Demo: GazeLock—Physical-World Adversarial Attacks on Camera-Based Driver Monitoring in Autonomous Vehicles

Mboutidem Mkpong, Halima Bouzidi, Mohammad Abdullah Al Faruque
University of California, Irvine, CA, USA
{mkpongm, hbouzidi, alfaruqu}@uci.edu

Abstract

Driver Monitoring Systems (DMS) are safety-critical components in Level 2–4 autonomous vehicles, responsible for verifying driver attentiveness and triggering takeover requests. These systems rely on camera-based gaze estimation models to infer where the driver is looking. However, the adversarial robustness of these models remains largely unexplored, raising the question of whether an attacker can manipulate gaze estimation to undermine DMS and vehicle safety guarantees. We present GazeLock, a novel physical-world attack framework that uses adversarial eyeglasses to force gaze estimators into outputting a targeted gaze prediction regardless of the driver's actual gaze direction. We explore two material modalities: traditional RGB-printed patterns and infrared-sensitive material patterns that are invisible to the human eye but effective against NIR DMS cameras. Our live demo allows attendees to wear the adversarial glasses and observe real-time gaze misdirection on a simulated DMS interface, with video demonstrations of the attack under varied conditions.

1 Introduction

Driver Monitoring Systems (DMS) play a critical role in the advancement of vehicle safety and the progression towards fully autonomous driving. DMS rely on RGB and/or NIR (Near-Infrared) cameras that capture images of a driver to infer information such as head pose, eyelid state, and gaze direction. These systems enhance driver safety by monitoring attention and alertness, and triggering alerts when necessary. They have been shown to reduce crashes and injuries by providing timely warnings regarding distracted or fatigued driving. The safety-criticality of these systems is further underscored by recent regulations mandating that new vehicles in the EU and US include DMS for detecting and warning drivers of unsafe driving conditions [1].

Despite this safety-critical role, the adversarial robustness of DMS gaze estimation has received minimal scrutiny. While adversarial attacks on external perception (e.g., LiDAR spoofing, camera-based object detection attacks) are well-studied, the in-cabin perception pipeline (which directly governs automation handoff decisions) remains largely unexamined.

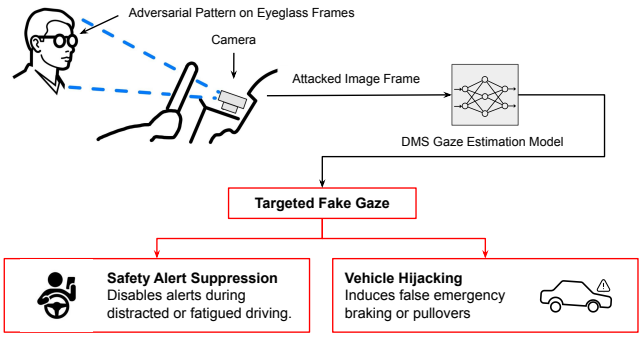


Figure 1: Attacker goals and downstream consequences.

To address this gap and validate the severity of these threats, we present GazeLock, the first adversarial attack framework targeting camera-based gaze estimation in DMS. GazeLock optimizes adversarial patterns on eyeglass frames that force gaze estimators to predict target gaze vector, regardless of the driver's actual gaze direction. We explore traditional RGB-printed patterns, as well as infrared-sensitive material patterns (retro-reflective and IR-absorbing materials) that are invisible to the human eye but effective against infrared DMS cameras. Our goal is to draw attention to the susceptibility of these safety-critical systems and demonstrate how adversarial attacks can undermine the effectiveness of DMS in flagging unsafe driving conditions.

2 Threat Model

Attack Goal. GazeLock is a targeted attack. Rather than simply degrading estimation accuracy, GazeLock steers predictions toward a specific gaze direction. As shown in Fig. 1, this gaze misdirection produces multiple cascading safety consequences in autonomous vehicle operation:

- **Safety Alert Suppression.** The attacker spoofs an "at-

tentive" gaze direction, so the DMS never triggers a takeover request even when the driver is intentionally distracted (e.g., looking away from the road or using a phone), or unintentionally distracted (e.g a fatigued or drowsy driver).

