulators, the subjects are asked to "shadow drive", i.e., to operate the controls to match what they see in the prerecorded driving footage [109, 105]. Such a mimicking task increases workload and can lead to the same attention allocation as during active vehicle control [106].

The effects of motion cues on attention distribution are understudied. However, further research is warranted, because simulators without motion bases led to more variation in lane position, indicating potential links with perception [102].

Experiment duration and subjects. *Duration* of the study can also affect transfer of findings to a real-world traffic environment. Most experiments we analyzed last just over 1 hour (Figure 2.3a). The actual driving time can be much shorter because most experiments include instruction, breaks, and post-experiment surveys. Although not often discussed,

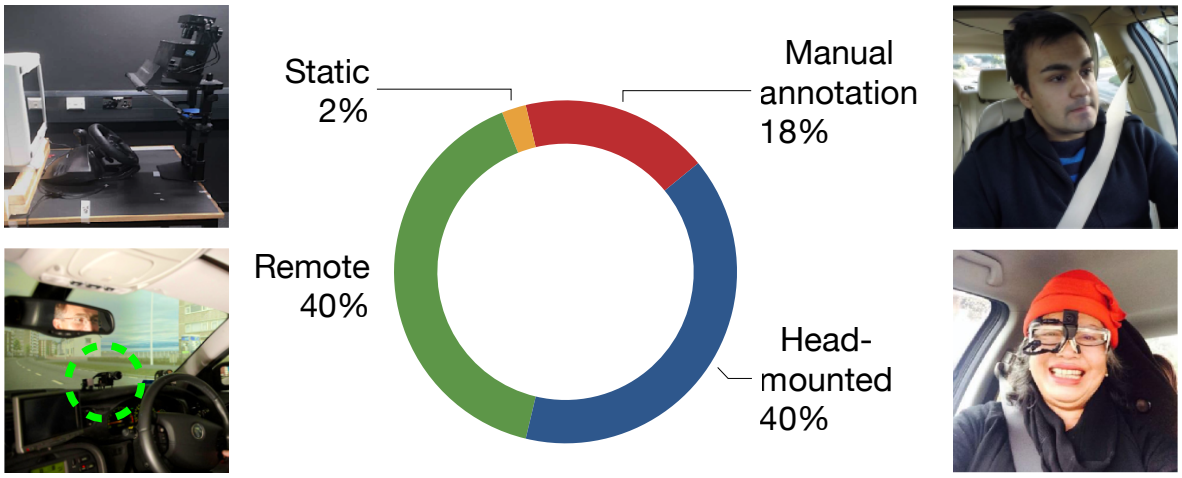


Figure 2.4: Eye trackers used in studies. Sources: static [108], manual annotation [113], head-mounted [114], remote [115].

lack of familiarity with the simulator may change the behaviors of the drivers, e.g., looking at the speedometer more frequently at the beginning of the experiment [110]. Thus, 5–10 minutes of familiarization, normally given to the participants, may not be sufficient to elicit normal driving behaviors. Care must be taken even when recording in a familiar vehicle and naturalistic conditions. For example, during the 100-Car Naturalistic Driving Study, participants were aware of being monitored in the first hour of the experiment. As a result, they were initially more attentive and were involved in fewer incidents, but later reverted to their normal driving style [111].

The majority of the studies we considered are single session, few involve multiple sessions, and even fewer are longitudinal, where information is collected over several weeks [92] or months [112] (Figure 2.3b).

The *number* and *diversity* of subjects are also an issue as studies involve 40 subjects on average, and only 30% are balanced across gender (Figures 2.3c and 2.3d). The majority of studies recruit students or car company employees, who may not be representative of the general population.

2.1.2 Recording equipment

Eye trackers. As mentioned in Chapter 1, eye-tracking technology is crucial for studying gaze patterns of drivers. By type, eye trackers are subdivided into static and head-mounted (see Figure 2.4). In the automotive domain, *type*, *sampling rate* (in Hz), *accuracy*, and *precision* of eye trackers matter the most. In general, static tower-mounted systems have the highest accuracy and precision while being the most restrictive. Head-mounted and remote

Figure 2.5: Images from driver-facing cameras. Drivers are looking towards left, right, rearview mirror and radio. Source: [117].

eye trackers allow normal head movement but produce higher errors [22]. As a result, only a few studies use static eye trackers since most driving tasks require head movement. Less invasive head-mounted and remote trackers are used equally often (each in $\approx 40\%$ of studies we reviewed), and the remaining 18% of works rely on the manual coding of gaze from videos.

The *sampling rate* of 30-60 Hz is the most common as this is also typical in video cameras. For some measures, e.g., fixation duration, low sampling rate can be mitigated by collecting more data, while for measures such as saccadic velocity and acceleration, a higher sampling rate is a must [22].

Accuracy and precision assess the quality of the eye tracking data. The former is the difference between the target and estimated gaze location, and the latter measures the reproducibility of repeated gaze measurements. Values reported in studies are often taken from the manufacturers' manuals and apply to the optimal recording conditions. However, both precision and accuracy degrade significantly under low illumination conditions and for gaze angles beyond 15° [116] that often occur during driving.

Manual coding of gaze. In some cases, the use of an eye tracker is impossible or cost-prohibitive. For example, during longitudinal naturalistic studies it is difficult to deploy eye-tracking equipment due to its cost and maintenance requirements. Thus, driver-facing cameras are used as an inexpensive and non-intrusive alternative. On the flip side, analysis of drivers' gaze patterns requires the labor-intensive process of annotating video frames from the driver-facing camera with text labels specifying in what approximate direction the driver is looking at the moment, e.g., rearview mirror, straight ahead, or eyes closed [118]. Often multiple trained annotators are needed to reduce subjective judgment [119, 1, 120] caused by the subtle changes in drivers' appearances and large individual differences between drivers (Figure 2.5). Because sampling rate of a driver-facing camera rarely exceeds 30Hz and can be as low as 10 Hz [121, 122, 118, 120, 123, 124], manual annotation may miss eye movements

(a) (b) (c)

Figure 2.6: Examples where gaze of the subject could not be estimated due to: a) reflection, b) blink, c) large head angle. Source: [126].

occurring at a higher rate [125], increasing bias towards longer glances.

Data loss. More than 12% of the reviewed behavioral studies report data loss from 10% up to 50% [127, 128, 129, 93, 130, 131] due to calibration issues [103, 132], large tracking errors [133], and faulty equipment. In particular, changes in the illumination, reflections, self-occlusions and blinks (Figure 2.6) affect studies that rely on manual annotation of gaze from videos [134, 123, 118, 135].

2.2 Measuring drivers' attention

While it is possible to conduct analysis of the raw eye-tracking data and use it in applications directly, the majority of studies we reviewed, further process and aggregate recordings. In this section, we describe how eye-tracking data is treated, what measures can be derived from it, and their meaning. Since driving is a visuomotor task, vehicle control information is often used to assess the effects of visual behavior on driving performance. We provide a brief overview of the common vehicle control measures and their relation to gaze measures.

2.2.1 Types of eye movements

Human eyes can make four types of eye movements: 1) *saccades* to move gaze from one location to another, 2) *stabilizing movements* against head and body motion (vestibulo-ocular reflex) and retinal slip (optokinetic reflex), 3) *smooth pursuit* to follow targets moving at speeds below 15° per second, and 4) *vergence* to focus on objects at different distances [136].

Research on driving and attention focuses predominantly on the episodes when the gaze is stationary (*fixations*) and on transitions between them (*saccades*). Other types of eye movements are rarely considered. Although vergence may reveal information about drivers' 3D location of gaze [136] or drowsy state [137], detecting it is challenging in the driving environment due to varying lighting conditions [138]. Likewise, microsaccades can be useful

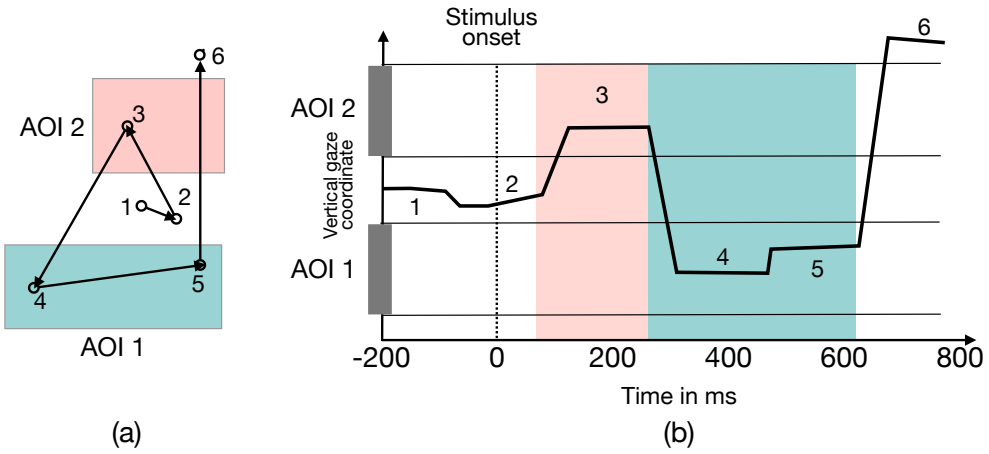


Figure 2.7: Schematic illustration of a glance: a) sequence of fixations between two areas of interest, b) space-time plot of fixations. Glances to AOI 1 and AOI 2 are shown in teal and pink respectively. Images modified from [22]).

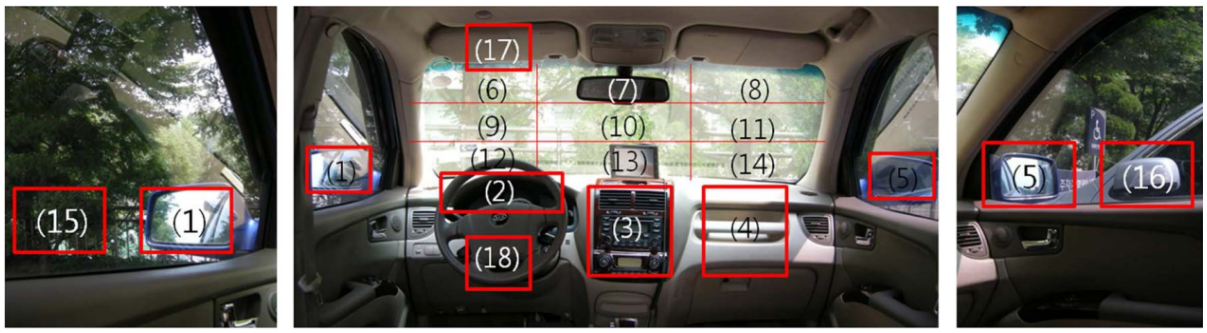
for detecting a distracted state of the driver [139] but are difficult to record outside of laboratory conditions [140].

2.2.2 Data pre-processing

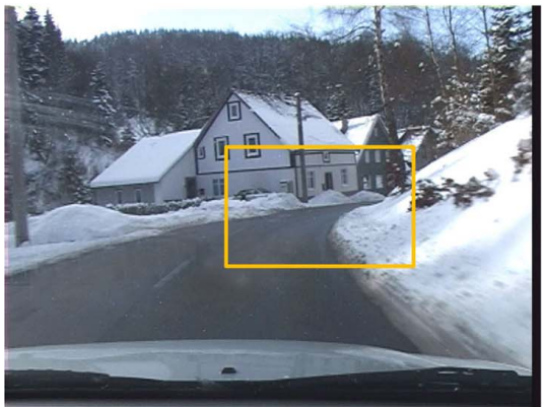
Fixations and glances. *Fixation* is a brief period of time lasting from a fraction of a second and up to several seconds when eyes are held steady at a single location. In reality, eyes rarely remain completely still due to small eye movements such as drifts, tremors, and microsaccades [136].

Fixation extraction from eye tracking data is done based on several parameters: *fixation duration threshold*, *gaze velocity*, and *dispersion* (the spread of gaze coordinates aggregated into a single fixation). However, in the literature, these parameters are infrequently and inconsistently specified. For example, duration cut-off varies from as low as 50ms [88] and up to 333ms [141], with justification rarely provided. Likewise, dispersion thresholds range from 0.5° [107] to 5° [142]. There is ample evidence that these parameters change data distribution thus affecting averages and variances used to compute most measures and variance-based tests commonly found in the literature [22]. Consequently, the under-specification of these settings makes the results difficult to reproduce and compare. Thus a proper justification of the chosen settings and analysis of their effect should be provided.

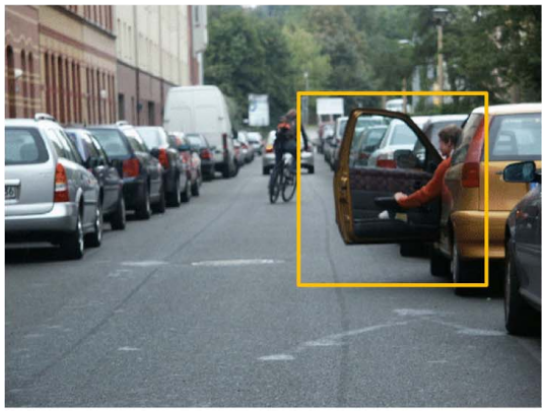
Identifying fixations is challenging in a driving context where everything is in relative motion. The presence of smooth pursuit common in such settings may result in inaccurate fixation data. To ensure correct fixation detection, one can modify the experiment to reduce such eye movements [143], manually inspect gaze data [144, 141] or use custom algorithms



(a)



(b)



(c)

Figure 2.8: Examples of static and dynamic areas of interest (AOIs). a) Static AOIs are defined with respect to the vehicle interior and remain unchanged for the duration of the study. Dynamic AOIs correspond to external areas or objects in the scene. Here, b) shows an AOI for a latent hazard (caused by poor visibility due to obstruction and road curving) and c) corresponds to a materialized hazard (an open vehicle door). Sources: a) [126], b) and c) [149].

[133].

Glance (called *dwell* in psychology [22]) refers to a single passage through an area of interest (AOI). According to ISO 15007 standard [145], glance duration is defined as the sum of transition time to the area and duration of fixations and saccades within the area (see Figure 2.7). Alternatively, transition time between the areas of interest may be counted towards the origin, instead of destination, or considered as a separate category. While the origin method is easier to annotate manually, all three approaches yield similar gaze-based measures [146]. When only the fixation durations are used, excluding saccades altogether, glance durations tend to be under-estimated [147, 148].

Areas of interest (AOI) serve to further aggregate fixations, glances or raw gaze data. *Static* AOIs are often defined with respect to the vehicle's interior or the driver's frame of reference. *On-* and *off-road* are the most basic gaze zones used in many studies, however, they

are not consistently defined (e.g., the whole [18, 150] or part of the windshield [122, 151, 152], or circular area with a radius ranging from 6° [153, 115] or 8° [121, 95, 154, 155]). The number of AOIs in studies varies significantly from 3 zones [156] to as many as 18 [157, 158] (Figure 2.8a). More fine-grained zoning allows for detailed driver behavior analysis and modeling.

Dynamic AOIs are primarily used for objects outside the vehicle. These areas are not fixed and may be updated due to the motion of the object or ego-motion of the vehicle. Dynamic AOIs are predominantly used in studies involving hazard perception. For example, to evaluate drivers' hazard perception skills, dynamic AOIs are defined around pedestrians [159, 160] or approaching vehicles [161, 74] (see Figure 2.8c). If a hazardous situation has not materialized or involves multiple objects, hazard zones are defined as bounding boxes around the relevant area in the scene [104, 162, 149, 163] (Figure 2.8b). The assumption is that drivers may not anticipate a hazard in time or will miss it entirely if they do not look at the predefined region. Since such hazard zones are tailor-made for each situation, they carry a degree of subjectivity.

2.2.3 Eye-tracking measures

Driver gaze data are further processed for statistical analysis and modeling. Eye-tracking measures help analyze different aspects of drivers' gaze allocation, such as duration of gaze on an object, frequency of visits to specific areas, and sequence and frequency of transitions between areas. Most eye-tracking measures aggregate basic units (fixations, glances, and saccades) over some time interval using summary statistics, e.g., mean or standard deviation. However, because there are several ways of assessing the same property of driver's attention, there is a staggering number of measures, many of which are used in only a handful of studies. We identified nearly a hundred different measures in the literature and organized them using the taxonomy proposed in [22]: *position*, *numerosity*, *movement*, and *latency/distance* (see Figure 2.9).

Position measures determine *where* and *for how long* the driver held gaze. *Gaze duration* is one of the most common measures that has been linked to depth [150, 164] and speed of processing [143, 165], task demand (shorter gaze during complex maneuvers [107]), visibility conditions (longer fixations when visibility is low [166]), and object importance (longer gaze when hazards are present [167, 165, 164]). *Gaze duration/location distribution* (e.g., histogram) shows a temporal and spatial allocation of gaze. For instance, in [168] distributions of off-road glance durations are shifted due to distractions. *Horizontal* and *vertical spread* (mean, standard deviation, and variance of gaze position) indicate the breadth of visual

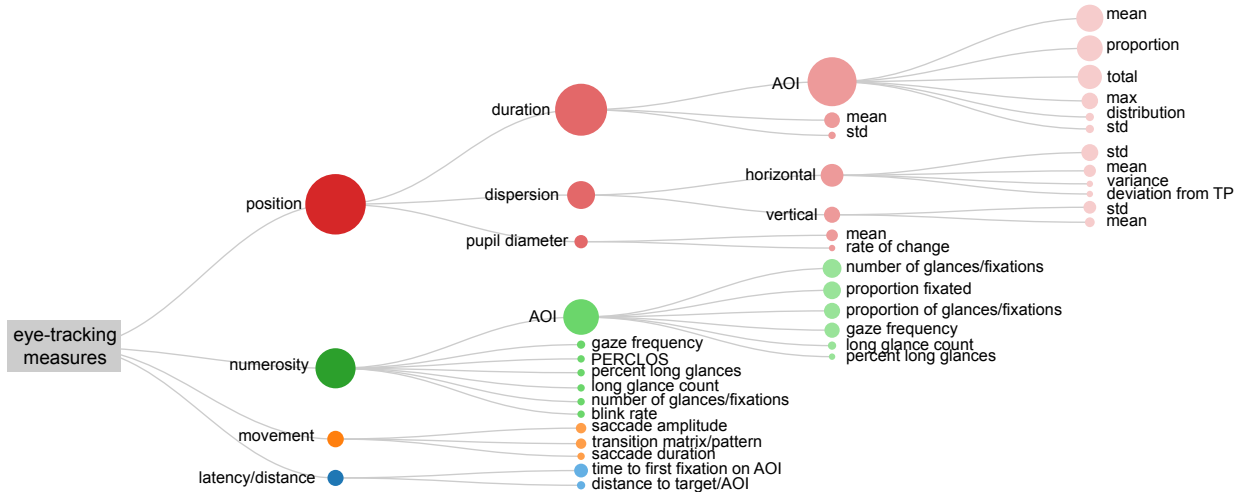


Figure 2.9: A dendrogram of eye-tracking measures. Node sizes are proportional to the number of studies using the corresponding measures. Only measures used in at least 3 studies are shown.

exploration of the scene and are associated primarily with driving experience and changes in workload. For instance, expert drivers have a larger horizontal spread of gaze than novices [166]. Cognitive tasks [100] and emotional conversations [169] were shown to decrease the horizontal gaze dispersion, whereas some visual tasks increased it [128].

Numerosity measures estimate *where* and *how often* the driver looked in the entire scene or within specific AOIs. The frequency of gaze indicates the rate at which the drivers sample the environment. For example, drivers check side mirrors more often when merging onto a highway [154], look around when prompted by navigation instructions [170] or when feeling drowsy [92], and fixate less on the road when distracted [99, 150]. The number of glances towards AOIs inside and outside the vehicle is indicative of scanning activities [171] during regular driving and when performing maneuvers [172].

Blink rate (frequency) (blinks per second [109, 173, 174] or minute [175]) is affected by changes in workload, but there are conflicting reports on *how* they are affected. In [100, 175] blink rates increased during a cognitive task, in [173] the opposite was shown and no effect was found in [109]. The nature of the secondary tasks could cause this discrepancy, however, more investigation is needed [100]. Increased blinking is associated with drowsiness [176] and has been used for fatigue detection in practice [177, 178, 179, 180]. *PERCLOS* (percentage of eye closure) is another drowsiness measure [137]. It is more reliable when computed over a long time interval [181] which delays drowsiness detection and makes the measure unsuitable for microsleep detection [182].

Movement measures focus on the *sequence* and *frequency of transitions* between different areas. The latter show visual exploration strategies of drivers and are used for modeling

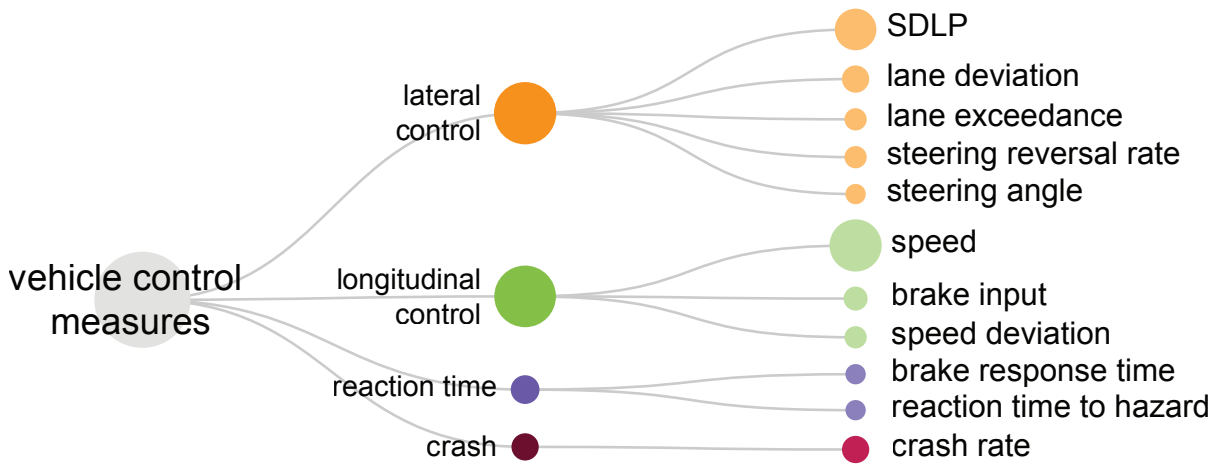


Figure 2.10: A dendrogram of vehicle control measures. Node sizes are proportional to the number of studies using the corresponding measures. Only measures used in at least 3 behavioral studies are shown.

driver's inattention. Transition matrices visualize changes in attention due to secondary task involvement [183], use of in-vehicle navigation systems [184], or the presence of digital billboards next to the road [185]. Transitions can help infer internal decision processes, e.g., at intersections while performing different maneuvers [186, 187, 172].

Latency/distance measures evaluate the *reaction times* and *distances* to fixated objects in the scene. *Time to the first fixation on AOI* indicates how fast drivers focus on objects or hazards. For instance, drivers fixate faster on actualized hazards rather than precursors [188], notice pedestrians intending to cross earlier [164] and look at the approaching cars sooner than motorcycles [165], suggesting that the severity of the hazard is linked to reaction times. Other studies indicate effects of workload. Naturally, drivers involved secondary tasks react to hazards slower [189]. Another experiment showed that even controlling the vehicle contributes to workload. In a simulation study, subjects who passively viewed driving footage were able to detect hazards faster than those who had to drive the vehicle [108].

2.2.4 Vehicle control measures

Many studies supplement gaze-based measures with measures of driving performance that look at aspects of vehicle control. Following the taxonomy in [101], vehicle control measures are divided into *longitudinal* and *lateral* control, *reaction time*, and *crash rate* (Figure 2.10).

Lateral control is typically measured in terms of a vehicle's lateral position within the ego-lane or by examining the frequency and magnitude of steering wheel rotations. *Lane deviation*, defined as mean (MLP) or standard deviation of lane position (SDLP) with respect

to the road center, is a standard measure for evaluating lateral control. There is conflicting evidence regarding the effects of secondary tasks on lateral control and their interpretation. Some studies show increased SDLP resulting from secondary tasks [123, 190, 191, 192, 193] or external distractions [185], whereas in others, SDLP reduced or did not change under increased visual [194, 195] and cognitive load [196, 197, 155, 128, 198]. Explanations include inhibition of top-down interferences with the automatized lane-keeping task [196, 167] and rigid steering [128]. Other findings suggest that lateral control is more demanding than longitudinal, and therefore high cognitive load makes the drivers focus more on the lateral control diverting resources from other tasks [198]. Lateral control may depend on the type and complexity of the secondary task [195] and driving experience (professional drivers retain their lateral control even when performing demanding tasks [199]).

Longitudinal control refers to changes in ego-vehicle speed and distance to the vehicle ahead. Many papers report reduction in speed as a compensatory strategy in high workload conditions such as reduced visibility [200], performing secondary tasks [127, 194] while navigating curves [128], or due to age-related deficiencies [201, 202, 203]. Speed reduction has also been associated with hazard anticipation, e.g., scanning at level crossings [204] or slowing down to notice potentially hazardous objects sooner [188, 103].

Headway to the lead vehicle (LV) is measured in terms of distance [99], time [205, 206] or time-to-collision (TTC) [168, 115]. When the safety margin is low, a stronger response is required to avoid a crash [99]. In some studies, inverse TTC is used to operationalize the change rate of LV looming and has been linked to crash risk in rear-end collisions where braking was used as an evasive maneuver [1, 168]. When the LV is far away, drivers are more likely to engage in secondary tasks [1]. Drivers already engaged in secondary visual tasks tend to compensate for distraction by increasing the headway [99]. At the same time, a higher workload causes the variation of headway distance [191] or time [192]. Engaging automation (e.g., adaptive cruise control) improves headway compared to manual control [115, 206, 121].

Reaction time measures assess how quickly the driver can react to hazardous events by applying brakes or steering. Specifically, *brake response time* is the time from the hazard appearance to brake onset. As with other vehicle control measures, it takes the drivers longer to react if they are distracted (e.g., texting [150]) or sleep-deprived [199]). The response time [98, 208], distance [209] and TTC [210] to the target may improve with early warning provided by driver assistance systems or better visibility of the scene.

Crash rate is a direct measure of driving safety, infrequently reported because of the rarity and severity of such events. Even in large-scale naturalistic datasets, crashes are too infrequent, thus actual crashes are often combined with near-crash events (where an accident

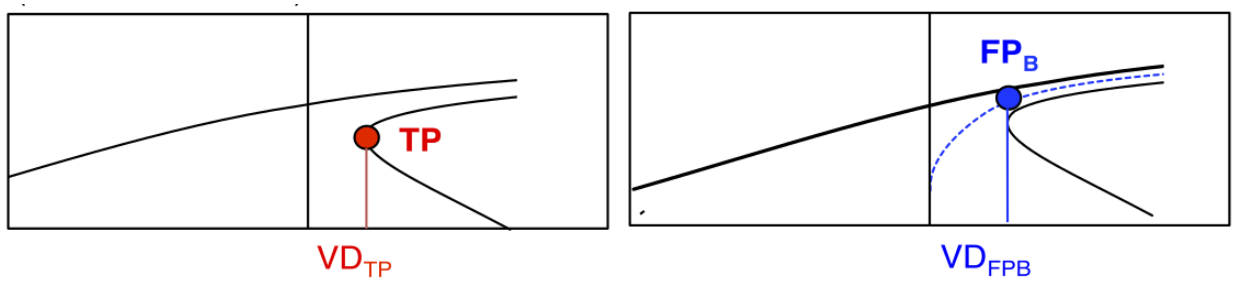


Figure 2.11: A schematic illustration of steering points that drivers use when driving on the curved roads. Abbreviations: TP—tangent point, FP_B—future path, VD—the egocentric visual direction. Source: [207]

was avoided by performing an evasive maneuver) to increase the statistical significance of the results [168, 211]. Crash risk has been directly linked to insufficient visual information intake due to long off-road glances [112, 168, 211], sleep deprivation [212], and reduced visibility [171, 103].

2.2.5 Visuomotor coordination

Road geometry, straight or curved, gives rise to different visual strategies. Driving on curved roads is perhaps one of the most studied visually guided tasks with 25 years of research accumulated to date. One of the long-standing debates is where drivers look to steer the vehicles (see detailed review in [207]). Land and Lee addressed it in their seminal work [213] where they proposed the tangent point (TP), i.e., the point on the image plane where the inside road edge reverses its direction, as a possible candidate, and established the relationship between the TP, the road curvature, and steering input. Wann and Swapp [214] proposed an alternative future path (FP) point (see illustrations of both points in Figure 2.11). Despite many other alternatives, TP remains the default theory for visually guided steering on curved roads. However, its status has been recently questioned. Due to gaze recording imprecision and spatial proximity of TP and FP, it is difficult to distinguish them [215, 96], especially while entering the curve since the area spanned by the curve in the field of view is too small [133]. Perhaps, the drivers rely on multiple reference points to navigate the curve, suggesting that the dichotomy between TP and FP may be false [96].

2.3 Endogenous factors

2.3.1 Driving tasks

A driving task is essentially any driving-related activity that the driver performs on the way to a destination or as part of the experimental study. Besides vehicle control, the most investigated driving tasks are *responding to hazards*, *vehicle following*, and *taking over control* (from automation) (see also Section 2.4.2). Other driving activities, such as maintaining speed [88], racing [216], and parking [125, 135, 217], but they are rarely considered in the context of drivers' attention.

For obvious reasons, virtually all experiments involving *hazard anticipation*, *detection*, and *response*, are conducted in simulators or with props [125, 95]. Hazardous scenarios differ by the *hazard instigator* and *type* (actualized or potential). Common instigators are pedestrians or cyclists [98, 209, 218, 72, 106, 164, 159, 219], stationary obstacles [208, 18, 220, 221, 222, 108], lead vehicle braking [208, 221, 19] and parked vehicles pulling out [222, 223]. Actualized (critical [224] or materialized [225]) hazards are fully visible, directly endanger the ego-vehicle, and require immediate response, e.g., pedestrians or stalled vehicles in the path of the vehicle. Latent hazards [226] present potential danger that may eventually actualize. Some examples could be a partially visible vehicle approaching the intersection from the side or a pedestrian about to step onto the road between two parked vehicles.

Hazard response is measured from the hazard onset and until the driver reacts (e.g., by applying brakes or performing evasive maneuver). Hazard perception, or how long it takes to find the target, is measured from the hazard onset until the first fixation on the hazard [165]. For latent hazards, fixations towards areas with cues and areas where the hazard may become visible are analyzed instead [227, 226, 104].

Findings across hazards studies are difficult to generalize because of the different scenarios presented to the subjects and other related factors being manipulated simultaneously. For example, hazard detection depends on the driving experience (Section 2.3.4). The safety impact of the hazard may also affect how fast it is identified [149, 70], whereas saliency and eccentricity [227, 149] have little effect.

During *vehicle following*, drivers often look at the lead vehicle (LV) and near road region before LV to adjust speed and maintain lane [128, 121, 228]. The LV presence also increases the risk of rear-end collision, especially if the driver is distracted [168, 229]. When an accident is imminent, drivers fixate longer on the LV during avoidance maneuvers [173]. Driver assistance systems such as forward collision warning (FCW) help direct their attention to the LV [205, 206]. Lead vehicles are often used to direct drivers [141, 230, 95], control the timing of events [141] and examine the extent of drivers' spatial attention [231, 232].

2.3.2 Non-driving-related tasks

Besides following the rules of the road, drivers must be fully aware of their surroundings to cooperate with other road users and respond to sudden events. In practice, according to surveys [233], crash reports [234], and naturalistic studies [1], a significant proportion of drivers engage in secondary non-driving related tasks (NDRT), i.e., tasks not related to vehicle control or improving forward, indirect or rearward visibility [235]. Typical activities include using a cell phone, in-car navigation and entertainment systems, talking to passengers, eating or drinking, and smoking [134].

Taxonomy and effects of secondary tasks. Unlike crash reports that group secondary tasks by type, a modality-based taxonomy provides a better understanding of what cognitive functions are affected [236]:

- Visual—require averting gaze off the road (e.g., cell phone);
- Cognitive—require thinking (e.g., talking, solving puzzles);
- Manual—require moving hands off the wheel (e.g., smoking).

Alternative taxonomies were also considered [237].

Visual tasks require the driver to look away from the road to read messages on roadside signs or cell phones. *Visual search* is a common task where the subject is asked to find the target and verbally report the result, e.g., looking for logos on the highway exit signs or checking the distance to the destination on the guiding signs [223, 238, 239]. The latter is an easy task with detection rates in the $96-98\%$ range [239, 238] and not as visually demanding as reading vendor signs [238, 239]. Artificial tasks, e.g., searching for target characters in random text strings [192, 227], allow control of task complexity. Sometimes, both approaches are combined, e.g., searching for characters in roadside signs [230].

Visual search affects attention as evidenced by longer and more frequent glances off-road [192, 223, 239, 203, 202]; however resulting changes to driving performance are inconsequential for road safety [223, 238, 239]. More important is the degradation of hazard perception and anticipation, e.g., fast-moving pedestrians [230] and latent hazards [227, 230] failing to attract attention in the presence of distractions, although these deficiencies are not directly linked to crash risk.

Text reading. Drivers occasionally read road message signs with reminders (e.g., fasten the seatbelts) or notifications (e.g., traffic conditions). In experiments where subjects were expected to read and report messages' contents, more prolonged off-road glances were observed [142, 240, 195, 241] and increased non-linearly as a function of the message length and vehicle speed [195]. Similarly, reading text from the cell phone or in-car screen reduces

the amount of time looking at the road [242, 155], however not as drastically as during text entry.

Visual-manual tasks are linked to random scanning patterns, a larger number of glances, shorter mean glance durations, and shorter on-road glance durations [183]. Arguably, visual-manual tasks have a high impact since they compete with cognitive and attentional resources needed for driving [243], however, as will be discussed later, even auditory cognitive tasks can cause dual-task interference.

Surrogate Reference Task (SuRT) is an artificial visual-manual task where the subject is periodically asked to look at a display, find a target among distractors, and select it (via touchscreen or button) [244]. The task is often used to simulate interaction with in-vehicle infotainment or smartphone usage [245, 246, 220, 208] and leads to more frequent and shorter glances towards the road [18].

Text entry is more disruptive than reading since it leads to longer off-road glances and fewer glances on the road [124, 155, 118, 242, 150]. The variability of the off-road durations across drivers is high due to the different compensatory strategies they employ [155]. Length of the message affects text entry more than reading [155], and entry errors further disrupt dual-tasking by increasing task time and total time looking off-road [191]. A voice interface does not solve the problem completely because drivers still check the screen for confirmation [247]. *Dialing a cell phone number* requires fewer long glances than texting [118] and can be somewhat mitigated by a tactile keyboard for those used to it [194].

Radio tuning is a benchmark task for a reasonable level of distraction during driving [248] involving three steps—selecting radio function, band, and tuning to the station using a knob. Like other NDRTs, it primarily affects long glances off-road [247, 168] and total eyes off-road time [211]. *Playlist search*, another common task [249, 190, 250], is comparable to a tuning task for short lists (≈20 songs). Larger lists of 80+ items significantly undermined driving performance and required more glances to the device [190].

Cognitive tasks. States like mind-wandering cannot be ruled out during the experiment but are difficult to identify [251]. Thus, cognitive tasks induce such a state or increase drivers' cognitive load in general.

Delayed digit recall (N-back task) is an artificial task initially proposed for measuring working memory capacity [252]. The subjects listen to digits being named and repeat them as presented (0-back), a previous digit (1-back), or a digit two steps before (2-back). As expected, task performance for the 0-back task is very high at 99% but quickly degrades to 85% for the 2-back task [201] (similar results are reported without driving [253]). The effects of this task are somewhat inconsistent. In some cases, it caused a decrease in horizontal gaze spread, suggesting reduced peripheral awareness [127, 201]. In a more recent experiment,

the distribution of attention was reduced uniformly across the visual field, affecting foveal vision [232]. Another study found an effect on blink rate but not other gaze measures or driving performance [175].

Conversation with passengers or on the handsfree phone, despite being common and socially acceptable, can still lead to distraction [254, 173], especially if the topic is emotional [169]. When talking to someone, drivers fixate less on critical objects [167] and are more likely to make "looked but failed to see" errors [72].

Visual-cognitive tasks such as *watching videos* are frequently requested and performed during automated driving. They may mitigate some effects of automation fatigue [115] at the cost of significantly reducing drivers' attention towards the forward roadway [193, 115, 251, 109].

There is little work on the joint effects of visual, cognitive, and manual distractions. Several studies [99, 255, 210] reveal a complex interplay between the effects of different types of distractions that may result in compounding effects that are difficult to anticipate. Likewise, only a few works investigate effects of different secondary tasks in the same conditions [193, 134], the same tasks in different conditions [109], or effects of the task complexity [192].

2.3.3 Demographics

Age is correlated with crash risk, which is highest for teen drivers, reaches a minimum for drivers in their 60s, and increases again for drivers in their 70s and beyond [256], caused by age-related functional decline despite having decades of driving experience. Since nearly 20% of drivers in Canada [257], the USA [258], and Japan [259] are older, understanding the effects of age on driving is a pressing concern.

Older is not defined in the literature, but a 60-64 years cut-off is common [260, 261, 262]. Even though the functional decline is minimal in the late 60s [114, 226], many studies focus on healthy older drivers in their 60s and 70s, and even drivers 50+ years old [123, 90, 263], whereas drivers aged 80+ are rarely considered [141, 131], which may explain the inconsistent and small effects of age across studies.

Effects of aging manifest themselves more in situations with complex road geometry and traffic, especially when compounded by distractions. Overall, older drivers have more conservative visual attention strategies [203], making fewer and longer fixations on important objects at intersections than younger groups, suggesting slowed processing and deficits in selective attention capacity [143, 114, 264, 141, 186]. In simulation studies, when exposed to external distractors, such as billboards, older drivers were not as accurate at detecting relevant signs [203], slowed down to compensate for the reduced visual function [202, 203],

and had lower off-road glance duration and frequency [203, 241, 132]. However, in an on-road experiment, older drivers made more long glances at billboards than younger age groups [158].

Older drivers showed compromised ability to multitask results in the prioritization of glances necessary for vehicle control [226]. Similarly, older drivers have trouble lane-keeping and attending to the road when using a cellphone compared to younger groups [124]. Difficulty to adjust search patterns in the presence of distractors was shown in a hazard detection study where no vehicle control was required [170].

Automation and ADAS offer improved mobility for the elderly, however, the interaction of older drivers with new technology is still not well understood. Several recent studies looked into gaze allocation changes when taking over from automation after being engaged in a secondary task, but none found significant age-related effects [254, 263, 131], possibly due to small sample size and inclusion of drivers in their 50s.

Gender is primarily associated with differences in traffic violations. Male drivers are more often involved in accidents, receive more traffic fines and make more insurance claims [265], likely due to the propensity towards risk-taking [266] and overconfidence in their driving ability [267]. Gender-related differences in attention allocation are not well-studied, even though most experiments include both genders.

No gender-specific gaze patterns were found in rear-end collision avoidance scenarios [173] and other near-crash events [90] in the presence of distractions. Drivers of both genders were equally likely to engage in secondary activities during partially automated driving and continued to make long off-road glances after returning to manual control [263]. Overall, female drivers were more attentive when approaching intersections, looked around more often, focused longer on the conflict vehicles, and were involved in fewer collisions in a simulation study [171]. In a different study, young male drivers were more prone to distractions—when given a cell phone, they looked less at the road and made more continuous long glances ($>$ 2s) towards the phone compared to female participants [194].

2.3.4 Driving experience

The vast majority of the studies we reviewed use subjects with driving experience, except for 5% of studies where it was not stated clearly. We have found only one example in our selection where non-drivers were selected on purpose [105]. Driver experience is usually defined as a combination of valid license, frequency of driving, or annual mileage (see Figure 2.12). Other aspects, such as *location* (where the experience was gained), *other vehicle types* (e.g., motorcycle [165]), *racing* [216], *driving style* [268], or *exposure to automation* [251], are less

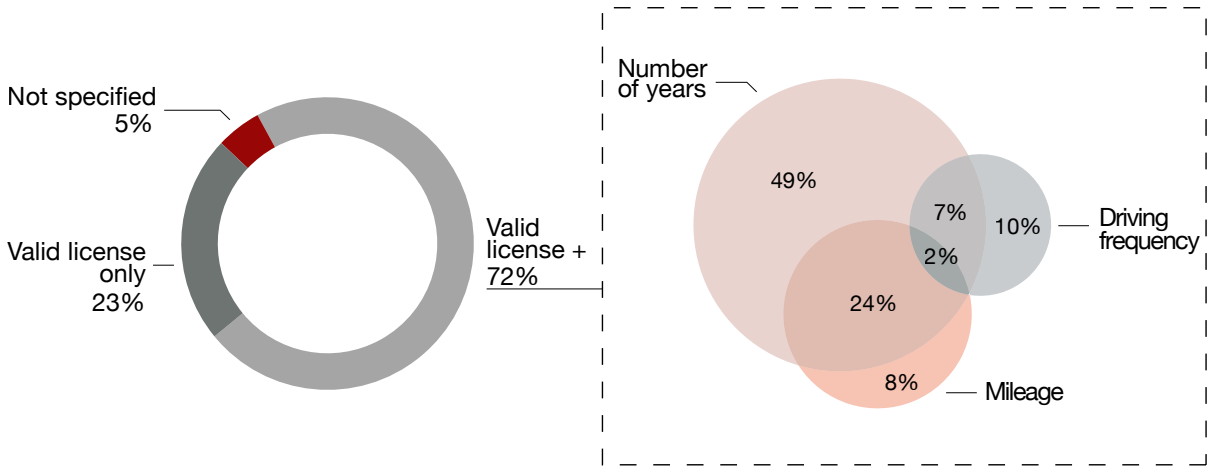


Figure 2.12: Definitions of driving experience in behavioral studies. In addition to having valid license, most studies require the subjects to specify other aspects of driving experience, such as number of years, frequency of driving (e.g., once per week), and annual mileage.

studied.

Qualitative descriptors, *novice*, *experienced* and *professional*, are not consistently defined in the literature. Two years of driving is the most common cut-off for a driver to be referred to as experienced [173, 269, 270, 149, 271], although justification is rarely provided. Often, to show larger effects, subjects with ≤ 1 year of driving experience are compared to those with $4-15$ years [224, 103, 186, 272, 164, 173, 225].

Differences between novice and experienced drivers. Novice drivers are over-represented in crash statistics, which is often attributed to limitations in their visual search strategies compared to the experienced drivers. Differences between the two groups were found in fixation durations [271, 166], saccade amplitudes [271], number of fixations [166, 225], vertical [225, 144] and horizontal [103, 166] spread of search. However, recent meta-analysis [6] established that only horizontal gaze spread remains statistically significant when all studies are pooled together.

The effects of driving experience are apparent in hazard anticipation performance. Although all drivers agree on what is considered a hazard [271, 219], more experienced ones are better at anticipating hazards (up to 30% more than novices [104]) and taking proactive measures [273, 221, 188, 219, 225]. Despite missing the early cues, novices may detect the same number of hazards of certain types, albeit slower than more experienced drivers [188]. The hazard's saliency has little to no effect on detection [271, 149, 230]; however, more imminent hazards consistently attract more gazes [149]. Novice drivers are more affected by distractions, detecting fewer hazards [166], reacting slower [241], and making more lane departures [230].

These differences may be caused by poor visual strategies of novice drivers [225] or longer time to process and respond to the information [271, 149]. Research on driver training techniques suggests that novice drivers' visual scanning patterns can be improved. While simply viewing recordings of experts' eye movements was not sufficient [163], hazard perception training in a simulator followed by instruction sessions appeared to have an effect [274, 163]. Improvements after hazard training persisted for 6-8 months but still did not reach experienced drivers' performance [275]. The safety effects of such interventions are unknown since visual behaviors of novice drivers were not linked to crash statistics.

Familiarity of the environment. Most drivers have little experience in different countries and spend most of their time commuting to work [276, 277], presumably traveling the same routes multiple times. Existing evidence agrees with the intuition that familiarity encourages inattention to the road ahead and driving-related objects such as signs while increasing dwell time on scenery and non-relevant areas [71, 278, 91]. However, a more recent study revealed that inattention might be caused by familiarity but ultimately depends on the context. Specifically, the effects of road type on attention to hazards were larger than those of familiarity regardless of prior driving experience [224]. A similar conclusion was reached in a study involving drivers from the UK and Malaysia who viewed driving clips from both locations [279]. Visual behaviors of both groups were similar irrespective of the clip origin, suggesting that visual strategy is moderated more by the driving environment rather than familiarity.

