InSight: Monitoring the State of the Driver in Low-Light Using Smartphones

ISHANI JANVEJA[*], AKSHAY NAMBI, SHRUTHI BANNUR[*], SANCHIT GUPTA[*], and VENKAT PADMANABHAN, Microsoft Research, India

Road safety is a major public health issue across the globe and over two-thirds of the road accidents occur at nighttime under low-light conditions or darkness. The state of the driver and her/his actions are the key factors impacting road safety. How can we monitor these in a cost-effective manner and in low-light conditions? RGB cameras present in smartphones perform poorly in low-lighting conditions due to lack of information captured. Hence, existing monitoring solutions rely upon specialized hardware such as infrared cameras or thermal cameras in low-light conditions, but are limited to only high-end vehicles owing to the cost of the hardware. We present InSight, a windshield-mounted smartphone-based system that can be retrofitted to the vehicle to monitor the state of the driver, specifically driver fatigue (based on frequent yawning and eye closure) and driver distraction (based on their direction of gaze). Challenges arise from designing an accurate, yet low-cost and non-intrusive system to continuously monitor the state of the driver.

In this paper, we present two novel and practical approaches for continuous driver monitoring in low-light conditions: (i) Image synthesis: enabling monitoring in low-light conditions using just the smartphone RGB camera by synthesizing a thermal image from RGB with a Generative Adversarial Network, and (ii) Near-IR LED: using a low-cost near-IR (NIR) LED attachment to the smartphone, where the NIR LED acts as a light source to illuminate the driver's face, which is not visible to the human eyes, but can be captured by standard smartphone cameras without any specialized hardware. We show that the proposed techniques can capture the driver's face accurately in low-lighting conditions to monitor driver's state. Further, since NIR and thermal imagery is significantly different than RGB images, we present a systematic approach to generate labelled data, which is used to train existing computer vision models. We present an extensive evaluation of both the approaches with data collected from 15 drivers in controlled basement area and on real roads in low-light conditions. The proposed NIR LED setup has an accuracy (F1-score) of 85% and 93.8% in detecting driver fatigue and distraction, respectively in low-light.

CCS Concepts: • **Computing methodologies** → *Artificial intelligence*; • **Human-centered computing** → *Ubiquitous and mobile computing systems and tools.*

Additional Key Words and Phrases: Mobile Systems, Edge Processing, Video Analytics, Driver Monitoring and Low-light.

ACM Reference Format:
Ishani Janveja, Akshay Nambi, Shruthi Bannur, Sanchit Gupta, and Venkat Padmanabhan. 2020. InSight: Monitoring the State of the Driver in Low-Light Using Smartphones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3, Article 83 (September 2020), 29 pages. https://doi.org/10.1145/3411819

[*] Work done while the author was an intern at Microsoft Research India.

Authors' address: Ishani Janveja; Akshay Nambi; Shruthi Bannur; Sanchit Gupta; Venkat Padmanabhan, Microsoft Research, India, https://aka.ms/hams, Contact:akshayn@microsoft.com.

2474-9567/2020/9-ART83 $15.00
https://doi.org/10.1145/3411819

1 INTRODUCTION

Road safety is a major public health issue the world over, with road accidents causing an estimated 1.35 million fatalities and many more injuries each year [7]. Over two-thirds of the accidents occur at nighttime under low-light conditions or darkness [4]. Many studies have found that the primary factors responsible for road accidents center on the driver [8], with key risk factors such as distracted driving, driver fatigue/drowsiness, and drinking and driving. Hence, monitoring the state of the driver and providing actionable feedback is key towards improving road safety, especially at nighttime or under low-light conditions.