- **Vehicle Hijacking.** The attacker spoofs an "inattentive" gaze direction, causing the DMS to escalate through warnings and potentially trigger emergency maneuvers such as minimum risk condition braking or pullover.

Attacker Capability. We assume a white-box setting in which the attacker has full knowledge of the DMS gaze estimation model architecture and weights. This is achievable by identifying the vehicle's DMS supplier and reverse engineering the deployed model. While our method is developed under white-box assumptions, we also evaluate adversarial transferability across architectures.

Design Challenges. Designing effective physical-world attacks against DMS gaze estimation poses three key challenges: *C1: Spatial Consistency.* The adversarial patch must sustain gaze misdirection across variation in head pose, lighting, and driver-to-camera distance, not just a single static frame. *C2: Temporal Consistency.* DMS decisions require sustained detection of a gaze state over a time window. The spoofed gaze must persist long enough to cross the DMS decision threshold. *C3: Stealth.* RGB-printed patterns on eyeglass frames are visually conspicuous, motivating the use of IR-sensitive materials that are invisible to the human eye but effective under NIR illumination.

3 Attack Design

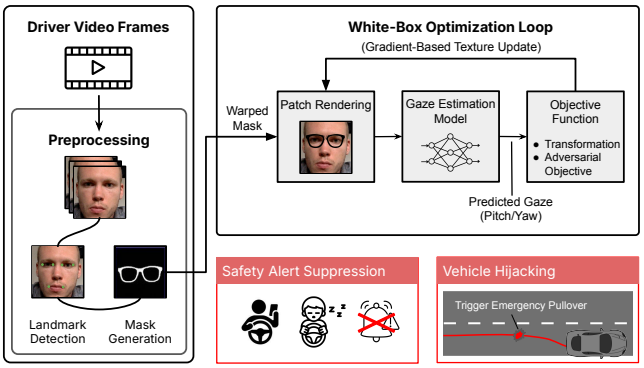


Figure 2: Walk-through of GazeLock Optimization Pipeline.

Fig. 2 illustrates the GazeLock pipeline. Given a face image, the framework composites an adversarial patch onto the eyeglass frame region and optimizes it to minimize the angular error between the model's predicted gaze and a target gaze vector. To address spatial consistency (*C1*), stochastic transformations are applied during optimization to simulate variation in head pose, lighting, and distance. To address temporal consistency (*C2*), we evaluate misdirection persistence across continuous video sequences rather than isolated frames. We explore two material modalities: an *RGB variant*, and an *IR-sensitive variant* using infrared-reflective materials that are invisible to the human eye but effective under 850 nm NIR illumination (*C3*).



Figure 3: Eyeglass Frames with GazeLock attack patterns. (a) A pair of glasses with RGB optimized patch. (b) A pair of glasses with stealthy NIR reflective patches when viewed with regular camera (left) vs an infrared camera (right).

4 Experiments

We evaluate GazeLock on the Driver Monitoring Dataset [2]. Our primary metric is angular misdirection: the angular deviation between the model's prediction under attack and the ground-truth gaze direction, measured in degrees.

Initial results indicate that the RGB adversarial glasses achieve significant gaze misdirection on gaze estimation models under white-box conditions. In the future, we intend on testing our proposed GazeLock on commercial/blackbox systems to validate the severity of these threats.

5 Demonstration Plan

Our demonstration provides a live, interactive experience of the GazeLock attack. The setup consists of a PC equipped with a USB webcam mounted at a driver-facing angle, and a display screen simulating a driver's view of the road. Attendees can sit in a chair positioned at typical driver-to-camera distance (~0.5 m). A custom software consumes the camera feed and outputs a DMS attention dashboard indicating the system's assessment of driver attentiveness (attentive / distracted / drowsy). The attendees would have access to use the physical pairs of eyeglasses with RGB or NIR reflective patterns (such as in 3) to observe the impact on DMS.

References

[1] European Union. Regulation (EU) 2019/2144 (General Safety Regulation), 2019.

[2] 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. In *ECCV Workshops*, pages 387–405, 2020.