Experience with other vehicle types. Motorcyclists and bicyclists represent a small fraction of all road users but are involved in a disproportionate number of accidents, particularly at intersections. According to studies, dual drivers experienced in both car and motorcycle use are more cautious towards vulnerable road users (VRUs) [74, 165] and fixate on them more often [74] and for a longer time [165]. Since it takes dual and regular drivers the same amount of time to notice conflicting VRUs [165], it may be a processing rather than a perceptual issue. Note, however, that these results obtained in low-fidelity simulators were not confirmed in a more recent study that did not find significant effects of dual experience in a high-fidelity simulator [75].

2.3.5 Driver's state

A driver's internal state can be characterized in terms of physiological (e.g., drowsy, drunk) and emotional factors (e.g., stressed). The effects of internal state on driving performance are well-studied; fewer works examine changes in attention.

Drunk driving is illegal because it increases collision risk due to the adverse influence of

alcohol on neurocognitive mechanisms. For instance, alcohol causes more dispersed and random distribution of gaze due to reduced top-down regulation of gaze control [280], deficiencies in visuomotor coordination, and increasing distractibility (see review in [281]).

Drowsiness after sleep deprivation or fatigue caused by the monotony of driving has been linked to many adverse effects on driving performance [282], but its impact on attention is less studied. Unlike alcohol intoxication, drowsiness is more difficult to induce and detect. Self-reported drowsiness (e.g., measured using the Stanford Sleepiness Scale [283] or the Karolinska Sleepiness Scale [284]) often does not match objective physiological measures.

Sleep deprivation, even for one night, has been shown to alter attention allocation significantly. A naturalistic study of shift workers showed that 62% of trips were drowsy according to the PERCLOS metric [92]. In lab conditions, PERCLOS of sleep-deprived drivers during the 30 min drive was more than double that of the control group [212]. Drowsy drivers spend more time with their eyes closed due to the reduced rate and duration of blinks and fewer fixations. At the same time, larger saccade amplitudes and gaze entropy measures indicate more dispersed and random scanning patterns. Together, these effects point to impairments of top-down modulation of visual attention [176]. Besides losing visual information due to long blinks, drowsy drivers make frequent long off-road glances, possibly seeking novel stimuli to offset drowsiness [92].

The *monotony* of driving has been studied primarily in the context of automation since it takes longer to manifest during manual driving [285]. For instance, drivers reached a self-reported medium level of fatigue after $45-50$ minutes of manual driving. In contrast, when automation was engaged, half of the subjects reached that level after only 20 min [19]. It has been shown that reduced arousal levels during automation induce drowsiness [115] and micro-sleeps [286]. Drowsiness may also result from complacency, as extreme drowsiness was more prevalent in cases where subjects were not informed about the system's limitations and were not given warnings for system failures [95]. An obvious consequence is that bored or tired drivers react slower when the automated system fails and may lack sufficient situational awareness to take appropriate actions [19]. However, there is no clear path towards solving this problem yet. For example, an experiment where driving was gamified as a sequence of coasting challenges increased drivers' engagement, but had a negative effect on hazard perception [287]. Still, more work is needed to understand how to effectively counteract monotony, especially with partial automation (see Section 2.4.2).

Effects of **emotional state** on attention are not well-known. One study showed that affective imagery (positive and negative) overlaid on top of still images of hazards changed subjects' perception of risk and reduced fixation durations [288], suggesting an inhibitory influence of the emotional state on hazard perception. In another study, drivers with arachnophobia

demonstrated cognitive and visual tunneling indicative of decreased visual awareness and vehicle control ability when having a conversation about spiders [169].

2.4 Exogenous factors

2.4.1 Environment

Some visual information from the environment is important (traffic signs), and some should be ignored (billboards) despite being conspicuous by design. Other factors, e.g., road, traffic and weather conditions, may affect a driver's workload.

Intersections are more visually taxing than driving on straight roads [143] and where more accidents happen [289, 290]. The *geometric structure* of the intersection determines the approach angle of the conflict vehicle. It is easier to estimate TTC at the perpendicular angle of approach than at obtuse angles because lateral movement provides better optical flow than the movement towards the observer. As a result, fixation durations are highest at obtuse angles suggesting increased visual processing load [161]. Likewise, obstructions by vegetation or buildings increase visual scanning [171]. The *type* of intersection changes the focus of attention, e.g., at stop signs, drivers observe the intersecting zone less and focus on the ego-lane instead [291, 187, 172] but when the drivers have priority, they scan the intersecting areas to anticipate conflict vehicles [187]. *Conflicting road user type* also matters. For example, at intersections, drivers prioritize vehicles compared to other road user categories: motorcycles may be more difficult to detect due to their smaller size [74], bicyclists may be perceived as less risky [75], and pedestrians may be overlooked when merging into high-density traffic [292].

Signs and billboards. Drivers are expected to look at signs for information, e.g., speed limit, but many do not [70]. According to [94], only 1 in 4 signs was looked at, and a mere 7% were recalled later, possibly because signs are hard to identify with peripheral vision (except those with distinct shapes or colors [50]), especially if the driver is distracted [167].

Reading text from signs does not put a strain on the drivers nor does it affect vehicle control [203, 239, 195], however, it is unclear how the visual complexity of the sign affects driver attention. One study found no differences between standard 6- and 9-panel signs [238], another reported an effect of the number of panels and target location on off-road glance duration [223], and yet another showed increased attentional demand but only for unfamiliar targets in the 9-panel layout [202]. The location of the billboard also matters as overhead billboards had a larger impact than roadside ones [293].

Unlike signs, billboards are irrelevant for driving but are salient by design (see Fig-

(a) (b) (c)

Figure 2.13: Examples of typical roadside signs and billboards used in experiments: a) a guiding sign showing distances to the destinations, b) a 9-panel sign with logos of the vendors available at the next highway exit, c) an advertisement installed on the sidewalk (red color on the overlaid heatmap shows which areas of the ad attracted subject' gaze the most). Sources: a) and b) [239], c) [294].

ure 2.13c), thus attracting gaze more than any other roadside object (e.g., provocative [294] or visually complex ads [295]), resulting in higher fixation rates, dwell times, and proportion of long off-road glances [93, 295, 241]. Digital advertising with frequent transitions is particularly distracting as drivers react to motion in peripheral vision [158, 132, 93]. Although long off-road glances increase accident rates [1], recent meta-studies [296, 297] found no link between changes in visual behavior caused by the billboard presence and crash risk.

Traffic. In dense traffic or construction zones, drivers brake more often and have shorter braking affordances (i.e., room for maneuver at current braking capability [298]). As a result, they pay more attention to the rear-view mirrors [299] as maintaining situation awareness becomes more difficult [300]. Drivers hesitate to start secondary tasks [299, 190] or compensate for distractions by making shorter gazes away from the road [190, 118, 242]. When approaching an intersection or merging on a highway, drivers tend to scan more to find a gap and to monitor the approaching vehicles [301, 162]

Road properties and visibility. As discussed in Section 2.2.5, road geometry affects gaze allocation. Likewise, *road type*, rural or urban, can prime drivers for certain events, e.g., in residential areas, drivers identify more pedestrians and fixate on them longer than in sparsely populated suburban areas [159]. On rural roads, drivers are more sensitive to oncoming vehicles, either because they pose a larger threat or because there are few moving objects that could attract attention [118]. Visually demanding urban locations require more frequent, short, and widely distributed fixations and may lead to an increase of LBFTS errors compared to rural areas [70].

Less is known about the effect of lighting and weather on drivers' visual behavior since

only a handful of in-vehicle experiments were conducted at night [185, 121, 183, 90] and none in adverse weather conditions. In naturalistic studies, only a fraction of the data was recorded during bad weather [118, 302]. Existing evidence suggests that environmental conditions matter and deserve further investigation. For instance, in bad visibility conditions, drivers tend to focus more on the road [118], their sampling rates are reduced, and fixations are longer, suggesting increased processing time [166]. Driving at night affects mean fixation durations [183] and visual exploration. Since the preview distance of the road is shorter, and detection is degraded, drivers focus more on the forward path and driving-relevant objects [264, 121].

2.4.2 Driving assistance and automation

Over the past decade, safety and driver-assistive technology have become widely available in commercial vehicles, thus research on how new features affect attention allocation has seen a significant increase. Specifically of interest are systems corresponding to SAE [15] Level 2 (where two or more driving functions are automated) and Level 3 (autonomous but requires intervention), although Level 1 systems where only one aspect of driving is automated (e.g., steering) are also considered.

Effects of automated driving. Significant changes in drivers' attention allocation occur as various aspects of driving are automated and as the driver switches from being an operator to a supervisory role. For example, automated steering decreases visuomotor coordination [153] and reduces gaze to near road section suggesting that the driver is not engaged in the lane position adjustment control loop [228]. With park assist activated, drivers stop observing the environment and instead focus on the rear-camera view [135]. Multiple studies report that when Level 1-3 automation is enabled, drivers divert their gaze from the road and are more likely to engage in non-driving activities [18, 254, 109, 121, 205, 303, 205, 16, 147] and are less distracted when traffic or visibility conditions degrade [115, 229, 16]. Overall, drivers who use automation and pay attention to the road have better situation awareness (SA) because they are not engaged in manual driving [17], but there is a consistent negative association between monitoring frequency and trust in the system [304].

Switching back to manual control in case of automation failure will be required until reliable driverless technology becomes available. After exposure to automation, drivers take longer to regain SA, especially if they were distracted [18] or fatigued [19], therefore take-over request (TOR) must ensure smooth and safe transition to manual driving.

Quiet automation failure is the worst-case scenario as drivers may not realize that they need to intervene or may do so late. For example, 75% of subjects left their lane when

Figure 2.14: Augmented reality visualizations used to guide the driver's gaze. Left: "virtual shadow" is a less intrusive visualization that leads attention to the pedestrian about to cross. Right: bounding boxes drawn around every pedestrian create visual noise, so the pedestrian about to cross is not immediately apparent. Source: [307].

automation disengaged without warning [303], and in another experiment even drivers who were looking at the road made lane excursions [142]. On a closed track, 38% of subjects did nothing when the vehicle drifted to another lane and nearly 30% crashed into stationary objects despite being briefed about the limitations of technology and keeping their eyes on the road and hands on the wheel [95]. In a simulation study, some drivers who were told to expect hazards but did not receive a TOR, hit pedestrians [98]. Although more research is needed to link attention, automation failure, and event criticality, it appears that observing the road is insufficient to prompt timely and adequate response from drivers.

Explicit take-over requests indicate failure and elicit a better reaction by forcing the driver to focus on the front scenario and mirrors [109], especially if issued in multiple modalities. Combinations of acoustic and visual signals [98] with seat vibration [131] have been tested. The amount of time needed to observe the scene and think of a course of action depends on the complexity and severity of the situation, but longer is better [254]. For instance, drivers detected 49% of hazards when given 6s to take over compared to only 29% with 3s [305]. An early warning [98] and audio guidance towards relevant areas of interest [131] improve drivers' SA, as do visual guides, e.g., LED lights under the windshield. The latter are more useful when they show how to avoid the hazard [210] rather than point at the obstacle itself [208].

Guidance and warning. Although various safety features are common in vehicles, most research focuses on driving performance, and less is known about their effect on drivers' attention. One field test showed no significant differences in drivers' gaze distributions with and without the monitoring function enabled, although some positive changes were observed, such as increased glancing towards the road and fewer long off-road glances [306]. Data from the four-year European Field Operational Test (EuroFOT) concluded in 2012 indicate that a forward collision warning (FCW) prompted drivers to redirect their glances to medium and high eccentricity locations, presumably to restore their situation awareness [205].

Another approach is to guide drivers' gaze to help them notice hazards or important objects sooner and brake smoother in response [209, 218]. Similarly, highlighting maneuver-relevant cues leads to a more optimal allocation of attention, such as increased number of fixations to relevant areas and reduced dwell time on salient irrelevant zones [308]. How this guidance is realized matters: too many cues may be detrimental, e.g., highlighting all pedestrians in the scene [307]. Cues should not be too conspicuous, e.g., bounding boxes may attract attention to the pedestrians but obscure their posture and subtle movements and distract the driver from other driving-related objects. A more subtle cue such as virtual shadow [307] (Figure 2.14) or thin lines converging on the pedestrian [218] is preferable. These techniques have been tested only in simulation due to the difficulty of implementing them in real vehicles.

2.5 Limitations of behavioral studies

Numerous studies discussed in this chapter have greatly improved our understanding of what drivers perceive in a traffic scene and how they respond to the surrounding road users and events. However, there are still some remaining methodological issues and open problems that should be addressed moving forward.

The limitations of using gaze as a proxy for attention. Since the utility of gaze is limited for visual processing beyond foveal vision and for covert changes of attention, most of the literature focuses on central vision. Only a few experimental and theoretical works consider the role of the periphery in driving, although interest towards this topic has increased recently [28, 309, 310]. Likewise, decision-making processes are difficult to infer from gaze as it is a product of data- and task-driven factors whose relative contributions and interaction are not well understood.

Lack of multi-modal sensory perception. In addition to the visual signal, the driver receives sensory information from other modalities including sound, and tactile feedback (e.g., wheel traction). While it is universally agreed that vision is the main source of information for driving, the contributions of the other sensory modalities and their effect on attention are understudied.

Inconsistent definitions and procedures. The definitions of the basic attention units, data recording, and analysis procedures in the literature are often inconsistent or under-specified. For instance, fixation detection thresholds reported in the literature vary significantly, in some cases by nearly an order of magnitude. The resulting changes in the data distribution affect subsequent computations of eye-tracking measures and statistical tests. Some of the commonly used terms, such as *forward roadway*, *novice* or *older* driver, have no

formal definitions. As a result, gathered evidence cannot be aggregated across studies, and conclusions regarding some of the factors are contradictory. Some reviews point to the same conclusion, showing that only a fraction of papers use comparable experimental setups and approaches to statistical analysis of data [311, 312, 313]. Thus, the effects of definitions and other settings on the outcomes of the studies should be analyzed and reported.

Ecological validity and lack of validation studies. The majority of the behavioral works are conducted in low- to mid-fidelity driving simulators rarely validated against on-road data. Several studies indicate significant changes in attention allocation caused by simulators, especially when subjects are not actively controlling the vehicle. Thus conclusions and models based on such data may not apply in real-world conditions. Note, however, that even precise gaze recording in ecologically valid conditions does not reduce the fundamental limitations of gaze as a tool for studying human attention.

Results are rarely tied to safety. Even though most of the research on driving and attention is motivated by safety, the results of the studies are rarely tied directly to crash risk. Currently, only naturalistic driving studies provide accident statistics with relevant attention data for limited scenarios, e.g., rear-end collisions. Accident data from simulated and closed track studies is difficult to aggregate due to the specificity of the scenarios and validation issues.

Focus on correlations. The bulk of the behavioral literature is dedicated to establishing correlations between attention and various factors, with conclusions indicating safety concerns that may be used to inform policy-making. Relatively few works attempt to infer internal decision-making processes or attentional modulation that lead to the observed gaze distribution and subsequent motor actions.

Uneven coverage of factors that affect attention. Factors related to environment, demographics, driving experience, and certain types of distractions are understudied. Conditions during most studies are optimal for driving. Environmental factors such as bad weather, unfamiliar environment, or heavy traffic are avoided in on-road studies and rarely modeled in driving simulators. Common in-vehicle distractions such as passengers requiring attention (small children or pets) or driver's emotional state (distress or excitement) are not well-studied. Some cognitive distractions such as mind-wandering or anxiety are difficult to study, and it is unclear whether artificial tasks used to induce such states (e.g., n-back or sound counting) are adequate substitutes. Finally, the diversity of the subject pool in terms of age, gender, driving experience, and socio-economic factors is lacking in many studies.

Fragmented nature of behavioral research. Many disciplines are involved in studying attention and driving, e.g., psychology, transportation, human factors, and ergonomics, each with its own goals, approaches, and terminology. Within a single discipline, the majority

of works are narrowly focused on several factors considered in isolation. Including a more comprehensive array of factors and conditions will enable direct comparisons between them and lead to more robust conclusions. Furthermore, increasing interdisciplinary connections is necessary to achieve a unified understanding of drivers' attention allocation.

Risk of research becoming obsolete. Driver education, infrastructure design, vehicle, and consumer electronics technology continuously improve and cause shifts in driver behavior. For example, touchscreens have recently become the primary input method for many in-vehicle interfaces that should be reflected in the existing benchmarks for secondary tasks. These changes should be taken into account, particularly when dealing with data from naturalistic studies that take a significant amount of time to record and process. For example, the data from 100-Car NDS recorded in 2004 may not reflect drivers' behaviors a decade and a half later.

2.6 Summary

In this chapter, we provided a broad overview of behavioral research on how drivers attend to the scene and what affects where they look. The main conclusion that motivates the rest of this dissertation is that drivers' actions and surrounding context are the most significant influences on drivers' gaze allocation. Other factors, such as age, gender, and driving experience, do not fundamentally change drivers' behaviors but rather amplify or inhibit certain aspects of attention allocation.

Most transportation psychology studies focus on scenarios possessing imminent hazards, either caused by the drivers performing other non-driving-related tasks or by the unexpected actions of other road users. Behaviors of attentive drivers during routine uneventful trips are relatively understudied in comparison. Nevertheless, there is sufficient evidence that a connection exists between drivers' attention, their actions, and properties of the environment. For example, drivers exhibit specific observation patterns depending on the task (e.g., following a lead vehicle, driving on a curved road), look at different areas of the road on approach to various types of intersections, and behave differently in familiar and unfamiliar places. Thus any monitoring or assistive system must include a representation of the drivers' task and surrounding context to handle any situation that may arise during driving. For example, an assistive system must anticipate upcoming maneuvers (e.g., lane change) and be able to detect drivers' insufficient attention, even if no safety concerns are present (e.g., vehicle in the neighboring lane). This is difficult to achieve and any mistakes may lead to safety issues or mistrust in the system.

In this chapter we also discussed how data should be recorded and processed to adequately

capture drivers' actions and surrounding environment. We identified the choice between on-road and in-lab recordings as a trade-off between ecological validity, reproducibility, and precision. In other words, on-road conditions lead to more natural behaviors but cannot be repeated exactly for multiple subjects, whereas driving simulators offer better scalability and accuracy at the expense of subjects' engagement in the task. Although the significant effects of task and context are consistently found in both conditions, there remain some differences. As most research is conducted in driving simulators, further investigations are needed to validate this data against real on-road conditions.

Chapter 3

Driver monitoring and assistance systems

Behavioral studies discussed in the previous chapter provide a wealth of evidence that attention allocation is tightly linked to vehicle control and is affected by multiple exogenous and endogenous factors. As a result, practical solutions are developed under the assumption that changes in gaze patterns and driving performance can be detected and associated with changes in the environment, driver's behavior, and state. Knowing where the driver is looking and what they are doing (or about to do) can, in turn, be used for various driver monitoring applications, such as inattention and drowsiness detection. More advanced scenarios include understanding what the driver is or is not aware of and issuing active warnings to direct their gaze. High-quality behavioral data collected in diverse conditions from a large population of drivers is crucial for developing algorithms, particularly those that rely on supervised machine learning.

This chapter is structured as follows. In Section 3.1 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 3.2 covers algorithms for modeling drivers' attention allocation in the traffic scene for assistive and autonomous driving applications. Section 3.3 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.

3.1 Attention estimation for driver monitoring

Vision-based driver monitoring systems (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 3.1.1), inattention detection (Sections 3.1.2 and 3.1.3), and action anticipation (Sections 3.1.4) framed as classification problems. Methods for driver awareness estimation are discussed in Section 3.1.5.

3.1.1 In-vehicle gaze estimation

In-vehicle gaze estimation is commonly framed as finding to which area of interest (AOI) or gaze zone within the vehicle the driver is paying attention, rather than a specific point of fixation. Computationally, the problem is usually formulated as a multi-class classification, i.e., categorizing various features (e.g., eye-tracking data, head pose, or features extracted from images of the drivers) as belonging to one of the predefined AOIs.

While it is possible to analytically determine where the driver is looking by computing their gaze direction in 3D and its intersection with the 3D model of the vehicle interior, only a few approaches do so [157, 117]. Thus most research in this field focuses on finding the best combination of features and classifiers. Figure 3.1 shows reviewed algorithms for in-vehicle gaze estimation grouped by features they use.

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 [314, 315]). 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 [316] or combined with a visible light camera [317], to make the system suitable for night driving conditions.

A feature extraction pipeline should ideally satisfy the following criteria: real-time operation, 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 [157]: 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.

Classifiers. In the literature, the minimal set of features used for this problem consists of facial landmarks [318, 319] or head poses extracted from the landmarks [320, 316]. 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

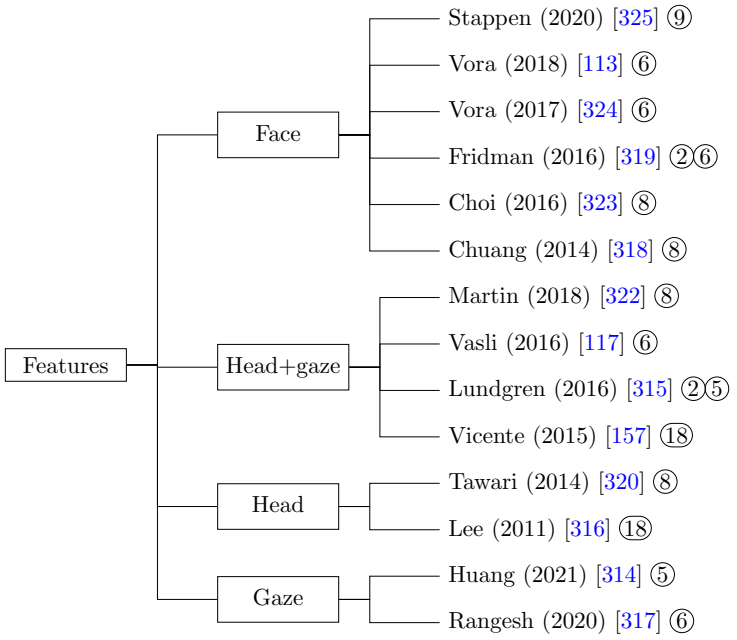


Figure 3.1: In-vehicle gaze estimation models organized by the visual features: *Gaze*—gaze coordinates or eye crop, *face*—face image/landmarks, *head*—3D head pose. Number of AOIs for each model is shown in circles.

and 3D head position features cannot reliably differentiate between neighboring zones, such as a speedometer and windshield [320] or side mirror and the window next to it [316]. In both cases, the drivers either made small head movements or did not move their heads and performed eye movements instead. Temporal filtering [318], aggregating features over time [320], and feature normalization [319] led to improved performance but ultimately did not resolve the issue, leading to the conclusion that eye features are also necessary. Fridman et al. [321] 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 [151].

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 [323, 324, 113, 325]. 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. [113] 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 [324], thus saving computational costs associated with face detection and tracking.

Evaluation and limitations. Because in-vehicle gaze estimation is treated as 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 is a wide range of AOIs defined inside the vehicle, but justification for the specific choice is rarely given. The minimum number of zones is 2 (driving- and non-driving related as in [315, 321]), but can go as high as 18 [316, 157] (e.g., see Figure 2.8a), with 6-8 being the most common. Naturally, more fine-grained zoning is challenging and leads to worse performance since it is more difficult to localize gaze within a smaller area [319]. As an alternative to fixed AOIs, Huang et al. [314] 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 [316, 318] and more thoroughly investigated in [321]. Such individual differences can be captured by training the models on user-specific data rather than data aggregated across many drivers [318, 319, 314], 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 [151, 320]. Nevertheless, 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 [151, 157, 321, 117, 315]. More recent CNN-based models appear to suffer less from misclassifications of these kinds [113]. The presence of eyewear causes another challenge. Glasses introduce glare and occlusions, making it difficult to estimate gaze direction [324]. In [317], a pre-processing step is added to remove eyewear via a gaze-preserving generative network. However, this approach only works if the glasses are relatively thin and likely cannot be applied to removing sunglasses.

3.1.2 Distraction 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 3.2 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 is formulated as a multi-class classification problem. Because of differences in behavioral

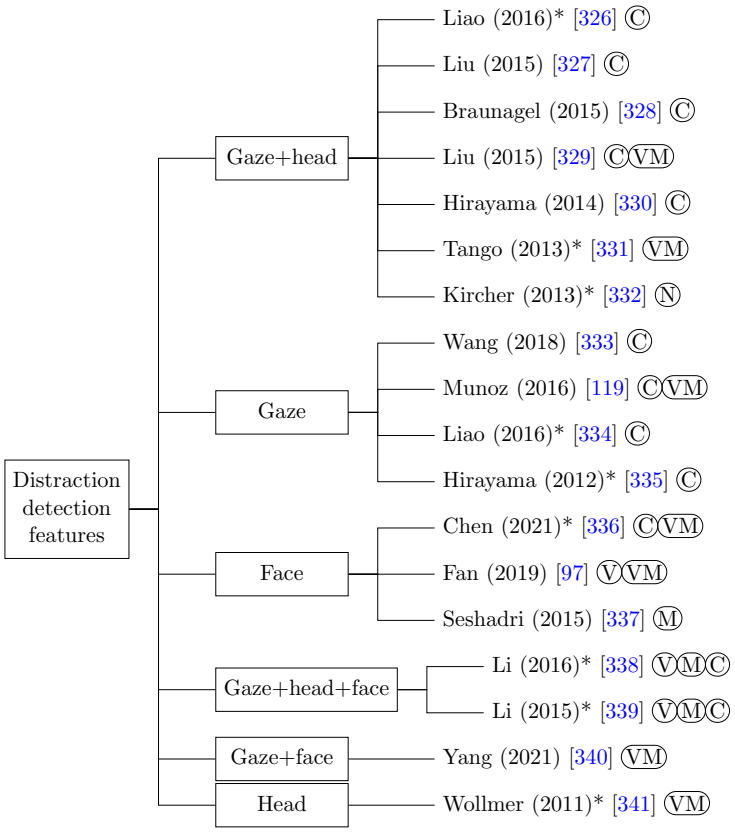


Figure 3.2: Summary of the distraction detection algorithms grouped by the visual features and types of distractions they can detect. Features: *gaze*—gaze coordinates or eye crop, *face*—face crop/landmarks, *head*—3D head pose. Distraction types: Ⓒ—cognitive, Ⓥ—visual, Ⓜ—manual, ⓋⓂ—visuo-manual, Ⓝ—non-specific. Algorithms marked with * use additional features, e.g., vehicle telemetry.

changes depending on the kind of distraction, the majority of the algorithms focus either on detecting one specific distraction type or distinguishing between different distractions.

Types of distractions. Cognitive and visuo-manual distractions (see Section 2.3.2) are the 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 [338] and voice-based playlist retrieval [330], and some include artificial tasks such as math quizzes [336] and n-back tasks [333]. Visual-manual tasks considered in the studies are everyday activities, such as using a cell phone for reading [340, 97] or sending messages [336, 97], radio tuning [119, 339, 341], and selecting a song from the playlist [341]. 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 [340, 97, 341].

Input sources and feature extraction. Features such as gaze coordinates or coarse AOIs (see Section 3.1.1) are naturally indicative of visual distractions. According to the evidence from psychological studies, longitudinal [127, 169, 201] and lateral [123, 190, 191, 192, 193] 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 [336, 334, 338, 326, 339, 341]. 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 [333, 334, 328, 329, 337, 332] or manually annotated [119, 330, 331, 342]. Fewer approaches incorporate explicit feature extraction pipelines (such as those discussed in Section 3.1.1). For instance, [340, 97, 337] rely on facial landmarks and [341] 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. [341] showed that head rotation angle and its derivatives are sensitive to visual-manual tasks, whereas Liao et al. [326] determined that gaze locations are useful for detecting cognitive distractions.

In addition to gaze and visual features, information such as vehicle data [334, 338, 326, 343], physiological signals [336], audio (to detect conversations) [338], detected objects in the scene [340, 337], and mirror-checking actions [338] have been shown to improve the accuracy of distraction detection.

Classifiers. The Support Vector Machine (SVM) is the most commonly used classifier for this problem, used in nearly half of the reviewed algorithms [334, 338, 326, 327, 328, 329, 337, 331, 332]. Other approaches, such as boosting [344], extreme learning machines [327, 329], K-means [339], and Hidden Markov Models (HMM) [119] have demonstrated high performance on the distraction detection task. Recent works that use deep learning methods, such as recurrent [333, 97, 341] and convolutional neural networks [336, 340, 97], 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 [332] 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 [248], this model uses a 2s buffer to track dis-

tractions 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 1 s to prevent issuing unnecessary warnings.

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, often over 90%, some even close to 100% [97]. 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 [326, 328], number of video frames [337], or number of events [335]. Direct comparisons are also affected by inconsistent recording conditions across studies, e.g., in-lab driving simulators [326], parked vehicles [340], or on-road settings [97].

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 [334], therefore thorough testing in various environments is desirable [341, 331]. Driving tasks, such as vehicle following and passing, also affect gaze distribution patterns and have to be taken into account when modeling distraction [343]. Although there are commonalities in how distractions manifest themselves in different drivers [119], the system should consider individual user characteristics for the best results [331, 338].

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 [97] 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 the AttenD algorithm [306]. 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 [345, 346] it is important to consider HCI issues when developing monitoring

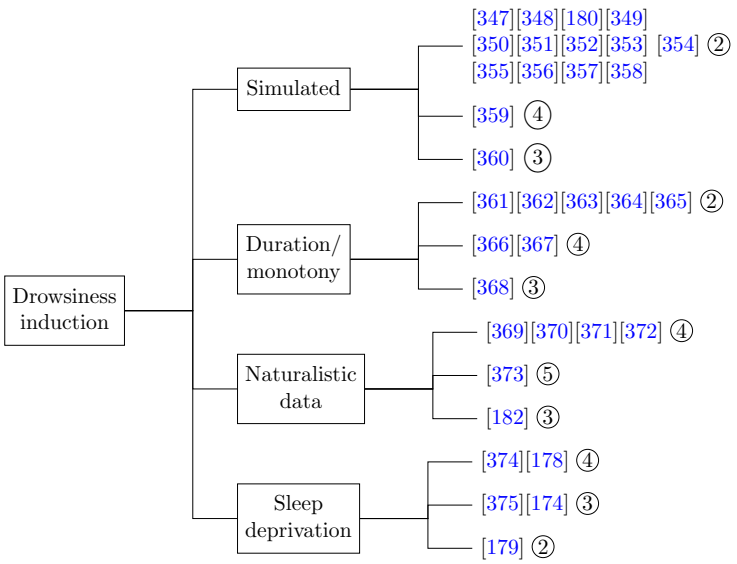


Figure 3.3: A summary of the drowsiness detection algorithms grouped by the type of drowsiness induction method and number of drowsiness levels (specified in circles).

algorithms and conduct user studies besides evaluation on datasets. Refer to Section 3.4 for further discussion.

3.1.3 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. A summary of drowsiness detection algorithms is shown in Figure 3.3

Input and feature extraction. Most algorithms rely on driver-facing video cameras to detect drowsiness. Near-infrared imaging (NIR) cameras are often used (as in [359, 349, 351] or combined with visible imaging color cameras [361]) 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 [375, 182]. These features are detected using methods described in Section 3.1.1, such as face detection [352, 375] and tracking[180], locating facial landmarks [375, 354], detecting blinks [355] and eye closures [354], recognizing mouth drooping and yawning [354], and measuring head pose and head movement [355]. Eye features (blinks and closures) are further processed into standard drowsiness measures such as percentage of eye closure (PERCLOS [181]) [178, 371, 182] and blink frequency [179, 174, 182].

Detection of many drowsiness symptoms, such as slow blinking and yawning, is further

improved by aggregating features across time [348]. Longer time intervals up to several minutes for the blink and eye closure features typically work best [181] but also increase the risk of missing microsleeps and "blank stares" [376], thus alerting the driver too late. Using multiple measures is an option as well, however, they might have different trends (e.g., increasing or decreasing at varying rates), therefore the algorithm design should account for that [366]. Drowsiness detection can be further improved by inclusion of global context features, such as continuous driving time, temperature, current time, and sleep duration [362, 368].

Classifiers. Rule-based approaches and thresholding are computationally the simplest methods for detecting drowsiness [179, 371, 363, 372], 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 [178, 366, 368], boosting [359], or HMMs [355], are more robust in practice.

Recently, deep learning methods have been applied to drowsiness detection. For instance, Zhao et al. [375] 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. [354] 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 the temporal dimension, Shih et al. [352] 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. [349] 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.

Evaluation and limitations. A significant limitation of research in this field is the prevalence of private datasets for developing models. The only widely used public dataset is Driver Drowsiness Detection [354], however, as discussed later in Section 3.3, it is recorded in lab conditions while subjects act drowsy. Many private datasets are also recorded in artificial conditions, e.g., lab [350, 364, 358] or parked vehicles [179, 355], and only a few are captured in on-road conditions, typically highways and rural roads with little traffic [372, 363, 368]. One study used recordings of drowsy passengers instead of drivers due to safety concerns [375].

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. Despite reporting excellent results, many algorithms still struggle with certain aspects

of the problem, notably extreme head angles [374, 375] and glare and occlusion from eyewear [180, 182, 349]. 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 [372] and difficult to detect when participants have smaller eyes [179]. Vehicle vibration and variability of driver positions with respect to cameras further exacerbate these problems [374]. Furthermore, some drivers do not show visible signs of drowsiness even when fatigued [361]. More diverse publicly available multi-modal datasets collected in naturalistic conditions are part of the solution to the problems listed above [361, 355]. 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 [361, 179].

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 [360, 350, 355, 357, 358], however, there is a risk that such data is not representative of more realistic conditions. Alternatively, drowsiness and fatigue can be induced by extended [363, 368] or monotonous [362] driving sessions, driving late at night [361], or reducing sleeping hours prior to experiment [178]. Some studies also employ night shift workers for collecting naturalistic data [374, 174].

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) [284] 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, [361] collapses the KSS scale into 2 levels, [374] uses 4, and [368] 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 [360], 4 [368], or 6 [182]. Intermediate levels of drowsiness may include KSS levels 6-7 [360, 182], 5-7 [368], or only 7 [174]. Only extreme drowsiness is consistently associated with KSS levels 8-9 across studies [360, 368, 174, 182]. Sleepiness scales other than KSS have also been used, e.g., custom 3-level scale [375] or Zilberg 4-level scale [377] in [366], although simple binary drowsy/alert assessment is the most common [348, 179, 180, 356, 357, 373, 365, 371]. The latter approach may be justified since intermediate drowsiness levels are easily misclassified as drowsy [360].

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 [361, 174, 182], observer-rated sleepiness [366, 374], or both [368]. According to evidence from psychological exper-

iments, neither is bias-free : observer ratings may not be reliable [378, 379] and self-rated sleepiness does not always correlate with driving performance [380]. Simulated data where drivers acted drowsy is devoid of issues with assigning ground truth and is a preferred choice for most studies (Figure 3.3). 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.

3.1.4 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 [381], it can be exploited to recognize and anticipate drivers' maneuvers.

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 [382]. The processing pipeline for obtaining gaze features usually includes face detection and tracking, followed by facial landmark detection, extraction of gaze zones [322, 383, 344], gaze duration, frequency, and blinks [344, 322]. Alternatively, an implicit representation of gaze can be used, such as tracked facial landmarks aggregated over time as proposed in [384, 385] or mirror-checking actions [344].

A variety of methods have been proposed to classify actions based on the temporal features above. For example, Li et al. [344] use boosting [386] with a combination of mirror-checking actions and vehicle dynamics features. Martin et al. [383, 322] 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. [384] and Akai et al. [382] 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 combining multi-modal data [385] but lack the explainability of HMMs.

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 [383]. Besides issues with face

detection and tracking due to illumination changes [384], 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 [383]) 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 [384]). Inclusion of scene and route information may help handle such cases.

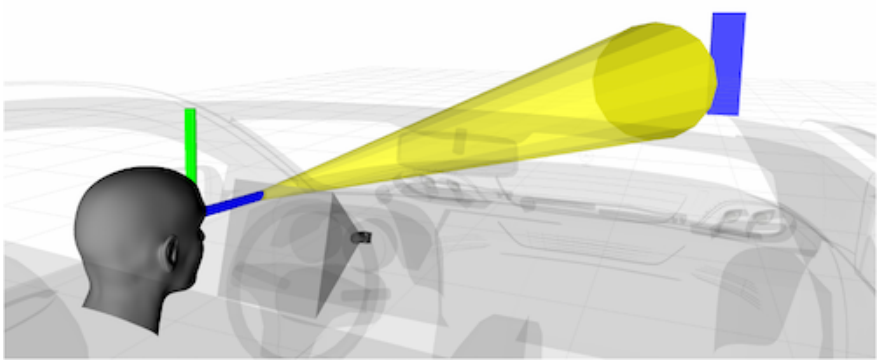
3.1.5 Driver awareness estimation

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.

Visual features and processing. Algorithms discussed in this section determine whether the driver was aware of vulnerable road users, signs, and traffic signals. Most approaches proceed as follows: 1) detect the driver's 3D gaze direction, 2) convert gaze to vehicle's frame of reference, 3) detect objects in the scene and their properties, and 4) match driver's gaze with objects to identify whether they were fixated. See Figure 3.4a for a schematic illustration.

Drivers gaze direction estimation uses approaches discussed in Section 3.1.1, whereas detecting objects in the scene relies on off-the-shelf algorithms [387], classical vision pipelines [388, 389, 390], or manually annotated bounding boxes [391]. Distances to the detected objects and their relative velocities may be inferred from a stereo camera [389], provided by range sensors [392, 389, 393, 394, 395], or determined using simple heuristics [387].

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 [387, 391, 394]. Since inaccurate measurement of 3D gaze direction can result in large errors, especially for objects far away, the authors of [389, 390] propose to treat attention as a cone projected from the drivers' eyes towards the windshield (Figure 3.4a). Some algorithms also take into account that drivers retain information about objects for some time after looking at them [395, 387, 397], as well as other properties of the scene, such as weather and proximity of other road users [397]. For example, Schwehr et al. [398, 392] 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. [397] modify the AttenD algorithm (described earlier in Section 3.1.2) to include elements of context via additional buffers for targets of relevance



(a)

(b)

Figure 3.4: Visualization of the driver's 3D gaze direction: a) view from inside the vehicle, b) projection of the attention cone onto the traffic scene to determine whether the road sign appeared within the driver's field of view. Sources: a) [396], b) [390].

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. [399] use SAGAT [400], 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.

Evaluation and limitations. Evaluating driver awareness models is not trivial. Quantitative performance is usually reported for the individual modules, such as gaze estimation [401, 402], road user trajectory prediction [396], object detection [389, 403], 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 [392, 396, 389, 393, 388], 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 [404], routes on a university campus [387], and tests in driving simulators [397, 405].

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. [402]. 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 is used in many studies instead of PoG to alleviate some of the imprecision but often does not localize the targets well (Figure 3.4b). An assumption that the drivers detect all objects within the intersection area [389, 390] may not be accurate. According to Kim et al. [406], gaze is generally correlated with lower levels of situation awareness (as defined in [407]), 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 [405] and braking intention [408], as well as detectability of signs [409, 410, 411] and pedestrians [412] in traffic scenes depending on the visual properties of the objects and the scene, e.g., illumination, visual clutter, visibility, etc.

3.2 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 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.

3.2.1 Modeling visual attention in a 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 3.5a). Fewer algorithms assign importance scores to specific objects. In this case, higher scores are

(a) (b)

Figure 3.5: 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) [422], b) [425].

associated with objects relevant to the ego-vehicle (e.g., lead vehicle shown in red in Figure 3.5b).

Bottom-up and top-down influences. As discussed in Section 1, 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 [413, 414, 415, 416, 417]. Top-down features are task-specific and hence more varied. For example, [413, 414] use vanishing point because drivers often focus on it to get an optimal view of the road ahead [213, 291]. Other options include drivers' actions [415, 416, 417], current driving task [418, 419], previous fixation locations [415, 416, 417], and vehicle telemetry [415, 416, 417, 420]. Other common features include semantic segmentation maps [421, 422, 423] and detected objects [420, 424, 425, 426].

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 [427, 428, 422, 429]. More recently, following successful applications in video action recognition problems [430], 3D convolutional networks have become a popular choice for encoding spatiotemporal data [431, 418, 423, 426]. Some approaches use recurrent networks, combined with individually encoded frames [432, 433] or with a set of frames processed via 3D convolutional layers [421].

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 [422, 434, 428, 435, 427], 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. [413, 414]. They later extended this method by learning the weights for individual features using a random forest [414]. Tavakoli et al. [427] 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 [415, 416, 417] 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 the bottom-up ones (e.g., saliency maps), however combination of both types of features performed best.

Two recent models, HammerDrive [418] and MEDIRL [419], 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).

Evaluation and limitations. Evaluation of driver gaze models follows the procedure and metrics established in free-viewing saliency research (see [436] for a review). These metrics assess how similar predicted saliency maps are to those of human drivers in terms of saliency value magnitudes, statistical distribution properties, and salient locations [437].

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 [422] or was more consistent with good driving practices [432]. 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 the lab.

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 [422], 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. [432] propose to mitigate the prevalence of common driving scenarios using two strate-

gies. 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 [431, 432].

3.2.2 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 [438]). Decision-making for motion planning, behavior selection, and vehicle control rely on precise mapping and localization [439] as well as understanding the behaviors of vulnerable road users [440].

Current self-driving systems that tackle these issues can be broadly subdivided into modular and end-to-end [441]. 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) [442]. 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.

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 (Figure 3.6b) 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 [443], multiple attention modules are inserted after intermediate and last layers of CNN to gradually refine features. Kim et al. [444] insert the attentional module after the convolutional feature extraction and also condition spatial attention weights on the hidden state from the previous timestep.

(a) (b)

Figure 3.6: Visualized spatial attention weights: a) object-level, b) pixel-wise. Sources: a) [446], b) [444].

Computing attention weights for image regions and specific objects in the scene instead of pixels or individual features is also possible (Figure 3.6a). Cultrera et al. [445] 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, 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 [446], for each frame of the input, features corresponding to individual objects 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. [447] and Wei et al. [448] 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 [449] has been shown effective for many vision tasks [450]. 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. The 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.

Transformers have become popular in the self-driving domain recently due to their flexibility with respect to various input modalities [450]. For example, Chitta et al. [451] 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. [452] 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. [453] 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 agent all other relevant agents. Experiments on naturalistic driving data show that focusing on the most relevant agents helps reduce the number of collisions. In addition, it is possible to visualize the learned attention weights to interpret model output [454, 455]

Some approaches leverage human data for training attention modules. For example, in [456] 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 [457], 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 human gaze, Kim et al. in a series of works, leverage human textual annotations using visuo-linguistic techniques. In [458] 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 [459] the authors show that textual advice is helpful for generating vehicle control commands (future waypoints and speed) and locating areas in the image that influenced the decision (represented as attention maps). An extension of this approach is presented in [460]. Here, textual explanations and spatial attention maps are generated to simultaneously explain and generate vehicle control commands from visual observations.

Evaluation and limitations. The driving performance of self-driving algorithms is evaluated in simulated environments [451, 452, 461], using pre-recorded datasets [444, 460, 462, 448], or both [446]. In simulations, safety metrics measure the number of collisions or infractions and the distance traveled between them [451, 446], 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 [458, 462, 453].

Performance of attention is evaluated quantitatively via ablation studies to show im-

proved vehicle control [456, 460, 453, 448, 461], reduction of traffic incidents [452, 446, 448, 461], and match with human data, such as gaze [456] or textual annotations [460].

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 (Figure 3.6a), where importance scores are visualized for individual objects, is easier to interpret and is more common [451, 452, 453, 448, 446]. Pixel-wise attention scores (Figure 3.6b) used in some studies [456, 462, 444] often do not match specific objects in the scene and, in some instances, do not adequately reflect the decisions made by the system [444].

3.3 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 a sufficiently large and diverse pool of participants. Thus large-scale data are a must for adequate evaluation of models and benchmarking overall progress. This section covers a number of public datasets for a range of applications and properties of drivers' attention they represent (Table 3.1).