While vehicles endowed with Advanced Driver Assistance Systems (ADAS) [3, 5] offer safety features such as driver drowsiness and distraction detection, there is a large installed base of vehicles that lacks such sensing capabilities. This is especially so in the developing regions, where such advanced features in vehicles are often unaffordable but road safety is an acute issue. Motivated by these observations researchers have developed several systems to detect signs of fatigue/distraction and these systems can be classified based on the level of intrusion. (i) Intrusive techniques: These include sensing electrophysiological signals using EEG sensors and tracking body movements using motion sensors [42, 63]. Specifically, wearable devices like smart glasses to track eye movements [63] or magnetic tags to track body movements [35] are used to detect signs of fatigue driving. While wearable devices may yield more accurate results and are independent of lighting conditions, the downside is that these are intrusive methods, drivers may find it hard to adapt to wearing these sensors and hard to scale, as these wearables are expensive. (ii) Non-invasive techniques: Recently, cameras are becoming the most sought-out ubiquitous sensors due to their affordable cost and low-intrusion during deployment. Several works have used cameras to develop vision-based systems to detect blinks, yawns or other signs of drowsiness [29, 53]. Recent works have used standalone cameras associated with custom processing devices to detect driver fatigue and distraction [23, 86]. For instance, in HAMS [53] a wind-shield mounted smartphone is used to monitor both the state of the driver and their driving using the front and back camera imagery.

Challenges and limitations. Current computer vision techniques are catered to only good lighting conditions and perform poorly in low-light [62]. "Low-light" conditions correspond to low/no ambient illuminance. To verify this we conducted an experiment with 1000 RGB images captured in low-light using a standard smartphone, and the existing models was able to detect faces in only 98 of them (<1%). This is due to two limitations:

(i) Lack of information in RGB imagery: The problem at its core is that the RGB image is lacking necessary visual information in low-light conditions [62]. Hence, a different imaging technique such as infrared (IR) or near-IR (NIR) imaging is required for low-light conditions, which illuminates the low-light scene with invisible IR light. Existing approaches [41, 51, 52] use specialized commercial cameras such as IR cameras [13] and FLIR thermal cameras [9] to capture good quality images in low-light conditions. However, they require installation of an expensive, specialized camera limiting the applicability of the system at scale.

(ii) Lack of labelled data for NIR imaging techniques: Due to the inherent limitations of RGB camera, the only feasible solution for monitoring the driver robustly is to rely on other imaging techniques such as NIR. However, existing face detectors and facial landmark models are not designed to work on these images as the imagery collected from the smartphone for NIR images are significantly different as compared to a standard RGB or IR imagery. In order to develop detectors for monitoring the state of the driver in low-light conditions, computer vision algorithms rely on plenty of labelled data, i.e., face boxes and 68 landmarks. However, to the best of our knowledge, there do not exist any pre-trained models or public datasets with such detailed ground truth labels for NIR images to train computer vision models. In addition, it is non-trivial to collect such a labeled dataset; for example, identifying reliably 68 landmarks in an NIR image in low-light is challenging and time consuming.

Proposed solution and contributions.

In this paper, we present two novel and practical approaches that address the above limitations for monitoring in low-light conditions. Our goal is to develop an accurate, low-cost, non-intrusive system to continuously monitor the state of the driver (fatigue and distraction) in low-light. To keep costs low, we aim to re-use the existing wind-shield mounted smartphones for driver monitoring. Figure 1 depicts the overall `InSight` system and we now present our key contributions:

(i) Thermal Image Synthesis: Given the recent advances in deep learning and its applications to computer vision [19, 46, 75, 76], we present a technique that employs Generative Adversarial Networks (GANs) to synthesize a thermal image based just on a low-light RGB image. The key insight for choosing thermal image as the representation for a low-light scene hinges on the observation that the thermal image is largely invariant and captures all the face attributes accurately across good and poor lighting conditions. The objective here is to translate an input low-light RGB image into a corresponding target thermal image given sufficient training data of input-target pairs (low light RGB image and its corresponding thermal image). This can be posed as an image-to-image translation task that aims to translate one representation into another. This solution does require a specialized camera such as a FLIR thermal camera, but only during training the GANs, hence supports the low-cost objective. The key novelty here is the development of an end to end GAN network, that can synthesize a thermal image in low-light without any additional hardware or specialized cameras. (Section 5)