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 [463] and BDD-A [432] focus on hazardous scenarios, C42CN [464] on secondary non-driving tasks, and the rest (MAAD [429], TrafficSaliency [435], DR(eye)VE [422], C42CN [464], TETD [413], and 3DDS [415]) 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 2.1, 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. [432] 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

Table 3.1: Public datasets for studying drivers' attention and inattention with links to the corresponding project pages and data properties.
The datasets are sorted by the 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 [429]	+		S^{RGB}	+	+	TL	simulator	-	23	60K	8
TrafficSaliency [435]			S^{RGB}	+			simulator	-	28	77K*	16
DADA-2000 [463]		+	S^{RGB}	+		BB, TL	simulator	-	20	658K	2000
DR(eye)VE [422]	+		S^{RGB}	+	+		on-road	+	8	555K	74
BBD-A [432]		+	S^{RGB}	+	+		simulator	-	45	378K*	1232
C42CN [464]	+	+	S^{RGB}	+			simulator	+	68	-	456
TETD [413]			S^{RGB}	+			simulator	-	20	100	
3DDS [415]			S^{RGB}	+			simulator	+	10	192K	108
LISA v2 [317]	+		$D^{RGB, IR}$				simulator	+	13	3.4M	-
DGAZE [465]			S^{RGB}, D^{RGB}			BB	simulator	-	20	100K	20
DGW [466]	+		D^{RGB}			TL	on-road	-	338	-	586
DMD [467]	+		$D^{RGB, DEPTH, IR}$			TL, BB	on-road, simulator	+	37	4.4M*	-
Drive&Act [468]	+		S^{RGB}			SEM, TL	simulator	+	15	9.6M	29
HAD [459]			S^{RGB}		+	TL	simulator	-	-	3.4M*	5675
RLDD [360]	+		D^{RGB}			TL	simulator	-	60	3.2M*	180
BBD-X [458]			S^{RGB}		+	TL	simulator	-	-	8.4M	6984
HDD [469]			S^{RGB}		+	BB, TL	simulator	-	-	1.2M*	137
DDD [354]	+		$D^{RGB, IR}$			TL	simulator	+	36	486K*	360
DAD [470]		+	S^{RGB}			BB, TL	simulator	-	-	175K	620
Brain4Cars [384]	+		S^{RGB}, D^{RGB}		+	TL	on-road	+	10	2M	-
DROZY [471]	+		$D^{RGB, DEPTH}$			TL, FL	simulator	-	14	500K	-
DIPLECS [472]			S^{RGB}		+		on-road	+	1	54K*	1
YawDD [473]	+		D^{RGB}			TL, BB	simulator	-	107	-	342

are significant and how they affect safety.

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. LISA v2 [317], DGW [466], and DMD [467] datasets were collected following this approach. For 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 was gathered similarly but with 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 vehi-

cle and speak the corresponding zone number, which was transcribed using a speech-to-text network. For the DGAZE dataset, drivers sat against the wall imitating a vehicle interior and looked at the annotated objects in the traffic scene projected on the screen. Capturing videos of the drivers looking at predefined objects helped associate drivers' appearance with points of interest in the image plane. However, it is yet to be determined whether such an in-lab setup can 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 [469], HAD [459], and BDD-X [458] 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 [470] 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 [471], YawDD [473], DDD [354], and RLDD [360] 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 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 [384] and DMD [467]. 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.

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 [474]. The largest NDS

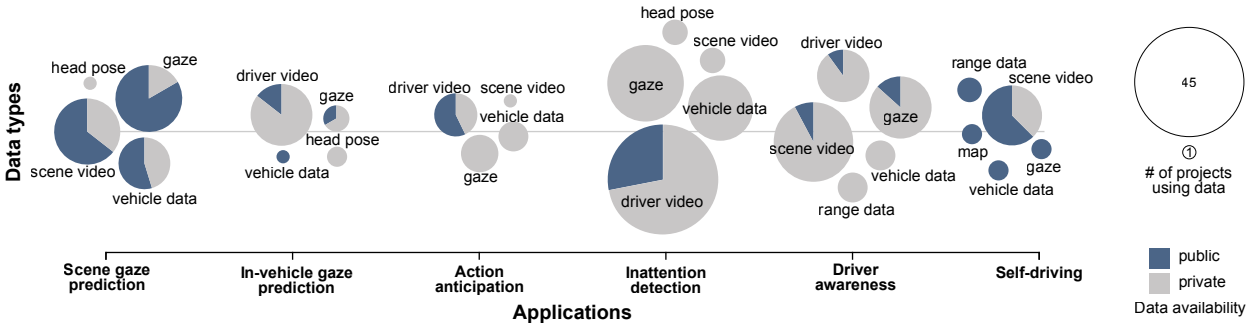


Figure 3.7: Types of data, data availability, and use in applications. The size of the circles reflect the number of publications using the corresponding data type for a specific application. Within each circle, the proportion of public and private data is shown.

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 [475]. Unlike datasets listed in the Table 3.1, NDS data is not freely accessible due to privacy concerns. For example, researchers interested in obtaining data from SHRP2[1] are required to complete training and provide a research proposal. Fees may also be charged depending on the data requested.

Data availability and properties. Overall, the datasets listed in Table 3.1 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 3.7, 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.

Another limitation of many public datasets is that attention-related data is often recorded in laboratory conditions. As discussed in Section 2.1, for eye-tracking data, such conditions have not been validated. 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 3.7 shows different kinds of annotations and their respective usage for various purposes. For example, to the best of our knowledge, there are no

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

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.

3.4 Discussion

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. Below, we group them into three categories: issues related to data availability and quality, evaluation, and approaches to modeling attention.

3.4.1 Data availability and quality

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 3.7, 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 largely 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 there is no complete information on the recording conditions, characteristics of the subjects, and tasks they performed for many unpublished datasets. 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 detection, self-driving, and drowsiness detection.

Improving data diversity and fidelity. Regardless of data accessibility, diversity and fidelity of most datasets is insufficient. Recording conditions significantly affect the quality of

the data and usability of the models in practice. Naturalistic recordings of drivers' behaviors are generally difficult to collect and analyze due to the 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 [285]. 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.

Both on-road and in driving simulators, highway and rural road scenarios, often with low traffic volume, are more common, 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 city streets, drivers interact with many other agents, which leads to more complex attention strategies. Conditions such as bad weather, unfamiliar environment, or heavy traffic are rarely modeled. Even in large naturalistic studies, these conditions are not well represented [118, 302].

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 3.1), recruiting more subjects with diverse demographic characteristics is highly desirable.

Besides artificial conditions, the actions of the drivers are often staged. It is common practice to induce distraction by asking the drivers to perform tasks at timed intervals (see Section 3.1). In reality, however, secondary task engagement depends on many factors, including experience, environmental, and situational (e.g., child or pet requiring immediate attention), as well as characteristics of the secondary task itself [476]. Performing meaningless tasks on-demand and incentivizing high performance produces detectable changes in gaze allocation and driving performance, but such behaviors may differ from inattention caused by distractions that occur naturally during driving.

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 [168] 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 [477, 478].

3.4.2 Evaluation

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. [479] 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. [402] 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 3.1.3, 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 3.1.2). Physiological indicators are more suitable for these purposes [480] but require additional sensors, making data collection more costly and the use of such systems less desirable in practice.

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 [5].

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 available online (see Section 3.3) to determine whether models perform adequately in such situations.

Another is testing the models in simulation to estimate crash risks (Section 3.2.1). 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 [342, 211]. 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 [332]. Incidentally, it is the only model that captures the duration of the inattention, whereas the rest focus on instantaneous detection.

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 [418, 419] and detect inattention [342, 481]. Likewise, gaze and appearance features are useful for detecting both distraction and drowsiness (Section 3.1) but relatively few works investigate these problems together [482, 483].

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 [484]). 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 3.2.1) 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 [209, 308, 218], how this guidance is realized is important. It has been shown that providing too many cues is detrimental and may obscure other important information [308, 307]. Specifically for hazards, indicating the path to avoid them [210] is more effective than pointing at the obstacle itself [208]. 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 [485], therefore context must be modeled as well. Individual characteristics of the drivers discussed in Section 3.1 or their driving preferences [486, 346] should also be factored in

when setting thresholds for warnings.

3.4.3 Limitations of attention models

Using gaze as a proxy for attention. As discussed in Chapter 1, the 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 [72] and change blindness [68, 69, 70] occur during driving). Second, drivers extensively use peripheral vision for vehicle control [44, 46] and hazard detection [47, 48, 49], 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.

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 [241, 264, 132, 203] and driving experience [6] affect drivers' attention allocation. Besides driver characteristics, external conditions matter. Effects of driving through intersections [162, 187, 172], on curved roads [207], in dense traffic [299], as well as the presence of outside distractors (e.g., billboards) [294, 158] have been investigated in numerous behavioral studies but are not taken into account in many implementations.

Incorporating explicit task representation. Recalling Chapter 1, 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 an 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 allocation 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 [381]. 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.

Modeling attention beyond selective and explanatory functions. In most models of drivers' attention reviewed in Section 3.2.1, 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 [487], 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.

3.5 Summary

In this chapter, we provided a broad overview of applied research on driving and attention. We identified two core applications, drivers' gaze prediction inside and outside the vehicle, and a number of downstream applications, such as distraction and drowsiness detection, action anticipation, and awareness estimation. Similar to behavioral research, there is more work on drivers' inattention, whereas behaviors of attentive drivers are studied less. There is, however, increasing interest to the latter, partly due to the introduction of several new large-scale datasets and relevance to autonomous driving.

Despite recent improvements, there remains a need for publicly available high-quality datasets and models. Many application categories continue to rely on private unpublished datasets, hindering benchmarking and replication efforts. Drivers' gaze prediction is, perhaps, the most open subfield, with nearly two-thirds of models based on the publicly available data.

Lastly, there is a disconnect between behavioral and applied research. For example, only effects of secondary tasks on gaze are considered explicitly, e.g., in models for detecting drivers' inattention and anticipating their actions. At the same time, the majority of models for scene gaze prediction are bottom-up and do not take task into account explicitly. Similarly, effects of context (road geometry, street structure, traffic density) are not included in most models, despite evidence of their effects on drivers' attention allocation and potential engagement in non-driving activities.

Chapter 4

Analysis of datasets for drivers' gaze prediction

From this chapter onward, we will concentrate on the practical issues of predicting drivers' gaze towards the scene. This problem was identified in Section 1.3 as one of the core problems for developing assistive and monitoring applications. To recap, in Chapter 2 we presented driving as a visuomotor task where perception of the environment and motor actions are tightly intertwined. Yet, according to our review in Chapter 3, many of the state-of-the-art (SOTA) models for drivers' gaze prediction are bottom-up. In other words, these models map the driving footage to human ground truth without considering the underlying decision-making processes and actions performed by the drivers.

Three issues are related to this state of research in the field: 1) the contents and quality of private datasets for drivers' gaze prediction cannot be examined and such analysis has not been done for public data, thus, 2) it is impossible to assess whether bottom-up models capture the driving task implicitly, and 3) it is difficult to train models incorporating explicit drivers' actions and scene attributes. All these issues are due in part to lack of appropriate annotations for private and public benchmarks.

In this chapter, we focus on the first two issues to gain a better understanding of the available data and model performance. We propose a new set of annotations for drivers' actions and traffic context for three large-scale public datasets. We then analyze the datasets to determine how well they capture what the driver perceived and the actions they performed. Lastly, we propose a set of recommendations to guide future data collection efforts.

Table 4.1: Properties of the selected datasets that contain driving footage and eye-tracking data.

		DR(eye)VE [422]	**BDD-A** [432]	**MAAD**[a] [429]	**LBW** [489]
Video data	# videos	74	1435	106	74
	# frames	555K	405K	794K	122K
	Continuous	yes	no	yes	no
	Video len (s)	300	11 (2)	299 (5)	329 (76)
	Segment len (s)	300	11 (2)	299 (5)	2 (3)
	Scene view	yes	yes	yes	yes
	Driver's view	yes	no	no	no
	In-cabin view	no	no	no	yes
	Traffic	light-med	light-heavy	light-med	light
Scene camera	Image size	1920×1080	1280×720	1920×1080	942×489
	Hz	25	30	25	5
	Field of view	<120	<120	<120	<120
Eye-tracking data	Where recorded	on-road	in-lab	in-lab	on-road
	Hz	60	1000	1000	5
	# subj. per video	1	≥4	1 – 11	1
	# of subjects	8	45	23	28
	Subjects' age	20–40	-	20-55	>18
	Subjects' gender	7M, 1F	-	22M, 6F	-
	Licensure (years)	-	>1	>2	-
Vehicle data	GPS (Hz)	1	1	1	-
	Speed (Hz)	25	1	25	-
	Heading (Hz)	25	1	25	-
Labels	Action	partial	no	no	no
	Video attributes	yes	no	yes	no

[a] MAAD dataset is composed of 74 original DR(eye)VE scene videos and their copies with various manipulations applied (e.g., blurred, rotated). Eye-tracking data was recorded as subjects viewed the videos on the monitor.

4.1 Review of relevant works

Datasets for drivers' gaze prediction. Large-scale public datasets stimulated innovation in many areas of artificial intelligence and computer vision, and driving attention research is no exception. Based on our estimates, there are over forty public datasets that capture aspects of drivers' attention [488]. Here, we focus on a subset of datasets that contain driving footage and human eye-tracking data. Specifically, we analyze DR(eye)VE [422], BDD-A [432], MAAD [429], and LBW [489][1]. All four datasets are comparable in volume; each contains 4–6 hours of video footage and corresponding eye-tracking data. BDD-A and LBW were recorded in North America, and DR(eye)VE in Europe. MAAD is based on the driving footage in DR(eye)VE.

These datasets represent different approaches to gathering data. DR(eye)VE and LBW are collected in on-road conditions, while the drivers whose gaze was recorded controlled the

[1] DADA-2000 [421] is another dataset that fits our selection criteria. However, it is hosted on Baidu Wangpan, which is not accessible in Canada.

vehicle. BDD-A and MAAD are collected in the lab while human subjects passively viewed driving footage recorded with dashboard cameras on the computer screen. The datasets feature diverse locations, visibility, and weather conditions. DR(eye)VE and LBW consist of long clips recorded in low- to mid-density traffic, BDD-A contains short clips sampled around evasive or braking maneuvers in the Berkeley DeepDrive Video (BDDV) dataset [490] and features low to dense traffic. Table 4.1 summarizes dataset properties.

Additional annotations. DR(eye)VE and BDD-A, as some of the early datasets for studying drivers' attention, have been used to develop and evaluate dozens of models. However, neither these datasets nor the more recent MAAD and LBW contain annotations relevant to the driving task and context as defined earlier. Only DR(eye)VE offers coarse textual annotations: video-level labels for weather, time of day, and location. In addition, some frames where saliency maps deviate from the average gaze distribution (approx. 20% of the data) are marked as *recording errors*, *inattentive*, *subjective*, and *acting*. The latter category refers to some unspecified actions performed by the driver. Similarly, in BDD-A, frames deviating from the average are given higher weight during training since they are associated with drivers' responses to traffic signals or actions of other road users. However, neither the implementation of per-frame weights nor corresponding labels are available.

Evaluation of top-down effects. Several past works estimated the impact of different drivers' actions and scenarios on model performance. For example, *acting* annotations in DR(eye)VE mentioned earlier have been used for evaluation [422, 432]. Other approaches to partitioning the data have also been attempted. In [419], test sets of DR(eye)VE and BDD-A were further split into categories corresponding to three tasks—merging-in, lane-keeping, and braking. The authors of [433] used yaw rate (derived from GPS) as a proxy for lateral actions in DR(eye)VE. However, neither of the works makes the labels or methods for generating them available.

4.2 Task and context annotations

We restate the definitions of the task and context from Section 1.3 for the reader's convenience. A drivers' *task* is an action or a sequence of actions that lead to maintaining speed/lane, remaining stationary, moving the vehicle laterally (turns or lane changes), or longitudinally (accelerating and decelerating). *Context* refers to the physical and temporal scenario for the task. Here, we focus on two aspects of context: intersection type (signalized, unsignalized, highway ramp) and driver's priority with respect to other road users (right-of-way, yield). Both have been shown to affect drivers' gaze distribution, as discussed earlier in Chapter 2.

Table 4.2: Action statistics across the datasets. U-turns and reversing are excluded due to their rarity (<1% of the data). *Turn/change lane*—actions performed while maintaining speed, *lat/lon actions*—both lateral and longitudinal changes in motion are present (e.g., braking while making a turn).

Action type	DR(eye)VE / MAAD[a]	BDD-A[b]	LBW[c]
Speed up	14.00	14.05	9.63
Slow down	12.46	46.63	14.43
Turn/change lane	2.99	2.27	3.21
Lateral/longitudinal actions	3.78	4.66	3.66
Maintain speed/lane	60.59	27.52	46.45
Stopped	6.18	4.86	22.62

[a] Action counts are the same for MAAD since it uses DR(eye)VE videos. Video # 6 was excluded due to mismatched vehicle data.
[b] Excluded 135 videos: the clips with severe image artifacts, missing gaze and vehicle data, and duplicates.
[c] Excluded two underexposed videos (Subject # 26).

We will use these definitions to annotate DR(eye)VE [422], BDD-A [432], MAAD [429], and LBW [489], and analyze their contents with respect to task and context.

4.2.1 Driving task

Labels for longitudinal actions in DR(eye)VE and BDD-A were extracted automatically based on the provided speed data. We first removed outliers, applied a moving median filter with window size of 20 frames, and identified the points where speed changed using the method proposed in [491]. We then computed acceleration between change points and identified braking and acceleration events using a threshold of 0.4 m/s^2. All episodes with acceleration below the threshold were labeled as *maintain speed*. The *stopped* label was assigned to the frames where speed did not exceed 1 km/h. Due to the limitations of the provided GPS and heading data, all lateral actions were labeled manually.

Since LBW has no vehicle data, we used the provided intrinsic and extrinsic camera matrices to infer camera movement via ORB-SLAM [492]. However, due to the missing frames and exposure issues (see Section 4.3.1), only portions of the data could be processed. The resulting noisy speed and direction estimates were smoothed with a median filter but were of sufficient length to extract only the quick relative changes in speed, which likely inflated the proportion of *maintain speed* episodes. The lateral actions (turns and lane changes) were labeled by hand.

Based on the summary of action statistics in Table 4.2, all three datasets are dominated

Table 4.3: Number of instances of different types of intersections and the ego-vehicle right-of-way across the datasets.

Intersection type	Priority	DR(eye)VE / MAAD	BDD-A	LBW
Unsignalized	Right-of-way	436	211	463
	Yield	105	15	238
Signalized	Right-of-way	85	96	479
	Yield	13	124	53
Roundabout	Right-of-way	29	-	5
	Yield	29	-	5
Highway ramp	Right-of-way	147	-	-
	Yield	12	-	-

* Excluded videos are the same as in Table 4.2.

by scenarios where the driver is *maintaining speed/lane*, the vehicle is *stopped*, or *speeding up*, within its lane, usually after a stop. In DR(eye)VE, LBW, and BDD-A, these three actions comprise 81%, 78%, and 46% of frames, respectively. Non-trivial scenarios, where drivers respond to traffic signals or interact with other road users, typically involve braking, turns, or lane changes. In BDD-A, only braking events are well-represented by design and comprise nearly half of the data. Lateral actions (turns and lane changes) in each dataset do not exceed 7%.

4.2.2 Context

We annotated the following context elements: intersections and the ego-vehicle priority when passing through them. For DR(eye)VE and BDD-A, we used the OpenStreetMap API [493] to extract the street network and identify intersections along the driven route for each video. We then visually inspected each intersection and assigned the following labels based on [494]: *signalized* (if a traffic signal was present), *unsignalized* (controlled by stop sign, yield sign, or no sign), *roundabout*, and *highway ramp*. We then marked the frame at which the vehicle passed through the middle of the intersection and ego-vehicle *priority* (right-of-way or yielding) with respect to other vehicles based on the rules of the road in the recording location. Interactions with pedestrians were not considered because there were too few instances across all datasets. In LBW, context was annotated manually based on video recordings, since no GPS data was provided.

Table 4.3 shows the summary statistics of intersections and the ego-vehicle priority in different datasets. With regard to context, the distribution across different types of intersections is largely location- and route-specific. For example, roundabouts are more common in

Europe but not in North America, therefore in DR(eye)VE there are more such instances. In BDD-A, the videos are short and often truncated before the vehicles reaches an intersection, thus the number of intersections is comparatively low. Overall, the episodes where the vehicle is passing through some kind of intersection comprise about 7% of the data in each dataset. In 70-80% of the cases, the driver is traveling on the main road or passes through an intersection on a green traffic light and does not have to negotiate with other traffic participants.

4.3 Analysis of recording and processing pipelines

In this section, we analyze the recording and processing pipelines of DR(eye)VE [422], BDD-A [432], MAAD [429], and LBW [489] with respect to capturing driving task and context that we defined and annotated earlier. In particular, we focus on the common shortcomings of the recorded video, vehicle telemetry, and eye-tracking data. In addition, we identify issues with processing and aggregation of eye-tracking data for training and evaluation.

4.3.1 Video data collection

Video footage should capture what the driver sees, ideally, as closely as possible. Therefore, the following properties of video data are most relevant: spatial context (camera angle of view and position), temporal context (duration and continuity), and video quality.

Limitations of spatial context for capturing drivers' view. Driving footage in DR(eye)VE and BDD-A is recorded with monocular dashboard cameras. LBW is captured with a stereo pair of Azure Kinect RGB-D cameras, but only one camera's output is made available. Cameras used across all these datasets have at most 120° horizontal field of view, which falls short of the human binocular visual field spanning 200° horizontally and 125° vertically [32]. Therefore, only a portion of what the driver perceives (even discounting head movements) can be recorded. Consequently, this may be detrimental to data analysis and model training. For example, drivers can turn their head to look out of the side windows, but these areas and the events that happened there will be absent in the recording (see Figure 4.1). The same applies to the view behind the vehicle, visible in the side and rearview mirrors.

Camera locations differ across datasets. In DR(eye)VE, a camera is installed on the rooftop of the vehicle. Due to being at a higher vantage point, the recording often does not match what the driver was seeing, which causes issues with data processing (see Section 4.3.4). In LBW and BDD-A, the camera is installed inside the car above the dashboard

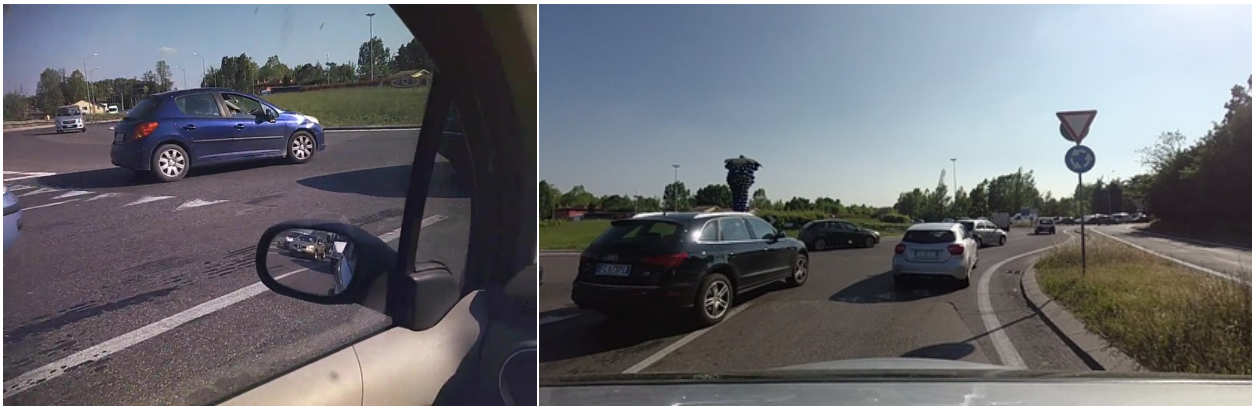


Figure 4.1: An example from DR(eye)VE, showing the view of the same roundabout scene from two perspectives: driver's view from the head-mounted camera (left) and scene view from the camera installed on the rooftop (right). Note that the scene camera does not capture what the driver can see from the side window by turning their head.

and better approximates the driver's view. While in LBW and DR(eye)VE the camera position is fixed, in BDD-A, positioning of the camera varies widely across videos, with more than 150 videos (10% of the dataset) recorded with the camera off-center, tilted, or partially obstructed (see examples in Figure 4.2). These issues may affect how observers in the lab view the videos.

Limited temporal context. Only DR(eye)VE contains continuous recordings spanning 5 minutes each. This duration allows capture of drivers' quick reactions and deliberate actions, as well as events preceding and following them.

LBW videos likewise span $\approx$ 5 minutes, but recordings are not continuous. Using the indices of the first and last frame of the videos to infer their duration, we estimated that 26.5% of the frames are missing, fracturing 74 videos into 11K segments with average duration of 2s (10 frames). Crucially, data loss is distributed unequally: 16–19% of frames are missing from episodes with longitudinal actions and 32–35% during lateral maneuvers (based on the annotations from Section 4.2.1). Thus, events or objects that led to drivers' behavior may be missing from the recording.

In contrast, BDD-A consists of 10 s clips extracted from the larger BDDV dataset [490] around braking events (6.5 s prior to and 3.5 s after). Because of this design decision, many videos end abruptly, e.g., before the intersection or in the middle of the turn. Furthermore, multiple videos contain repeated or dropped frames. Regardless of the data loss and cropping, the short duration of videos itself poses a concern. On one hand, it is likely that not all observers had sufficient time to adequately process the scene in front of them, since at least several seconds may be necessary to gain situation awareness. For instance, one study found that subjects took between 7 to 12 s to be able to reproduce a traffic situation upon viewing

Figure 4.2: Examples of BDD-A videos with image quality issues: A) reflection on the windshield, B) image out of focus, C) camera tilted and partially obstructed, D) camera tilted up, blurry artifacts.

a short video clip [495]. Likewise, hazard-perception tests that are part of standard licensure process in the UK use videos about 60 s long. On the other hand, use of short videos may also artificially stimulate observers' attention, thus recent proposals suggest increasing the duration of the hazard tests to 10 min to better approximate conditions of the real-world driving [496].

Uneven video quality. Lastly, video quality varies across datasets. In BDD-A, which is based on recordings crowdsourced from hundreds of drivers, nearly 37% of all videos have issues, such as blurriness, reflections, obstructions, and compression artifacts (see Figure 4.2). The videos in DR(eye)VE and LBW are recorded in more controlled environments and are thus more consistent. However, in LBW, about 6K frames (5% of all data) are overexposed and two nighttime videos (comprising ≈2K frames) are underexposed, making it difficult to discern road markings, signs, signals, and vehicles (see Figure 4.3). In DR(eye)VE, the placement of the camera outside the vehicle is problematic for rainy scenarios since water drops accumulate on the camera lens and lead to issues with data processing, as described later in Section 4.3.4.

4.3.2 Vehicle data collection

Vehicle telemetry is a direct reflection of drivers' actions and intentions. Current speed, status of the brake or acceleration pedals, steering wheel rotation, IMU and GPS signals are

Figure 4.3: Overexposed (left) and underexposed (right) frames in LBW.

commonly used in the literature as a proxy for what the driver is doing [497], whereas turn signals can indicate drivers' intention to change lane or make a turn.

Minimal, coarsely sampled, and noisy vehicle data. DR(eye)VE and BDD-A provide only a small set of vehicle data, namely, GPS coordinates, heading, and speed, most of which are coarsely sampled at 1Hz (see Table 4.1). While this data can be used to infer general route information and longitudinal actions, detection of lateral maneuvers, such as lane changes, requires an IMU signal.

Moreover, some vehicle data is either missing or noisy. In BDD-A, it is absent for 90 videos (approx. half of the validation set). In several of the DR(eye)VE videos, vehicle data lags noticeably (in one instance, by nearly a minute). Lastly, LBW does not provide any vehicle information.

4.3.3 Recording eye-tracking data

Eye-tracking data represents (with caveats discussed in Chapter 1) what the driver looked at, and can shed light on the unobservable decision-making processes. However, there is a trade-off between quality and ecological validity of eye-tracking data acquisition, depending on the recording equipment (mobile or static eye tracker) and conditions (on-road or in-lab).

Sparsity of on-road gaze recordings. On-road setup is the most ecologically valid, since the driver controls the car and their decisions have consequences. At the same time, rides cannot be replicated across multiple participants and only mobile or remote eye trackers can be used to allow for free head and body movement. As a result, recorded gaze data is sparse and has lower precision and sampling rate than what can be achieved in the lab. For example, gaze data for DR(eye)VE and LBW recorded in actual vehicles on the road is sampled at 50Hz, resulting in few data points per frame. In LBW, it is further downsampled to 5Hz and a single gaze coordinate per frame.

Lack of engagement and limited context in the laboratory conditions. In the lab,

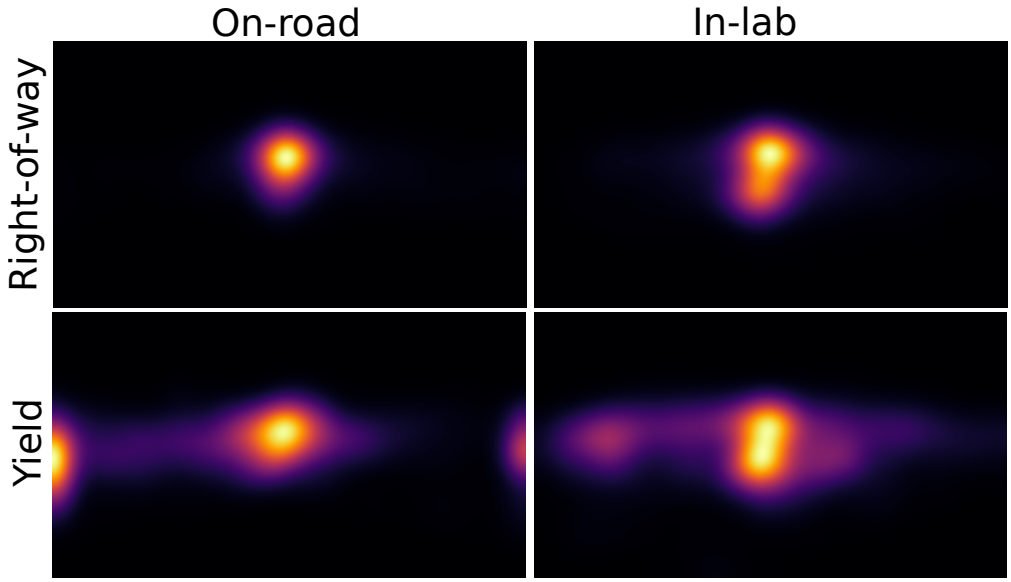


Figure 4.4: Gaze recorded on-road (from DR(eye)VE) and in-lab (from MAAD) for right-of-way and yielding episodes near intersections.

Figure 4.5: An example of gaze recorded on-road and in-lab aggregated over one yielding scenario.

it is possible to recreate the same conditions across participants and use more restrictive but accurate recording equipment. For example, to record ground truth for BDD-A and MAAD, multiple subjects passively viewed the same videos on the computer monitor while their gaze was registered with static eye trackers at 1000Hz.

The in-lab setup also has many known limitations [498]. Here, the main ones are lack of full context and engagement in the task, which impact how human subjects observe the scene. We will demonstrate this using DR(eye)VE and MAAD, that provide eye-tracking data recorded on-road and in-lab, respectively, for the same videos. Figure 4.4 shows on-road and in-lab fixations aggregated across all scenarios where driver had right-of-way or yielded. Note that right-of-way scenarios look similar, but on yielding episodes there are differences. Specifically, in data recorded on the road, there are peaks at either side of the image frame, corresponding to driver's gaze falling on side mirrors or side windows[2]. In comparison, subjects in the lab only have access to what the scene camera has recorded.

[2]We manually labeled fixations that fall outside of bounds of the forward facing camera and pushed them to the edge of the image to preserve the elevation and direction of the driver's gaze.

Since they cannot see beyond the image frame, their viewing patterns are more center-biased. Figure 4.5 shows a qualitative example of a three-way junction, where the driver is repeatedly checking the intersecting road, whereas in-lab subjects focus more on the center of the frame or passing vehicles.

Similar differences in gaze patterns recorded on-road and in-lab are observed for lane changes, where recordings do not show parts of the scene that were visible to the driver through side windows and mirrors.

4.3.4 Processing eye-tracking data

For the gaze prediction task, raw eye-tracking data is processed and aggregated into per-frame heatmaps (or saliency maps) representing areas where human subjects fixated. Below we discuss the processing pipelines adopted in different datasets. Since all raw video and eye-tracking data is available only for the DR(eye)VE dataset, we were able to analyze it in more detail.

DR(eye)VE

As mentioned earlier, DR(eye)VE consists of 6 hours (0.55M frames) of driving footage and gaze data recorded on-road in an actual vehicle. Eight participating drivers wore eye-tracking glasses which allowed free head movement. Due to the insufficient quality of the eye tracker world-facing camera, its limited view, and drivers' head movements, an additional wide-angle Garmin camera was installed on top of the vehicle's roof. The original ground truth saliency maps were created following a two-stage procedure as described in [422]:

1. Drivers' gaze coordinates were mapped from the driver's view (D) to the rooftop camera view (S) via a geometric homography transformation;
2. Eye-tracking data from a single participant resulted in only a handful of data points recorded per frame. Thus, to reduce sparsity and individual driver's biases, data points were aggregated over a sliding window of 25 frames (1 s)

As we examined the raw eye tracker output and ground truth data in DR(eye)VE, several irregularities were identified, which will be discussed in detail:

- temporal misalignment of the world-view eye-tracking glasses camera (D) and rooftop camera (S) video streams;
- incorrect inclusion of saccades and blinks in the ground truth;
- incorrect inclusion of drivers' gaze towards the vehicle interior and gaze to areas outside the rooftop camera viewing span;

Figure 4.6: Examples of temporal misalignment between drivers' and scene-facing camera views. The left column shows the driver's view, the middle column shows the original alignment, and the right column shows the correct alignment. The amount of misalignment in all cases is below 0.5 s (from top to bottom, 14, 8, and 13 frames, respectively), but the effect is significant. In the original alignment, vehicles visible in the driver's view are either not visible (top row) or only partially visible (middle and bottom rows) in the scene view. As a result, fixations on approaching vehicles will not be registered correctly. Temporal misalignment also adds noise due to homography projections.

- noisy homography transformations for mapping gaze from driver's (D) to scene (S) views and aggregating gaze across scene (S) frames;
- artifacts caused by the Gaussian response function used to compute the original ground truth.

Temporal misalignment of video streams. Due to different frame rates of the cameras, each driver (D) and scene (S) video contains 9000 and 7500 frames, respectively. The original mapping between D and S frames is consistent with linear interpolation. However, upon closer inspection, we found that videos were not aligned temporally (Figure 4.6); often, the scene camera continued recording up to 15–20 frames (0.5-0.8s) after the driver's camera stopped.

Incorrect inclusion of saccades and blinks. As mentioned earlier in Section 2.2.3,

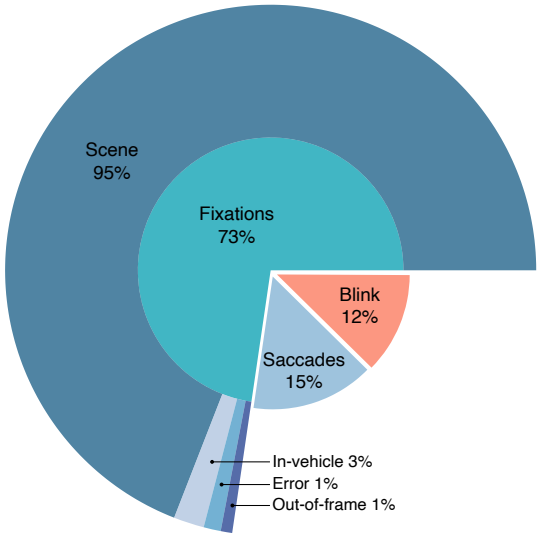


Figure 4.7: Composition of eye-tracking data in DR(eye)VE. Inner circle: % of data points labeled as fixations, blinks, and saccades. Outer circle: spatial distribution of fixations.

human eyes make several types of movements, however, only some are detected by eye-trackers. The eye-tracking glasses used for recording DR(eye)VE, provide labels for fixations, saccades, blinks, and recording errors. The distribution of these events in the dataset is shown in Figure 4.7. Since during blinks and saccades vision is largely suppressed [499], it is common to filter out saccades and blinks from the ground truth, however it was not done in the case of DR(eye)VE. Inclusion of such data effectively biases the dataset by implying that the driver perceived something at those locations. Saccades also produce "trails" in the ground truth maps due to temporal aggregation (see Figure 4.8, top row).

Incorrect inclusion of in-vehicle and out-of-bounds fixations. In DR(eye)VE, 95% of fixations fall on the traffic scene visible through the windshield and only a small fraction is directed at the vehicle interior or towards the side windows (Figure 4.7):

- *In-vehicle gaze.* Gaze towards the car interior is not automatically labeled, but when projected onto the scene, these gaze points will incorrectly reflect what the driver was looking at. This is not a concern for datasets, such as BDD-A and MAAD, that are recorded in the lab, since the videos do not show much of the cars' interior. But in DR(eye)VE, which is recorded on the road, 3% of all fixations are directed towards the vehicle's cabin, varying from 0% to 10% per video, depending on the drivers' individual preferences and the environment.

  At the same time, in-vehicle gaze may also contain useful information. Drivers' glances towards elements of the vehicle interior, such as rearview mirrors, dashboard, and infotainment screen, can be used to infer their intentions and situation awareness. For example,

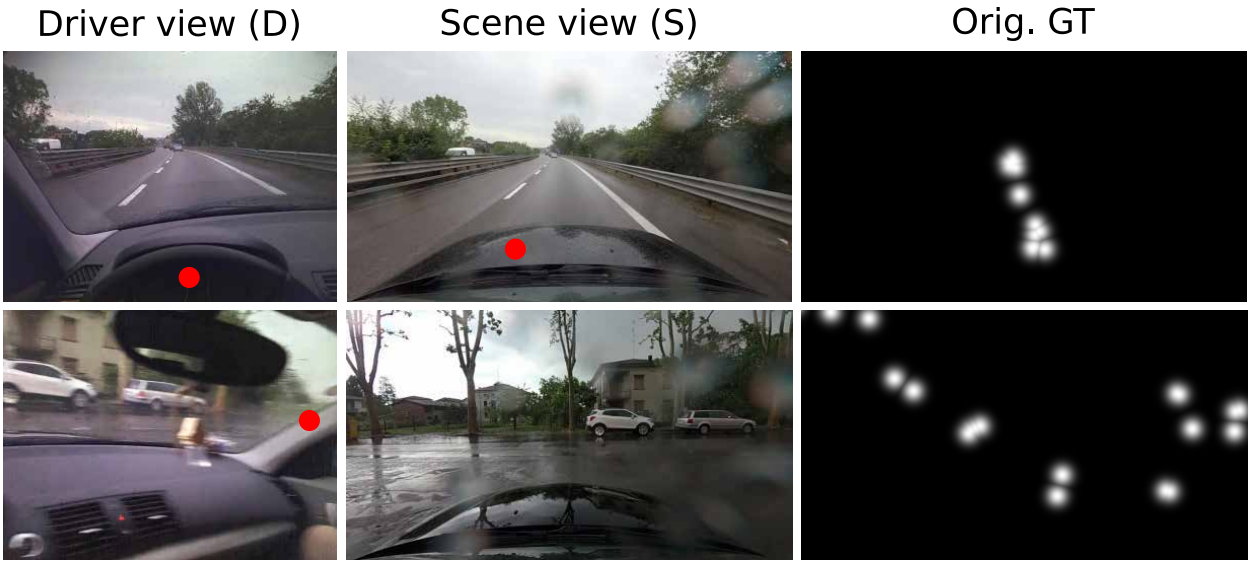


Figure 4.8: Examples of irregularities in the original DR(eye)VE ground truth. Fixations (red circles) are shown in driver and scene views at time t. Ground truth (GT) shows fixations aggregated over frames $t \pm 12$ (1 s). Top row: The "trail" of points in the original GT corresponds to a saccade towards the speedometer. Bottom row: Driver's gaze outside the scene view cannot be mapped to anything in the driver's view, resulting in a random pattern of fixations in the original ground truth.

before braking or making lane changes, drivers usually check the mirrors.

- *Errors.* Fixations that were far out of bounds in the driver's camera view (1% of the total) were removed as they could not be reliably mapped to GAR view.
- *Gaze outside the camera viewpoint.* When the driver turns their head, e.g., to look through side windows or over the shoulder, their gaze often falls outside the rooftop camera (S) view (see Figure 4.8, bottom row).

Noisy gaze transformation from the head-mounted eye tracker view to scene image. The homography calculation is inherently noisy since it depends on feature detection and selection. We measured the quality of the homographies between S and D cameras. Analysis of the homography matrices provided by the authors revealed that 18% of them result in anomalous rescaling ($> 3\times$), 17% are missing, 10% are singular, and another 2% do not preserve orientation (i.e., result in mirror images). For the night videos, however, the rates were worse, on average only 35% homographies were acceptable, with multiple videos only having $10 - 15\%$ usable homographies. Due to central bias of fixations, even poor homographies generally preserved them, however, this also exacerbated errors on off-center fixations.

The theoretical upper bound for S-D homography transformation is estimated as 200 px (for 1920×1080 image) in [422]. We verified this error empirically using the following

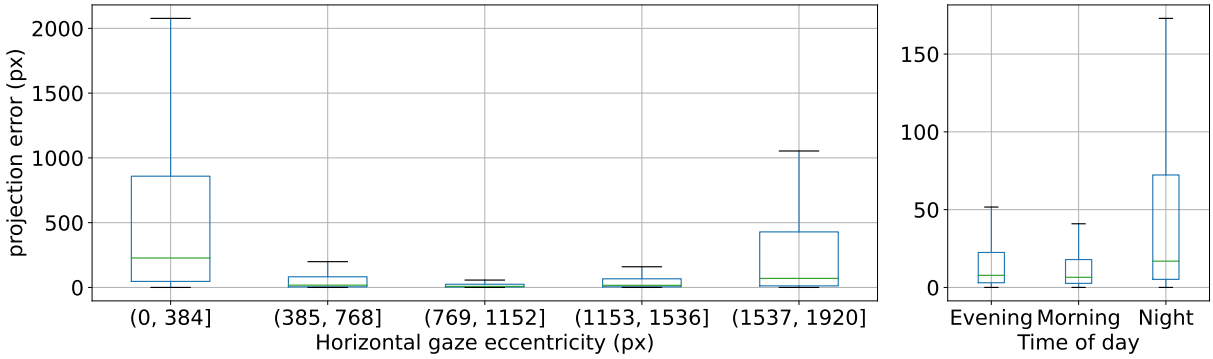


Figure 4.9: A box plot of DR(eye)VE projection errors from the drivers' to scene view for different horizontal gaze eccentricity (left) and time of day (right). The boxes are drawn from first to third quartile, horizontal lines within the boxes show the median values, and whiskers are based on 1.5 IQR value. Outliers are not plotted for clarity.

procedure. We first selected from each video 1000 random pairs of driver and scene frames with at least one valid fixation (74,000 frames in total or 13% of the data) and manually verified the correspondences, correcting when necessary. We then ran a homography transformation 10 times on each pair of frames and measured the deviation from the verified fixations as Euclidean distance in pixels. We found that median projection error for most videos is within 20 px, but over 13% of the frames contain outliers, half of which exceed 200 px, particularly in night and/or rainy conditions. Errors also increase with horizontal gaze eccentricity (Figure 4.9), i.e., fixations closer to the edges of the scene are more difficult to map or fall outside the camera viewpoint. This occurs when the driver rotates their head, which may cause motion blur and reduced overlap between the drivers' and scene views. To further illustrate these issues, several qualitative examples from the DR(eye)VE dataset are shown in Figure 4.10.

LBW

Similar to DR(eye)VE, LBW is recorded on the road and contains a single driver's gaze associated with each video. In addition to scene-facing and head-mounted cameras, a third driver-facing camera was installed inside the vehicle. The same homography transformation approach as in DR(eye)VE is used to project the drivers' gaze from the eye-tracking glasses onto the scene. We examined the LBW data and found many frames where gaze was projected incorrectly, particularly, in cases with the large head rotations (e.g., Figure 4.11). However, the exact errors could not be estimated since neither the raw eye tracker output nor the driver's eye-tracking camera view is available. It is not reported whether in-vehicle

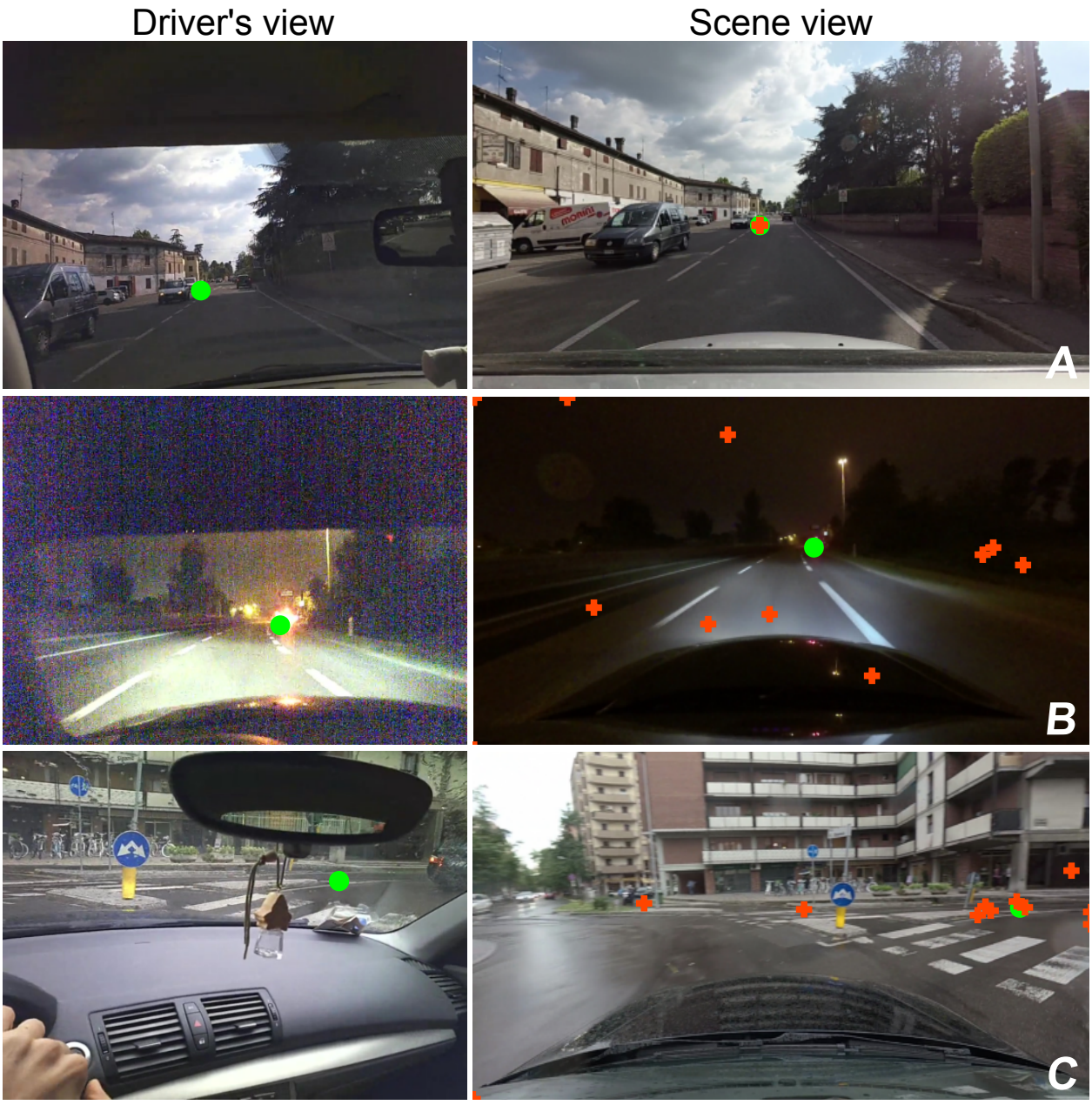


Figure 4.10: Gaze projection errors in DR(eye)VE. Errors are low when projections (red crosses) are close to true gaze location (green circles) in the scene view. Note that errors increase from day (A) to night (B) scene, even in the simple case when the driver maintains speed and looks straight ahead. During turns (C), matches are noisy due to the driver's head rotation, which reduces the overlap between two views.

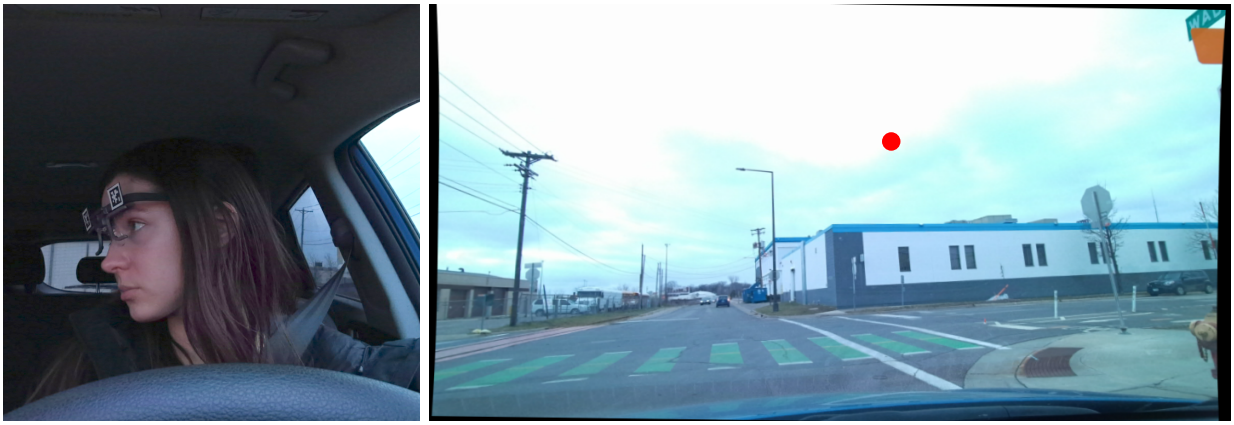


Figure 4.11: A frame from the LBW dataset with incorrectly projected driver's gaze. Since the video from eye-tracking glasses is not available, we show a frame from the driver-facing camera (left) and a corresponding scene view (right) with the driver's gaze location shown as a red circle.

fixations were excluded.

The authors of LBW also did not specify whether they had filtered non-fixation eye movements. Since no raw eye-tracking data is available, we ran an eye-blink detector [500] on the driver-facing videos, however, the false positive rate was extremely high. Manual inspection of several videos revealed at least 1% of frames where drivers' eyes were closed. It was impossible to determine whether saccadic eye movements were included.

BDD-A

BDD-A provides no raw eye-tracking data. Instead, for each scene video clip, there is an associated ground truth video of saliency maps. However, there is a mismatch between the two. First, video clips are recorded at 30 fps and ground truth saliency maps are provided as 29 fps videos. Second, while the resolution of both driving footage and saliency map videos is 1024×768, the ground truth video has black bars on top and bottom of the frame. We address these issues as follows: we split both scene and saliency videos into frames assuming a 29 fps rate to get equal number of corresponding frames. We then crop out the black bars and resize ground truth maps to match the scene view.

MAAD

MAAD eye-tracking data, similar to BDD-A, is recorded in the lab using videos from the DR(eye)VE dataset, to some of which modifications (e.g., flipping, blurring, etc.) are applied. Out of 74 of the original DR(eye)VE videos, for 52 only one subject's recording is available, there are 14 videos with 252+ subjects, 6 videos with 6 subjects, and 2 videos with 11

subjects. No further processing is applied to the raw eye-tracking data.

4.3.5 Differences in ground truth saliency maps across datasets.

The final step in processing eye-tracking data is combining the fixations into saliency maps. In the saliency literature, a typical procedure is as follows: non-saccadic eye movements (fixations, pursuit) of multiple subjects observing the same stimuli are filtered (to remove outliers), aggregated per image or video frame, and convolved with a Gaussian filter with kernel roughly equal to the size of the fovea in the given setup [501, 502, 503, 504]. In each of the datasets we considered, there are some deviations from this procedure:

- DR(eye)VE: Since DR(eye)VE is recorded on the road, each video has only a single observer. Consequently, there are at most few fixations per frame or none at all (if the driver blinked or moved their eyes), which in turn results in sparse saliency maps, if the usual methods for creating ground truth are applied.

  To resolve this problem, the authors of DR(eye)VE propose to aggregate fixations within a temporal window of 1 s (25 frames), i.e., for each frame t, the ground truth saliency map contains fixations from 12 preceding and 12 following frames. To account for motion in the scene, a homography computation is applied to map all fixations within the temporal window $[t-12, t+12]$ to the key frame t.

  Once the fixations were aggregated across the temporal window for every frame, the original ground truth was computed as a max over Gaussian response function over these fixations.

- LBW: Since in LBW, both videos and eye-tracking data are subsampled to 5Hz, the gaze data is even sparser than in DR(eye)VE. There is no additional aggregation step performed, thus each saliency map is based on the 3D direction of a *single* fixation. To generate the gaze map, the fixation location is combined with the depth map of the scene and smoothed with the large Gaussian-like kernel which covers some of the peripheral field of view. This is different from most saliency datasets, including DR(eye)VE and BDD-A, that aim to represent foveal vision.

- BDD-A: The ground truth for BDD-A is recorded from multiple participants who viewed the video clips on the monitor. To form the ground truth, the authors removed non-fixation events, aggregated fixations from all subjects for each frame, and applied a 25 px Gaussian smoothing kernel. Differently from other similar saliency datasets, an additional Gaussian smoothing step is applied temporally across 4 consecutive frames.

Figure 4.12: Scene images from each dataset and corresponding ground truth saliency maps. Note that in DR(eye)VE and BDD-A, saliency maps are more representative of foveal vision and combine multiple fixations. In LBW, a single fixation is used to generate a saliency map which covers part of the peripheral visual field.

Since raw eye-tracking data is not provided, it is not possible to estimate the differences resulting from the different frame rates of videos and eye-tracking data and temporal smoothing.

- MAAD: There are no ground truth saliency maps associated with videos, likely due to unbalanced number of subjects across videos.

These differences are further illustrated in Figure 4.12, which shows examples of the frames and corresponding ground truth saliency maps for each dataset.

4.4 New DR(eye)VE ground truth

Out of the datasets we considered, only DR(eye)VE provided raw eye-tracking data, driver's video, and code used to generate the original saliency maps. Thus, we were able to examine and correct some of the errors discovered earlier as follows:

- We first re-aligned driver's (D) and scene (S) videos temporally. As mentioned in the previous section, the two cameras used to capture data did not start and finish recording at the same time. Since none of the videos contained timestamps, we manually aligned them using events visible in both camera views, such as windshield wiper movement, traffic light changes, etc. We then compared the new and original alignments and found that most D-S pairs were misaligned by 7 ± 5 frames on average, 40% of the correspondences by up to 15 frames (0.5 s), and 10% by over 25 frames (1 s). The effects of asynchrony are naturally more noticeable in busy urban scenes with other vehicles passing the ego-vehicle, but are also apparent in the relatively "empty" rural and night sequences too.
- Video re-alignment changed $\approx 90\%$ of the correspondences between driver's and scene frames, therefore we could no longer use the pre-computed homographies supplied with the dataset to map gaze from D to S view. We used the VLFeat library [505] with the same parameters and thresholds as in the original implementation to re-compute homography transformations for the new alignment, inspected the results, and manually corrected all outliers.
- We used labels produced by the eye-tracker to filter out non-fixations (saccades and blinks).
- We manually annotated fixations towards the vehicle interior and removed those irrelevant to the driving scene (e.g., speedometer, passengers) from the ground truth. Nearly 3% of all fixations are directed towards the speedometer, mirrors, and passengers in the vehicle. In the original dataset, these fixations were erroneously mapped onto the scene (causing random patterns in ground truth) or manually removed (resulting in blank ground truth) but not marked in annotations.

  We also noted fixations towards areas that were not visible in the scene view. Since they were present in virtually all safety-critical scenarios (e.g., intersections), it was important to avoid complete data loss. Therefore, we pushed such out-of-frame fixations towards the edge of the image to preserve the direction and elevation of driver's gaze.

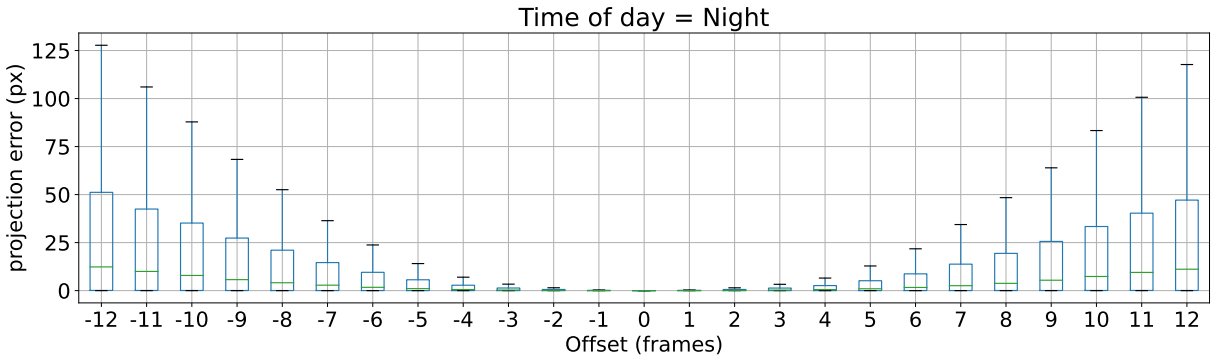


Figure 4.13: A box plot of homography projection errors in DR(eye)VE night videos within the 25 frame (1 s) temporal window, centered on the 0-th key frame. The boxes are drawn from first to third quartile, horizontal lines within the boxes show the median values, and whiskers are based on 1.5 IQR value. Outliers are not plotted for clarity. Outliers are not plotted for clarity.

- Although the use of homography is justified in case of S-D transformation, it is less so for aggregating scene frames across a 1 s temporal window. Even within a short time window, changes in the scene can still be significant due to interactions between moving objects and the camera ego-motion.

  We estimated median error for the temporal homography similarly to S-D described earlier (see Figure 4.13). Since manual ground truth is prohibitively labor-intensive to generate for each frame, we instead compute the mean coordinate of all 10 projected points as a reference. As a result, the median projection error increases towards the edges of the temporal window, but does not exceed 10 px. At the same time, 12–14% of outlier errors exceed 100 px. In the night videos, the median error increases and although the percentage of outliers remains the same, most are >200 px.

  Since the main purpose of this transformation is to compensate for motion in the scene, we instead apply a technique from [506] that relies on optical flow. We use RAFT [507] to compute optical flow and then use its magnitude and direction to trace locations of fixations from frames $t-12, ..., t-1$ and $t+1, ..., t+12$ to the key frame t.

  The use of Gaussian response function for smoothing the original ground truth was also problematic since it resulted in unwanted artifacts and did not reflect different densities of fixations across regions. We recomputed the ground truth maps by setting the fixated locations to 1 and applying an isotropic Gaussian filter with the size of 60 px, following the common procedure for generating ground truth described earlier.

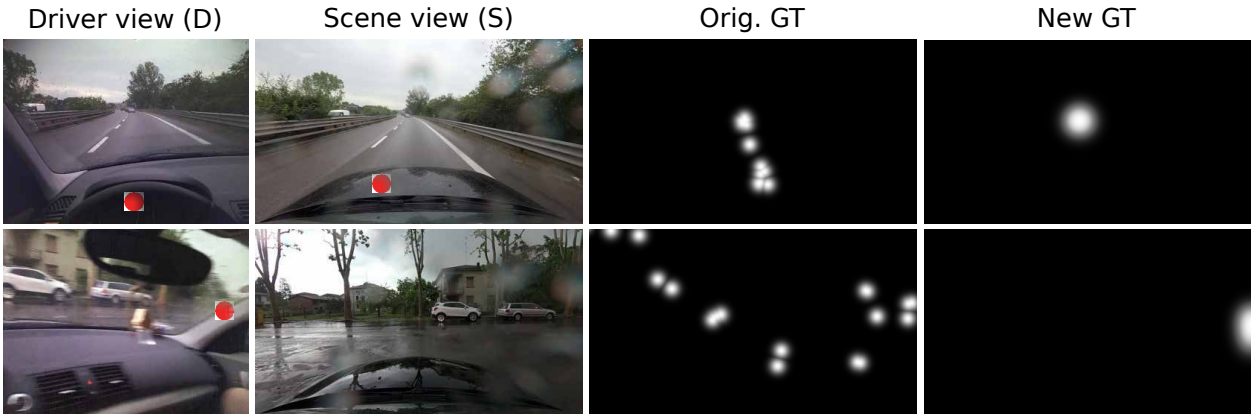


Figure 4.14: Irregularities in the original DR(eye)VE ground truth and proposed corrections. Fixations (red circles) are shown in driver and scene views at time t. Original and new ground truth (GT) show fixations aggregated over frames $t \pm 12$ (1 s). Top row: The "trail" of points is absent in the new GT after removal of saccades and in-vehicle fixations. Bottom row: In the new ground truth, out-of-frame fixations are pushed towards the edge of the frame to preserve the direction and elevation of the driver's gaze.

The final result is shown in Figure 4.14. It can be seen that the new ground truth maps are less noisy, particularly around lateral maneuvers. However, it is difficult to quantify noise reduction effects resulting from different corrections that we introduced in the processing pipeline.

4.5 Recommendations for future data collection

Based on the issues identified in this chapter, there are a number of suggestions that will improve the quality of the collected data. The following four desiderata are the most important:

1. **Sufficient temporal and spatial context.** Recorded videos should better represent what the drivers perceived. We suggest using several synchronized cameras that cover $\geq 180°$ forward view and at least one other camera looking backwards. Videos should be several minutes long to capture enough temporal context for observed events. If these videos are later used for human experiments in the lab, there should be at least 15 s prior to events of interest as observers will need some time to gain situation awareness.
2. **Comprehensive vehicle telemetry.** To adequately represent what the driver was doing, vehicle telemetry should include accurate information on maneuvers and location. The datasets we considered in this chapter provide only GPS data and speed, which is not sufficient for determining current maneuvers or anticipating drivers' upcoming ac-

tion. As a result, we manually labeled these events. To avoid manual annotation and achieve better precision in detecting lateral and longitudinal maneuvers, outputs from other sensors should be considered, such as inertial measurement unit (IMU), position of the brake/throttle pedal, and steering wheel rotations. In addition, drivers' use of turn signals is an indication of their intention, which is also useful for assistance and monitoring applications. On most vehicle models produced in the last decade, this additional information does not require extensive instrumentation and can be acquired through the standard on-board diagnostic (OBD) port or Controller Area Network (CAN) bus.

3. **Correct recording of the scene gaze.** One of the major issues with datasets recorded on the road is inaccurate projection of the driver's gaze onto the image of the scene captured with the forward-facing camera. In DR(eye)VE, where we could examine the raw data, we identified several areas of improvement and suggest the following solutions:

    - Night and adverse weather conditions can cause significant visual artifacts which make it difficult to match drivers' view from eye-tracking glasses to the scene view. There are multiple ways of addressing these issues, including use of a better quality cameras, installing the scene camera inside the vehicle so that it is not exposed to elements, and use of a remote eye tracker together and calibrated cameras to complement feature matching via homography.
    - Drivers often look at the dashboard, mirrors, and other elements of the car interior. These fixations are then incorrectly projected onto the scene. To avoid this, one can put physical markers on the different parts of the interior and use them to filter the data.
    - During turns and lane changes, drivers may look at objects and events that are not captured with the scene camera. Naturally, increasing the spatial extent of the recording would lead to an improvement since there will be more overlap between the driver's and scene views. Additionally, calibrating the cameras or using a remote eye tracker could be used to alleviate this problem.
    - While on-road data is more ecologically valid, only one subject's gaze can be recorded at a time since conditions cannot be reproduced exactly across multiple runs. We showed that aggregating gaze across multiple frames leads to noise, however, another consideration is also the potential delay that it may cause. In the case of the DR(eye)VE dataset, data aggregated over 1 s temporal window implies up to 0.5 s of delay, which is significant compared to the average driver reaction time of about 1 s in unexpected situations [508]. On the other hand, driving simulators allow drivers to repeat the same run, but reduce engagement. Future research should focus on im-

proving subjects' engagement in the task as well as on validating driving simulators to determine the scenarios that can be reliably replicated in the lab conditions.

4. **Availability of the raw data.** Ideally, all raw eye-tracking, video, and vehicle data should be made available for re-analysis and reproduction of the results. DR(eye)VE and MAAD provide sufficient information, however, the other two datasets do not. For example, for BDD-A, it is not specified which BDDV videos were used to extract the 10 s clips, there is no raw eye-tracking data from the human experiment, and the method for computing the weighted sampling cannot be reproduced. LBW is the most processed of the four datasets we considered: there is no video from the driver's eye-tracking glasses to verify the accuracy of gaze projection, all data is subsampled to a low frame rate, raw eye-tracking data is not available, and a quarter of the data is missing for unexplained reasons.

Detailed analysis of the datasets revealed a number of other issues that were minor but likely contributed to noisy or missing data and should be taken into account for future data collection efforts. For example, in DR(eye)VE, all vehicles used for recording the data had rearview mirror hanging accessories, which limited the view of the road, especially during turns. There are numerous cases of road violations by drivers in DR(eye)VE and BDD-A, most commonly ignoring the stop signs, changing lane without checking mirrors, and exceeding the posted speed limit. In both LBW and DR(eye)VE the drivers are frequently seen talking to the passenger in the vehicle. Conversation is considered a cognitive distraction and has been shown to change how drivers observe the scene. In addition to interacting with passengers, some drivers were seen checking their cell phones, sometimes for several seconds. Given that the purpose of these datasets is to provide a baseline of attentive competent driver, road rule violations and distractions should be avoided or at the very least clearly labeled to avoid contaminating the data.

4.6 Summary

In this chapter, we analyzed common datasets for drivers' gaze prediction. Most datasets provide very little information in addition to eye-tracking data and driving footage, thus the contents of the datasets is unknown. We extended the annotations to include information on the task and context and showed that all datasets lack variety and diversity of environments and scenarios. The most common scenarios are maintaining lane and speed or remaining stationary, followed by slowing down before a stop sign or red traffic light. There is little

interaction with other road users and even fewer deliberate actions performed by the driver, such as crossing intersections or making lateral maneuvers.

We also found a number of limitations in the data collection and processing pipelines which contribute to data loss and amplify noise in the ground truth. Temporal, spatial, and quality characteristics of the video recordings do not accurately reflect what the driver has perceived. Because all recordings are made with a single camera, videos do not capture important objects and areas to the sides or behind the vehicle that are visible to the driver. Many videos also end abruptly in the middle of a maneuver or event or interrupted due to data loss. The quality of many recordings suffers due to incorrect positioning of the camera and visual artifacts resulting from exposure to elements (e.g., raindrops on the lens), vibration, etc. Finally, vehicle telemetry provides only the most basic information on drivers' actions, which limits data analysis and model development. Data recording and processing issues especially affect data quality of safety-critical scenarios, such as lateral maneuvers and intersections, which are already the least represented in the data.

For DR(eye)VE, which provides raw eye-tracking data, we were able to conduct a more detailed investigation of the processing pipeline. There, we were able to identify additional issues with alignment of data streams, incorrect inclusion of non-fixation events, and noisy data processing and aggregation procedures. Indirect evidence for similar problems was found for other datasets as well.

We then created a new ground truth for DR(eye)VE that corrected some of the issues found earlier. Specifically, we manually re-aligned the videos, recomputed and manually corrected scene gaze from the driver's view, excluded blinks, saccades, and in-vehicle gaze, adjusted out-of-bounds gaze, and re-aggregated fixations into saliency maps using a more appropriate method.

Lastly, we compiled a set of recommendations for future data collection efforts that emphasize sufficient temporal and spatial context, comprehensive vehicle telemetry, and raw data availability.

Chapter 5

Benchmarking models for drivers' gaze prediction

Accurate modeling of drivers' gaze is necessary for vision-based monitoring and driver assistance systems. Understanding whether attentive drivers adequately observe the scene also has the potential to improve road safety and may be used in driver training. Although there is no complete model of visual attention during driving, several categories of factors affecting how and when drivers look have been identified (see Chapter 2). These various influences on attention are subdivided into bottom-up and top-down [487]. Both play a role during driving; bottom-up cues help react to signals and signs (salient by design) [89] and potential hazards [509, 510], and top-down strategies are important for safely steering the vehicle along the planned route. For example, at intersections, drivers display different gaze patterns depending on maneuver, presence, and status of traffic signals, road layout, and actions of other road users [162, 292, 187, 172].

Even though there is significant evidence for top-down effects on directing drivers' gaze, most of the existing models do not explicitly include them. Furthermore, driving datasets with attention-related data do not contain labels relevant to the task and context, thus making training models and evaluating their performance with respect to these factors difficult. In the previous Chapter 4, we augmented several public datasets with task- and context-relevant annotations. Here, we will use these new labels to examine performance of the SOTA bottom-up models on subsets of the data involving non-trivial drivers' actions using the annotations created earlier. Furthermore, we will discuss how data limitations identified earlier affect model performance.

5.1 Related works

Visual attention. Research on estimating and predicting driver's gaze is rooted in the field of vision science studying bottom-up (data-driven) and top-down (task-driven) influences on attention [487, 511]. The majority of the models proposed to date perform bottom-up saliency prediction in static images and video sequences by identifying properties of rare or out-of-context visual features that attract gaze [512, 513]. Top-down attention, on the other hand, is relatively less studied and is typically investigated within the paradigm of visual search involved in many common activities [514]. In the early works [515, 516], characteristics of the search target (e.g., color) were used to adjust weights of bottom-up features in the output. Modern deep learning models continue to explore modulation techniques during feedforward [517, 518, 519] and backward passes [520, 521, 522] for a range of applications beyond visual search, e.g., object detection, segmentation, and visuo-linguistic tasks.

Driver gaze prediction. Models for predicting drivers' gaze can likewise be subdivided by the types of attentional mechanisms they model and stimuli they use. The majority of deep learning methods achieve promising results on drivers' gaze prediction benchmarks by learning associations between images of the scene and human saliency maps in a bottom-up fashion [523]. Image-based models use CNNs pretrained on an image classification task to encode visual information and then apply convolution and upsampling layers to generate gaze maps [428, 524]. Video-based models aim to extract spatio-temporal patterns from stacks of frames; DR(eye)VENet [422], SCAFNet [525], and MAAD [429] use a 3D CNN pretrained on action classification task as an encoder, whereas BDD-ANet [432] and ADA [526] extract features from each frame via a pretrained CNN and then feed them into a ConvLSTM. Many models use features in addition to visual input, such as optical flow [428, 422, 429], segmentation maps [422, 420, 525], and depth [420].

Top-down influences are often modeled by blending bottom-up saliency outputs with driving-task-related features, such as vanishing point [527, 427, 528], where drivers often maintain gaze [213], drivers' actions derived from vehicle telemetry [415, 417], or gaze location priors associated with certain actions [434]. A recent HammerDrive model [418] recognizes maneuvers, e.g., lane changes, from vehicle data and uses this information to assign weights to bottom-up saliency predictors trained on each task. MEDIRL [419], learns an attention policy via inverse reinforcement learning, where the agent state is represented by the vehicle telemetry, as well as local and global context extracted from visual input.

5.1.1 Problem formulation

The problem of drivers' gaze prediction was formulated earlier in Section 1.3 as predicting where the drivers are more likely to look at time t_N given a stack of N RGB images representing observation up to that point. Since only bottom-up models are considered for the benchmark, no other inputs are provided.

5.1.2 Experiment setup

Models. We select a representative set of saliency models with publicly available code. These include generic bottom-up saliency algorithms (DeepGaze II [529], UNISAL [530], ViNet [531]) and drivers' gaze prediction algorithms (CDNN [527], DReyeVENet [422], BDD-ANet [432], and ADA [526]) that operate on image and video data (UNISAL works on both).
Metrics. We report the evaluation results using common saliency metrics: Kullback-Leibler divergence (KLD), Pearson's correlation coefficient (CC), histogram similarity (SIM), and Normalized Scanpath Saliency (NSS) [436]. We compute all metrics using code in [532]. In particular, we found that this version of KLD is more numerically stable than implementation provided with the DR(eye)VE dataset [422], which is also broadly used in the literature to report the results. In Appendix B, we show the discrepancies between the two implementations. Note that as a result of this change only the absolute KLD values change, the relative ranking of the models is preserved.

Results for different action and context subsets are reported with respect to KLD since it is the metric for which most models optimize. Results on other metrics show similar trends and are provided in Appendix C.
Implementation. All models that have training code, are trained on each dataset using the official code, train/test splits, and default hyperparameters. Because LBW has no official train/test split, we randomly selected videos of subjects 15, 19, 20, 22, and 25 for testing, 16 and 17 for validation, and the rest for training.

Input is either a single frame (for CDNN, image-based version of UNISAL, and DeepGaze II) or a stack of frames representing 0.5 s observation (for DReyeVENet, BDD-ANet, ViNet, video-based UNISAL, and ADA). Note that LBW data has a much lower framerate (5Hz) compared to DR(eye)VE and BDD-A, with 25 Hz and 29 Hz, respectively. To maintain the same observation length across all datasets, for LBW, we use 3 consecutive frames which approximately correspond to 0.5 s. Since most models we test expect a fixed input of 16 frames, we fill the first 13 frames with zeros and append the 3 observation frames at the end. Prediction is made for a single input frame or the last frame in a stack.

Frames that contain U-turns and reversals are excluded from training and evaluation

Table 5.1: Results on the original and new DR(eye)VE ground truth. Arrows next to metrics indicate that larger (↑) and smaller (↓) values are better. Best and second-best values are shown as **bold** and underlined.

	Model	Original GT				New GT			
		KLD↓	CC↑	NSS↑	SIM↑	KLD↓	CC↑	NSS↑	SIM↑
Image	CDNN†	5.33	0.36	1.75	0.24	4.22	0.44	1.51	0.32
	CDNN†*	2.69	0.54	**2.45**	0.35	2.00	0.67	2.03	0.54
	DeepGaze II	3.10	0.25	1.25	0.13	2.52	0.30	1.13	0.16
	UNISAL	2.78	0.34	1.82	0.16	2.40	0.40	1.73	0.21
Video	UNISAL	3.10	0.45	2.26	0.30	2.16	0.54	2.05	0.39
	BDD-ANet†	3.15	0.47	2.08	0.32	2.07	0.57	1.72	0.41
	BDD-ANet†*	2.55	0.51	2.31	0.35	1.61	0.64	2.05	0.48
	DreyeVENet†*	2.63	0.54	2.43	**0.39**	2.13	0.67	2.01	**0.56**
	ADA†	2.33	**0.56**	2.41	**0.39**	1.33	**0.68**	2.00	0.51
	ViNet	1.97	0.48	2.42	0.30	1.37	0.59	**2.21**	0.40
	ViNet*	**1.81**	0.54	2.34	0.34	**1.21**	0.66	1.99	0.48

*—model is trained on data; †—model for drivers' gaze prediction.

because there are only few instances of each, all in dead ends with no traffic, and most gaze information is lost since the driver looks at the mirrors or over the shoulder. All other instances of blank ground truth maps due to blinks/saccades/errors are also discarded.

5.2 Benchmark results

5.2.1 Original and new DR(eye)VE

Overall performance. We first compare the model performance on the original and new DR(eye)VE ground truth to examine the effect of cleaning up the data in Chapter 4. Models with training code are trained on each ground truth, others are only tested.

Results presented in Table 5.1 lead to several observations. There is a noticeable performance improvement of models on the new data, even for those models that were not trained on it (marked with *). For example, even without training on the new data, tested models perform up to 50% better on the new ground truth. At the same time, training on the new ground truth further improves the results by additional 5-10% and also generalizes well to the old ground truth. This can be attributed mainly to the significant reduction of noise due to removal of a large number of blinks and saccades as well as artifacts introduced by the original fixation aggregation procedure (see Section 4.3.4).

There is also no specific difference between generic and driving-specific gaze prediction

Table 5.2: Results on the **new** DR(eye)VE ground truth for different types of actions: None—maintain speed/lane, Acc—accelerate, Dec—decelerate, Lat—lateral actions only, Lat/lon—simultaneous lateral and longitudinal actions, Stop—ego-vehicle is stopped. Green and red colors indicate the best and worst values. Best and second-best values are **bold** and underlined.

		Actions (KLD↓)					
	Model	None	Acc	Dec	Lat	Lat/Lon	Stop
Image	CDNN†	3.07	4.76	5.24	7.01	8.81	7.81
	CDNN†*	1.39	1.53	2.09	3.40	3.88	5.78
	DeepGaze II	2.37	2.56	2.55	3.24	3.52	3.04
	UNISAL	2.13	2.22	2.55	3.98	4.44	3.02
Video	UNISAL	1.64	1.73	2.49	4.81	5.92	4.27
	BDD-ANet†	1.66	1.96	2.26	3.62	4.08	3.34
	BDD-ANet†*	1.22	1.44	1.77	2.70	3.09	3.19
	DreyeVENet†*	1.58	1.83	2.68	4.18	4.27	3.38
	ADA†	1.09	1.17	1.43	2.65	2.57	2.40
	ViNet	1.13	1.28	1.51	2.04	2.54	**2.28**
	ViNet*	**0.96**	**1.11**	**1.33**	**1.76**	**2.08**	2.43

*—model is trained on data; †—model for drivers' gaze prediction.

models. Since generic saliency models (particularly image-based ones) are trained on static and dynamic stimuli that are very different from the driving datasets, they do not generalize well to driving data. However, when trained, they can perform on par or better than driving-specific gaze prediction models. For example, ViNet outperforms driving-specific models on some metrics despite not being trained on multiple driving datasets (like ADA), using any additional information (e.g., semantic segmentation and optical flow as in DReyeVENet), or weighted sampling (as BDD-ANet). This is expected since architectures of all saliency models share many similarities and the same loss functions are used during training.

Next, using the corrected ground truth and new task and context annotations, we analyze performance of the models with respect to various aspects of task and context, namely, driver's actions, intersection types, and ego-vehicle priority when crossing them them.

Results on different types of actions. In Table 5.2, the results are aggregated over different types of actions, as well as times when the ego-vehicle is not moving or is maintaining lane/speed. Naturally, predicting *stop* sequences is more challenging, since drivers' gaze tends to wander around the scene. In sequences where drivers maintain speed and lane (*None*), they predominantly look at the vanishing point or vehicle ahead. As a result, the gaze is very centered and performance of all models is high across all metrics. Longitudinal actions (*Acc* and *Dec*) pose less challenge than lateral actions since they still require the driver to primarily attend to the objects located in front, such as the lead vehicle and traffic

Table 5.3: Results on the **new** DR(eye)VE ground truth for different types of intersections and priorities (RoW—right-of-way). Green and red colors indicate the best and worst values. Best and second-best values are **bold** and underlined.

		Context (KLD↓)							
		Roundabout		Highway ramp		Signalized		Unsignalized	
	Model	RoW	Yield	RoW	Yield	RoW	Yield	RoW	Yield
Image	CDNN†	6.66	10.06	1.66	8.17	5.39	5.99	5.12	13.81
	CDNN†*	2.66	4.72	1.23	7.29	2.43	3.23	1.76	6.17
	DeepGaze II	2.85	3.22	2.35	3.87	2.73	3.09	2.73	4.81
	UNISAL	2.75	4.34	2.06	7.87	2.57	3.68	2.57	8.73
Video	UNISAL	2.35	7.48	1.33	7.43	2.81	4.91	2.86	11.69
	BDD-ANet†	2.18	4.89	1.41	7.96	2.31	4.39	2.13	6.96
	BDD-ANet†*	2.56	3.96	1.04	5.96	1.91	2.42	1.69	4.28
	DreyeVENet†*	3.15	7.59	1.37	8.02	2.42	4.48	2.30	7.68
	ADA†	**1.78**	2.72	0.91	6.77	**1.43**	**2.09**	1.45	4.07
	ViNet	1.87	3.05	1.00	2.57	1.62	2.17	1.50	3.95
	ViNet*	**1.78**	**2.69**	**0.89**	**1.97**	1.46	2.11	**1.30**	**2.58**

*—model is trained on data; †—model for drivers' gaze prediction.

signals or signs. Lateral actions (*Lat*) often involve checking mirrors and scanning across the scene, therefore are more difficult to predict. Finally, both lateral and longitudinal maneuvers (*Lat/Lon*) are performed when crossing intersections or interacting with other road users. In such cases, gaze patterns are more complex because drivers need to look out for potential conflicting vehicles and pedestrians. If areas of interest are not visible to the scene camera due to its limited field of view, models lack visual cues to associate with the recorded drivers' gaze (see Figure 4.14). Since such scenarios are both highly variable and the least represented in the training set, thus performance of all models on this subset is reduced across all metrics.

Results on different types of context. Table 5.3 shows results aggregated by four intersection types (roundabouts, highway ramps, signalized, unsignalized) and two priority options (right-of-way and yield). Here, we observe that sequences where an ego-vehicle must yield to other road participants leads to a dramatic drop in performance for all models. Yielding scenarios are generally more complex than right-of-way ones, since the driver must check different areas of the road (e.g., look to the left and right for incoming traffic), often in advance, and irrespective of presence of other road users. Although in DR(eye)VE traffic density is light, there is still a lot of variability in terms of possible scenarios in each context and a small number of samples in the training data, the bottom-up learning strategy used in all tested models struggles to capture such gaze patterns.

Table 5.4: Results of gaze prediction algorithms on BDD-A and LBW. ↑ and ↓ indicate that larger and smaller values are better. Best values are shown in **bold**. For subsets of data corresponding to different actions and context, cells are colored in the shades of green and orange to highlight the best and the worst performance, respectively. Abbreviations: None—maintain speed/lane, Stop—ego-vehicle is stopped, Lat—lateral actions only, Lon—longitudinal only, Lat/Lon–simultaneous lateral and longitudinal, RoW—ego-vehicle has right-of-way.

		All				Actions (KLD↓)						Context (KLD↓)			
												Signalized		Unsignalized	
Dataset	Model	KLD↓	CC↑	NSS↑	SIM↑	None	Acc	Dec	Lat	Lat/Lon	Stop	RoW	Yield	RoW	Yield
BDD-A	CDNN	1.84	0.62	**1.61**	0.46	1.81	1.79	1.71	2.91	2.25	2.13	1.88	3.19	1.78	3.07
	BDD-ANet	1.84	0.51	1.30	0.39	1.85	1.78	1.79	1.97	1.85	2.16	1.81	1.96	1.60	2.01
	DReyeVENet	1.43	0.64	1.53	**0.50**	1.42	1.39	1.39	1.61	1.54	1.60	1.57	2.26	1.43	1.83
	ViNet	**1.04**	**0.66**	1.58	0.47	**0.97**	**1.00**	**1.03**	**1.17**	**1.15**	**1.25**	**1.15**	**1.30**	**1.03**	**1.27**
LBW	CDNN	0.33	**0.81**	**0.27**	**0.73**	0.27	0.26	**0.34**	0.69	0.53	0.46	0.28	**0.36**	0.27	0.73
	ViNet	**0.31**	**0.81**	0.26	0.74	**0.24**	**0.23**	**0.34**	**0.54**	**0.48**	**0.41**	**0.27**	0.44	**0.25**	**0.69**

[1]Context samples are gathered 1 s prior to entering the intersection.

[2]Official code for BDD-ANet and DReyeVENet could not be adapted to training on LBW, therefore only CDNN and ViNet are evaluated.

Not all intersection types are equally challenging. Roundabouts and unsignalized intersections are more difficult since they always require the driver to check the surroundings and look out for other road users. Whereas when passing signalized intersections or merging on a highway, gaze patterns are simpler since drivers tend to observe the same areas, e.g., looking towards the lane immediately to the left when entering highway on-ramp or checking the traffic approaching from the opposite direction when making a left turn.

5.2.2 BDD-A and LBW

Overall performance. Results in Table 5.4 demonstrate that the overall performance of the models is comparable on DR(eye)VE and BDD-A but very different on LBW. On the latter, the three tested models perform much more uniformly and achieve better results on distribution-based metrics KLD, CC, and SIM. However, location-based NSS is reduced almost by one order of magnitude for some models. The most likely explanation is the nature of the LBW saliency maps, which are based on a single fixation and cover a large part of the image (see Figure 4.12). Because of this, even minor mismatches are heavily penalized by NSS, since it is sensitive to false positives. At the same time, the broad extent of the saliency maps is beneficial for the distribution-based metrics that measure intersection (SIM), correlation (CC), or difference (KLD) between ground truth and prediction. In other words, the models do not need to be precise to achieve good results.

Scenarios with lateral actions and yielding. All models across all three datasets per-

form better on action categories *maintain*, *accelerate* and *decelerate* than on sequences with lateral actions and stops. Similarly, performance is significantly worse on yielding scenarios compared to right-of-way ones. Although it is tempting to attribute these differences to the distribution of training data, the correlation between the results and data composition in Tables 4.2 and 4.3 is not statistically significant.

Complex interactions with other road users are likely not a factor for the poor performance of yielding and lateral actions. We estimated that only 27%, 24% and 22% of such scenarios in DR(eye)VE, LBW, and BDD-A involve at least one conflict vehicle, the rest are on empty roads.

5.3 Discussion

We argue that the limited spatial context and loss of eye-tracking data discussed in Chapter 4 contribute the most to the observed performance trends across all datasets. To recap, near intersections, drivers often look at areas and objects that are not visible in the scene camera view. Some drivers' gaze information can be preserved, for example, in DR(eye)VE, fixations that fall outside the camera view were pushed towards the edges of the frame to indicate direction and elevation of gaze. However, there are no cues in the images to associate actual gaze with.

When videos from the DR(eye)VE dataset were shown to observers in the lab [429], this lack of context translated to different gaze patterns (see Section 4.3.3 and Figures 4.4,4.5). When yielding at the intersections or changing lanes, many subjects did not check the intersecting road since only a small part of it was visible on the monitor and did not see conflicting vehicles until they appeared, thus their gaze was more centered. We speculate that the same is true for BDD-A, which is recorded in the lab and where many scenarios occur near intersections. This may explain why the gap between right-of-way and yielding sequences and lateral and longitudinal actions in BDD-A is smaller than in DR(eye)VE, although further investigation is warranted.

While these issues are more apparent in intersections, lane changes and deceleration scenarios are affected as well. Here, the front-facing camera also does not provide sufficient context, since the driver must be aware of road users on either side and behind the ego-vehicle. In the on-road datasets, the drivers' fixations towards side and rearview mirrors are either lost or erroneously projected onto the forward view. In the datasets recorded in the lab, observers do not have any way of anticipating the upcoming lane change or checking for vehicles outside the view, therefore their gaze patterns simply follow the ego-vehicle's motion.

Neither use of additional features nor weighting techniques can compensate for these data limitations. For instance, DReyeVENet relies on optical flow and semantic segmentation maps for scene analysis and BDD-ANet implements sample weighting to penalize errors on frames with "unusual" gaze patterns. However, both models are consistently outperformed by ViNet, which uses only images and a similar spatio-temporal feature extraction.

5.4 Summary

In this chapter, we benchmarked a number of baseline and state-of-the-art bottom-up models on the datasets we analyzed and annotated earlier in Chapter 4. We first demonstrated that corrections introduced to the processing pipeline to create a new ground truth for DR(eye)VE dataset reduced noise and resulted in better performance even for models not trained on the cleaned-up data. We then showed that the overall performance of the bottom-up models is dominated by the trivial "driving straight" scenarios and that the results drop off significantly on all scenarios involving longitudinal and lateral actions, and crossing intersections. Since the magnitude of performance reduction does not correlate neither with the complexity of the data nor with its composition, we argue that simply increasing the proportion of non-trivial scenarios in data or more fine-grained analysis of the driving scene is not sufficient. Instead two solutions should be considered: 1) collection of the new data to correct existing issues in data collection, recording, processing, and annotation procedures, and 2) model design that includes explicit information about drivers' actions and environment.

Chapter 6

Task- and context-aware gaze prediction

Drivers' gaze prediction is a subfield of a larger research area dedicated to computational modeling of human visual attention. One of the long-standing problems is understanding the mechanisms that govern how people observe the scene: on what areas they focus (fixate) and why do they make overt eye movements (saccades) to look at something else. As mentioned earlier in Chapter 1, factors causing these changes in attention are subdivided into bottom-up (data-driven) and top-down (task-driven). In driving, as in most other everyday visually-guided activities, both types of influences are involved, but there is no comprehensive theory of the complex interactions between them.