(ii) Near-IR imagery: Instead of using expensive specialized cameras such as IR or FLIR cameras, we present an alternative solution that relies upon a near-IR (NIR) LED attached to the smartphone. The NIR LED acts as a light source to illuminate the driver's face, which is not visible to the human eyes. However, most smartphone cameras are equipped with IR filters that are designed to block out light outside of the visible spectrum. Due to the complexity involved in developing such accurate IR filters, they typically do not have a single cut-off frequency and hence allow a range of spectrum between 700nm to 1000nm. Our solution relies on this key insight and uses LEDs that emit near-infrared light with wavelength of 810-850nm. This NIR light is able to pass through the IR filters present in the standard smartphones by exploiting the variation in cut-off frequency of IR filters. Thus, our key novelty compared to previous systems that employ NIR LEDs [32, 41, 48, 52] is that, we use off-the-shelf standard smartphone cameras to capture the NIR illuminated driver's face without any additional hardware or custom camera setup. Since the setup relies on just a NIR LED (typically <5$) with minimal electronics to support robust illumination of the NIR LED, *the setup is extremely low-cost*. Furthermore, the near-IR light emitted by the LED is safe and invisible to the human eyes, making *the system non-intrusive*. (Section 6)

(iii) Generating labelled data for training: The features extracted from the NIR images using the smartphone are significantly different compared to a standard RGB imagery. Hence, existing face detectors and facial landmark models are not designed to work on these images. To the best of our knowledge, there exists no pre-trained models or public datasets to re-train the models and collecting ground truth data on low-light imagery is a challenging task. To address these issues, we present a simple, automated and robust way to generate labelled NIR data, which is then used to train existing face detectors and facial landmarks. Specifically, we mount an IR camera in tandem with the proposed NIR cameras, where the labels from IR imagery is transferred precisely to the NIR imagery by applying a suitable transformation. Thus generating labelled training data for NIR imagery. Instead of re-inventing the computer vision models, we extend state of the art detectors by re-training them on the generated low-light labelled data and in Section 3 we highlight specific extensions performed on state of the art models to cater to `InSight` requirements. (Section 7)

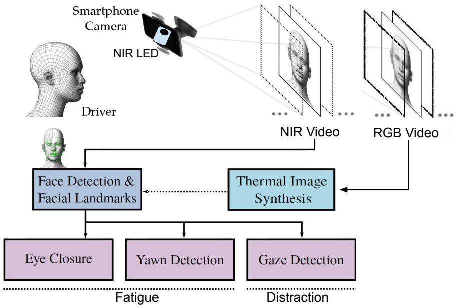


Fig. 1. Overview of `InSight` system.

(iv) Real-world evaluation: We have collected over 25 hours of data from 15 drivers in both controlled basement area and on real roads. The data on real roads were collected between 8 PM to 4 AM capturing both realistic traffic conditions and low-light variations. We conduct extensive evaluation to determine the efficacy of the proposed computer vision models highlighting its applicability on unseen data from new drivers. Finally, we compare the performance of the proposed InSight system with NIR imagery, synthesized imagery and imagery from specialized cameras such as FLIR cameras for driver fatigue and distraction monitoring. (Section 8)

2 RELATED WORK

Prevalent research [53, 54, 78] have used cameras to monitor driver's state such as driver fatigue and distraction, mostly in good lighting conditions. In this section, we will restrict our focus on techniques that are applicable to low-light conditions only and are broadly classified into hardware-based and software-based solutions.