In the previous chapter, we demonstrated that the bottom-up approach alone is not sufficient and linked observed trends in the model performance to data limitations discussed in Chapter 4. Here, we test whether including explicit information related to the actions of the driver and state of the environment is helpful for predicting gaze. To this end, we develop a model of task- and context-modulated attention that utilizes vehicle data and augmented annotations to represent top-down influences. We show that this additional information is useful and even compensates for the data loss in some scenarios. We then test the viability of the more accessible map and route information as a representation for task and context. We evaluate our approach on two popular datasets for gaze prediction against state-of-the-art models.

6.1 SCOUT: Task- and context-modulated attention

We start by briefly recalling the problem definition from Section 1.3. Given a set of RGB images recorded from ego-vehicle perspective, predict the distribution of drivers' gaze as a

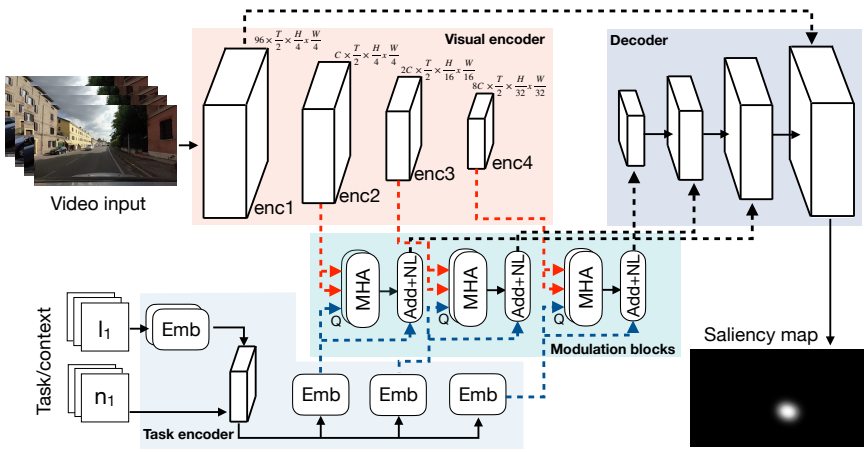


Figure 6.1: Architecture of SCOUT model: two encoders for visual and task/context data, multi-head attention (MHA) modulation blocks, and a decoder. Text labels $l_1, l_2, \ldots l_n$ and numeric features $n_1, n_2, \ldots, n_m$ represent task and context information. Dashed arrows indicate possible connections between modules depending on how many modulation blocks are used for fusing visual and task/context features.

saliency map for the last frame in the input set. In addition to images, other inputs may be provided, such as task and context information.

6.1.1 Model architecture

The architecture of SCOUT comprises three modules: an encoder for visual and task/context features, modulation blocks, and a decoder (as shown in Figure 6.1).

Encoder. To encode video data, we use Video Swin Transformer (VST) [533] as a backbone for spatio-temporal feature extraction. Compared to C3D [534] and S3D [535] backbones used in other saliency models [422, 531], VST is lightweight and performs better on action recognition benchmarks, which is useful for transfer learning. VST takes a set of 16 frames as input and outputs features from each of its four blocks.

Task and context input consists of per-sample text labels and per-frame numeric values (indicated as $l_1, \ldots l_n$ and $n_1, \ldots, n_m$ in Figure 6.4), representing *global context* (weather, time of day, and location labels provided with DR(eye)VE), *local context* (next lateral action label, distance to the next intersection in meters derived from GPS, and priority label), and *current action* (speed in km/h, acceleration in m/s^2, and lateral action label). More details on how task and context are represented is given in Section 6.1.2.

We use a linear embedding layer to encode text labels and downsample the numeric values along the temporal dimension to match the temporal dimension of VST outputs. We then stack the encoded features, apply an affine transformation to match the feature dimension

Table 6.1: Task and context representation for the SCOUT model

Category	Features	Values
Global context	Weather	text label: sunny, cloudy, or rainy
	Time of day	text label: morning, evening, night
	Location	text label: highway, urban, suburban
Local context	Distance to intersection (m)	numeric value
	Priority	right-of-way, yield
	Next action	text label: turn right, turn left, drive straight
Current action	Speed (km/h)	numeric array
	Acceleration (m/s2)	numeric array
	Action	text label: turn right/left, lane change right/left, drive straight

C of the corresponding encoder block and replicate the features across that block's spatial dimensions.

Modulation blocks. We modulate spatio-temporal visual features with task and context information via cross-attention using multi-head attention (MHA) [449], following the success of this technique for modulating attention with visual [519] and multi-modal features [536]. Encoded task features are passed to the MHA as queries, and visual features are used as keys and values. Intuitively, this increases the contribution of visual features that best correspond to the task encoding. We then add the initial task features to the MHA output and apply layer normalization (LN).

Decoder. The decoder gradually fuses information from encoder and modulation blocks, while increasing their spatial dimension and reducing the channel and temporal dimensions. Each of the decoder blocks consists of one upsampling layer, ReLU layer, and 3D convolution layer. The output of each encoder or modulation block is first upsampled and concatenated with the preceding block. The result is fed through ReLU layer and then into the 3D convolutional layer.

6.1.2 Task and context representation

This section details how task and context labels were used for training the proposed SCOUT model. Table 6.1 lists all task and context information grouped into three sets: *global context* represents global weather, time of day, and location, *local context* provides information about the upcoming intersections and actions to be taken along the route, *current action* contains information about the current longitudinal and lateral actions (speed, acceleration, turns, lane changes).

Global context labels were provided with the dataset and are unchanged. Local context

information relies on the new manual annotations and GPS data provided with the dataset. The DR(eye)VE dataset description in [422] does not specify whether the drivers knew the complete route in advance or were given step-by-step instructions. In the latter case, it is also unknown how well in advance such instructions were given and in what form. Therefore, we use guidelines designed for navigation systems that provide voice navigation commands to drivers [537]. We first compute the distance in meters from the current ego-vehicle location to the nearest intersection on the traveled route using available GPS coordinates of the ego-vehicle. If that distance is larger than the max lead distance defined in [537] as $(\text{speed(km/h)} * 2.22) + 37.144$, we set it to *inf* to indicate that the intersection is too far, otherwise, we use the distance directly.

To represent current longitudinal action, we use the sequence of speed and acceleration values. Raw speed values provided with the dataset are interpolated, passed through a median filter to remove outliers, and then used to compute acceleration. For lateral action, we use manually assigned text labels since heading angle is too coarse and inaccurate and yaw information is not available. If there is more than one label in the sample, the most frequently occurring label is chosen.

6.1.3 Implementation details

We use the Swin-S Video Swin Transformer model [533] pre-trained on the Kinetics-400 dataset [538] as encoder backbone and continue to update its weights during training. The input to the network is a set of 16 consecutive frames ($\approx$ 0.5s observation) resized to $224 \times 224 \times 3$. MHA blocks used for modulating visual data with task and context information each have 2 heads and embedding sizes set to match the C dimension of the corresponding encoder blocks.

We use the first 34 videos from DR(eye)VE dataset for training, videos 35–37 for validation, and the remaining videos for testing. New ground truth introduced in Section 4.4 is used to train and evaluate all models. Training samples are generated by splitting the videos into 16-frame segments with 8 frame overlap. Sequences with blank ground truth and U-turns are excluded from training, validation, and testing, following benchmark in Section 5.2.

The model is trained on a single NVIDIA Titan 1080Ti GPU for 20 epochs with the KLD loss, Adam optimizer [539], early stopping based on validation loss, a constant learning rate of $1e-4$, and a batch size of 4.

Table 6.2: Performance of the SCOUT model without task, with weighting, and with task and context information on the **new** ground truth for DR(eye)VE dataset. ↑ and ↓ indicate that larger and smaller values are better. Best and second-best values are shown as **bold** and underlined, respectively. Abbreviations: None—maintain speed/lane, Acc—longitudinal acceleration, Dec—longitudinal deceleration, Lat—lateral actions only, Lat/Lon–simultaneous lateral and longitudinal, Stop—ego-vehicle is stopped, RoW—ego-vehicle has right-of-way.

		All				Actions (KLD↓)						Context (KLD↓)							
												Roundabout		Highway ramp		Signalized		Unsignalized	
	Model	KLD↓	CC↑	NSS↑	SIM↑	None	Acc	Dec	Lat	Lat/Lon	Stop	RoW	Yield	RoW	Yield	RoW	Yield	RoW	Yield
	ADA†	1.33	0.68	2.00	0.51	1.09	1.17	1.43	2.65	2.57	2.40	1.78	2.72	0.91	6.77	1.43	2.09	1.45	4.07
	ViNet*	1.21	0.66	1.99	0.48	0.96	1.11	1.33	1.76	2.08	2.43	1.78	2.69	0.89	1.97	1.46	2.11	1.30	2.58
SCOUT	w/o task	1.08	0.63	3.89	0.52	0.81	0.99	1.25	1.62	1.99	2.43	**1.26**	2.89	0.74	2.24	1.15	2.46	1.10	3.11
SCOUT	w/o task+weight. loss	1.01	0.65	4.04	0.51	0.77	0.93	1.15	1.55	1.88	2.05	1.41	2.60	0.72	2.26	1.09	1.88	1.02	2.74
SCOUT	w/o task+weight. sampl.	1.01	0.65	4.02	0.52	0.76	0.94	1.14	1.55	1.86	2.24	1.51	2.61	0.66	2.12	1.16	1.77	1.01	2.62
SCOUT	w/task CA [3]	0.97	0.66	4.06	0.52	0.73	0.91	1.10	1.54	1.83	2.05	1.29	2.43	0.64	2.29	1.08	1.90	1.04	2.42
SCOUT	w/task GC+LC+CA [2,3,4]	0.95	0.66	4.09	0.54	0.73	0.88	1.06	1.45	1.75	1.95	1.58	1.99	0.64	2.21	1.03	2.12	0.93	2.13
SCOUT	w/task GC+LC+CA [3]	0.94	**0.67**	4.13	0.54	**0.70**	0.89	**1.05**	1.44	1.67	2.09	1.52	**1.90**	**0.57**	2.07	1.01	**1.50**	0.97	2.14
SCOUT	w/task LC [2]	**0.92**	**0.67**	**4.17**	**0.55**	**0.70**	**0.86**	**1.05**	**1.40**	**1.62**	**2.03**	1.43	2.03	0.67	**1.95**	**0.99**	1.60	**0.89**	**2.09**

*—model is trained on the data; †—model for drivers' gaze prediction. Task/context input: GC—global context, LC—local context, CA—current action. []—numbers in square brackets are indices of encoder blocks (shown in Figure 6.1) modulated with task/context information.

6.1.4 Results

Table 6.2 shows the evaluation results of the SCOUT model variants with and without task information on the entire DR(eye)VE dataset, as well as subsets of it corresponding to different types of actions and context. ADA and ViNet, models with the strongest results on action/context from Chapter 5, are shown for reference. As in the previous chapter, results for different action and context subsets are reported with respect to KLD. Results on other metrics show similar trends and are provided in Appendix C.

Performance without task information. We begin by establishing the baseline performance of the proposed model without task information. SCOUT without task improves SOTA results on all metrics, except CC, by a large margin. In particular, it improves NSS (highly sensitive to false positives) by 76% and KLD (that penalizes false negatives) by 11% overall, and achieves better results on some actions and context subsets (by up to 5–30% KLD).

It has been hypothesized that whenever drivers' gaze deviates from mean distribution, the corresponding samples may contain important events [432, 422, 419] and therefore it may be beneficial to focus on them more. We thus investigate two common techniques that emphasize rare samples during training, namely, weighted sampling (as in [432]) and weighted loss. For both, we first compute per-sample weight as a mean KLD score between the sample ground truth and the average saliency map for the video from which the sample is extracted. For weighted sampling, we use the normalized weights as a probability of selecting the sample during training, so that rarer samples are chosen more frequently. However, this

method has the disadvantage that not all samples may be seen during training. Weighted loss allows avoiding this issue by instead using weights to penalize errors on rare samples more.

According to Table 6.2, both techniques lead to an additional improvement of 6% KLD and 4% NSS on the entire dataset. However, when looking at the specific action and context subsets, the improvements are not as significant and sometimes a degradation in performance can be seen (e.g., on lateral actions). We speculate that deviations of gaze distributions from the average are not always correlated with drivers' actions and can be a result of distraction or recording errors, which will be unnecessarily amplified during training.

Effects of task and context input. Some of the improvements discussed earlier could be attributed to the new ground truth for DR(eye)VE, which has less noise due to removal of saccades and blinks and preserves some information when the driver looks beyond the field of view of the forward-facing camera. We can observe this by comparing models that were trained on the new data vs. those that were only tested. For instance, ViNet trained on the new data performs better on all action and context subsets according to the results reported earlier in Tables 5.2 and 5.3. Upon closer examination, the model remains largely reactive and does not consistently check intersections, instead the model tends to flicker between elements of the scene. The bottom-up version of SCOUT performs better, likely attributable to the stronger backbone, but also lacks consistency in its predictions.

In the following experiments, we show that further improvements can be achieved with task and context input to SCOUT: we vary the types of task/context features and whether they are incorporated into a single encoder block or several. The results in Table 6.2 indicate that any additional information is beneficial overall (89% better on NSS and 24% on KLD compared to SOTA, and 2–8% across all metrics with respect to SCOUT w/o task) and especially for action and context subsets (10–30% KLD over SOTA and SCOUT w/o task). The overall best performance is reached when task/context features are added to the second and third encoder blocks, presumably because the information is more compressed at these blocks and contains fewer redundancies and noise. Significant performance gains (20–25% KLD) are also achieved on the yielding scenarios at roundabouts and unsignalized intersections. This is demonstrated in Figure 6.2 rows 1, 2, and 4. Note that SCOUT using task information accurately highlights the left side of the upcoming intersection where conflict vehicles may appear, whereas other models remain centered. At the same time, some right-of-way subsets (roundabout and highway) see a slight decline in performance, caused by the model learning to always check the edges of the road near the intersections (see row 3 in Figure 6.2) even if the ego-vehicle has the right-of-way.

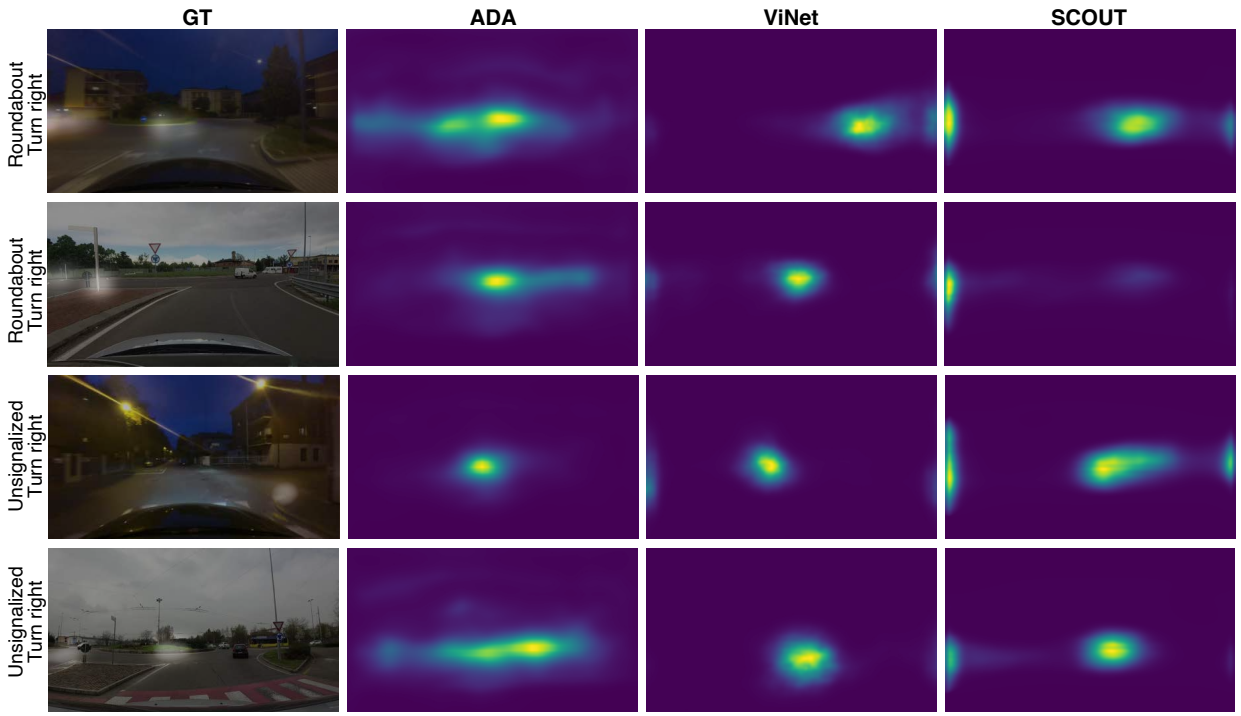


Figure 6.2: Examples of yielding scenarios at different intersections types while performing a turn. The ground truth (in white) is overlaid on top of the images in the first column.

6.2 SCOUT+: Towards practical top-down gaze prediction

Although explicit task models often show improvement over bottom-up approaches, representing the task (even coarsely) requires additional information, which may be privileged or difficult to obtain. Towards making these systems more practical, here, we investigate usefulness of map and route information as a compact representation for surrounding context (upcoming intersections and road curves) and drivers actions (speed and direction of travel). An additional advantage is that this information can be inferred from GPS, which is already available in some datasets and most consumer car models.

6.2.1 Map and route generation

Street network extraction. Map and route information is inferred from the GPS data associated with each video. We first obtain a street network for the geographical region using OpenStreetMap [493] and OSMnx library [540]. To reduce the size of the map, we plot only the driving network (omitting sidewalks and bike lanes) and crop the street network to the bounding box containing the route endpoints. To avoid loss of relevant street nodes, a small

Figure 6.3: Example of the map and route inferred from DR(eye)VE GPS data. Left: street network with overlaid route traveled by the vehicle. Right: enlarged portion of the map shows map matching the original noisy GPS coordinates (shown in red) to the street network (green). Map colors are inverted here for readability. For use in training and inference, street network and route are rendered as white lines on a black background.

offset (500 to 1500 m) is added to the corners of the bounding box.

GPS map matching. GPS positional data is often inaccurate due to measurement and sampling errors [541]. As a result, route coordinates may deviate from the road boundaries (see Figure 6.3). Map matching is a common technique that reduces positional errors by "snapping" GPS location coordinates onto the road network. Here, we use the implementation of the Hidden Markov map matching algorithm [542] provided in the Valhalla[1] routing engine. Since GPS is sampled at 1 Hz, we interpolate the map matched coordinates to obtain a value for each frame in the video.

Map and route rasterization. The final step is to rasterize the street network and route for use in training. We plot the street graph as white lines on a black background. To make sampling easier, we adjust the size of the grayscale map image so that 1 px at 100 dpi resolution approximately corresponds to 1 m in the final image. Using the image size in pixels and corresponding coordinates of the edges in terms of latitude and longitude, we convert the GPS route coordinates to pixel locations in the image plane.

Sampling map and route. During training and testing, we use a portion of the rasterized map and route. For each observation, we crop a square patch of the map centered at the last observed ego-vehicle location and with sides at a predefined lookahead distance away. We normalize orientations of all patches by rotating them such that the direction of travel for the given observation points up. This matches the map to the view of the scene and is similar to how the drivers might view maps on the navigation screen. To represent route,

[1] https://github.com/valhalla/

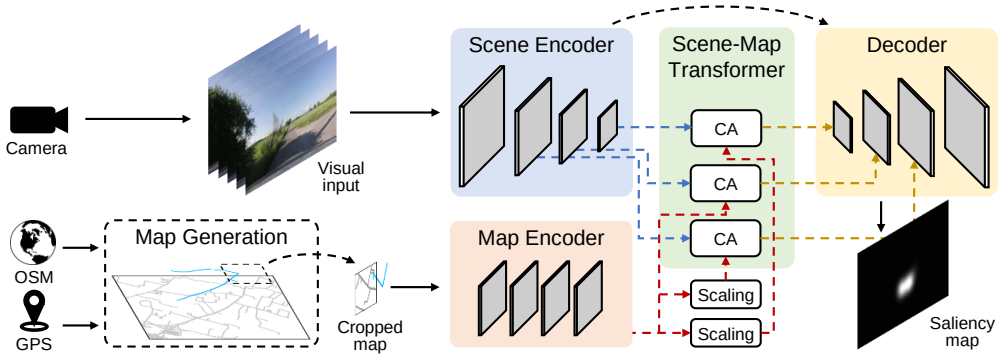


Figure 6.4: Diagram of the SCOUT+ architecture. *Scene encoder* takes in a set of images and outputs 3D spatio-temporal features. To obtain map and route features, the *map generation* module extracts the map and route information from the GPS data using OpenStreetMap (OSM) API. Then, a patch of the map corresponding to the location in the visual input is cropped and fed into the *map encoder*, which produces 2D map and route features via a shallow CNN network. *Scene-map transformer* applies cross-attention (CA) to fuse map and visual features. The *decoder* receives input either directly from the encoder or from the CA blocks (shown as dashed arrows) and gradually mixes and upscales features to obtain the final saliency map.

we create an empty patch of the same size and plot past observed locations of the vehicle corresponding to the input frames as a white line. Then we similarly rotate the patch with the plotted route to match the orientation of the patch with the map. For training and inference, route and map patches are concatenated along the depth dimension.

6.2.2 Model architecture

The proposed SCOUT+ model (shown in Figure 6.4) consists of scene and map encoders, for visual spatio-temporal feature extraction and one for map/route encoding, respectively, a scene-map cross-attention block, and a decoder that gradually fuses and upscales the features to produce the output. Below we discuss each module in detail.

Scene encoder. As in SCOUT we use Video Swin Transformer (VST) [533] for spatio-temporal feature extraction from a stack of RGB input frames. Given a video clip input $\mathcal{I} = \{I_t\}_{t=1}^{T}$, the VST visual backbone produces N visual feature maps $\mathcal{F}^v = \{f_n^v\}_{n=1}^{N}$. Here, we use the modified VST proposed in [543], which does not downsample the input temporally. The outputs of all four encoder blocks are used.

Map encoder. To efficiently encode map and route information, we design a lightweight CNN, following similar approaches in the self-driving domain [544, 545]. The network is composed of four 2D convolutional layers with 10, 20, 10, and 1 kernels of sizes 5×5, 3×3, 3×3, and 1×1, respectively, each followed by a leaky ReLU activation [546]. The network

has a fixed input of size $M \times H_m \times W_m$, where the number of channels corresponds to the number of map features. This 2D encoding is then rescaled via bilinear interpolation and replicated along other dimensions to match the sizes of the visual encoder outputs, resulting in N map features $\mathcal{F}^m = \{f_n^m\}_{n=1}^N$.

Scene-map transformer. We use cross-attention (CA) to fuse visual features with map and route information as follows. The encoded visual features f_n^v are used as key and value, and the corresponding map encoder output f_n^m as query:

$$CA(f_n^v, f_n^m) = \text{Softmax}(\frac{f_n^m W^Q \cdot (f_n^v W^K)^T}{\sqrt{d_{head}}}) f_n^v W^V,$$

where W^Q, W^K, and $W^V \in \mathcal{R}^{c \times d_{head}}$ are learnable parameters.

Decoder. The decoder structure is the same as in SCOUT and consists of four blocks of upsampling, ReLU and 3D convolutional layers. Decoder blocks receive outputs of the encoder or modulation blocks, increase their spatial dimension via upsampling, pass them through ReLU activation layers, and reduce their depth using a 3D convolution layer. The final output is passed through the sigmoid activation function.

6.2.3 Evaluation setup

The evaluation setup remains unchanged. We evaluate our model on two datasets: DR(eye)VE [422] and BDD-A [432]. For training and testing models on DR(eye)VE, we use the new ground truth introduced in Section 4.4.

Performance comparisons are made to a representative set of the models for which official training code is available. Specifically, we use the driver gaze prediction models that take as input a single image (CDNN [527]) or a stack of images (DReyeVENet [422], BDD-ANet [432], and SCOUT [547]). In addition, we use a bottom-up generic video saliency model, ViNet [531]. The models are trained with their default hyperparameters on each dataset.

The results are reported with respect to the following common saliency metrics: Kullback-Leibler divergence (KLD), Pearson's correlation coefficient (CC), normalized scanpath saliency (NSS), and histogram similarity (SIM) [436]. The Python implementations of all metrics are provided in [532].

6.2.4 Implementation

The input to the proposed model is a set of RGB video frames and a single map/route patch. The size of the RGB input is fixed to 16 consecutive frames ($\approx$ 0.5s observation) each resized to $224 \times 224 \times 3$ px. We use Swin-S VST model pre-trained on the Kinetics-400 dataset [538].

As described earlier in Section 6.2.1, map/route input is cropped based on the map radius setting (between 25 and 200 m), rotated, and resized to a fixed size of 128×128 px. Route features are appended to the map patch along the channel dimension. The weights of the pre-trained VST backbone are fine-tuned during training. The map-scene cross-attention modules have 2 attention heads and embedding sizes set to the channel dimensions of the corresponding VST outputs.

As before, we use the first 34 videos from the DR(eye)VE dataset for training, videos 35–37 for validation, and the remaining videos for testing. From BDD-A, we exclude 135 videos, those with quality issues (e.g., unfocused or tilted camera) and ones without gaze or GPS data. Training samples are generated by splitting the videos into 16-frame segments with 8 frame overlap. Samples containing U-turns and reversing are removed from training, validation, and testing. The model is trained on a single NVIDIA Titan 1080Ti GPU for 20 epochs with the KLD loss, Adam optimizer [539], early stopping based on validation loss, a constant learning rate of $1e-4$, and the batch size of 4.

6.2.5 Evaluation results

We first established the optimal map size by running tests with different amounts of lookahead, from 25 and up to 200 meters, to represent different planning horizons. This range was chosen following evidence from the literature on how drivers make decisions. For example, navigation devices are designed to give sufficient time for the driver to prepare for their next action and usually 100-150 m at the city speed of 50 km/h is considered as ideal for alerting the driver to the approaching turn [537, 548]. While drivers often plan their actions several seconds in advance, they do not begin to actively check the intersection until sufficiently close ($<$ 66 m away) [549].

We found that adding maps of any size was beneficial compared to using only the visual information. Gradually expanding the map around the ego-vehicle improved the results, but saturated at 150 m, and started to decline at 200 m. The best improvements were achieved at 100 m, which is also the optimal warning distance established in the literature for the urban speed limit, therefore the map size was fixed at this value for the rest of the experiments.

Results on the DR(eye)VE dataset.

Overall performance. Table 6.3 shows the results of SCOUT+ on DR(eye)VE compared to performance of the SOTA models and two versions of SCOUT—with and without task information. Overall, we observe that all versions of SCOUT+ with map information result in improvement over the SCOUT model without task and are comparable to SCOUT that has access to ground truth task and context information. For example, SCOUT+ performs

Table 6.3: Evaluation results on the DR(eye)VE dataset with **new** ground truth. ↑ and ↓ indicate that larger and smaller values are better. Best and second-best values are shown as **bold** and underlined, respectively. *Task*—whether task/context features are used, *Map*—whether map is used, *Enc.*—at which encoder block(s) map/task features are fused with visual information. The following abbreviations are used for actions and context subsets: *None*—maintain speed/lane, *Stop*—ego-vehicle is stopped, *Lat*—lateral actions only, *Lon*—longitudinal actions only, *Lat/Lon*—simultaneous lateral and longitudinal actions, *RoW*—ego-vehicle has right-of-way.

				All				Actions (KLD↓)						Context (KLD↓)							
														Signalized		Unsignalized		Roundabout		Highway	
	Task	Map	Enc	KLD↓	CC↑	NSS↑	SIM↑	None	Acc	Dec	Lat	Lat/Lon	Stop	RoW	Yield	RoW	Yield	RoW	Yield	RoW	Yield
DReyeVENet	-	-	-	2.13	**0.67**	2.01	**0.56**	1.58	1.83	2.68	4.18	4.27	3.38	2.42	4.49	2.30	7.72	3.16	7.50	1.39	8.10
CDNN	-	-	-	2.00	**0.67**	2.03	0.54	1.39	1.53	2.09	3.40	3.88	5.78	2.43	3.23	1.77	6.19	2.66	4.69	1.24	7.37
BDD-ANet	-	-	-	1.61	0.64	2.05	0.48	1.22	1.44	1.77	2.70	3.09	3.19	1.91	2.42	1.69	4.28	2.56	3.92	1.05	6.01
ViNet	-	-	-	1.21	0.66	1.99	0.48	0.96	1.11	1.33	1.76	2.08	2.43	1.46	2.11	1.29	2.59	1.78	2.67	0.89	1.99
SCOUT	-	-		1.07	0.63	3.89	0.52	0.81	0.99	1.25	1.61	1.99	2.43	1.15	2.46	1.10	3.11	**1.26**	2.72	0.74	2.24
SCOUT	+	-	2	**0.92**	**0.67**	**4.17**	0.55	0.70	0.86	**1.05**	**1.40**	**1.62**	2.03	**0.99**	1.60	**0.89**	**2.09**	1.43	**2.03**	0.67	**1.95**
SCOUT+	-	+	2,3,4	0.96	0.66	4.09	0.53	0.73	0.87	1.11	1.54	1.95	**1.93**	1.07	1.79	1.00	2.82	1.71	2.49	**0.65**	2.22
	-	+	2	0.99	0.66	4.06	0.54	0.74	0.92	1.18	1.55	1.92	2.04	1.13	1.85	1.03	2.81	1.60	2.82	0.69	2.25
	-	+	3	0.94	**0.67**	4.15	0.55	**0.69**	**0.85**	1.08	1.55	1.86	2.00	1.09	**1.57**	0.97	2.56	1.53	2.43	**0.65**	2.16
	-	+	4	0.95	**0.67**	4.12	0.53	0.70	0.86	1.08	1.46	1.73	2.22	1.07	1.66	0.93	2.45	1.53	2.38	0.67	2.10

Table 6.4: Effect of different inputs on SCOUT+ performance on DR(eye)VE. All models are trained with 100 m context. Models that use visual and map information, fuse it with the last encoder block.

Scene	Map	Route	KLD↓	CC↑	NSS↑	SIM↑
	+		1.14	0.62	3.84	0.47
+	+		**0.95**	**0.67**	**4.12**	0.53
+	+	+	0.97	0.66	4.10	**0.54**

as well as SCOUT with task on SIM and CC, and is comparable on NSS and KLD, which are more sensitive to false positives.

In addition, several versions of SCOUT+ are shown where the map is combined with outputs from different blocks of the visual encoder. Although differences in performance on the entire dataset are relatively small (up to 2%) across versions, it is evident that adding map to features from later encoder blocks is better and that fusing features from all blocks is not as effective. We speculate that spatio-temporal information in the early encoder blocks may be too low-level, thus its exclusion reduces noise. At the same time, regardless of how fusion is done, maps lead to significant improvements on challenging subsets corresponding to lateral actions and intersections. For example, decelerations and lateral maneuvers often occur near intersections, for which map information is the most useful. In particular, yielding scenarios and lateral actions see the largest gains (up to 30% KLD) compared to the no-task version of SCOUT. SCOUT+ accurately predicts gaze of drivers who enter roundabouts or intersections (Figure 6.5 rows 1,2) and check for conflicting traffic, sometimes, even if the driver did not check (row 3). Not all action and context categories see equal improvements,

Table 6.5: Evaluation results on the BDD-A dataset. ↑ and ↓ indicate that larger and smaller values are better. Best and second-best values are shown as **bold** and underlined, respectively. *Task*—whether task/context features are used, *Map*—whether map is used, *Enc.*—at which encoder block(s) map/task features are fused with visual information. The following abbreviations are used for actions and context subsets: *None*—maintain speed/lane, *Stop*—ego-vehicle is stopped, *Lat*—lateral actions only, *Lon*—longitudinal actions only, *Lat/Lon*—simultaneous lateral and longitudinal actions, *RoW*—ego-vehicle has right-of-way.

				All				Actions (KLD↓)						Context (KLD↓) Signalized		Unsignalized	
	Task	Map	Enc.	KLD↓	CC↑	NSS↑	SIM↑	None	Acc	Dec	Lat	Lat/Lon	Stop	RoW	Yield	RoW	Yield
DReyeVENet	-	-	-	1.84	0.62	1.61	0.46	1.81	1.79	1.71	2.91	2.25	2.13	1.88	3.19	1.78	3.07
CDNN	-	-	-	1.84	0.51	1.30	0.39	1.85	1.78	1.79	1.97	1.85	2.16	1.81	1.96	1.60	2.01
BDD-ANet	-	-	-	1.43	0.64	1.53	**0.50**	1.42	1.39	1.39	1.61	1.54	1.60	1.57	2.26	1.43	1.83
ViNet	-	-	-	1.04	**0.66**	1.58	0.47	0.97	1.00	1.03	1.17	1.15	**1.25**	**1.15**	1.30	1.03	1.27
SCOUT	-	-	-	1.04	0.63	3.06	0.48	0.99	0.98	1.03	1.11	1.07	1.29	1.18	1.49	1.01	1.25
SCOUT	+	-	2	1.02	0.63	3.08	0.49	0.99	**0.95**	1.02	**1.05**	**1.05**	1.30	1.18	**1.26**	0.97	**1.20**
SCOUT+	-	+	2,3,4	1.10	0.60	2.90	0.45	1.06	1.05	1.08	1.29	1.19	1.34	1.24	1.53	1.07	1.37
	-	+	2	1.05	0.62	3.01	0.47	1.02	1.00	1.04	1.08	1.08	1.31	1.28	**1.26**	1.05	1.25
	-	+	3	**1.01**	0.63	3.08	0.49	0.98	**0.95**	**1.00**	**1.05**	1.06	1.27	1.17	1.33	**0.94**	**1.20**
	-	+	4	**1.01**	0.64	**3.10**	0.48	**0.96**	0.96	**1.00**	1.11	1.14	1.26	1.19	1.34	1.01	1.24

but it is difficult to associate them with specific map or route features due to diversity of the scenarios in each subset.

Effect of different map features. To further investigate the effect of maps, we train the model on only map information, i.e., with no visual input. As can be seen in Table 6.4, despite a predictable performance decrease across all metrics, the results are still competitive with some bottom-up models. Besides confirming the usefulness of maps, it is likely that the properties of the DR(eye)VE dataset also contribute to this outcome. Since nearly half of the videos in DR(eye)VE are collected in rural or nighttime scenes with little traffic, the properties of the road, which are sufficiently represented by the street network, become more dominant. On the other hand, addition of the observed trajectory (route), which provides representation for ego-vehicle direction and speed of motion, leads to only minor improvements. We speculate, that to achieve more significant performance gains, a more precise and complete map representation is needed. For example, for maneuvers, one needs to know what lane is occupied by the ego-vehicle and whether other road users are present.

Results on the BDD-A dataset We report evaluation results on BDD-A in Table 6.5. Note that here even using ground truth task information (SCOUT w/task) does not lead to improvements of the same magnitude as in the case of DR(eye)VE. There are several reasons for this. First, BDD-A is composed of short 10 s clips cropped around braking events from the BDDV dataset [490], where drivers' actions are reactive rather than planned. Second, many clips terminate before or in the middle of the planned actions, e.g., lane changes

Table 6.6: Effect of different inputs on SCOUT+ performance on BDD-A. All models are trained with 100 m context. Models that use visual and map information, fuse it with the last encoder block.

Scene	Map	Route	KLD↓	CC↑	NSS↑	SIM↑
	+		1.57	0.43	2.08	0.33
+	+		**1.01**	**0.63**	**3.08**	**0.49**
+	+	+	1.03	**0.63**	3.07	**0.49**

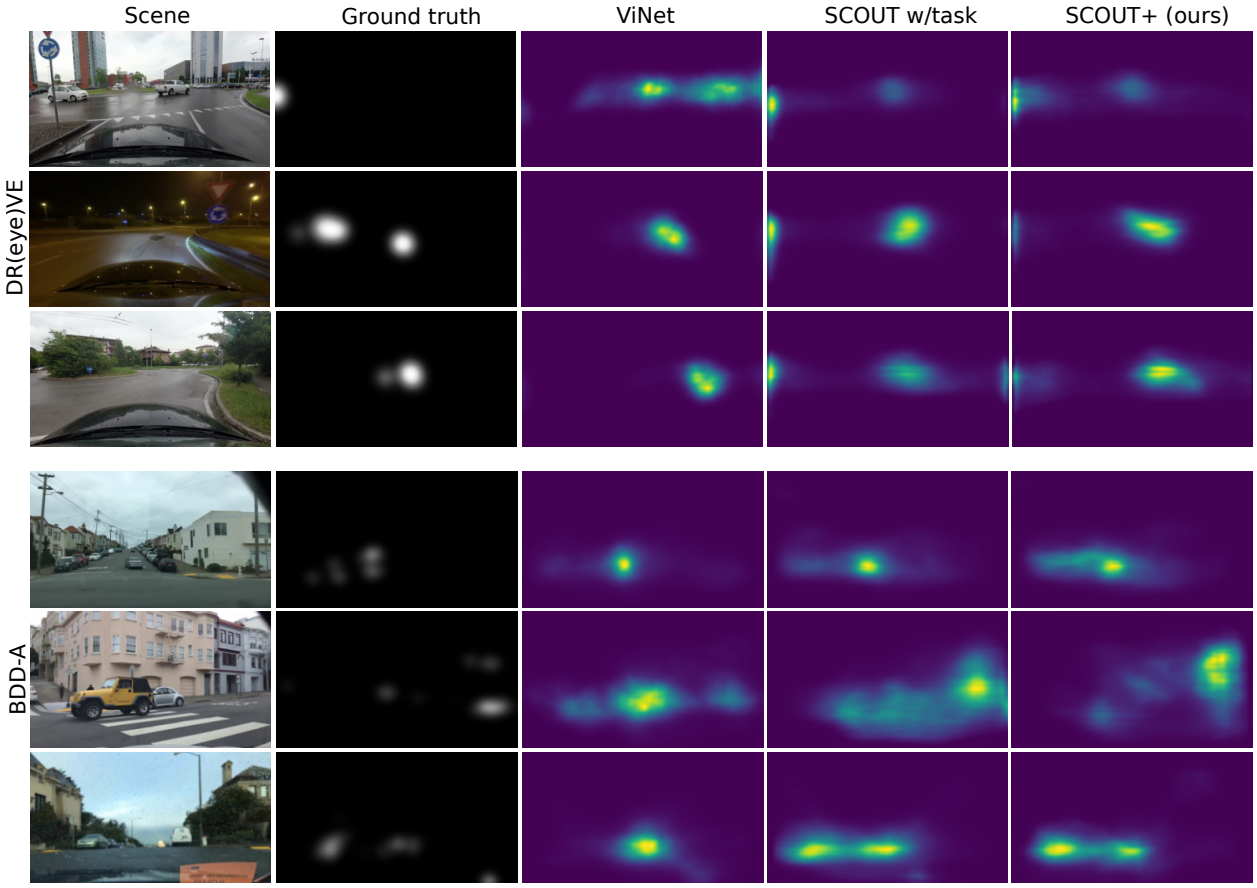


Figure 6.5: Qualitative samples showing performance of SCOUT+ on challenging scenarios near intersections in DR(eye)VE (top 3 rows) and BDD-A (bottom 3 rows) against the bottom-up model ViNet and SCOUT which uses privileged task information.

and crossing the intersections. Third, unlike DR(eye)VE, eye-tracking data for BDD-A was collected in the lab with observers viewing driving footage on the monitor. Finally, BDD-A ground truth combines gaze of at least four subjects, whereas in DR(eye)VE there is only one subject per video. A combination of these factors leads to qualitatively different gaze distributions in BDD-A. For example, gaps in performance between *maintain* and *lateral* actions and different types of intersections are not as large for this dataset compared to DR(eye)VE, which warrants further investigation.

We repeat the same set of experiments as before to evaluate the proposed SCOUT+ model. Here also, map information leads to performance improvement, particularly, when it is fused with later encoder blocks (Table 6.5). As before, relying only on maps without any visual input hinders performance (Table 6.6). However, the decrease is more significant in relative terms, likely because in BDD-A there are more interactions with other road users, thus road features alone are not as helpful. Nevertheless, as shown in Figure 6.5, SCOUT+ learns to anticipate drivers' actions, such as left (rows 4,6) and right (row 5) turns at unsignalized intersections, similar to SCOUT that uses ground truth task information. Note that in these scenarios, the bottom-up model (ViNet) remains more centered.

6.3 Summary

In this chapter, we investigated modeling task and context influences on drivers' attention. We proposed a new model, SCOUT, with an encoder-decoder architecture and cross-attention fusion blocks that modulate visual inputs with explicit task and context information derived from the vehicle data and annotations introduced in Chapter 4. Even considering data limitations pointed out earlier, our model was much more effective than previous state-of-the-art relying on bottom-up learning. These results confirm the usefulness of task and context information for the problem of drivers' gaze prediction. The performance improvements reached 80% overall, and up to 30% on challenging subsets of the data involving crossing intersections and performing lateral maneuvers. We also evaluated other techniques, such as use of additional features and weighted sampling strategies, but none lead to the same performance improvement, likely because they are not directly related to the driving task.

To reduce the reliance on manually annotated task labels, we created a modified version of the model, SCOUT+, to use a more practical and readily available map and route information derived from the GPS data. Extensive experimentation on two datasets against several state-of-the-art models showed that SCOUT+ achieves competitive results, even compared to the original SCOUT model that used annotations directly. Importantly, SCOUT+ with only map information, performs as well on the most challenging subsets with lateral actions and intersection scenarios.

Chapter 7

Final remarks

7.1 Dissertation summary

Driving is a classical example of a common everyday visuomotor task. What distinguishes it from many others is that it possesses inherent risk to both the driver and other traffic participants. For decades, drivers' attention, or lack of it, has been known as one of the main contributors to road accidents. This in turn prompted further research and development of a number of technological solutions for monitoring and assisting drivers to mitigate this problem. Much of this research seeks deeper understanding of how drivers distribute their attention, in other words, where, when, and why, do they attend to different areas in the traffic scene. Therefore, in this dissertation, we focused on the core problem of predicting where the drivers are more likely to look depending on their goals and properties of the traffic environment.

We began by discussing how to study human attention in a dynamic and complex setting that is the traffic environment. We conducted a meta-analysis of the large body of over 200 behavioral studies published in the last decade to identify best practices for recording and analyzing human eye-tracking data as it is predominantly used as a proxy for attention. We examined the independent variables defined in the studies and grouped them into endogenous and exogenous influences on drivers' attention. The results clearly indicate that drivers' involvement in driving- and non-driving-related tasks and both static and dynamic properties of the environment (e.g., road structure and traffic density) affect how drivers observe the scene.

Furthermore, we conducted a meta-study on the applied research in this domain published within the same timeframe. Specifically, we focused on the algorithms for detecting and modeling drivers' alertness and attention, anticipating their maneuvers, and estimating their situation awareness. We found that correctly identifying gaze towards the elements of the

vehicle cabin or traffic scene is key for estimating drivers' state, e.g., drowsiness, distraction, and situation awareness. Although significant progress has been made over the years, several challenges remain unaddressed: 1) limited availability of high-quality publicly available large-scale data and code for models that hinders the ability to reproduce and compare the results; 2) insufficient quality and diversity of the available data, and 3) ignoring the active nature of driving in the data and most models.

The last two challenges were the focus of the remainder of this dissertation. To address the first challenge, we conducted a thorough analysis of the available data, identified issues with processing, and corrected them to the extent possible. In particular, due to the importance of high-quality data for applied research, we analyzed four large-scale public datasets containing eye-tracking data and driving footage to determine how well the data represent what the drivers observed and did. We started with the in-depth analysis of the properties of these datasets, including recording conditions, recording equipment, and data processing pipelines. Specifically, we showed evidence that subjects whose gaze was recorded in the lab did not observe the scene as the drivers did on the road. We attributed the differences to the limited spatial and temporal context, caused largely by the chosen sensors that do not fully represent what the driver perceived and missing data due to cropping and sampling. We then re-analyzed the eye-tracking data processing pipeline and found several issues with the inappropriate inclusion of non-fixation eye-movements, transformation, and aggregation of fixations, which amplify noise in the resulting ground truth used for model development. Based on our observations, we proposed procedures for reducing these artifacts for future data collection efforts. Finally, we cleaned-up and re-aggregated existing data and showed a significant reduction in noise based on the model results.

To link eye-tracking data to the tasks that the drivers were performing, we augmented the datasets with relevant annotations, such as maneuvers and road structure. This enabled us to examine the contents of the data, quantify how well purely data-driven (or bottom-up) models can capture the task, and to develop a new task- and context-aware model for drivers' gaze prediction. In particular, we showed that data is dominated by trivial scenarios where the driver is maintaining lane and speed, or is stationary. Likewise, model performance is significantly higher on these common scenarios and drops-off sharply in cases when maneuvers or potential interactions with other road users are needed (e.g., at intersections). We argued that these performance issues cannot be solved simply by increasing the diversity of the data or considering other features (e.g., object detection, semantic maps, optical flow). We then related some of the observed trends in performance to data limitations and others to the lack of explicit task and context representation.

To further demonstrate the need for more explicit task representations, we developed

two versions of the model for drivers' gaze prediction. The first model, SCOUT, uses the raw visual data (i.e., without additional features such as semantic segmentation or optical flow) and leverages the annotations that we created to modulate the output based on the current vehicle information, street structure, and vehicle right-of-way information. Addition of this information led to large overall improvements of up to 80% on the test set and up to 30% improvements on the challenging subsets involving maneuvers. Despite demonstrating the usefulness of the explicit task representation, the main limitation of this approach was reliance on the extensive annotations. We addressed this problem by developing a modified version of the model, SCOUT+, that uses map and route features derived from GPS to represent task and context. Similar performance reached by the new model on two common datasets showed that these features are a viable option.

In sum, the main contributions of this dissertation are:

- Meta-analysis of the past decade of behavioral and applied research on driving and attention that demonstrates the link between how drivers observe the environment and tasks they perform.
- Analysis of the four publicly available datasets to determine how well they capture what the drivers perceived and actions they performed.
- Cleaned-up ground truth and extended annotations for tasks and context for the public datasets.
- Analysis of the contents of the datasets showing uneven distribution of tasks and dominance of the trivial scenarios (e.g., driving straight).
- Benchmark of the models for drivers' gaze prediction without explicit task representation on the new annotations showing their inability to capture scenarios involving drivers' actions.
- A novel model that in addition to visual data uses explicit task and context representation based on the new annotations to modulate output for different scenarios and establishes state-of-the-art performance on two datasets.
- A version of the proposed model that uses more readily available map and route information to represent task and context and achieves similar performance.
- Recommendations for future behavioral studies, data collection, and model development based on our findings.

7.2 Future directions

Our analysis of the literature, datasets, and models raised a number of questions that could lead to new future research directions. We will split them into three groups: validation studies, cognitively-inspired model development, and safety-focused evaluation.

Given the evidence in Chapter 4, there is a clear need for more experimental validation of the effects that recording conditions have on the transferability of the gaze behaviors from simulators to on-road conditions. In particular, it is important to validate low-fidelity simulators since they are the main source for the majority of the data used in applied research (see Table 3.1). For example, it would be valuable to establish whether in-lab gaze recordings could be made more realistic by improving the spatial and temporal context available to observers (e.g., wider field of view around the vehicle and longer duration of stimuli). Another study could examine the effects of making the task more active, e.g., by giving the passive observers the same instructions that the driver received or by utilizing "shadow driving" technique to engage them in the task (see Section 4.3.3).

In our view, the main issue regarding model development is its lack of cognitive realism. It is likely that improving and scaling data further will lead to immediate improvements even without major changes to the existing model architectures and learning methods. However, based on the trends in other areas of computer science, scaling alone would not achieve human-like behaviors. Thus, we advocate for more cognitively-inspired models that include foveal and peripheral vision as well as working memory to enable continuous processing of the traffic scene, rather than splitting the observations into discrete independent chunks. Another direction of interest is incorporation of explicit domain knowledge and rules of the road and use of reinforcement learning to better link vision to motor actions.

Evaluation of models for assistive and monitoring applications should be more directly related to safety. A first step towards this goal would require establishing an evaluation protocol that adequately represents real-world driving conditions, as well as drivers' perception and actions. Analysis and recommendations in Chapter 4 provide a good starting point for gathering data with such properties that would be useful both for training and evaluation purposes. In addition, evaluation should better reflect the intended applications. For example, to design and test systems for driver assistance, it would be helpful to record or construct in simulation a range of scenarios with potential hazards accompanied by recordings of drivers observing the scene and performing actions. Furthermore, it would be helpful to include both correct and incorrect actions of drivers to test the assistive and monitoring functions of the algorithms, rather than mere similarity to human performance as it is done now.

Appendix A

Publications

Listed below are papers that have resulted from this dissertation:

- **Chapter 2 What affects where drivers look?**

  I. Kotseruba and J. K. Tsotsos, "Behavioral research and practical models of drivers' attention," *arXiv:2104.05677*, 2021.

- **Chapter 3 Driver monitoring and assistance systems**

  I. Kotseruba and J. K. Tsotsos, "Attention for vision-based assistive and automated driving: A review of algorithms and datasets," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 11, pp. 19 907–19 928, 2022.

- **Chapter 4 Analysis of publicly available datasets**

  I. Kotseruba and J. K. Tsotsos, "Understanding and modeling the effects of task and context on drivers' gaze allocation," *arXiv:2310.09275*, 2023, (accepted for presentation at the Intelligent Vehicles Symposium 2024).

  I. Kotseruba and J. K. Tsotsos, "Data limitations for modeling top-down effects on drivers' attention", (accepted for presentation at the Intelligent Vehicles Symposium 2024).

- **Chapter 6 Task- and context-aware gaze prediction**

  I. Kotseruba and J. K. Tsotsos, "Understanding and modeling the effects of task and context on drivers' gaze allocation," *arXiv:2310.09275*, 2023, (accepted for presentation at the Intelligent Vehicles Symposium 2024).

  I. Kotseruba and J. K. Tsotsos, "SCOUT+: Towards Practical Task-Driven Drivers' Gaze Prediction", (accepted for presentation at the Intelligent Vehicles Symposium 2024).

In addition, for the past 3 years, I have maintained and regularly updated a curated list of papers, datasets, and code repositories related to various theoretical and practical aspects of drivers' attention at https://github.com/ykotseruba/attention_and_driving.

Below I list publications that were done during my doctoral studies, but were not directly related to the topic of this dissertation:

- I. Kotseruba and J.K. Tsotsos, "Computational evolution of cognitive architectures", *Oxford University Press* (in production).
- A. Rasouli and I. Kotseruba, "PedFormer: Pedestrian behavior prediction via cross-modal attention modulation and gated multitask learning", *IEEE International Conference on Robotics and Auomation (ICRA)*, 2023.
- I. Kotseruba and A. Rasouli, "Intend-Wait-Perceive-Cross: Exploring the effects of perceptual limitations on pedestrian decision-making", *IEEE Intelligent Vehicles Symposium (IV)*, 2023.
- A. Rasouli and I. Kotseruba, "Intend-Wait-Cross: Towards modeling realistic pedestrian crossing behavior", *IEEE Intelligent Vehicles Symposium (IV)*, 2022.
- I. Kotseruba, M. Papagelis, J.K. Tsotsos, "Industry and academic research in computer vision", *arXiv:2107.04902*, 2021.
- I. Kotseruba, A. Rasouli, J.K. Tsotsos, "Benchmark for evaluating pedestrian action prediction", *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021.
- I. Kotseruba, A. Rasouli, J.K. Tsotsos, "Do they want to cross? Understanding pedestrian intention for behavior prediction", *IEEE Intelligent Vehicles Symposium (IV)*, 2020.

Appendix B

Implementation of Kullback-Leibler divergence

Kullback-Leibler divergence (KLD) is the key dissimilarity metric commonly reported when evaluating saliency algorithms [436]. KLD is defined as $P\|Q = \sum p_i \ln \frac{p_i}{q_i}$, where P and Q are two distributions being compared. Since saliency maps are usually sparse, cases when non-zero value in one image ($p_i > 0$) corresponds to zero value in another ($q_i = 0$) occur frequently. To prevent division by zero, a small constant ϵ is added to the denominator in the equation above. Consequently, if the saliency values are small, the result can be dominated by this parameter. Depending on the implementation, ϵ may be set differently. Below are three typical values used:

- MATLAB machine epsilon: $\epsilon = 1.1920929\text{e}-07$;
- `numpy` machine epsilon: $\epsilon = 2.2204\text{e}-16$[1] (default in Python implementations);
- A small constant: $\epsilon = 0.0001$.

The Python implementation of KLD used in the DR(eye)VE evaluation code[2] follows the widely used Matlab implementation [550] for the MIT saliency benchmark[3], but this Python version is numerically unstable for some inputs. We compared the DR(eye)VE version of KLD to two other Python implementations: Fahimi & Bruce [532, 551] and `pysaliency` [552, 553], the official evaluation code for the MIT/Tuebingen saliency benchmark [4].

Table B.1 shows the results of testing different implementations with several values of ϵ on three test cases:

[1] `np.finfo(np.float32).eps`
[2] https://github.com/ndrplz/dreyeve/blob/master/experiments/metrics/metrics.py
[3] saliency.mit.edu
[4] saliency.tuebingen.ai

Table B.1: Results of testing different KLD implementations with several standard values of ϵ on three scenarios. Diverging results of DR(eye)VE implementation are highlighted in orange.

		ϵ		
		MATLAB	Numpy	Small constant
Data	Implementation	1.19E-07	2.22E-16	0.0001
Identical images	DR(eye)VE	-0.029013	0	-2.771208
	Fahimi & Bruce	0	0	0
	pysaliency	0	0	0
Random images	DR(eye)VE	0.087539	0.499305	-5.205599
	Fahimi & Bruce	0.499304	0.499305	0.498398
	pysaliency	0.499304	0.499306	0.498398
Salmaps from DR(eye)VE	DR(eye)VE	2.322985	7.277252	-2.70738
	Fahimi & Bruce	5.135229	9.888739	3.525612
	pysaliency	5.135229	9.888739	3.525612

1. *Two identical randomly generated images.* KLD for two identical images is expected to be 0, however, DR(eye)VE implementation fails to produce this value with MATLAB and small constant ϵ for two randomly generated identical images (1000×1000 px).

2. *Two different randomly generated images.* Similarly, DR(eye)VE implementation of KLD diverges significantly for MATLAB and small constant ϵ when tested on two *different* randomly generated images. Two other implementations remain stable with small differences in the sixth decimal.

3. *Randomly selected ground truth and predicted saliency maps from DR(eye)VE.* DR(eye)VE implementation diverges from others on a pair of ground truth and predicted saliency maps randomly selected from the dataset. Here, even using the default ϵ generates results different from other implementations.

Table B.2: Evaluation results on the original DR(eye)VE dataset ground truth using DR(eye)VE implementation with MATLAB ϵ and Fahimi & Bruce code with Numpy ϵ.

Model	KLD (DR(eye)VE implementation)	KLD (Fahimi & Bruce)
BDD-ANet	1.92	3.15
DReyeVENet	1.59	2.63
ADA	1.56	2.33

Table B.2 shows the evaluation results of BDD-ANet [432], DReyeVENet [422], and ADA [526], using DR(eye)VE implementation with MATLAB ϵ and Fahimi & Bruce code

with Numpy ϵ. The results obtained with the DR(eye)VE implementation with MATLAB ϵ replicate the KLD values reported in the literature. Fahimi & Bruce implementation produces different absolute values, but preserves the relative ranking of the models.

These results have implications for the published benchmarks as KLD is one of the most popular metrics. It is likely that the majority of results in the literature are reported using the DR(eye)VE implementation, however it cannot be verified for closed-source models. Because the DR(eye)VE implementation tends to produce significantly lower KLD values, it may put models that are evaluated using the correct implementation at a disadvantage. In this dissertation, all reported results are computed using the more accurate Fahimi & Bruce implementation with Python ϵ.

Appendix C

Additional results

In Chapters 5 and 6, we reported model results on action and context sequences using only the KLD metric. The reason for that decision was that majority of models (including our own) were trained with KLD loss since empirically it has been shown to result in the best performance across all metrics. Here, for completeness we provide additional evaluation results on CC, SIM, and NSS metrics.

Additional benchmark results on DR(eye)VE in Table C.1 demonstrate consistent effects of action and context on the performance of the bottom-up models of drivers gaze. Specifically, actions involving lateral maneuvers as well as yielding at roundabouts and unsignalized intersections are the most challenging for all models.

Tables C.2, C.3, and C.4 show that the proposed SCOUT and SCOUT+ models achieve competitive results on the most challenging scenarios (in DReyeVE and BDD-A) with respect to other metrics as well.

Table C.1: Additional benchmark results on the **new** DR(eye)VE ground truth with respect to CC, NSS, and SIM metrics.↑ indicates that larger values are better. Green and red colors indicate better and worse performance, respectively. Best and second-best values are shown as **bold** and underlined. Abbreviations: None—maintain speed/lane, Acc—longitudinal acceleration, Dec—longitudinal deceleration, Lat—lateral actions only, Lat/Lon–simultaneous lateral and longitudinal, Stop—ego-vehicle is stopped, RoW—ego-vehicle has right-of-way.

		Actions (CC↑)						Context (CC↑)							
								Roundabout		Highway ramp		Signalized		Unsignalized	
	Model	None	Acc	Dec	Lat	Lat/Lon	Stop	RoW	Yield	RoW	Yield	RoW	Yield	RoW	Yield
Image	CDNN†	0.54	0.35	0.33	0.31	0.20	0.13	0.18	0.07	0.66	0.24	0.30	0.32	0.37	0.06
	CDNN†*	**0.75**	0.69	0.61	**0.50**	**0.42**	0.27	**0.57**	0.19	0.77	0.50	0.57	0.40	**0.66**	0.21
	DeepGaze II	0.33	0.27	0.28	0.23	0.15	0.19	0.13	0.08	0.35	0.23	0.23	0.18	0.26	0.06
	UNISAL	0.44	0.38	0.38	0.31	0.22	0.33	0.22	0.17	0.45	0.30	0.33	0.30	0.34	0.08
Video	UNISAL	0.59	0.54	0.52	0.42	0.29	0.32	0.36	0.22	0.64	0.38	0.50	0.42	0.52	0.14
	BDD-ANet†	0.65	0.55	0.53	0.40	0.26	0.34	0.35	0.21	0.70	0.36	0.48	0.33	0.52	0.11
	BDD-ANet†*	0.72	0.64	0.59	0.46	**0.33**	0.31	0.37	0.23	0.75	0.44	0.53	0.42	0.61	0.20
	DreyeVENet†*	**0.75**	0.68	0.61	**0.50**	0.42	0.35	0.52	0.18	**0.78**	0.46	0.62	**0.47**	0.65	0.23
	ADA	**0.75**	**0.70**	**0.63**	0.49	0.39	**0.39**	0.52	**0.26**	0.77	0.48	**0.59**	0.45	0.63	0.19
	ViNet	0.65	0.60	0.55	0.46	0.33	0.34	0.37	0.21	0.68	0.48	0.51	0.44	0.57	0.16
	ViNet*	0.74	0.67	0.61	**0.50**	0.40	0.32	0.43	0.25	0.75	**0.51**	0.58	0.38	0.64	**0.30**

		Actions (NSS↑)						Context (NSS↑)							
								Roundabout		Highway ramp		Signalized		Unsignalized	
	Model	None	Acc	Dec	Lat	Lat/Lon	Stop	RoW	Yield	RoW	Yield	RoW	Yield	RoW	Yield
Image	CDNN†	1.77	1.29	1.24	1.10	0.76	0.55	0.75	0.26	2.16	0.88	1.16	1.21	1.28	0.20
	CDNN†*	2.20	2.09	1.95	1.56	1.34	1.14	1.74	0.67	2.27	1.43	1.77	1.36	1.97	0.76
	DeepGaze II	1.23	1.07	1.12	0.91	0.64	0.78	0.58	0.42	1.24	0.87	0.82	0.82	1.03	0.25
	UNISAL	1.87	1.68	1.63	1.31	0.95	1.44	1.01	0.71	1.92	1.22	1.36	1.33	1.51	0.31
Video	UNISAL	2.22	2.09	1.94	1.51	1.12	1.35	1.50	0.79	2.38	1.48	1.80	1.54	1.93	0.50
	BDD-ANet†	1.86	1.74	1.67	1.27	0.89	1.36	1.25	0.71	1.94	1.06	1.53	1.14	1.61	0.34
	BDD-ANet†*	2.22	2.10	1.97	1.49	1.15	1.32	1.39	0.78	2.31	1.38	1.74	1.46	1.98	0.68
	DreyeVENet†*	2.15	2.06	1.90	1.54	**1.42**	1.38	1.71	0.64	2.21	1.37	1.86	1.53	1.96	0.85
	ADA	2.11	2.09	1.95	1.50	1.29	**1.57**	**1.83**	**0.89**	2.16	1.35	1.81	1.59	1.86	0.62
	ViNet	**2.40**	**2.26**	**2.09**	**1.67**	1.26	1.43	1.48	0.77	**2.54**	**1.66**	**1.93**	**1.66**	**2.10**	0.55
	ViNet*	2.13	2.09	1.93	1.56	1.31	1.27	1.50	0.86	2.17	1.45	1.83	1.34	1.94	**1.03**

		Actions (SIM↑)						Context (SIM↑)							
								Roundabout		Highway ramp		Signalized		Unsignalized	
	Model	None	Acc	Dec	Lat	Lat/Lon	Stop	RoW	Yield	RoW	Yield	RoW	Yield	RoW	Yield
Image	CDNN†	0.38	0.26	0.24	0.25	0.17	0.12	0.16	0.10	0.46	0.22	0.22	0.23	0.27	0.08
	CDNN†*	**0.61**	0.55	0.48	0.41	**0.34**	0.23	**0.42**	0.19	0.63	**0.41**	0.42	0.31	0.51	0.18
	DeepGaze II	0.17	0.15	0.16	0.15	0.12	0.13	0.11	0.11	0.17	0.13	0.13	0.12	0.14	0.08
	UNISAL	0.22	0.20	0.20	0.18	0.15	0.19	0.16	0.14	0.22	0.16	0.18	0.17	0.18	0.09
Video	UNISAL	0.42	0.39	0.38	0.32	0.24	0.26	0.28	0.19	0.45	0.29	0.36	0.30	0.37	0.14
	BDD-ANet†	0.46	0.37	0.38	0.30	0.20	0.24	0.26	0.18	0.51	0.29	0.32	0.24	0.35	0.12
	BDD-ANet†*	0.54	0.47	0.43	0.36	0.25	0.23	0.27	0.20	0.57	0.36	0.37	0.30	0.44	0.16
	DreyeVENet†*	0.62	**0.56**	**0.50**	**0.42**	0.34	0.27	**0.43**	0.17	**0.65**	0.38	**0.50**	**0.38**	**0.54**	0.19
	ADA	0.57	0.51	0.47	0.38	0.28	**0.28**	0.30	**0.21**	0.57	0.39	0.41	0.32	0.46	0.16
	ViNet	0.43	0.39	0.37	0.32	0.24	0.25	0.26	0.19	0.45	0.34	0.34	0.29	0.37	0.14
	ViNet*	0.53	0.48	0.44	0.37	0.29	0.24	0.31	0.20	0.53	0.37	0.41	0.27	0.45	**0.20**

*—model is trained on the data; †—model for drivers' gaze prediction.

Table C.2: Additional results for variants of the SCOUT model on the **new** DR(eye)VE ground truth with respect to CC, NSS, and SIM metrics. ↑ indicates that larger values are better. Best and second-best values are shown as **bold** and underlined. Abbreviations: None—maintain speed/lane, Acc—longitudinal acceleration, Dec—longitudinal deceleration, Lat—lateral actions only, Lat/Lon–simultaneous lateral and longitudinal, Stop—ego-vehicle is stopped, RoW—ego-vehicle has right-of-way.

		Actions (CC↑)						Context (CC↑)							
								Roundabout		Highway ramp		Signalized		Unsignalized	
	Model	None	Acc	Dec	Lat	Lat/Lon	Stop	RoW	Yield	RoW	Yield	RoW	Yield	RoW	Yield
	ADA	**0.75**	**0.70**	**0.63**	0.49	0.39	**0.39**	0.52	0.26	**0.77**	0.48	0.59	0.45	0.63	0.19
	ViNet	0.74	0.67	0.61	0.50	0.40	0.32	0.43	0.25	0.75	**0.51**	0.58	0.38	0.64	0.30
SCOUT	w/o task	0.71	0.64	0.57	0.49	0.37	0.27	**0.53**	0.16	0.72	0.48	0.59	0.25	0.61	0.17
SCOUT	w/o task + weighted loss	0.72	0.67	0.61	0.50	0.39	0.35	0.47	0.22	0.73	0.45	0.62	0.40	0.64	0.24
SCOUT	w/o task + weighted sampling	0.72	0.66	0.60	0.50	0.40	0.30	0.45	0.21	0.75	0.46	0.60	0.41	0.64	0.25
SCOUT	w/ task CA[3]	0.73	0.66	0.61	0.50	0.39	0.33	0.41	0.32	0.75	0.46	0.61	0.44	0.63	0.33
SCOUT	w/ task GC+LC+CA [2, 3, 4]	0.73	0.67	0.62	0.52	0.41	0.37	0.38	0.30	0.75	0.46	0.63	0.25	0.66	0.33
SCOUT	w/ task GC+LC+CA [3]	0.74	0.67	**0.63**	0.52	0.43	0.34	0.46	0.26	0.75	0.46	0.62	0.39	0.63	0.28
SCOUT	w/ task LC [2]	0.74	0.68	0.62	**0.55**	**0.45**	0.36	0.42	**0.34**	0.75	0.49	**0.64**	**0.46**	**0.67**	**0.34**

		Actions (NSS↑)						Context (NSS↑)							
								Roundabout		Highway ramp		Signalized		Unsignalized	
	Model	None	Acc	Dec	Lat	Lat/Lon	Stop	RoW	Yield	RoW	Yield	RoW	Yield	RoW	Yield
	ADA	2.11	2.09	1.95	1.50	1.29	1.57	1.83	0.89	2.16	1.35	1.81	1.59	1.86	0.62
	ViNet	2.13	2.09	1.93	1.56	1.31	1.27	1.50	0.86	2.17	1.45	1.83	1.34	1.94	1.03
SCOUT	w/o task	4.40	3.91	3.51	2.97	2.18	1.63	3.07	0.92	4.53	2.97	3.55	1.54	3.76	0.98
SCOUT	w/o task + weighted loss	4.51	4.07	3.70	3.03	2.31	2.07	**2.70**	1.17	4.58	2.77	3.72	2.47	3.91	1.46
SCOUT	w/o task + weighted sampling	4.52	4.02	3.68	3.00	2.38	1.80	2.59	1.17	**4.73**	2.77	3.57	2.57	3.91	1.65
SCOUT	w/ task CA[3]	4.55	4.07	3.73	3.04	2.35	1.97	2.39	1.99	4.70	2.81	3.68	2.76	3.90	2.34
SCOUT	w/ task GC+LC+CA [2, 3, 4]	4.56	4.09	3.79	3.20	2.52	2.23	2.18	1.88	4.70	2.78	3.77	1.59	4.02	2.34
SCOUT	w/ task GC+LC+CA [3]	4.61	4.11	**3.83**	3.18	2.61	2.01	2.65	1.49	4.72	2.80	3.68	2.40	3.86	1.81
SCOUT	w/ task LC [2]	**4.64**	**4.16**	**3.83**	**3.36**	**2.76**	**2.17**	2.40	**2.09**	4.71	**3.00**	**3.80**	**2.91**	**4.12**	**2.43**

		Actions (SIM↑)						Context (SIM↑)							
								Roundabout		Highway ramp		Signalized		Unsignalized	
	Model	None	Acc	Dec	Lat	Lat/Lon	Stop	RoW	Yield	RoW	Yield	RoW	Yield	RoW	Yield
	ADA	0.57	0.51	0.47	0.38	0.28	0.28	0.30	0.21	0.57	0.39	0.41	0.32	0.46	0.16
	ViNet	0.53	0.48	0.44	0.37	0.29	0.24	0.31	0.20	0.53	0.37	0.41	0.27	0.45	0.20
SCOUT	w/o task	0.58	0.52	0.47	0.41	0.31	0.23	**0.41**	0.16	0.59	**0.41**	0.47	0.23	0.50	0.16
SCOUT	w/o task + weighted loss	0.57	0.52	0.48	0.40	0.30	0.26	0.36	0.20	0.57	0.36	0.48	0.29	0.49	0.18
SCOUT	w/o task + weighted sampling	0.58	0.52	0.48	0.41	0.32	0.24	0.36	0.19	0.60	0.37	0.47	0.32	0.51	0.20
SCOUT	w/ task CA[3]	0.59	0.52	0.48	0.40	0.29	0.24	0.33	0.25	**0.61**	0.38	0.49	**0.33**	0.51	0.22
SCOUT	w/ task GC+LC+CA [2, 3, 4]	0.60	0.54	0.49	0.42	0.32	**0.30**	0.30	0.25	**0.61**	0.38	0.48	0.21	0.51	**0.23**
SCOUT	w/ task GC+LC+CA [3]	0.60	0.53	0.49	0.42	0.32	0.26	0.34	0.22	**0.61**	0.37	0.47	0.26	0.49	0.19
SCOUT	w/ task LC [2]	**0.61**	**0.55**	**0.51**	**0.44**	**0.35**	0.28	0.33	**0.26**	**0.61**	0.39	**0.51**	0.32	**0.54**	0.22

*—model is trained on the data; †—model for drivers' gaze prediction.

Table C.3: Additional results for variants of the SCOUT+ model on the **new** DR(eye)VE ground truth with respect to CC, NSS, and SIM metrics. ↑ indicates that larger values are better. Best and second-best values are shown as **bold** and underlined. Abbreviations: None—maintain speed/lane, Acc—longitudinal acceleration, Dec—longitudinal deceleration, Lat—lateral actions only, Lat/Lon–simultaneous lateral and longitudinal, Stop—ego-vehicle is stopped, RoW—ego-vehicle has right-of-way.

				Actions (CC↑)						Context (CC↑)							
										Signalized		Unsignalized		Roundabout		Highway	
Models	Task	Map	Enc.	None	Acc	Dec	Lat	Lat/Lon	Stop	RoW	Yield	RoW	Yield	RoW	Yield	RoW	Yield
DReyeVENet	-	-	-	**0.75**	0.68	0.61	0.50	0.42	0.35	0.62	0.47	0.66	0.23	0.52	0.18	**0.78**	0.46
CDNN	-	-	-	**0.75**	**0.69**	0.61	0.50	0.42	0.27	0.57	0.40	0.66	0.21	**0.57**	0.20	0.77	0.50
BDD-ANet	-	-	-	0.72	0.64	0.59	0.46	0.33	0.31	0.53	0.42	0.61	0.20	0.38	0.23	0.75	0.44
ViNet	-	-	-	0.74	0.67	0.61	0.50	0.40	0.32	0.58	0.38	0.64	0.30	0.43	0.25	0.75	**0.51**
SCOUT w/o task	-	-		0.71	0.64	0.57	0.49	0.37	0.27	0.59	0.25	0.61	0.17	0.53	0.16	0.72	0.48
SCOUT w/task	+	-	2	0.74	0.68	**0.62**	**0.55**	**0.45**	0.36	**0.64**	0.46	**0.67**	**0.34**	0.42	**0.34**	0.75	0.49
SCOUT+	-	+	2,3,4	0.73	0.68	0.61	0.50	0.36	**0.37**	0.62	0.42	0.64	0.21	0.38	0.22	0.75	0.47
	-	+	2	0.73	0.66	0.60	0.51	0.40	0.35	0.61	0.40	0.63	0.25	0.45	0.21	0.74	0.47
	-	+	3	0.74	0.68	**0.62**	0.50	0.39	**0.37**	0.63	**0.48**	0.65	0.26	0.44	0.24	0.75	0.50
	-	+	4	0.74	0.68	**0.62**	0.52	0.41	0.30	0.63	0.45	0.66	0.27	0.43	0.24	0.74	0.46

				Actions (SIM↑)						Context (SIM↑)							
										Signalized		Unsignalized		Roundabout		Highway	
Models	Task	Map	Enc.	None	Acc	Dec	Lat	Lat/Lon	Stop	RoW	Yield	RoW	Yield	RoW	Yield	RoW	Yield
DReyeVENet	-	-	-	**0.62**	**0.56**	0.50	0.42	0.34	0.27	0.50	**0.38**	**0.54**	0.19	**0.43**	0.17	**0.65**	0.38
CDNN	-	-	-	0.61	0.55	0.48	0.41	0.34	0.23	0.42	0.31	0.51	0.18	0.42	0.19	0.63	**0.41**
BDD-ANet	-	-	-	0.54	0.47	0.43	0.36	0.25	0.23	0.37	0.30	0.44	0.16	0.27	0.20	0.56	0.35
ViNet	-	-	-	0.53	0.48	0.44	0.37	0.29	0.24	0.41	0.27	0.45	0.20	0.31	0.20	0.53	0.37
SCOUT w/o task	-	-		0.58	0.52	0.47	0.41	0.31	0.23	0.47	0.23	0.50	0.16	0.41	0.16	0.59	**0.41**
SCOUT w/task	+	-	2	0.61	0.55	**0.51**	**0.44**	**0.35**	0.28	**0.51**	0.32	**0.54**	**0.22**	0.33	**0.26**	0.61	0.39
SCOUT+	-	+	2,3,4	0.60	0.54	0.49	0.41	0.29	0.28	0.49	0.31	0.51	0.18	0.32	0.21	0.61	0.39
	-	+	2	0.60	0.54	0.49	0.42	0.32	**0.29**	0.49	0.32	0.51	0.20	0.37	0.20	0.61	0.39
	-	+	3	0.61	0.55	0.50	0.42	0.31	**0.29**	**0.51**	0.37	0.53	0.21	0.35	0.21	0.62	**0.41**
	-	+	4	0.60	0.54	0.49	0.41	0.32	0.24	0.49	0.32	0.52	0.21	0.34	0.21	0.60	0.37

				Actions (NSS↑)						Context (NSS↑)							
										Signalized		Unsignalized		Roundabout		Highway	
Models	Task	Map	Enc.	None	Acc	Dec	Lat	Lat/Lon	Stop	RoW	Yield	RoW	Yield	RoW	Yield	RoW	Yield
DReyeVENet	-	-	-	2.15	2.06	1.90	1.54	1.42	1.38	1.86	1.52	1.96	0.84	1.71	0.65	2.20	1.36
CDNN	-	-	-	2.20	2.09	1.95	1.56	1.34	1.14	1.77	1.36	1.97	0.75	1.74	0.67	2.27	1.42
BDD-ANet	-	-	-	2.22	2.10	1.97	1.49	1.15	1.32	1.74	1.46	1.98	0.68	1.39	0.78	2.30	1.37
ViNet	-	-	-	2.13	2.09	1.93	1.56	1.31	1.27	1.83	1.34	1.94	1.02	1.50	0.86	2.17	1.45
SCOUT w/o task	-	-		4.40	3.91	3.51	2.97	2.18	1.63	3.55	1.54	3.76	0.98	**3.07**	0.92	4.53	2.97
SCOUT w/task	+	-	2	**4.64**	4.16	**3.83**	**3.36**	**2.76**	2.17	**3.80**	2.91	**4.12**	**2.43**	2.40	**2.09**	4.71	3.00
SCOUT+	-	+	2,3,4	4.58	4.14	3.72	3.03	2.12	2.19	3.76	2.59	3.96	1.24	2.22	1.18	**4.73**	2.91
	-	+	2	4.55	4.06	3.68	3.10	2.35	2.14	3.66	2.49	3.88	1.60	2.61	1.09	4.63	2.89
	-	+	3	**4.64**	**4.19**	3.80	3.06	2.32	**2.21**	**3.80**	**3.01**	4.02	1.68	2.51	1.38	4.69	**3.03**
	-	+	4	4.63	4.18	3.80	3.13	2.46	1.80	3.76	2.82	4.06	1.75	2.46	1.38	4.64	2.85

Table C.4: Additional results for variants of the SCOUT+ model on BDD-A with respect to CC, NSS, and SIM metrics. ↑ indicates that larger values are better. Best and second-best values are shown as **bold** and underlined. Abbreviations: None—maintain speed/lane, Acc—longitudinal acceleration, Dec—longitudinal deceleration, Lat—lateral actions only, Lat/Lon–simultaneous lateral and longitudinal, Stop—ego-vehicle is stopped, RoW—ego-vehicle has right-of-way.

				Actions (CC↑)						Context (CC↑)			
										Signalized		Unsignalized	
Models	Task	Map	Enc.	None	Acc	Dec	Lat	Lat/Lon	Stop	RoW	Yield	RoW	Yield
DReyeVENet	-	-	-	0.68	0.65	0.65	0.57	0.58	0.50	0.56	**0.53**	**0.73**	0.54
CDNN	-	-	-	0.66	0.65	0.63	0.51	0.52	0.52	0.55	0.41	0.67	0.48
BDD-ANet	-	-	-	0.54	0.54	0.52	0.42	0.44	0.38	0.43	0.35	0.55	0.42
ViNet	-	-	-	**0.70**	**0.68**	**0.66**	0.59	0.59	**0.54**	**0.60**	**0.53**	0.69	**0.56**
SCOUT w/o task	-	-	-	0.65	0.66	0.62	0.59	0.59	0.51	0.59	0.47	0.68	0.52
SCOUT w/task	+	-	2	0.66	0.66	0.64	0.60	**0.60**	0.52	0.57	0.48	0.70	0.55
SCOUT+	-	+	2,3,4	0.63	0.62	0.60	0.51	0.54	0.48	0.54	0.44	0.65	0.50
	-	+	2	0.64	0.65	0.62	0.59	0.58	0.50	0.53	0.43	0.66	0.53
	-	+	3	0.66	0.67	0.64	**0.61**	0.58	0.51	0.57	0.41	0.69	0.55
	-	+	4	0.67	0.67	0.64	0.59	0.57	0.52	0.57	0.43	0.69	0.54

				Actions (SIM↑)						Context (SIM↑)			
										Signalized		Unsignalized	
Models	Task	Map	Enc.	None	Acc	Dec	Lat	Lat/Lon	Stop	RoW	Yield	RoW	Yield
DReyeVENet	-	-	-	**0.52**	**0.51**	**0.50**	0.45	**0.47**	**0.41**	0.43	**0.43**	**0.54**	**0.44**
CDNN	-	-	-	0.49	0.48	0.46	0.40	0.41	**0.41**	**0.41**	0.37	0.48	0.39
BDD-ANet	-	-	-	0.41	0.41	0.40	0.34	0.36	0.33	0.34	0.31	0.41	0.35
ViNet	-	-	-	0.49	0.48	0.47	0.43	0.43	0.40	0.42	0.40	0.47	0.41
SCOUT w/o task	-	-	-	0.48	0.49	0.47	0.42	0.44	0.40	0.42	0.40	0.47	0.41
SCOUT w/task	+	-	2	0.51	0.50	0.49	**0.47**	**0.47**	**0.41**	**0.43**	0.41	0.51	0.42
SCOUT+	-	+	2,3,4	0.47	0.47	0.46	0.41	0.42	0.38	0.41	0.36	0.47	0.39
	-	+	2	0.49	0.48	0.48	0.46	0.45	0.39	0.41	0.39	0.48	0.41
	-	+	3	0.50	**0.51**	0.49	**0.47**	0.46	**0.41**	**0.43**	0.38	0.51	0.43
	-	+	4	0.50	0.50	0.49	0.46	0.45	0.40	**0.43**	0.39	0.50	0.42

				Actions (NSS↑)						Context (NSS↑)			
										Signalized		Unsignalized	
Models	Task	Map	Enc.	None	Acc	Dec	Lat	Lat/Lon	Stop	RoW	Yield	RoW	Yield
DReyeVENet	-	-	-	1.54	1.52	1.58	1.43	1.42	1.26	1.36	1.18	1.64	1.26
CDNN	-	-	-	1.63	1.60	1.67	1.35	1.39	1.38	1.43	1.18	1.73	1.16
BDD-ANet	-	-	-	1.31	1.29	1.36	1.12	1.14	1.01	1.13	0.86	1.40	0.98
ViNet	-	-	-	1.59	1.56	1.63	1.45	1.43	1.38	1.45	1.37	1.65	1.25
SCOUT w/o task	-	-	-	3.10	3.12	3.10	2.30	2.35	1.92	2.40	1.58	3.30	2.00
SCOUT w/task	+	-	2	3.26	3.16	3.17	2.78	**2.72**	**2.21**	**2.73**	**1.97**	**3.62**	2.00
SCOUT+	-	+	2,3,4	3.06	2.95	2.99	2.42	2.48	2.02	2.58	1.75	3.37	2.19
	-	+	2	3.15	3.08	3.09	2.77	2.61	2.13	2.54	1.79	3.40	2.31
	-	+	3	3.22	**3.20**	3.16	**2.83**	2.66	2.18	**2.73**	1.62	3.58	**2.42**
	-	+	4	**3.29**	3.18	**3.18**	2.76	2.60	2.18	2.72	1.69	3.55	2.36

Bibliography

[1] T. Victor, M. Dozza, J. Bärgman, C.-N. Boda, J. Engström, C. Flannagan, J. D. Lee, and G. Markkula, "Analysis of naturalistic driving study data: Safer glances, driver inattention, and crash risk," Transportation Research Board, Tech. Rep. SHRP 2 Report S2-S08A-RW-1, 2015.

[2] E. C. Olsen, S. E. Lee, and B. G. Simons-Morton, "Eye movement patterns for novice teen drivers: does 6 months of driving experience make a difference?" *Transportation Research Record*, vol. 2009, no. 1, pp. 8–14, 2007.

[3] J. J. Gibson and L. E. Crooks, "A theoretical field-analysis of automobile-driving," *The American Journal of Psychology*, vol. 51, no. 3, pp. 453–471, 1938.

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

[5] 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," National Highway Traffic Safety Administration (NHTSA), Tech. Rep. DOT HS 810 59, 2006.

[6] C. Robbins and P. Chapman, "How does drivers' visual search change as a function of experience? A systematic review and meta-analysis," *Accident Analysis & Prevention*, vol. 132, p. 105266, 2019.

[7] M. J. Daluge, M. DeLong, L. Hanig, H. Kalla, C. Klauer, K. Klein, S. Klekar, L. McMillan, C. Quiroga, J. Soule, M. Tracy, and B. Wessinger, "Outdoor advertising control practices in australia, europe, and japan," U.S. Federal Highway Administration. Office of International Programs, Tech. Rep. FHWA-PL-10-031, 2011.

[8] J. K. Zell, "Driver's eye movements as a function of driving experience," Ohio State University, Tech. Rep. IE-16, 1969.

[9] R. R. Mourant and T. H. Rockwell, "Visual information seeking of novice drivers," SAE International, Tech. Rep. 700397, 1970.

[10] M. Faus, F. A. Pla, C. E. Martínez, and S. A. U. Hernández, "Are adult driver education programs effective? A systematic review of evaluations of accident prevention training courses," *International Journal of Educational Psychology*, vol. 12, no. 1, pp. 62–91, 2023.

[11] "How Prevalent Are Advanced Driver Assistance Systems (ADAS)?" https://rts.i-car.com/collision-repair-news/how-prevalent-are-advanced-driver-assistance-systems-adas.html, 2017, accessed: 2024-03-22.

[12] S. Sinclair, "Guide to Cars With Advanced Safety Systems," *Consumer Reports*, 2021, https://www.consumerreports.org/car-safety/cars-with-advanced-safety-systems/, accessed: 2024-03-22.

[13] 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 Analysis & Prevention*, vol. 117, pp. 55–64, 2018.

[14] A. El Khatib, C. Ou, and F. Karray, "Driver inattention detection in the context of next-generation autonomous vehicles design: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 11, pp. 4483–4496, 2019.

[15] SAE, "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles," *SAE International, Standard J3016*, 2018.

[16] J. Gaspar and C. Carney, "The effect of partial automation on driver attention: a naturalistic driving study," *Human Factors*, vol. 61, no. 8, pp. 1261–1276, 2019.

[17] J. C. 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," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 27, pp. 196–217, 2014.

[18] 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*, C. M. Schlick, S. Duckwitz, F. Flemisch, M. Frenz, S. Kuz, A. Mertens, and S. Mütze-Niewöhner, Eds., 2017, pp. 309–318.

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

[20] Euro NCAP, "Euro NCAP 2025 Roadmap—In pursuit of Vision Zero," https://cdn.euroncap.com/media/30700/euroncap-roadmap-2025-v4.pdf, 2017, accessed: 2024-03-22.

[21] J. W. Senders, A. Kristofferson, W. Levison, C. Dietrich, and J. Ward, "The attentional demand of automobile driving," *Highway Research Record*, vol. 195, pp. 15–32, 1967.

[22] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer, *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press, 2011.

[23] D. A. Gordon, "Experimental isolation of the driver's visual input," *Human Factors*, vol. 8, no. 2, pp. 129–138, 1966.

[24] T. H. Rockwell, R. L. Ernst, and M. J. Rulon, "Visual requirements in night driving," Highway Research Board, Tech. Rep. 99, 1970.

[25] N. A. Kaluger and G. Smith Jr, "Driver eye-movement patterns under conditions of prolonged driving and sleep deprivation," *Highway Research Record*, vol. 336, pp. 92–106, 1969.

[26] W. S. Geisler and J. S. Perry, "Real-time foveated multiresolution system for low-bandwidth video communication," in *Human Vision and Electronic Imaging III*, vol. 3299, 1998, pp. 294–305.

[27] J. Tsotsos, I. Kotseruba, and C. Wloka, "A focus on selection for fixation," *Journal of Eye Movement Research*, vol. 9, pp. 1–34, 2016.

[28] B. Wolfe, J. Dobres, R. Rosenholtz, and B. Reimer, "More than the Useful Field: Considering peripheral vision in driving," *Applied Ergonomics*, vol. 65, pp. 316–325, 2017.

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

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

[31] R. A. Cohen, "Cortical magnification," in *Encyclopedia of Clinical Neuropsychology*, J. S. Kreutzer, J. DeLuca, and B. Caplan, Eds. Springer New York, 2011, pp. 718–719.

[32] H. Strasburger, "Seven myths on crowding and peripheral vision," *i-Perception*, vol. 11, no. 3, p. 2041669520913052, 2020.

[33] G. E. Legge and D. Kersten, "Contrast discrimination in peripheral vision," *Journal of the Optical Society of America A*, vol. 4, no. 8, pp. 1594–1598, 1987.

[34] T. Hansen, L. Pracejus, and K. R. Gegenfurtner, "Color perception in the intermediate periphery of the visual field," *Journal of Vision*, vol. 9, no. 4, pp. 26–26, 2009.

[35] W. H. Warren and K. J. Kurtz, "The role of central and peripheral vision in perceiving the direction of self-motion," *Perception & Psychophysics*, vol. 51, no. 5, pp. 443–454, 1992.

[36] A. Traschütz, W. Zinke, and D. Wegener, "Speed change detection in foveal and peripheral vision," *Vision Research*, vol. 72, pp. 1–13, 2012.

[37] J. Baldwin, A. Burleigh, R. Pepperell, and N. Ruta, "The perceived size and shape of objects in peripheral vision," *i-Perception*, vol. 7, no. 4, p. 2041669516661900, 2016.

[38] M. Boucart, Q. Lenoble, J. Quettelart, S. Szaffarczyk, P. Despretz, and S. J. Thorpe, "Finding faces, animals, and vehicles in far peripheral vision," *Journal of Vision*, vol. 16, no. 2, pp. 10–10, 2016.

[39] H. Strasburger, I. Rentschler, and M. Jüttner, "Peripheral vision and pattern recognition: A review," *Journal of Vision*, vol. 11, no. 5, pp. 13–13, 2011.

[40] D. Whitney and D. M. Levi, "Visual crowding: A fundamental limit on conscious perception and object recognition," *Trends in Cognitive Sciences*, vol. 15, no. 4, pp. 160–168, 2011.

[41] R. Rosenholtz, "Capabilities and limitations of peripheral vision," *Annual Review of Vision Science*, vol. 2, pp. 437–457, 2016.

[42] L. C. Loschky, A. Nuthmann, F. C. Fortenbaugh, and D. M. Levi, "Scene perception from central to peripheral vision," *Journal of Vision*, vol. 17, no. 1, pp. 6–6, 2017.