Specialized hardware. Majority of the solutions for monitoring in low-light employ a specialized hardware such as IR [13] or thermal cameras [10]. Unlike typical RGB cameras, which can just see some IR light, IR cameras generally include a built-in or external IR illuminator [13, 14]. Thus it can 'see' IR light and also shine their own light, which is invisible to human eyes and hence making the scene brighter in low-light [11]. However, good quality IR cameras are expensive and hence limits its applicability for deployment at scale, necessitating development of a low-cost system for effective monitoring at low-light.

Low-cost custom IR/NIR cameras. Several academic efforts aim to address the cost issue of IR cameras, by developing a low-cost IR camera setup [26, 41, 45, 52]. These cameras use an array of IR LEDs along with low-cost infrared sensitive CCD (charge-coupled device) cameras. While these cameras can illuminate and see the scene clearly using IR LEDs, they have several limitations [60]. Since an array of LEDs are used, the incident area is not uniformly illuminated leading to overexposed, noisy and blurry images. Furthermore, these are custom hardware solutions, which makes it challenging to integrate with existing driver monitoring systems. For instance, in [48] authors design a custom hardware for capturing low-light data, which includes NIR LEDs along with a long pass optical filter attached to the camera, i.e., requiring hardware modification to the camera. The long-pass filter ensures only NIR light is permitted, thus blocking all visible light and enabling capturing of sharp the low-light images without any artifacts. Similarly, in [32] authors use a setup with 8 NIR LEDs and a custom camera to capture low-light driving data. In contrast to all the prior works on low-cost IR/NIR cameras, we develop a non-invasive NIR setup, where an NIR LED is attached to the smartphone, which captures the NIR imagery without any hardware modification or custom camera setup. This makes the system deployment seamless and can be easily retrofitted to any off-the-shelf smartphone. Further, unlike prior approaches which use specialized cameras for capturing clear IR/NIR imagery, our NIR setup captures additional artifacts in the scene as the smartphone camera can see both visible and NIR spectrum. Hence, there is a need to develop custom computer vision detectors to detect face and facial landmarks on NIR imagery. We present an robust and automated way to generate labelled data, which is used to re-train state of the art computer vision models. Finally, unlike existing IR cameras, which acts as just sensors to capture IR imagery, we use the smartphone as both the sensing and compute platform. This makes the system easy to deploy at scale.

Software-based solutions. Given the hardware limitations, several software techniques are proposed in the literature to enhance low-light images. Some of the traditional methods employ exposure compensation [22, 77] or histogram equalisation [37, 70] strategies, which perform contrast adjustment to amplify dark regions and to prevent over saturation of bright regions. Other image enhancement techniques employ Gamma correction [36, 59], which is a nonlinear operation on images used to increase the brightness of images. However, Gamma correction is carried out on each pixel individually without considering the relationship of a certain pixel with its neighbours. Moreover, these methods assume that the image is taken in dim lighting as opposed to very low light conditions. Therefore, these methods add more noise and blur to the image, which degrades the performance of the computer vision detectors when applied. Recently, few research efforts have focused on improving the quality of RGB images taken in low-light conditions using deep neural networks and learning approaches. Chen et al. [21] employs a learning based approach to combine raw short-exposure low-light images, with corresponding long-exposure reference images to improve the quality of the low light images. Hasinoff et al. [34] presents a computational photography pipeline that captures, aligns, and merges a burst of frames to reduce noise and increase dynamic range for smartphone cameras. These features are now part of Google pixel phones called Night Sight [6]. However, a key limitation is that these approaches work only on an image (due to the requirement of capturing images at different exposure levels) and takes several seconds to derive an enhanced low-light image, which is not feasible in the case of continuous monitoring of the driver.

In contrast to existing approaches, in this paper we present two low-cost practical solutions to monitor the state of the driver in low-light with off-the-shelf smartphones. With extensive evaluation on real driving data, we show that the NIR LED based setup is robust and accurate in driver fatigue and distraction monitoring.