[43] R. R. Mourant and T. H. Rockwell, "Strategies of visual search by novice and experienced drivers," *Human Factors*, vol. 14, no. 4, pp. 325–335, 1972.

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

[45] H. Summala, T. Nieminen, and M. Punto, "Maintaining lane position with peripheral vision during in-vehicle tasks," *Human Factors*, vol. 38, no. 3, pp. 442–451, 1996.

[46] Crundall, David and Underwood, Geoffrey and Chapman, Peter, "Attending to the peripheral world while driving," *Applied Cognitive Psychology*, vol. 16, no. 4, pp. 459–475, 2002.

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

[48] 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, 2019.

[49] B. Wolfe, B. Seppelt, B. Mehler, B. Reimer, and R. Rosenholtz, "Rapid holistic perception and evasion of road hazards," *Journal of Experimental Psychology: General*, vol. 149, no. 3, p. 490, 2020.

[50] M. Costa, L. Bonetti, V. Vignali, C. Lantieri, and A. Simone, "The role of peripheral vision in vertical road sign identification and discrimination," *Ergonomics*, vol. 61, no. 12, pp. 1619–1634, 2018.

[51] H. Summala, D. Lamble, and M. Laakso, "Driving experience and perception of the lead car's braking when looking at in-car targets," *Accident Analysis & Prevention*, vol. 30, no. 4, pp. 401–407, 1998.

[52] M. A. Recarte and L. M. Nunes, "Mental workload while driving: effects on visual search, discrimination, and decision making," *Journal of Experimental Psychology: Applied*, vol. 9, no. 2, p. 119, 2003.

[53] A. F. Sanders, "Some aspects of the selective process in the functional visual field," *Ergonomics*, vol. 13, no. 1, pp. 101–117, 1970.

[54] C. Owsley and G. McGwin Jr, "Vision and driving," *Vision Research*, vol. 50, no. 23, pp. 2348–2361, 2010.

[55] B. Thorslund and N. Strand, "Vision measurability and its impact on safe driving: A literature review," *Scandinavian Journal of Optometry and Visual Science*, vol. 9, no. 1, pp. 1–9, 2016.

[56] K. Ball and C. Owsley, "Identifying correlates of accident involvement for the older driver," *Human Factors*, vol. 33, no. 5, pp. 583–595, 1991.

[57] C. Owsley, K. Ball, G. McGwin Jr, M. E. Sloane, D. L. Roenker, M. F. White, and E. T. Overley, "Visual processing impairment and risk of motor vehicle crash among older adults," *Journal of the American Medical Association*, vol. 279, no. 14, pp. 1083–1088, 1998.

[58] O. J. Clay, V. G. Wadley, J. D. Edwards, D. L. Roth, D. L. Roenker, and K. K. Ball, "Cumulative meta-analysis of the relationship between useful field of view and driving performance in older adults: Current and future implications," *Optometry and Vision Science*, vol. 82, no. 8, pp. 724–731, 2005.

[59] K. K. Ball, D. L. Roenker, V. G. Wadley, J. D. Edwards, D. L. Roth, G. McGwin Jr, R. Raleigh, J. J. Joyce, G. M. Cissell, and T. Dube, "Can high-risk older drivers be identified through performance-based measures in a department of motor vehicles setting?" *Journal of the American Geriatrics Society*, vol. 54, no. 1, pp. 77–84, 2006.

[60] A. Bélanger, S. Gagnon, and S. Yamin, "Capturing the serial nature of older drivers' responses towards challenging events: A simulator study," *Accident Analysis & Prevention*, vol. 42, no. 3, pp. 809–817, 2010.

[61] A. Cuenen, E. M. Jongen, T. Brijs, K. Brijs, M. Lutin, K. Van Vlierden, and G. Wets, "Does attention capacity moderate the effect of driver distraction in older drivers?" *Accident Analysis & Prevention*, vol. 77, pp. 12–20, 2015.

[62] K. K. Ball, B. L. Beard, D. L. Roenker, R. L. Miller, and D. S. Griggs, "Age and visual search: Expanding the useful field of view," *Journal of Optical Society of America*, vol. 5, no. 12, pp. 2210–2219, 1988.

[63] G. Underwood, P. Chapman, Z. Berger, and D. Crundall, "Driving experience, attentional focusing, and the recall of recently inspected events," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 6, no. 4, pp. 289–304, 2003.

[64] Z. Bian, J. J. Kang, and G. J. Andersen, "Changes in extent of spatial attention with increased workload in dual-task driving," *Transportation Research Record*, vol. 2185, no. 1, pp. 8–14, 2010.

[65] J. Rogé, T. Pébayle, E. Lambilliotte, F. Spitzenstetter, D. Giselbrecht, and A. Muzet, "Influence of age, speed and duration of monotonous driving task in traffic on the driver's useful visual field," *Vision Research*, vol. 44, no. 23, pp. 2737–2744, 2004.

[66] M. S. Jensen, R. Yao, W. N. Street, and D. J. Simons, "Change blindness and inattentional blindness," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 2, no. 5, pp. 529–546, 2011.

[67] R. Young, "Cognitive distraction while driving: A critical review of definitions and prevalence in crashes," *SAE International Journal of Passenger Cars-Electronic and Electrical Systems*, vol. 5, no. 1, pp. 326–342, 2012.

[68] B. Metz and H.-P. Krüger, "Do supplementary signs distract the driver?" *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 23, pp. 1–14, 2014.

[69] I. M. Harms and K. A. Brookhuis, "Dynamic traffic management on a familiar road: Failing to detect changes in variable speed limits," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 38, pp. 37–46, 2016.

[70] 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 Analysis and Prevention*, vol. 100, pp. 111–122, 2017.

[71] S. G. Charlton and N. J. Starkey, "Driving on familiar roads: Automaticity and inattention blindness," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 19, pp. 121–133, 2013.

[72] 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 Analysis & Prevention*, vol. 42, no. 6, pp. 1822–1830, 2010.

[73] D. Crundall, K. Humphrey, and D. Clarke, "Perception and appraisal of approaching motorcycles at junctions," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 11, no. 3, pp. 159–167, 2008.

[74] G. Underwood, K. Humphrey, and E. Van Loon, "Decisions about objects in real-world scenes are influenced by visual saliency before and during their inspection," *Vision Research*, vol. 51, no. 18, pp. 2031–2038, 2011.

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

[76] M. A. Recarte and L. M. Nunes, "Effects of verbal and spatial-imagery tasks on eye fixations while driving." *Journal of Experimental Psychology: Applied*, vol. 6, no. 1, p. 31, 2000.

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

[78] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[79] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 20, no. 11, pp. 1254–1259, 1998.

[80] A. L. Yarbus, "Eye movements during perception of complex objects," in *Eye movements and vision*. Springer, 1967, ch. 8, pp. 171–211.

[81] J. K. Tsotsos, I. Kotseruba, A. Rasouli, and M. D. Solbach, "Visual attention and its intimate links to spatial cognition," *Cognitive Processing*, vol. 19, pp. 121–130, 2018.

[82] J. K. Tsotsos, O. Abid, I. Kotseruba, and M. D. Solbach, "On the control of attentional processes in vision," *Cortex*, vol. 137, pp. 305–329, 2021.

[83] E. Awh, A. V. Belopolsky, and J. Theeuwes, "Top-down versus bottom-up attentional control: A failed theoretical dichotomy," *Trends in Cognitive Sciences*, vol. 16, no. 8, pp. 437–443, 2012.

[84] B. W. Tatler, M. M. Hayhoe, M. F. Land, and D. H. Ballard, "Eye guidance in natural vision: Reinterpreting salience," *Journal of Vision*, vol. 11, no. 5, pp. 5–5, 2011.

[85] J. B. Hutchinson and N. B. Turk-Browne, "Memory-guided attention: Control from multiple memory systems," *Trends in Cognitive Sciences*, vol. 16, no. 12, pp. 576–579, 2012.

[86] M. H. Martens and M. R. Fox, "Do familiarity and expectations change perception? drivers' glances and response to changes," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 10, no. 6, pp. 476–492, 2007.

[87] A. Borowsky, D. Shinar, and Y. Parmet, "The relation between driving experience and recognition of road signs relative to their locations," *Human Factors*, vol. 50, no. 2, pp. 173–182, 2008.

[88] 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," *Journal of Vision*, vol. 12, no. 13, pp. 19–19, 2012.

[89] J. S. McCarley, K. S. Steelman, and W. J. Horrey, "The View from the Driver's Seat: What Good Is Salience?" *Applied Cognitive Psychology*, vol. 28, no. 1, pp. 47–54, 2014.

[90] M. Dozza, "What factors influence drivers' response time for evasive maneuvers in real traffic?" *Accident Analysis & Prevention*, vol. 58, pp. 299–308, 2013.

[91] A. H. Young, A. K. Mackenzie, R. L. Davies, and D. Crundall, "Familiarity breeds contempt for the road ahead: The real-world effects of route repetition on visual attention in an expert driver," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 57, pp. 4–9, 2018.

[92] J. Kuo, M. G. Lenné, M. Mulhall, T. Sletten, C. Anderson, M. Howard, S. Rajaratnam, M. Magee, and A. Collins, "Continuous monitoring of visual distraction and drowsiness in shift-workers during naturalistic driving," *Safety Science*, vol. 119, pp. 112–116, 2019.

[93] T. Dukic, C. Ahlström, C. Patten, C. Kettwich, and K. Kircher, "Effects of electronic billboards on driver distraction," *Traffic Injury Prevention*, vol. 14, no. 5, pp. 469–476, 2013.

[94] M. Costa, A. Simone, V. Vignali, C. Lantieri, A. Bucchi, and G. Dondi, "Looking behavior for vertical road signs," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 23, pp. 147–155, 2014.

[95] T. W. Victor, E. Tivesten, P. Gustavsson, J. Johansson, F. Sangberg, and M. Ljung Aust, "Automation expectation mismatch: Incorrect prediction despite eyes on threat and hands on wheel," *Human Factors*, vol. 60, no. 8, pp. 1095–1116, 2018.

[96] O. Lappi, E. Lehtonen, J. Pekkanen, and T. Itkonen, "Beyond the tangent point: Gaze targets in naturalistic driving," *Journal of Vision*, vol. 13, no. 13, pp. 11–11, 2013.

[97] X. Fan, F. Wang, D. Song, Y. Lu, and J. Liu, "GazMon: Eye Gazing Enabled Driving Behavior Monitoring and Prediction," *IEEE Transactions on Mobile Computing*, vol. 20, no. 4, pp. 1420–1433, 2019.

[98] Z. Lu, B. Zhang, A. Feldhütter, R. Happee, M. Martens, and J. De Winter, "Beyond mere take-over requests: The effects of monitoring requests on driver attention, take-over perfor-

mance, and acceptance," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 63, pp. 22–37, 2019.

[99] D. B. Kaber, Y. Liang, Y. Zhang, M. L. Rogers, and S. Gangakhedkar, "Driver performance effects of simultaneous visual and cognitive distraction and adaptation behavior," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 15, no. 5, pp. 491–501, 2012.

[100] S. W. Savage, D. D. Potter, and B. W. Tatler, "Does preoccupation impair hazard perception? A simultaneous EEG and eye tracking study," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 17, pp. 52–62, 2013.

[101] 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*, D. L. Fisher, M. Rizzo, J. Caird, and J. D. Lee, Eds. CRC Press, 2011, pp. 1–16.

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

[103] C. F. Alberti, A. Shahar, and D. Crundall, "Are experienced drivers more likely than novice drivers to benefit from driving simulations with a wide field of view?" *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 27, pp. 124–132, 2014.

[104] G. Pai 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?" *Transportation Research Record*, vol. 2673, no. 10, pp. 455–464, 2019.

[105] M. R. Ciceri and D. Ruscio, "Does driving experience in video games count? Hazard anticipation and visual exploration of male gamers as function of driving experience," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 22, pp. 76–85, 2014.

[106] 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 *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 63, no. 1, 2019, pp. 2099–2103.

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

[108] A. K. Mackenzie and J. M. Harris, "Eye movements and hazard perception in active and passive driving," *Visual Cognition*, vol. 23, no. 6, pp. 736–757, 2015.

[109] X. Li, R. Schroeter, A. Rakotonirainy, J. Kuo, and M. G. Lenné, "Effects of different non-driving-related-task display modes on drivers' eye-movement patterns during take-over in an automated vehicle," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 70, pp. 135–148, 2020.

[110] C. D. Fitzpatrick, S. Samuel, and M. A. Knodler Jr, "Evaluating the effect of vegetation and clear zone width on driver behavior using a driving simulator," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 42, pp. 80–89, 2016.

[111] T. A. Dingus, S. G. Klauer, V. L. Neale, A. Petersen, S. E. Lee, J. Sudweeks, M. A. Perez, J. Hankey, D. Ramsey, S. Gupta, C. Bucher, Z. R. Doerzaph, J. Jermeland, and R. R. Knipling, "The 100-Car Naturalistic Driving Study. Phase II—Results of the 100-Car Field Experiment," NHTSA, Tech. Rep. DOT HS 810 593, 2006.

[112] B. G. Simons-Morton, F. Guo, S. G. Klauer, J. P. Ehsani, and A. K. Pradhan, "Keep your eyes on the road: Young driver crash risk increases according to duration of distraction," *Journal of Adolescent Health*, vol. 54, no. 5, pp. S61–S67, 2014.

[113] S. Vora, A. Rangesh, and M. M. Trivedi, "Driver gaze zone estimation using convolutional neural networks: A general framework and ablative analysis," *IEEE Transactions on Intelligent Vehicles*, vol. 3, no. 3, pp. 254–265, 2018.

[114] Q. C. Sun, J. C. Xia, J. He, J. Foster, T. Falkmer, and H. Lee, "Towards unpacking older drivers' visual-motor coordination: A gaze-based integrated driving assessment," *Accident Analysis & Prevention*, vol. 113, pp. 85–96, 2018.

[115] A. H. Jamson, N. Merat, O. M. Carsten, and F. C. Lai, "Behavioural changes in drivers experiencing highly-automated vehicle control in varying traffic conditions," *Transportation Research Part C: Emerging Technologies*, vol. 30, pp. 116–125, 2013.

[116] Tobii, "Eye tracker data quality report: Accuracy, precision and detected gaze under optimal conditions—controlled environment," Cached version available online at: https://web.archive.org/web/20220628220853/https://www.tobiipro.com/siteassets/tobii-pro/accuracy-and-precision-tests/tobii-pro-glasses-2-accuracy-and-precision-test-report.pdf, accessed: 2024-12-01.

[117] B. Vasli, S. Martin, and M. M. Trivedi, "On driver gaze estimation: Explorations and fusion of geometric and data driven approaches," in *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2016.

[118] E. Tivesten and M. Dozza, "Driving context and visual-manual phone tasks influence glance behavior in naturalistic driving," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 26, pp. 258–272, 2014.

[119] M. Muñoz, B. Reimer, J. Lee, B. Mehler, and L. Fridman, "Distinguishing patterns in drivers' visual attention allocation using Hidden Markov Models," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 43, pp. 90–103, 2016.

[120] Y. Liang, J. D. Lee, and W. J. Horrey, "A looming crisis: the distribution of off-road glance duration in moments leading up to crashes/near-crashes in naturalistic driving," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 58, no. 1, 2014, pp. 2102–2106.

[121] A. Morando, T. Victor, and M. Dozza, "A reference model for driver attention in automation: Glance behavior changes during lateral and longitudinal assistance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 8, pp. 2999–3009, 2018.

[122] B. D. Seppelt, S. Seaman, J. Lee, L. S. Angell, B. Mehler, and B. Reimer, "Glass half-full: On-road glance metrics differentiate crashes from near-crashes in the 100-car data," *Accident Analysis & Prevention*, vol. 107, pp. 48–62, 2017.

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

[124] J. M. Owens, S. B. McLaughlin, and J. Sudweeks, "Driver performance while text messaging using handheld and in-vehicle systems," *Accident Analysis & Prevention*, vol. 43, no. 3, pp. 939–947, 2011.

[125] R. Kim, R. Rauschenberger, G. Heckman, D. Young, and R. Lange, "Efficacy and usage patterns for three types of rearview camera displays during backing up," SAE International, Tech. Rep. 2012-01-0287, 2012.

[126] 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 Transactions on Intelligent Transportation Systems*, vol. 12, no. 1, pp. 254–267, 2011.

[127] 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 *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 54, no. 24, 2010, pp. 2052–2055.

[128] G. K. Kountouriotis and N. Merat, "Leading to distraction: Driver distraction, lead car, and road environment," *Accident Analysis & Prevention*, vol. 89, pp. 22–30, 2016.

[129] N. Merat, A. H. Jamson, F. C. Lai, M. Daly, and O. M. Carsten, "Transition to manual: Driver behaviour when resuming control from a highly automated vehicle," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 27, pp. 274–282, 2014.

[130] E. Lehtonen, O. Lappi, H. Kotkanen, and H. Summala, "Look-ahead fixations in curve driving," *Ergonomics*, vol. 56, no. 1, pp. 34–44, 2013.

[131] J. R. Clark, N. A. Stanton, and K. M. Revell, "Directability, eye-gaze, and the usage of visual displays during an automated vehicle handover task," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 67, pp. 29–42, 2019.

[132] D. Stavrinos, P. R. Mosley, S. M. Wittig, H. D. Johnson, J. S. Decker, V. P. Sisiopiku, and S. C. Welburn, "Visual behavior differences in drivers across the lifespan: A digital billboard simulator study," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 41, pp. 19–28, 2016.

[133] T. Itkonen, J. Pekkanen, and O. Lappi, "Driver gaze behavior is different in normal curve driving and when looking at the tangent point," *PloS one*, vol. 10, no. 8, p. e0135505, 2015.

[134] S. G. Klauer, F. Guo, J. Sudweeks, and T. A. Dingus, "An analysis of driver inattention using a case-crossover approach on 100-Car data: Final report," National Highway Traffic Safety Administration (NHTSA), Tech. Rep. DOT HS 811 334, 2010.

[135] D. G. Kidd, B. Reimer, J. Dobres, and B. Mehler, "Changes in driver glance behavior when using a system that automates steering to perform a low-speed parallel parking maneuver," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 58, pp. 629–639, 2018.

[136] S. Liversedge, I. Gilchrist, and S. Everling, *The Oxford handbook of eye movements*. Oxford University Press, 2011.

[137] W. W. Wierwille, S. Wreggit, C. Kirn, L. Ellsworth, and R. Fairbanks, "Research on vehicle-based driver status/performance monitoring; Development, validation, and refinement of algorithms for detection of driver drowsiness," National Highway Traffic Safety Administration (NHTSA), Tech. Rep. DOT HS 808 247, 1994.

[138] I. T. Hooge, R. S. Hessels, and M. Nyström, "Do pupil-based binocular video eye trackers reliably measure vergence?" *Vision Research*, vol. 156, pp. 1–9, 2019.

[139] A. S. Le, T. Suzuki, and H. Aoki, "Evaluating driver cognitive distraction by eye tracking: From simulator to driving," *Transportation Research Interdisciplinary Perspectives*, vol. 4, p. 100087, 2020.

[140] M. Poletti and M. Rucci, "A compact field guide to the study of microsaccades: Challenges and functions," *Vision Research*, vol. 118, pp. 83–97, 2016.

[141] M. R. Romoser, A. Pollatsek, D. L. Fisher, and C. C. Williams, "Comparing the glance patterns of older versus younger experienced drivers: Scanning for hazards while approaching and entering the intersection," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 16, pp. 104–116, 2013.

[142] T. Louw, J. Kuo, R. Romano, V. Radhakrishnan, M. G. Lenné, and N. Merat, "Engaging in NDRTs affects drivers' responses and glance patterns after silent automation failures," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 62, pp. 870–882, 2019.

[143] Q. Sun, j. Xia, T. Falkmer, and H. Lee, "Investigating the spatial pattern of older drivers eye fixation behaviour and associations with their visual capacity," *Journal of Eye Movement Research*, vol. 9, no. 6, 2016.

[144] E. Lehtonen, O. Lappi, I. Koirikivi, and H. Summala, "Effect of driving experience on anticipatory look-ahead fixations in real curve driving," *Accident Analysis & Prevention*, vol. 70, pp. 195–208, 2014.

[145] "ISO 15007. Road vehicles—Measurement of driver visual behaviour with respect to transport information and control systems," 2020.

[146] J. A. McClafferty and J. M. Hankey, "Comparing three methods of eye-glance transition coding," National Surface Transportation Safety Center for Excellence (NSTSCE), Tech. Rep. 16-UT-045, 2016.

[147] A.-K. Kraft, F. Naujoks, J. Wörle, and A. Neukum, "The impact of an in-vehicle display on glance distribution in partially automated driving in an on-road experiment," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 52, pp. 40–50, 2018.

[148] C. Ahlström and K. Kircher, "Changes in glance behaviour when using a visual eco-driving system–A field study," *Applied Ergonomics*, vol. 58, pp. 414–423, 2017.

[149] L. Huestegge and A. Böckler, "Out of the corner of the driver's eye: Peripheral processing of hazards in static traffic scenes," *Journal of Vision*, vol. 16, no. 2, pp. 11–11, 2016.

[150] M. Hashash, M. Abou Zeid, and N. M. Moacdieh, "Social media browsing while driving: effects on driver performance and attention allocation," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 63, pp. 67–82, 2019.

[151] A. Tawari, K. H. Chen, and M. M. Trivedi, "Where is the driver looking: Analysis of head, eye and iris for robust gaze zone estimation," in *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2014.

[152] T. Pech, P. Lindner, and G. Wanielik, "Head tracking based glance area estimation for driver behaviour modelling during lane change execution," in *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2014.

[153] Z. Wang, R. Zheng, T. Kaizuka, and K. Nakano, "Relationship between gaze behavior and steering performance for driver–automation shared control: A driving simulator study," *IEEE Transactions on Intelligent Vehicles*, vol. 4, no. 1, pp. 154–166, 2018.

[154] K. Kircher and C. Ahlström, "Evaluation of methods for the assessment of attention while driving," *Accident Analysis & Prevention*, vol. 114, pp. 40–47, 2018.

[155] K. L. Young, C. M. Rudin-Brown, C. Patten, R. Ceci, and M. G. Lenné, "Effects of phone type on driving and eye glance behaviour while text-messaging," *Safety Science*, vol. 68, pp. 47–54, 2014.

[156] B. Wortelen, M. Baumann, and A. Lüdtke, "Dynamic simulation and prediction of drivers' attention distribution," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 21, pp. 278–294, 2013.

[157] F. Vicente, Z. Huang, X. Xiong, F. De la Torre, W. Zhang, and D. Levi, "Driver gaze tracking and eyes off the road detection system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 2014–2027, 2015.

[158] 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 Analysis & Prevention*, vol. 88, pp. 88–96, 2016.

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

[160] S. S.-Y. Lee, J. M. Wood, and A. A. Black, "Blur, eye movements and performance on a driving visual recognition slide test," *Ophthalmic and Physiological Optics*, vol. 35, no. 5, pp. 522–529, 2015.

[161] E. M. Van Loon, F. Khashawi, and G. Underwood, "Visual strategies used for time-to-arrival judgments in driving," *Perception*, vol. 39, no. 9, pp. 1216–1229, 2010.

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

[163] Y. Yamani, P. Bıçaksız, D. B. Palmer, J. M. Cronauer, and S. Samuel, "Following expert's eyes: Evaluation of the effectiveness of a gaze-based training intervention on young drivers' latent hazard anticipation skills," in *Driving Assessment Conference*, 2017, pp. 129–135.

[164] W. Chen, X. Zhuang, Z. Cui, and G. Ma, "Drivers' recognition of pedestrian road-crossing intentions: Performance and process," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 64, pp. 552–564, 2019.

[165] D. Crundall, E. Crundall, D. Clarke, and A. Shahar, "Why do car drivers fail to give way to motorcycles at t-junctions?" *Accident Analysis & Prevention*, vol. 44, no. 1, pp. 88–96, 2012.

[166] 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 Analysis & Prevention*, vol. 42, no. 3, pp. 827–834, 2010.

[167] T. M. Garrison and C. C. Williams, "Impact of relevance and distraction on driving performance and visual attention in a simulated driving environment," *Applied Cognitive Psychology*, vol. 27, no. 3, pp. 396–405, 2013.

[168] 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," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 35, pp. 152–169, 2015.

[169] G. F. Briggs, G. J. Hole, and M. F. Land, "Emotionally involving telephone conversations lead to driver error and visual tunnelling," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 14, no. 4, pp. 313–323, 2011.

[170] S. S.-Y. Lee, A. A. Black, P. Lacherez, and J. M. Wood, "Eye movements and road hazard detection: effects of blur and distractors," *Optometry and Vision Science*, vol. 93, no. 9, pp. 1137–1146, 2016.

[171] X. Yan, X. Zhang, Y. Zhang, X. Li, and Z. Yang, "Changes in drivers' visual performance during the collision avoidance process as a function of different field of views at intersections," *PLoS one*, vol. 11, no. 10, p. e0164101, 2016.

[172] G. Li, Y. Wang, F. Zhu, X. Sui, N. Wang, X. Qu, and P. Green, "Drivers' visual scanning behavior at signalized and unsignalized intersections: A naturalistic driving study in China," *Journal of Safety Research*, vol. 71, pp. 219–229, 2019.

[173] X. Li, A. Rakotonirainy, X. Yan, and Y. Zhang, "Driver's visual performance in rear-end collision avoidance process under the influence of cell phone use," *Transportation Research Record*, vol. 2672, no. 37, pp. 55–63, 2018.

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

[175] M. Niezgoda, A. Tarnowski, M. Kruszewski, and T. Kamiński, "Towards testing auditory–vocal interfaces and detecting distraction while driving: A comparison of eye-movement measures in the assessment of cognitive workload," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 32, pp. 23–34, 2015.

[176] B. A. Shiferaw, L. A. Downey, J. Westlake, B. Stevens, S. M. Rajaratnam, D. J. Berlowitz, P. Swann, and M. E. Howard, "Stationary gaze entropy predicts lane departure events in sleep-deprived drivers," *Scientific Reports*, vol. 8, no. 1, pp. 1–10, 2018.

[177] J. Jo, S. J. Lee, J. Kim, H. G. Jung, and K. R. Park, "Vision-based method for detecting driver drowsiness and distraction in driver monitoring system," *Optical Engineering*, vol. 50, no. 12, p. 127202, 2011.

[178] L. Jin, Q. Niu, Y. Jiang, H. Xian, Y. Qin, and M. Xu, "Driver sleepiness detection system based on eye movements variables," *Advances in Mechanical Engineering*, vol. 5, p. 648431, 2013.

[179] C. Zhang, X. Wu, X. Zheng, and S. Yu, "Driver drowsiness detection using multi-channel second order blind identifications," *IEEE Access*, vol. 7, pp. 11 829–11 843, 2019.

[180] W. Deng and R. Wu, "Real-time driver-drowsiness detection system using facial features," *IEEE Access*, vol. 7, pp. 118 727–118 738, 2019.

[181] D. F. Dinges and R. Grace, "PERCLOS: A valid psychophysiological measure of alertness as assessed by psychomotor vigilance. FHWA-MCRT-98-006," 1998.

[182] F. Friedrichs and B. Yang, "Camera-based drowsiness reference for driver state classification under real driving conditions," in *IEEE Intelligent Vehicles Symposium (IV)*, 2010.

[183] Y. Wang, S. Bao, W. Du, Z. Ye, and J. R. Sayer, "Examining drivers' eye glance patterns during distracted driving: Insights from scanning randomness and glance transition matrix," *Journal of Safety Research*, vol. 63, pp. 149–155, 2017.

[184] S. A. Birrell and M. Fowkes, "Glance behaviours when using an in-vehicle smart driving aid: A real-world, on-road driving study," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 22, pp. 113–125, 2014.

[185] L. Zhang, J. Kong, B. Cui, and T. Fu, "Safety Effects of Freeway Roadside Electronic Billboards on Visual Properties of Drivers: Insights from Field Experiments," *Journal of Transportation Engineering, Part A: Systems*, vol. 146, no. 2, p. 04019071, 2020.

[186] H. Scott, L. Hall, D. Litchfield, and D. Westwood, "Visual information search in simulated junction negotiation: Gaze transitions of young novice, young experienced and older experienced drivers," *Journal of Safety Research*, vol. 45, pp. 111–116, 2013.

[187] S. Lemonnier, R. Brémond, and T. Baccino, "Gaze behavior when approaching an intersection: Dwell time distribution and comparison with a quantitative prediction," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 35, pp. 60–74, 2015.

[188] D. Crundall, P. Chapman, S. Trawley, L. Collins, E. Van Loon, B. Andrews, and G. Underwood, "Some hazards are more attractive than others: Drivers of varying experience respond differently to different types of hazard," *Accident Analysis & Prevention*, vol. 45, pp. 600–609, 2012.

[189] G. Wood, G. Hartley, P. Furley, and M. Wilson, "Working memory capacity, visual attention and hazard perception in driving," *Journal of Applied Research in Memory and Cognition*, vol. 5, no. 4, pp. 454–462, 2016.

[190] 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," *Human Factors*, vol. 54, no. 2, pp. 250–263, 2012.

[191] J. Y. Lee, M. C. Gibson, and J. D. Lee, "Error recovery in multitasking while driving," in *ACM Conference on Human Factors in Computing Systems (CHI)*, 2016.

[192] M. Smith, J. L. Gabbard, and C. Conley, "Head-up vs. head-down displays: examining traditional methods of display assessment while driving," in *International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI)*, 2016.

[193] 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 Analysis & Prevention*, vol. 92, pp. 230–239, 2016.

[194] B. Reimer, B. Mehler, and B. Donmez, "A study of young adults examining phone dialing while driving using a touchscreen vs. a button style flip-phone," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 23, pp. 57–68, 2014.

[195] F. Schieber, K. Limrick, R. McCall, and A. Beck, "Evaluation of the visual demands of digital billboards using a hybrid driving simulator," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 58, no. 1, 2014, pp. 2214–2218.

[196] J. M. Cooper, N. Medeiros-Ward, and D. L. Strayer, "The impact of eye movements and cognitive workload on lateral position variability in driving," *Human Factors*, vol. 55, no. 5, pp. 1001–1014, 2013.

[197] C. Lemercier, C. Pecher, G. Berthié, B. Valery, V. Vidal, P.-V. Paubel, M. Cour, A. Fort, C. Galéra, C. Gabaude, E. Lagarde, and B. Maury, "Inattention behind the wheel: How factual internal thoughts impact attentional control while driving," *Safety Science*, vol. 62, pp. 279–285, 2014.

[198] J. He, J. S. McCarley, and A. F. Kramer, "Lane keeping under cognitive load: performance changes and mechanisms," *Human Factors*, vol. 56, no. 2, pp. 414–426, 2014.

[199] M. Zahabi and D. Kaber, "Effect of police mobile computer terminal interface design on officer driving distraction," *Applied Ergonomics*, vol. 67, pp. 26–38, 2018.

[200] P. Van Leeuwen, R. Happee, and J. De Winter, "Vertical field of view restriction in driver training: A simulator-based evaluation," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 24, pp. 169–182, 2014.

[201] 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," *Human Factors*, vol. 54, no. 3, pp. 454–468, 2012.

[202] M. Zahabi, P. Machado, M. Y. Lau, Y. Deng, C. Pankok Jr, J. Hummer, W. Rasdorf, and D. B. Kaber, "Driver performance and attention allocation in use of logo signs on freeway exit ramps," *Applied Ergonomics*, vol. 65, pp. 70–80, 2017.

[203] M. Zahabi, P. Machado, M. Lau, Y. Deng, C. Pankok, J. Hummer, W. Rasdorf, and D. Kaber, "Effect of Driver Age and Distance Guide Sign Format on Driver Attention Allocation and Performance," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 62, no. 1, 2018, pp. 1903–1907.

[204] J. Grippenkoven and S. Dietsch, "Gaze direction and driving behavior of drivers at level crossings," *Journal of Transportation Safety & Security*, vol. 8, no. sup1, pp. 4–18, 2016.

[205] A. Morando, T. Victor, and M. Dozza, "Drivers anticipate lead-vehicle conflicts during automated longitudinal control: Sensory cues capture driver attention and promote appropriate and timely responses," *Accident Analysis & Prevention*, vol. 97, pp. 206–219, 2016.

[206] E. Tivesten, A. Morando, and T. Victor, "The timecourse of driver visual attention in naturalistic driving with adaptive cruise control and forward collision warning," in *International Driver Distraction and Inattention Conference*, 2015.

[207] O. Lappi, "Future path and tangent point models in the visual control of locomotion in curve driving," *Journal of Vision*, vol. 14, no. 12, pp. 21–21, 2014.

[208] 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 *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2018.

[209] E. Bozkir, D. Geisler, and E. Kasneci, "Assessment of driver attention during a safety critical situation in VR to generate VR-based training," in *ACM Symposium on Applied Perception (SAP)*, 2019.

[210] S. S. Borojeni, L. Chuang, W. Heuten, and S. Boll, "Assisting drivers with ambient take-over requests in highly automated driving," in *International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI)*, 2016.

[211] 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 Analysis & Prevention*, vol. 110, pp. 29–37, 2018.

[212] M. L. Jackson, S. Raj, R. J. Croft, A. C. Hayley, L. A. Downey, G. A. Kennedy, and M. E. Howard, "Slow eyelid closure as a measure of driver drowsiness and its relationship to performance," *Traffic Injury Prevention*, vol. 17, no. 3, pp. 251–257, 2016.

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

[214] J. P. Wann and D. K. Swapp, "Why you should look where you are going," *Nature Neuroscience*, vol. 3, no. 7, pp. 647–648, 2000.

[215] F. Mars and J. Navarro, "Where we look when we drive with or without active steering wheel control," *PloS one*, vol. 7, no. 8, p. e43858, 2012.

[216] P. M. Van Leeuwen, S. de Groot, R. Happee, and J. C. de Winter, "Differences between racing and non-racing drivers: A simulator study using eye-tracking," *PLoS one*, vol. 12, no. 11, p. e0186871, 2017.

[217] D. G. Kidd and A. T. McCartt, "Differences in glance behavior between drivers using a rearview camera, parking sensor system, both technologies, or no technology during low-speed parking maneuvers," *Accident Analysis & Prevention*, vol. 87, pp. 92–101, 2016.

[218] L. Pomarjanschi, M. Dorr, and E. Barth, "Gaze guidance reduces the number of collisions with pedestrians in a driving simulator," *ACM Transactions on Interactive Intelligent Systems*, vol. 1, no. 2, pp. 1–14, 2012.

[219] A. Borowsky, D. Shinar, and T. Oron-Gilad, "Age, skill, and hazard perception in driving," *Accident Analysis & Prevention*, vol. 42, no. 4, pp. 1240–1249, 2010.

[220] L. Lorenz, P. Kerschbaum, and J. Schumann, "Designing take over scenarios for automated driving: How does augmented reality support the driver to get back into the loop?" in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 58, no. 1, 2014, pp. 1681–1685.

[221] P. Stahl, B. Donmez, and G. A. Jamieson, "Eye glances towards conflict-relevant cues: the roles of anticipatory competence and driver experience," *Accident Analysis & Prevention*, vol. 132, p. 105255, 2019.

[222] B. Metz, N. Schömig, and H.-P. Krüger, "Attention during visual secondary tasks in driving: Adaptation to the demands of the driving task," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 14, no. 5, pp. 369–380, 2011.

[223] Y. Zhang, E. Harris, M. Rogers, D. Kaber, J. Hummer, W. Rasdorf, and J. Hu, "Driver distraction and performance effects of highway logo sign design," *Applied Ergonomics*, vol. 44, no. 3, pp. 472–479, 2013.

[224] V. Beanland and R. A. Wynne, "Does familiarity breed competence or contempt? Effects of driver experience, road type and familiarity on hazard perception," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 63, no. 1, 2019, pp. 2006–2010.

[225] A. Borowsky and T. Oron-Gilad, "Exploring the effects of driving experience on hazard awareness and risk perception via real-time hazard identification, hazard classification, and rating tasks," *Accident Analysis & Prevention*, vol. 59, pp. 548–565, 2013.

[226] Y. Yamani, W. J. Horrey, Y. Liang, and D. L. Fisher, "Age-related differences in vehicle control and eye movement patterns at intersections: Older and middle-aged drivers," *PLoS one*, vol. 11, no. 10, p. e0164124, 2016.

[227] S. Samuel and D. L. Fisher, "Evaluation of the minimum forward roadway glance duration," *Transportation Research Record*, vol. 2518, no. 1, pp. 9–17, 2015.

[228] J. Navarro, F. Osiurak, M. Ovigue, L. Charrier, and E. Reynaud, "Highly Automated Driving Impact on Drivers' Gaze Behaviors during a Car-Following Task," *International Journal of Human-Computer Interaction*, vol. 35, no. 11, pp. 1008–1017, 2019.

[229] T. Louw and N. Merat, "Are you in the loop? Using gaze dispersion to understand driver visual attention during vehicle automation," *Transportation Research Part C: Emerging Technologies*, vol. 76, pp. 35–50, 2017.

[230] G. Divekar, A. K. Pradhan, A. Pollatsek, and D. L. Fisher, "Effect of external distractions: Behavior and vehicle control of novice and experienced drivers evaluated," *Transportation Research Record*, vol. 2321, no. 1, pp. 15–22, 2012.

[231] G. J. Andersen, R. Ni, Z. Bian, and J. Kang, "Limits of spatial attention in three-dimensional space and dual-task driving performance," *Accident Analysis & Prevention*, vol. 43, no. 1, pp. 381–390, 2011.

[232] J. G. Gaspar, N. Ward, M. B. Neider, J. Crowell, R. Carbonari, H. Kaczmarski, R. V. Ringer, A. P. Johnson, A. F. Kramer, and L. C. Loschky, "Measuring the useful field of view during simulated driving with gaze-contingent displays," *Human Factors*, vol. 58, no. 4, pp. 630–641, 2016.

[233] AAA Foundation for Traffic Safety, "2018 Traffic Safety Culture Index," https://aaafoundation.org/wp-content/uploads/2019/06/2018-TSCI-FINAL-061819_updated.pdf, accessed: 2024-03-22.

[234] European Commission, "Study on good practices for reducing road safety risks caused by road user distractions. Final report," https://road-safety.transport.ec.europa.eu/system/files/2021-07/distraction_study.pdf, accessed: 2024-03-22.

[235] L. Angell, M. Perez, and W. R. Garrott, "Explanatory Material About the Definition of a Task Used in NHTSA's Driver Distraction Guidelines, and Task Examples," NHTSA, Tech. Rep., 2013.

[236] "Distraction," https://road-safety.transport.ec.europa.eu/european-road-safety-observatory/statistics-and-analysis-archive/young-people/distraction_en, accessed: 2024-03-22.

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

[238] D. Kaber, C. Pankok Jr, B. Corbett, W. Ma, J. Hummer, and W. Rasdorf, "Driver behavior in use of guide and logo signs under distraction and complex roadway conditions," *Applied Ergonomics*, vol. 47, pp. 99–106, 2015.

[239] C. Pankok Jr, D. Kaber, W. Rasdorf, and J. Hummer, "Effects of guide and logo signs on freeway driving behavior," *Transportation Research Record*, vol. 2518, no. 1, pp. 73–78, 2015.

[240] S. Hurtado and S. Chiasson, "An eye-tracking evaluation of driver distraction and unfamiliar road signs," in *International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI)*, 2016.

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

[242] Y. Peng and L. N. Boyle, "Driver's adaptive glance behavior to in-vehicle information systems," *Accident Analysis & Prevention*, vol. 85, pp. 93–101, 2015.

[243] K. Young, M. Regan, and M. Hammer, "Driver distraction: A review of the literature," in *Distracted driving*, I. Faulks, M. Regan, M. Stevenson, J. Brown, A. Porter, and J. Irwi, Eds. Sydney, NSW: Australasian College of Road Safety, 2007, pp. 379–405.

[244] ISO, "Road vehicles—ergonomic aspects of transport information and control systems—calibration tasks for methods which assess driver demand due to the use of in-vehicle systems," Tech. Rep. ISO/TS 14198:2012(E), 2012.

[245] M. Walch, D. Lehr, M. Colley, and M. Weber, "Don't you see them? Towards gaze-based interaction adaptation for driver-vehicle cooperation," in *International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI)*, 2019.

[246] S. Benedetto, M. Pedrotti, R. Bremond, and T. Baccino, "Leftward attentional bias in a simulated driving task," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 20, pp. 147–153, 2013.

[247] B. Reimer, B. Mehler, J. Dobres, H. McAnulty, A. Mehler, D. Munger, and A. Rumpold, "Effects of an 'expert mode' voice command system on task performance, glance behavior & driver physiology," in *International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI)*, 2014.

[248] NHTSA, "Visual-manual NHTSA Driver Distraction Guidelines for In-vehicle Devices," Tech. Rep. NHTSA-2010-0053, 2012.

[249] B. Donmez, L. N. Boyle, and J. D. Lee, "Differences in off-road glances: Effects on young drivers' performance," *Journal of Transportation Engineering*, vol. 136, no. 5, pp. 403–409, 2010.

[250] T. Kujala, "Browsing the information highway while driving: three in-vehicle touch screen scrolling methods and driver distraction," *Personal and Ubiquitous Computing*, vol. 17, no. 5, pp. 815–823, 2013.

[251] 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, 2020.

[252] W. K. Kirchner, "Age differences in short-term retention of rapidly changing information," *Journal of Experimental Psychology*, vol. 55, no. 4, p. 352, 1958.

[253] P. D. Gajewski, E. Hanisch, M. Falkenstein, S. Thönes, and E. Wascher, "What does the n-back task measure as we get older? relations between working-memory measures and other cognitive functions across the lifespan," *Frontiers in Psychology*, vol. 9, p. 2208, 2018.

[254] H. Clark and J. Feng, "Age differences in the takeover of vehicle control and engagement in non-driving-related activities in simulated driving with conditional automation," *Accident Analysis & Prevention*, vol. 106, pp. 468–479, 2017.

[255] L. Yekhshatyan and J. D. Lee, "Changes in the correlation between eye and steering movements indicate driver distraction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 1, pp. 136–145, 2012.

[256] B. C. Tefft, "Rates of Motor Vehicle Crashes, Injuries, and Deaths in Relation to Driver Age, United States, 2014–2015 (Research Brief)," http://aaafoundation.org/wp-content/uploads/2017/11/CrashesInjuriesDeathsInRelationToAge2014-2015Brief.pdf, 2017.

[257] Transport Canada, "Canadian Motor Vehicle Traffic Collision Statistics: 2016," https://tc.canada.ca/en/canadian-motor-vehicle-traffic-collision-statistics-2016, accessed: 2024-03-22.

[258] U.S. Department of Transportation. Federal Highway Administration, "Highway Statistics 2017. Distribution of Licensed Drivers by Sex and Percentage in each Age Group and Relation to Population," https://www.fhwa.dot.gov/policyinformation/statistics/2017/dl20.cfm, accessed: 2024-03-22.

[259] National Police Agency (Japan), "Statistics 5-14 Number of driver's license holders by age group and gender (2017)," accessed: 2024-03-22.

[260] P. F. Waller, "The older driver," *Human Factors*, vol. 33, no. 5, pp. 499–505, 1991.

[261] S. M. Retchin and J. Anapolle, "An overview of the older driver," *Clinics in Geriatric Medicine*, vol. 9, no. 2, pp. 279–296, 1993.

[262] K. L. Ratnapradipa, C. N. Pope, A. Nwosu, and M. Zhu, "Older Driver Crash Involvement and Fatalities, by Age and Sex, 2000–2017," *Journal of Applied Gerontology*, p. 0733464820956507, 2020.

[263] E. E. Miller and L. N. Boyle, "Adaptations in attention allocation: implications for takeover in an automated vehicle," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 66, pp. 101–110, 2019.

[264] P. Urwyler, N. Gruber, R. M. Müri, M. Jäger, R. Bieri, T. Nyffeler, U. P. Mosimann, and T. Nef, "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, 2015.

[265] Insurance Institute for Highway Safety, "Fatality Facts 2018. Gender," https://www.iihs.org/topics/fatality-statistics/detail/gender, accessed: 2023-03-22.

[266] N. Rhodes and K. Pivik, "Age and gender differences in risky driving: The roles of positive affect and risk perception," *Accident Analysis & Prevention*, vol. 43, no. 3, pp. 923–931, 2011.