3 MONITORING THE STATE OF THE DRIVER

We now present how to monitor the state of the driver using the smartphone camera imagery. In this paper, we restrict the driver monitoring definition to driver fatigue and driver distraction. In InSight we use behavioral measures such as eye blinks and yawn frequency to detect drowsiness [53]. Furthermore, we derive driver's gaze information to determine driver distraction. Specifically, we use mirror scanning behavior to detect distraction. The common substrate for detecting both fatigue and distraction is first to detect driver's face and then the corresponding facial landmarks accurately.

We now present the details of existing face detection and landmarks algorithms that are used in InSight. The techniques presented in this section are state-of-the-art techniques, but they work only in good lighting conditions. For each technique, we also present InSight specific extensions to make it work on NIR/thermal imagery and for robust driver monitoring. In Section 7 we present in detail how to re-train these models for NIR/thermal imagery.

3.1 Face Detection

To detect the drivers face from the smartphone imagery, we use the Histogram of Oriented Gradients (HOG) features [27] to train a Linear Support Vector Machine (SVM) model. Our choice of detector is informed by the requirement of a memory and compute efficient detection algorithm. The Histogram of Oriented Gradients (HOG) is a feature descriptor used for object detection in computer vision. The orientation of gradients of pixels for a particular object in an image can describe the shape of object. This information can in turn act as a template describing the object thereby allowing us to match this template with instances of that object occurring in other images. During training HOG descriptor is extracted for each face image, which is then used to train the SVM model. In the testing phase, a window slides across the image and at each position of the window a HOG descriptor is calculated and further classified by the SVM classification model. Thus the output of the face detection block results in a bounding box around the driver's face.

InSight-specific extension: The existing face detector models are trained on only good light conditions, resulting in a template of features that fit well for good lighting RGB images. Due to the domain bias (between good lighting RGB image and low-lighting NIR image) and change in feature descriptors for a low-light NIR/thermal image, the existing pre-trained models do not work on low-light NIR/thermal imagery. Hence we need to re-train these models with feature descriptors obtained from labelled NIR imagery for accurate detection (see Section 7 for more details).

3.2 Facial Landmarks Detection

Driver fatigue and distraction detection relies heavily upon the accurate prediction of facial landmarks on a driver's face. Facial landmark detection is a foundational problem for several computer vision tasks. It aims to localize facial feature points like eye corners, mouth corners, nose tip, etc., as shown in Figure 2 [61].

Over the past decade significant research has been performed in deriving accurate facial landmarks from explicitly modelling facial appearance [24] to the Constrained Local Model (CLM) methods [25] that rely on both the local and global facial appearance to deep learning [80], [44] and 3D based methods [85]. However, recent works on extracting facial landmarks in driving scenarios [29] have found that the cascaded regression-based methods show better performances in terms of both accuracy and time [73].

In InSight, we use the Ensemble of Regression Trees [43] for localizing 68 facial landmarks. A cascade of regressors iterate over the estimated key points to refine their initial positions. The regressors produce a new estimate from the previous one, trying to reduce the alignment error of the estimated points at each iteration. However, this landmark detector requires the facial region coordinates to estimate the landmarks and hence, it is heavily dependent upon the accuracy of the face detector. Thus the input to the facial landmark model is the face bounding box, resulting in 68 facial landmarks.

InSight-**specific extension:** Similar to face detection, the existing pre-trained facial landmark models do not work on NIR/thermal imagery due to domain change from RGB to NIR imagery. In addition, there is a lack of labelled data, i.e., 68 landmarks on NIR imagery for training. Hence, these models need to be re-trained with labelled NIR/thermal imagery (see Section 7).

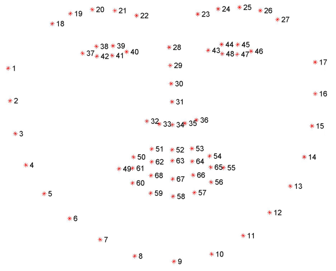


Fig. 2. 68 Facial Landmarks [61].