[267] J. Bergdahl, "Sex differences in attitudes toward driving: A survey," *Social Science Journal*, vol. 42, no. 4, pp. 595–601, 2005.

[268] D. Huo, J. Ma, and R. Chang, "Lane-changing-decision characteristics and the allocation of visual attention of drivers with an angry driving style," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 71, pp. 62–75, 2020.

[269] Y. Shinohara and Y. Nishizaki, "Where do drivers look when driving in a foreign country?" in *International Conference on Sofware Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, 2017.

[270] Y. Shinohara, R. Currano, W. Ju, and Y. Nishizaki, "Visual attention during simulated autonomous driving in the US and Japan," in *International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI)*, 2017.

[271] L. Huestegge, E.-M. Skottke, S. Anders, J. Müsseler, and G. Debus, "The development of hazard perception: Dissociation of visual orientation and hazard processing," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 13, no. 1, pp. 1–8, 2010.

[272] X. Zheng, Y. Yang, S. Easa, W. Lin, and E. Cherchi, "The effect of leftward bias on visual attention for driving tasks," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 70, pp. 199–207, 2020.

[273] D. He and B. Donmez, "The influence of visual-manual distractions on anticipatory driving," *Human Factors*, p. 0018720820938893, 2020.

[274] W. Vlakveld, M. R. Romoser, H. Mehranian, F. Diete, A. Pollatsek, and D. L. Fisher, "Do crashes and near crashes in simulator-based training enhance novice drivers' visual search for latent hazards?" *Transportation Research Record*, vol. 2265, no. 1, pp. 153–160, 2011.

[275] T. G. Taylor, K. M. Masserang, A. K. Pradhan, G. Divekar, S. Samuel, J. W. Muttart, A. Pollatsek, and D. L. Fisher, "Long term effects of hazard anticipation training on novice drivers measured on the open road," in *Proceedings of the International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design*, 2011.

[276] Statistics Canada, "Commuting to work," https://www12.statcan.gc.ca/nhs-enm/2011/assa/99-012-x/99-012-x2011003_1-eng.cfm, 2011, accessed: 2024-03-22.

[277] U.S. Department of Transportation. Bureau of Transportation Statistics, "National Household Travel Survey Daily Travel Quick Facts," https://www.bts.gov/statistical-products/surveys/national-household-travel-survey-daily-travel-quick-facts, accessed: 2024-03-22.

[278] B. R. Burdett, S. G. Charlton, and N. J. Starkey, "Not all minds wander equally: The influence of traits, states and road environment factors on self-reported mind wandering during everyday driving," *Accident Analysis & Prevention*, vol. 95, pp. 1–7, 2016.

[279] P. C. Lim, E. Sheppard, and D. Crundall, "Cross-cultural effects on drivers' hazard perception," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 21, pp. 194–206, 2013.

[280] B. A. Shiferaw, D. P. Crewther, and L. A. Downey, "Gaze entropy measures detect alcohol-induced driver impairment," *Drug and Alcohol Dependence*, vol. 204, p. 107519, 2019.

[281] B. Shiferaw, C. Stough, and L. Downey, "Drivers' visual scanning impairment under the influences of alcohol and distraction: A literature review," *Current Drug Abuse Reviews*, vol. 7, no. 3, pp. 174–182, 2014.

[282] P. Philip, P. Sagaspe, N. Moore, J. Taillard, A. Charles, C. Guilleminault, and B. Bioulac, "Fatigue, sleep restriction and driving performance," *Accident Analysis & Prevention*, vol. 37, no. 3, pp. 473–478, 2005.

[283] E. Hoddes, V. Zarcone, H. Smythe, R. Phillips, and W. C. Dement, "Quantification of sleepiness: a new approach," *Psychophysiology*, vol. 10, no. 4, pp. 431–436, 1973.

[284] T. Åkerstedt and M. Gillberg, "Subjective and objective sleepiness in the active individual," *International Journal of Neuroscience*, vol. 52, no. 1-2, pp. 29–37, 1990.

[285] 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 Transactions on Human-Machine Systems*, vol. 44, no. 4, pp. 524–530, 2014.

[286] J. Schmidt, C. Braunagel, W. Stolzmann, and K. Karrer-Gauß, "Driver drowsiness and behavior detection in prolonged conditionally automated drives," in *IEEE Intelligent Vehicles Symposium (IV)*, 2016.

[287] F. Steinberger, R. Schroeter, and C. N. Watling, "From road distraction to safe driving: Evaluating the effects of boredom and gamification on driving behaviour, physiological arousal, and subjective experience," *Computers in Human Behavior*, vol. 75, pp. 714–726, 2017.

[288] M. Jones, P. Chapman, and K. Bailey, "The influence of image valence on visual attention and perception of risk in drivers," *Accident Analysis & Prevention*, vol. 73, pp. 296–304, 2014.

[289] European Road Safety Observatory, "Traffic Safety Basic Facts," 2015, accessed: 2024-03-22.

[290] U.S. Department of Transportation. Federal Highway Administration, "Intersection Safety," https://safety.fhwa.dot.gov/intersection/about/index.cfm, accessed: 2024-03-22.

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

[292] J. Werneke and M. Vollrath, "How do environmental characteristics at intersections change in their relevance for drivers before entering an intersection: analysis of drivers' gaze and driving behavior in a driving simulator study," *Cognition, Technology & Work*, vol. 16, no. 2, pp. 157–169, 2014.

[293] L. Meuleners, P. Roberts, and M. Fraser, "Identifying the distracting aspects of electronic advertising billboards: a driving simulation study," *Accident Analysis & Prevention*, vol. 145, p. 105710, 2020.

[294] D. Topolšek, I. Areh, and T. Cvahte, "Examination of driver detection of roadside traffic signs and advertisements using eye tracking," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 43, pp. 212–224, 2016.

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

[296] J. S. Decker, S. J. Stannard, B. McManus, S. M. Wittig, V. P. Sisiopiku, and D. Stavrinos, "The impact of billboards on driver visual behavior: A systematic literature review," *Traffic Injury Prevention*, vol. 16, no. 3, pp. 234–239, 2015.

[297] O. Oviedo-Trespalacios, V. Truelove, B. Watson, and J. A. Hinton, "The impact of road advertising signs on driver behaviour and implications for road safety: A critical systematic review," *Transportation Research Part A: Policy and Practice*, vol. 122, pp. 85–98, 2019.

[298] B. R. Fajen, "Affordance-based control of visually guided action," *Ecological Psychology*, vol. 19, no. 4, pp. 383–410, 2007.

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

[300] Z. Lu, X. Coster, and J. De Winter, "How much time do drivers need to obtain situation awareness? A laboratory-based study of automated driving," *Applied Ergonomics*, vol. 60, pp. 293–304, 2017.

[301] Y. Cheng, L. Gao, Y. Zhao, and F. Du, "Drivers' visual characteristics when merging onto or exiting an urban expressway," *PloS one*, vol. 11, no. 9, p. e0162298, 2016.

[302] 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," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 59, pp. 244–259, 2018.

[303] A. Feldhütter, N. Härtwig, C. Kurpiers, J. M. Hernandez, and K. Bengler, "Effect on mode awareness when changing from conditionally to partially automated driving," in *Congress of the International Ergonomics Association*, 2018, pp. 314–324.

[304] S. Hergeth, L. Lorenz, R. Vilimek, and J. F. Krems, "Keep your scanners peeled: Gaze behavior as a measure of automation trust during highly automated driving," *Human Factors*, vol. 58, no. 3, pp. 509–519, 2016.

[305] W. Vlakveld, N. van Nes, J. de Bruin, L. Vissers, and M. van der Kroft, "Situation awareness increases when drivers have more time to take over the wheel in a Level 3 automated car: A simulator study," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 58, pp. 917–929, 2018.

[306] C. Ahlström, K. Kircher, and A. Kircher, "A gaze-based driver distraction warning system and its effect on visual behavior," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 2, pp. 965–973, 2013.

[307] H. Kim and J. L. Gabbard, "Assessing distraction potential of augmented reality head-up displays for vehicle drivers," *Human Factors*, p. 0018720819844845, 2019.

[308] R. Eyraud, E. Zibetti, and T. Baccino, "Allocation of visual attention while driving with simulated augmented reality," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 32, pp. 46–55, 2015.

[309] B. Wolfe, B. D. Sawyer, and R. Rosenholtz, "Toward a theory of visual information acquisition in driving," *Human Factors*, vol. 64, no. 4, pp. 694–713, 2022.

[310] C. Vater, B. Wolfe, and R. Rosenholtz, "Peripheral vision in real-world tasks: A systematic review," *Psychonomic Bulletin & Review*, vol. 29, no. 5, pp. 1531–1557, 2022.

[311] T. C. Lansdown, A. N. Stephens, and G. H. Walker, "Multiple driver distractions: A systemic transport problem," *Accident Analysis & Prevention*, vol. 74, pp. 360–367, 2015.

[312] O. Oviedo-Trespalacios, M. M. Haque, M. King, and S. Washington, "Understanding the impacts of mobile phone distraction on driving performance: A systematic review," *Transportation Research Part C: Emerging Technologies*, vol. 72, pp. 360–380, 2016.

[313] S. Cao, S. Samuel, Y. Murzello, W. Ding, X. Zhang, and J. Niu, "Hazard perception in driving: A systematic literature review," *Transportation Research Record*, vol. 2676, no. 12, pp. 666–690, 2022.

[314] J. Huang, Y. Long, and X. Zhao, "Driver glance behavior modeling based on semi-supervised clustering and piecewise aggregate representation," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[315] M. Lundgren, L. Hammarstrand, and T. McKelvey, "Driver-gaze zone estimation using bayesian filtering and gaussian processes," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 10, pp. 2739–2750, 2016.

[316] 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 Transactions on Intelligent Transportation Systems*, vol. 12, no. 1, pp. 254–267, 2011.

[317] A. Rangesh, B. Zhang, and M. M. Trivedi, "Driver gaze estimation in the real world: Overcoming the eyeglass challenge," in *IEEE Intelligent Vehicles Symposium (IV)*, 2020.

[318] M.-C. Chuang, R. Bala, E. A. Bernal, P. Paul, and A. Burry, "Estimating gaze direction of vehicle drivers using a smartphone camera," in *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, 2014.

[319] L. Fridman, P. Langhans, J. Lee, and B. Reimer, "Driver gaze region estimation without use of eye movement," *IEEE Intelligent Systems*, vol. 31, no. 3, pp. 49–56, 2016.

[320] A. Tawari and M. M. Trivedi, "Robust and continuous estimation of driver gaze zone by dynamic analysis of multiple face videos," in *IEEE Intelligent Vehicles Symposium (IV)*, 2014.

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

[322] S. Martin, S. Vora, K. Yuen, and M. M. Trivedi, "Dynamics of driver's gaze: Explorations in behavior modeling and maneuver prediction," *IEEE Transactions on Intelligent Vehicles*, vol. 3, no. 2, pp. 141–150, 2018.

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

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

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

[326] 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 *IEEE Intelligent Vehicles Symposium (IV)*, 2016.

[327] T. Liu, Y. Yang, G.-B. Huang, Y. K. Yeo, and Z. Lin, "Driver distraction detection using semi-supervised machine learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 4, pp. 1108–1120, 2015.

[328] C. Braunagel, E. Kasneci, W. Stolzmann, and W. Rosenstiel, "Driver-activity recognition in the context of conditionally autonomous driving," in *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2015.

[329] T. Liu, Y. Yang, G.-B. Huang, Z. Lin, F. Klanner, C. Denk, and R. H. Rasshofer, "Cluster regularized extreme learning machine for detecting mixed-type distraction in driving," in *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2015.

[330] 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 *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2014.

[331] F. Tango and M. Botta, "Real-time detection system of driver distraction using machine learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 2, pp. 894–905, 2013.

[332] K. Kircher and C. Ahlström, *The driver distraction detection algorithm AttenD*, J. D. Lee and M. A. Regan, Eds. CRC Press, 2017.

[333] R. Wang, P. V. Amadori, and Y. Demiris, "Real-time workload classification during driving using hypernetworks," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018.

[334] 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 Transactions on Intelligent Transportation Systems*, vol. 17, no. 6, pp. 1628–1637, 2016.

[335] T. Hirayama, K. Mase, and K. Takeda, "Detection of driver distraction based on temporal relationship between eye-gaze and peripheral vehicle behavior," in *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2012.

[336] J. Chen, Y. Jiang, Z. Huang, X. Guo, B. Wu, L. Sun, and T. Wu, "Fine-grained detection of driver distraction based on neural architecture search," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, pp. 5783–5801, 2021.

[337] 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 *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, 2015.

[338] N. Li and C. Busso, "Detecting drivers' mirror-checking actions and its application to maneuver and secondary task recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 4, pp. 980–992, 2015.

[339] N. Li and C. Busso, "Predicting perceived visual and cognitive distractions of drivers with multimodal features," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 1, pp. 51–65, 2014.

[340] 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 Recognition*, vol. 116, p. 107955, 2021.

[341] M. Wollmer, C. Blaschke, T. Schindl, B. Schuller, B. Farber, S. Mayer, and B. Trefflich, "Online driver distraction detection using long short-term memory," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 2, pp. 574–582, 2011.

[342] 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," *Human Factors*, vol. 54, no. 6, pp. 1104–1116, 2012.

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

[344] N. Li and C. Busso, "Detecting drivers' mirror-checking actions and its application to maneuver and secondary task recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 4, pp. 980–992, 2015.

[345] R. D. Sorkin, "Why are people turning off our alarms?" *The Journal of the Acoustical Society of America*, vol. 84, no. 3, pp. 1107–1108, 1988.

[346] 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 Transactions on Intelligent Transportation Systems*, vol. 21, no. 10, pp. 4196–4208, 2019.

[347] M. Ahmed, S. Masood, M. Ahmad, and A. A. Abd El-Latif, "Intelligent Driver Drowsiness Detection for Traffic Safety Based on Multi CNN Deep Model and Facial Subsampling," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, pp. 19 743–19 752, 2021.

[348] 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 Transactions on Intelligent Transportation Systems*, 2020.

[349] J. Yu, S. Park, S. Lee, and M. Jeon, "Driver drowsiness detection using condition-adaptive representation learning framework," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 11, pp. 4206–4218, 2018.

[350] 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 *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, 2017.

[351] J. Yu, S. Park, S. Lee, and M. Jeon, "Representation learning, scene understanding, and feature fusion for drowsiness detection," in *Asian Conference on Computer Vision (ACCV)*, 2016.

[352] T.-H. Shih and C.-T. Hsu, "MSTN: Multistage spatial-temporal network for driver drowsiness detection," in *Asian Conference on Computer Vision (ACCV) Workshops*, 2016.

[353] 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 *Asian Conference on Computer Vision (ACCV)*, 2016.

[354] C.-H. Weng, Y.-H. Lai, and S.-H. Lai, "Driver drowsiness detection via a hierarchical temporal deep belief network," in *Asian Conference on Computer Vision (ACCV)*, 2016.

[355] 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," *Applied Sciences*, vol. 6, no. 5, p. 137, 2016.

[356] S. Park, F. Pan, S. Kang, and C. D. Yoo, "Driver drowsiness detection system based on feature representation learning using various deep networks," in *Asian Conference on Computer Vision (ACCV)*, 2016.

[357] E. Tadesse, W. Sheng, and M. Liu, "Driver drowsiness detection through hmm based dynamic modeling," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2014.

[358] B. Akrout and W. Mahdi, "A visual based approach for drowsiness detection," in *IEEE Intelligent Vehicles Symposium (IV)*, 2013.

[359] M. Vijay, N. N. Vinayak, M. Nunna, and S. Natarajan, "Real-time driver drowsiness detection using facial action units," in *International Conference on Pattern Recognition (ICPR)*, 2021.

[360] R. Ghoddoosian, M. Galib, and V. Athitsos, "A realistic dataset and baseline temporal model for early drowsiness detection," in *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, 2019.

[361] B. Bakker, B. Zabłocki, A. Baker, V. Riethmeister, B. Marx, G. Iyer, A. Anund, and C. Ahlström, "A multi-stage, multi-feature machine learning approach to detect driver sleepiness in naturalistic road driving conditions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, pp. 4791–4800, 2021.

[362] K. Qian, T. Koike, T. Nakamura, B. Schuller, and Y. Yamamoto, "Learning multimodal representations for drowsiness detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, pp. 11 539–11 548, 2021.

[363] A. Dasgupta, D. Rahman, and A. Routray, "A smartphone-based drowsiness detection and warning system for automotive drivers," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 11, pp. 4045–4054, 2018.

[364] H. Yin, Y. Su, Y. Liu, and D. Zhao, "A driver fatigue detection method based on multi-sensor signals," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2016.

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

[366] 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 *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2018.

[367] H. Matsuo and A. Khiat, "Prediction of drowsy driving by monitoring driver's behavior," in *International Conference on Pattern Recognition (ICPR)*, 2012.

[368] 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 Transactions on Intelligent Transportation Systems*, vol. 18, no. 12, pp. 3408–3420, 2017.

[369] T.-H. Chang and Y.-R. Chen, "Driver fatigue surveillance via eye detection," in *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2014.

[370] I. Garcia, S. Bronte, L. M. Bergasa, J. Almazán, and J. Yebes, "Vision-based drowsiness detector for real driving conditions," in *IEEE Intelligent Vehicles Symposium (IV)*, 2012.

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

[372] S. Dari, N. Epple, and V. Protschky, "Unsupervised blink detection and driver drowsiness metrics on naturalistic driving data," in *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2020.

[373] X. Li, E. Seignez, and P. Loonis, "Vision-based estimation of driver drowsiness with ord model using evidence theory," in *IEEE Intelligent Vehicles Symposium (IV)*, 2013.

[374] A. Joshi, S. Kyal, S. Banerjee, and T. Mishra, "In-the-wild drowsiness detection from facial expressions," in *IEEE Intelligent Vehicles Symposium (IV)*, 2020.

[375] L. Zhao, Z. Wang, X. Wang, and Q. Liu, "Driver drowsiness detection using facial dynamic fusion information and a dbn," *IET Intelligent Transport Systems*, vol. 12, no. 2, pp. 127–133, 2017.

[376] L. N. Boyle, J. Tippin, A. Paul, and M. Rizzo, "Driver performance in the moments surrounding a microsleep," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 11, no. 2, pp. 126–136, 2008.

[377] 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 *International Conference on Wireless Broadband and Ultra Wideband Communications*, 2007.

[378] 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, p. e64782, 2013.

[379] C. Ahlström, C. Fors, A. Anund, and D. Hallvig, "Video-based observer rated sleepiness versus self-reported subjective sleepiness in real road driving," *European Transport Research Review*, vol. 7, no. 4, pp. 1–9, 2015.

[380] P. Philip, P. Sagaspe, J. Taillard, N. Moore, C. Guilleminault, M. Sanchez-Ortuno, T.-b. Åkerstedt, and B. Bioulac, "Fatigue, sleep restriction, and performance in automobile drivers: a controlled study in a natural environment," *Sleep*, vol. 26, no. 3, pp. 277–280, 2003.

[381] O. Lappi, P. Rinkkala, and J. Pekkanen, "Systematic observation of an expert driver's gaze strategy – an on-road case study," *Frontiers in Psychology*, vol. 8, p. 620, 2017.

[382] 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 *IEEE Intelligent Vehicles Symposium (IV)*, 2019.

[383] S. Martin and M. M. Trivedi, "Gaze fixations and dynamics for behavior modeling and prediction of on-road driving maneuvers," in *IEEE Intelligent Vehicles Symposium (IV)*, 2017.

[384] 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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015.

[385] A. Jain, A. Singh, H. S. Koppula, S. Soh, and A. Saxena, "Recurrent neural networks for driver activity anticipation via sensory-fusion architecture," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2016.

[386] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost: A hybrid approach to alleviating class imbalance," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 40, no. 1, pp. 185–197, 2009.

[387] Y. Ma, J. Wu, and C. Long, "GazeFCW: Filter Collision Warning Triggers by Detecting Driver's Gaze Area," in *Proceedings of the Workshop on Machine Learning on Edge in Sensor Systems*, 2019.

[388] M. Rezaei and R. Klette, "Look at the driver, look at the road: No distraction! No accident!" in *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2014.

[389] 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 *IEEE International Conference on Robotics and Automation (ICRA)*, 2016.

[390] 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 *IEEE Intelligent Vehicles Symposium (IV)*, 2017.

[391] A. Tawari, A. Møgelmose, S. Martin, T. B. Moeslund, and M. M. Trivedi, "Attention estimation by simultaneous analysis of viewer and view," in *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2014.

[392] J. Schwehr and V. Willert, "Multi-hypothesis multi-model driver's gaze target tracking," in *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2018.

[393] 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 *IEEE Intelligent Vehicles Symposium (IV)*, 2014.

[394] T. Bär, D. Linke, D. Nienhüser, and J. M. Zöllner, "Seen and missed traffic objects: A traffic object-specific awareness estimation," in *IEEE Intelligent Vehicles Symposium (IV)*, 2013.

[395] M. Mori, C. Miyajima, P. Angkititrakul, T. Hirayama, Y. Li, N. Kitaoka, and K. Takeda, "Measuring driver awareness based on correlation between gaze behavior and risks of surrounding vehicles," in *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2012.

[396] M. Roth, F. Flohr, and D. M. Gavrila, "Driver and pedestrian awareness-based collision risk analysis," in *IEEE Intelligent Vehicles Symposium (IV)*, 2016.

[397] C. Ahlström, G. Georgoulas, and K. Kircher, "Towards a context-dependent multi-buffer driver distraction detection algorithm," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[398] J. Schwehr and V. Willert, "Driver's gaze prediction in dynamic automotive scenes," in *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2017.

[399] H. Zhu, T. Misu, S. Martin, X. Wu, and K. Akash, "Improving driver situation awareness prediction using human visual sensory and memory mechanism," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 6210–6216.

[400] M. R. Endsley, "Situation awareness global assessment technique (sagat)," in *Proceedings of National Aerospace and Electronics Conference*, 1988, pp. 789–795.

[401] 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 Transactions on Intelligent Transportation Systems*, vol. 21, no. 10, pp. 4318–4327, 2019.

[402] J. Schwehr, M. Knaust, and V. Willert, "How to evaluate object-of-fixation detection," in *IEEE Intelligent Vehicles Symposium (IV)*, 2019.

[403] S. Zabihi, S. S. Beauchemin, E. De Medeiros, and M. A. Bauer, "Frame-rate vehicle detection within the attentional visual area of drivers," in *IEEE Intelligent Vehicles Symposium (IV)*, 2014.

[404] K. Gillmeier, F. Diederichs, and D. Spath, "Prediction of ego vehicle trajectories based on driver intention and environmental context," in *IEEE Intelligent Vehicles Symposium (IV)*, 2019.

[405] D. Tran, J. Du, W. Sheng, D. Osipychev, Y. Sun, and H. Bai, "A human-vehicle collaborative driving framework for driver assistance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 9, pp. 3470–3485, 2018.

[406] 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 *IEEE Intelligent Vehicles Symposium (IV)*, 2020.

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

[408] F. Diederichs, T. Schüttke, and D. Spath, "Driver intention algorithm for pedestrian protection and automated emergency braking systems," in *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2015.

[409] K. Doman, D. Deguchi, T. Takahashi, Y. Mekada, I. Ide, H. Murase, and U. Sakai, "Estimation of traffic sign visibility considering local and global features in a driving environment," in *IEEE Intelligent Vehicles Symposium (IV)*, 2014.

[410] K. Doman, D. Deguchi, T. Takahashi, Y. Mekada, I. Ide, H. Murase, and Y. Tamatsu, "Estimation of traffic sign visibility considering temporal environmental changes for smart driver assistance," in *IEEE Intelligent Vehicles Symposium (IV)*, 2011.

[411] K. Doman, D. Deguchi, T. Takahashi, Y. Mekada, I. Ide, H. Murase, and Y. Tamatsu, "Estimation of traffic sign visibility toward smart driver assistance," in *IEEE Intelligent Vehicles Symposium (IV)*, 2010.

[412] D. Engel and C. Curio, "Detectability prediction in dynamic scenes for enhanced environment perception," in *IEEE Intelligent Vehicles Symposium (IV)*, 2012.

[413] 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 Transactions on Intelligent Transportation Systems*, vol. 17, no. 7, pp. 2051–2062, 2016.

[414] T. Deng, H. Yan, and Y.-J. Li, "Learning to boost bottom-up fixation prediction in driving environments via random forest," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 9, pp. 3059–3067, 2017.

[415] A. Borji, D. N. Sihite, and L. Itti, "Computational modeling of top-down visual attention in interactive environments." in *British Machine Vision Conference (BMVC)*, 2011.

[416] A. Borji, D. N. Sihite, and L. Itti, "Probabilistic learning of task-specific visual attention," in *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2012.

[417] A. Borji, D. N. Sihite, and L. Itti, "What/where to look next? Modeling top-down visual attention in complex interactive environments," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no. 5, pp. 523–538, 2013.

[418] P. V. Amadori, T. Fischer, and Y. Demiris, "HammerDrive: A Task-Aware Driving Visual Attention Model," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, pp. 5573–5585, 2021.

[419] 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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[420] A. Pal, S. Mondal, and H. I. Christensen, "'Looking at the Right Stuff'—Guided Semantic-Gaze for Autonomous Driving," in *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2020.

[421] J. Fang, D. Yan, J. Qiao, J. Xue, and H. Yu, "Dada: Driver attention prediction in driving accident scenarios," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[422] A. Palazzi, D. Abati, S. Calderara, F. Solera, and R. Cucchiara, "Predicting the Driver's Focus of Attention: the DR (eye) VE Project," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 41, no. 7, pp. 1720–1733, 2018.

[423] L. Palmer, A. Bialkowski, G. J. Brostow, J. Ambeck-Madsen, and N. Lavie, "Predicting the perceptual demands of urban driving with video regression," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2017.

[424] M. Gao, A. Tawari, and S. Martin, "Goal-oriented object importance estimation in on-road driving videos," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

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

[426] Z. Zhang, A. Tawari, S. Martin, and D. Crandall, "Interaction graphs for object importance estimation in on-road driving videos," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[427] H. R. Tavakoli, E. Rahtu, J. Kannala, and A. Borji, "Digging deeper into egocentric gaze prediction," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2019.

[428] M. Ning, C. Lu, and J. Gong, "An Efficient Model for Driving Focus of Attention Prediction using Deep Learning," in *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019.

[429] D. Gopinath, G. Rosman, S. Stent, K. Terahata, L. Fletcher, B. Argall, and J. Leonard, "Maad: A model and dataset for" attended awareness" in driving," in *ICCVW*, 2021, pp. 3426–3436.

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

[431] A. Palazzi, F. Solera, S. Calderara, S. Alletto, and R. Cucchiara, "Learning where to attend like a human driver," in *IEEE Intelligent Vehicles Symposium (IV)*, 2017.

[432] Y. Xia, D. Zhang, J. Kim, K. Nakayama, K. Zipser, and D. Whitney, "Predicting driver attention in critical situations," in *Asian Conference on Computer Vision (ACCV)*, 2018.

[433] A. Tawari, P. Mallela, and S. Martin, "Learning to attend to salient targets in driving videos using fully convolutional RNN," in *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2018.

[434] A. Tawari and B. Kang, "A computational framework for driver's visual attention using a fully convolutional architecture," in *IEEE Intelligent Vehicles Symposium (IV)*, 2017.

[435] T. Deng, H. Yan, L. Qin, T. Ngo, and B. Manjunath, "How do drivers allocate their potential attention? Driving fixation prediction via convolutional neural networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 5, pp. 2146–2154, 2019.

[436] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 41, no. 3, pp. 740–757, 2018.

[437] 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 *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2013.

[438] J. Janai, F. Güney, A. Behl, and A. Geiger, "Computer vision for autonomous vehicles: Problems, datasets and state of the art," *Foundations and Trends in Computer Graphics and Vision*, vol. 12, no. 1–3, pp. 1–308, 2020.

[439] C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, L. Jesus, R. Berriel, T. M. Paixão, F. Mutz, L. de Paula Veronese, T. Oliveira-Santos, and A. F. De Souza, "Self-driving cars: A survey," *Expert Systems with Applications*, vol. 165, p. 113816, 2020.

[440] A. Rasouli and J. K. Tsotsos, "Autonomous vehicles that interact with pedestrians: A survey of theory and practice," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 900–918, 2019.

[441] A. Tampuu, M. Semikin, N. Muhammad, D. Fishman, and T. Matiisen, "A Survey of End-to-End Driving: Architectures and Training Methods," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, pp. 1364–1384, 2020.

[442] G. Brauwers and F. Frasincar, "A general survey on attention mechanisms in deep learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, pp. 3279–3298, 2021.

[443] K. Ishihara, A. Kanervisto, J. Miura, and V. Hautamaki, "Multi-task learning with attention for end-to-end autonomous driving," in *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, 2021.

[444] J. Kim and J. Canny, "Interpretable learning for self-driving cars by visualizing causal attention," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

[445] L. Cultrera, L. Seidenari, F. Becattini, P. Pala, and A. Del Bimbo, "Explaining Autonomous Driving by Learning End-to-End Visual Attention," in *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, 2020.

[446] D. Wang, C. Devin, Q.-Z. Cai, F. Yu, and T. Darrell, "Deep object-centric policies for autonomous driving," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[447] S. He, D. Kangin, Y. Mi, and N. Pugeault, "Aggregated Sparse Attention for Steering Angle Prediction," in *International Conference on Pattern Recognition (ICPR)*, 2018.

[448] 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 *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[449] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.

[450] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Computing Surveys*, vol. 54, no. 10, pp. 1–41, 2021.

[451] K. Chitta, A. Prakash, and A. Geiger, "NEAT: Neural Attention Fields for End-to-End Autonomous Driving," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[452] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2021.

[453] L. L. Li, B. Yang, M. Liang, W. Zeng, M. Ren, S. Segal, and R. Urtasun, "End-to-end contextual perception and prediction with interaction transformer," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[454] J. Vig, "A multiscale visualization of attention in the transformer model," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2019.

[455] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2021.

[456] 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 *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[457] Y. Chen, C. Liu, L. Tai, M. Liu, and B. E. Shi, "Gaze training by modulated dropout improves imitation learning," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[458] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual explanations for self-driving vehicles," in *European Conference on Computer Vision (ECCV)*, 2018.

[459] J. Kim, T. Misu, Y.-T. Chen, A. Tawari, and J. Canny, "Grounding human-to-vehicle advice for self-driving vehicles," in *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2019.

[460] J. Kim, S. Moon, A. Rohrbach, T. Darrell, and J. Canny, "Advisable learning for self-driving vehicles by internalizing observation-to-action rules," in *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2020.

[461] 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 *IEEE Intelligent Vehicles Symposium (IV)*, 2019.

[462] S. Mund, R. Frank, G. Varisteas, and R. State, "Visualizing the Learning Progress of Self-Driving Cars," in *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2018.

[463] J. Fang, D. Yan, J. Qiao, J. Xue, H. Wang, and S. Li, "DADA-2000: Can Driving Accident be Predicted by Driver Attentionf Analyzed by A Benchmark," in *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019.

[464] S. Taamneh, P. Tsiamyrtzis, M. Dcosta, P. Buddharaju, A. Khatri, M. Manser, T. Ferris, R. Wunderlich, and I. Pavlidis, "A multimodal dataset for various forms of distracted driving," *Scientific Data*, vol. 4, p. 170110, 2017.

[465] I. Dua, T. A. John, R. Gupta, and C. Jawahar, "Dgaze: Driver gaze mapping on road," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[466] 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 *IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2021.

[467] J. D. Ortega, N. Kose, P. Cañas, M.-A. Chao, A. Unnervik, M. Nieto, O. Otaegui, and L. Salgado, "DMD: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis," in *European Conference on Computer Vision (ECCV)*, 2020.

[468] M. Martin, A. Roitberg, M. Haurilet, M. Horne, S. Reiß, M. Voit, and R. Stiefelhagen, "Drive&Act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[469] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko, "Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning," in *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2018.

[470] F.-H. Chan, Y.-T. Chen, Y. Xiang, and M. Sun, "Anticipating accidents in dashcam videos," in *Asian Conference on Computer Vision (ACCV)*, 2016.

[471] Q. Massoz, T. Langohr, C. François, and J. G. Verly, "The ULg multimodality drowsiness database (called DROZY) and examples of use," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2016.

[472] N. Pugeault and R. Bowden, "How much of driving is preattentive?" *IEEE Transactions on Vehicular Technology*, vol. 64, no. 12, pp. 5424–5438, 2015.

[473] S. Abtahi, M. Omidyeganeh, S. Shirmohammadi, and B. Hariri, "YawDD: A yawning detection dataset," in *Proceedings of the ACM Multimedia Systems Conference*, 2014.

[474] V. L. Neale, T. A. Dingus, S. G. Klauer, J. Sudweeks, and M. Goodman, "An overview of the 100-Car naturalistic study and findings," in *Proceedings of International Technical Conference on the Enhanced Safety of Vehicles*, 2005.

[475] T. A. Dingus, J. M. Hankey, J. F. Antin, S. E. Lee, L. Eichelberger, K. E. Stulce, D. McGraw, M. Perez, and L. Stowe, *Naturalistic driving study: Technical coordination and quality control*, 2015, no. SHRP 2 Report S2-S06-RW-1.

[476] T. A. Ranney, W. R. Garrott, and M. J. Goodman, "NHTSA driver distraction research: Past, present, and future," in *International Technical Conference on Enhanced Safety of Vehicles*, 2001.

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

[478] S. W. Kim, J. Philion, A. Torralba, and S. Fidler, "Drivegan: Towards a controllable high-quality neural simulation," in *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2021.

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

[480] G. Sikander and S. Anwar, "Driver fatigue detection systems: A review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 6, pp. 2339–2352, 2018.

[481] J. Kim, K. Kim, D. Yoon, Y. Koo, and W. Han, "Fusion of driver-information based driver status recognition for co-pilot system," in *IEEE Intelligent Vehicles Symposium (IV)*, 2016.

[482] D. Tran, H. M. Do, J. Lu, and W. Sheng, "Real-time detection of distracted driving using dual cameras," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[483] 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 Transactions on Intelligent Transportation Systems*, vol. 21, no. 1, pp. 346–361, 2019.

[484] L. Viktorová and M. Sucha, "Drivers' acceptance of advanced driver assistance systems—what to consider," *Journal of Traffic and Transportation Engineering*, vol. 8, pp. 320–333, 2018.

[485] J. F. May and C. L. Baldwin, "Driver fatigue: The importance of identifying causal factors of fatigue when considering detection and countermeasure technologies," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 12, no. 3, pp. 218–224, 2009.

[486] 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," *Safety Science*, vol. 119, pp. 76–83, 2019.

[487] J. K. Tsotsos, *A computational perspective on visual attention*. MIT Press, 2011.

[488] I. Kotseruba, "Attention and driving," https://github.com/ykotseruba/attention_and_driving/blob/2.0/datasets.md, 2022.

[489] I. Kasahara, S. Stent, and H. S. Park, "Look Both Ways: Self-supervising driver gaze estimation and road scene saliency," in *European Conference on Computer Vision (ECCV)*, 2022.

[490] H. Xu, Y. Gao, F. Yu, and T. Darrell, "End-to-end learning of driving models from large-scale video datasets," in *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2017.

[491] R. Killick, P. Fearnhead, and I. A. Eckley, "Optimal detection of changepoints with a linear computational cost," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1590–1598, 2012.

[492] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[493] OpenStreetMap contributors, "Planet dump retrieved from https://planet.osm.org ," https://www.openstreetmap.org, 2017.

[494] R. P. Roess and E. S. Prassas, *The highway capacity manual: A conceptual and research history*. Springer, 2014.

[495] Z. Lu, X. Coster, and J. De Winter, "How much time do drivers need to obtain situation awareness? a laboratory-based study of automated driving," *Applied Ergonomics*, vol. 60, pp. 293–304, 2017.

[496] D. Crundall, E. Van Loon, T. Baguley, and V. Kroll, "A novel driving assessment combining hazard perception, hazard prediction and theory questions," *Accident Analysis & Prevention*, vol. 149, p. 105847, 2021.

[497] I. Kotseruba and J. K. Tsotsos, "Behavioral research and practical models of drivers' attention," *arXiv:2104.05677*, 2021.

[498] 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 Transactions on Human-Machine Systems*, vol. 44, no. 4, pp. 524–530, 2014.

[499] J. Ross, M. C. Morrone, M. E. Goldberg, and D. C. Burr, "Changes in visual perception at the time of saccades," *Trends in Neurosciences*, vol. 24, no. 2, pp. 113–121, 2001.

[500] W. Zeng, Y. Xiao, S. Wei, J. Gan, X. Zhang, Z. Cao, Z. Fang, and J. T. Zhou, "Real-time multi-person eyeblink detection in the wild for untrimmed video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 854–13 863.

[501] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, "MIT Saliency Benchmark," http://saliency.mit.edu/.

[502] S. Mathe and C. Sminchisescu, "Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 37, no. 7, pp. 1408–1424, 2014.

[503] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao, "Predicting human gaze beyond pixels," *Journal of Vision*, vol. 14, no. 1, pp. 28–28, 2014.

[504] A. Nguyen and Z. Yan, "A saliency dataset for 360-degree videos," in *ACM Multimedia Systems Conference*, 2019, pp. 279–284.

[505] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," in *Proceedings of the ACM International Conference on Multimedia*, 2010, pp. 1469–1472.

[506] K. Kurzhals and D. Weiskopf, "Space-time visual analytics of eye-tracking data for dynamic stimuli," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2129–2138, 2013.

[507] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *European Conference on Computer Vision (ECCV)*, 2020.

[508] H. Summala, "Brake reaction times and driver behavior analysis," *Transportation Human Factors*, vol. 2, no. 3, pp. 217–226, 2000.

[509] P. R. Chapman and G. Underwood, "Visual search of driving situations: Danger and experience," *Perception*, vol. 27, no. 8, pp. 951–964, 1998.

[510] G. Underwood, P. Chapman, Z. Berger, and D. Crundall, "Driving experience, attentional focusing, and the recall of recently inspected events," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 6, no. 4, pp. 289–304, 2003.

[511] K. Nobre and S. Kastner, *The Oxford Handbook of Attention*. Oxford University Press, 2014.

[512] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 35, no. 1, pp. 185–207, 2012.

[513] A. Borji, "Saliency prediction in the deep learning era: Successes and limitations," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 43, no. 2, pp. 679–700, 2021.

[514] J. M. Wolfe and T. S. Horowitz, "Five factors that guide attention in visual search," *Nature Human Behaviour*, vol. 1, no. 3, p. 0058, 2017.

[515] V. Navalpakkam and L. Itti, "Modeling the influence of task on attention," *Vision Research*, vol. 45, no. 2, pp. 205–231, 2005.

[516] S. Frintrop, *VOCUS: A visual attention system for object detection and goal-directed search*. Springer, 2006.

[517] A. Rosenfeld, M. Biparva, and J. K. Tsotsos, "Priming neural networks," in *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, 2018.

[518] V. Ramanishka, A. Das, J. Zhang, and K. Saenko, "Top-down visual saliency guided by captions," in *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2017.

[519] Z. Ding, X. Ren, E. David, M. Vo, G. Kreiman, and M. Zhang, "Efficient zero-shot visual search via target and context-aware transformer," *arXiv:2211.13470*, 2022.

[520] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu, D. Ramanan, and T. S. Huang, "Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015.

[521] M. Biparva and J. Tsotsos, "STNet: Selective Tuning of convolutional networks for object localization," in *IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2017.

[522] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," *International Journal of Computer Vision*, vol. 126, no. 10, pp. 1084–1102, 2018.

[523] I. Kotseruba and J. K. Tsotsos, "Attention for vision-based assistive and automated driving: A review of algorithms and datasets," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 11, pp. 19 907–19 928, 2022.

[524] T. Deng, H. Yan, L. Qin, T. Ngo, and B. Manjunath, "How do drivers allocate their potential attention? driving fixation prediction via convolutional neural networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 5, pp. 2146–2154, 2019.

[525] J. Fang, D. Yan, J. Qiao, J. Xue, and H. Yu, "DADA: Driver attention prediction in driving accident scenarios," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 4959–4971, 2021.

[526] S. Gan, X. Pei, Y. Ge, Q. Wang, S. Shang, S. E. Li, and B. Nie, "Multisource adaption for driver attention prediction in arbitrary driving scenes," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 11, pp. 20 912–20 925, 2022.

[527] 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 Transactions on Intelligent Transportation Systems*, vol. 17, no. 7, pp. 2051–2062, 2016.

[528] T. Deng, H. Yan, and Y.-J. Li, "Learning to boost bottom-up fixation prediction in driving environments via random forest," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 9, pp. 3059–3067, 2017.

[529] M. Kümmerer, T. S. Wallis, and M. Bethge, "DeepGaze II: Reading fixations from deep features trained on object recognition," *arXiv:1610.01563*, 2016.

[530] R. Droste, J. Jiao, and J. A. Noble, "Unified image and video saliency modeling," in *European Conference on Computer Vision (ECCV)*, 2020.

[531] S. Jain, P. Yarlagadda, S. Jyoti, S. Karthik, R. Subramanian, and V. Gandhi, "ViNet: Pushing the limits of visual modality for audio-visual saliency prediction," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021.

[532] R. Fahimi and N. D. Bruce, "On metrics for measuring scanpath similarity," *Behavior Research Methods*, vol. 53, no. 2, pp. 609–628, 2021.

[533] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video Swin Transformer," in *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2022, pp. 3202–3211.

[534] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015.

[535] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning for video understanding," in *European Conference on Computer Vision (ECCV)*, 2018.

[536] A. Rasouli and I. Kotseruba, "PedFormer: Pedestrian Behavior Prediction via Cross-Modal Attention Modulation and Gated Multitask Learning," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[537] J. L. Campbell, C. Carney, and B. H. Kantowitz, "Human factors design guidelines for advanced traveler information systems (ATIS) and commercial vehicle operations (CVO)," US Department of Transportation. Federal Highway Administration, Tech. Rep. FHWA-RD-98-057, 1998.

[538] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2017.

[539] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.

[540] G. Boeing, "OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks," *Computers, Environment and Urban Systems*, vol. 65, pp. 126–139, 2017.

[541] P. Chao, W. Hua, R. Mao, J. Xu, and X. Zhou, "A survey and quantitative study on map inference algorithms from GPS trajectories," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 1, pp. 15–28, 2020.

[542] P. Newson and J. Krumm, "Hidden Markov map matching through noise and sparseness," in *ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL)*, 2009.

[543] A. Botach, E. Zheltonozhskii, and C. Baskin, "End-to-end referring video object segmentation with multimodal transformers," in *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2022.

[544] E. Amirloo, A. Rasouli, P. Lakner, M. Rohani, and J. Luo, "LatentFormer: Multi-agent transformer-based interaction modeling and trajectory prediction," *arXiv:2203.01880*, 2022.

[545] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," in *European Conference on Computer Vision (ECCV)*, 2020.

[546] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *International Conference on Machine Learning (ICML)*, 2013.

[547] I. Kotseruba and J. K. Tsotsos, "Understanding and modeling the effects of task and context on drivers' gaze allocation," *arXiv:2310.09275*, 2023.

[548] S. Boll, A. Asif, and W. Heuten, "Feel your route: A tactile display for car navigation," *IEEE Pervasive Computing*, vol. 10, no. 3, pp. 35–42, 2011.

[549] S. Lemonnier, L. Désiré, R. Brémond, and T. Baccino, "Drivers' visual attention: A field study at intersections," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 69, pp. 206–221, 2020.

[550] Z. Bylinski, "MATLAB implementation of saliency metrics," https://github.com/cvzoya/saliency/blob/master/code_forMetrics.

[551] R. Fahimi and N. D. Bruce, "Code for "On metrics for measuring scanpath similarity"," https://github.com/rAm1n/saliency.

[552] M. Kümmerer, Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, "MIT/Tübingen Saliency Benchmark," https://saliency.tuebingen.ai/.

[553] M. Kümmerer, T. S. A. Wallis, and M. Bethge, "Saliency Benchmarking Made Easy: Separating Models, Maps and Metrics," in *European Conference on Computer Vision (ECCV)*, 2018.