3.3 Detecting Driver Fatigue

We use behavioural measures such as eye blinks and yawn frequency to detect driver fatigue. In this paper, we use terms fatigue and drowsiness interchangeably. The detection of both these events is based on visual queues provided by the facial landmarks and is explained in the following sub-sections.

3.3.1 Eye State Detection. Several studies have shown that **PER**centage of **CLOS**ure of the eye (PERCLOS) is a reliable metric for driver alertness [57], [30]. It is defined as the proportion of time in a minute when the eyes are closed. Robust and real-time eye state detection on real-world videos is challenging due to the diversity in lighting conditions. Prior work has considered template matching and pupil identification, which are prone to mis-classification due to change in illumination and head pose variation [79], [82]. In contrast, we leverage the facial landmarks to determine the state of the eye.

Eye state: Six landmarks (32-42) represent each eye (see Figure 2). We use the eye aspect ratio metric (EAR) [64] to measure the state of the eye, which is the ratio of the height and width of the eye:

$$EAR = \frac{||p_{42} - p_{38}|| + ||p_{41} - p_{39}||}{2 \times ||p_{40} - p_{37}||} \tag{1}$$

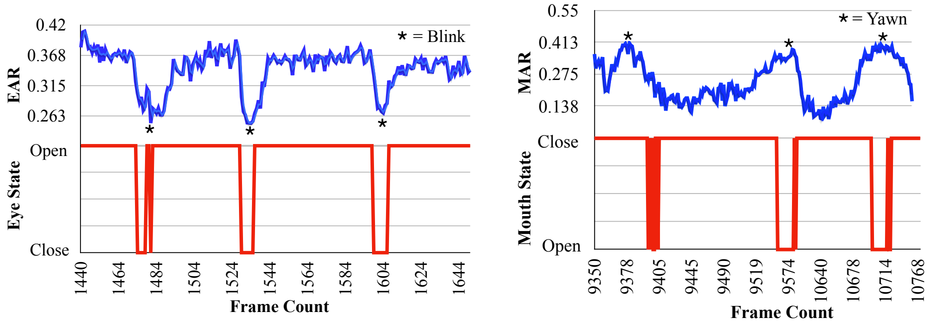
where p_37 , ..., p_42 are the eye landmarks. The EAR values of both the eyes are averaged to get the final measure of EAR. The value of EAR drops drastically (close to zero) when the eyes are closed.

InSight-**specific extension for detecting Blinks:** The measure of EAR varies from person-to-person depending on how wide a person opens her/his eyes. We present two approaches to define EAR threshold for detecting blinks:

(i) Traditional fixed calibration: One way to derive EAR threshold is to measure EAR values for a specific driver during a calibration phase, say the first few minutes of the drive and use this mean value as a hard threshold to detect when the eye is closed. However, this method relies on the assumption that the average EAR will remain the same for all types of facial expressions, and gaze directions, leading to false positives as demonstrated in [65].

(ii) Proposed auto-calibration (Moving average): To overcome the issue of fixed EAR threshold, we detect blinks using a moving average of EAR combined with standard Z-score [72]. The idea is that when a person blinks, his/her EAR value drops significantly from the average EAR values. Thus, if the average EAR value is beyond 3 standard deviation, i.e., negative peak, then we detect it as a blink. This way there is no hard threshold for each person and the system automatically adjusts the threshold based on the mean EAR values making it robust to

across person and facial expressions. Figure 3a shows the EAR and eye state derived using the proposed moving average based EAR threshold. In Section 9.1 we compare the accuracy of detecting blinks using both the fixed threshold and moving average.



(a)

(b)

Fig. 3. (a)Detecting Blinks using EAR Waveform, and (b) Detecting Yawns using MAR Waveform.

PERCLOS: We compute PERCLOS, which is the ratio of the number of frames in which eyes are closed to the total number of frames in a minute. When the value of PERCLOS exceeds 30%, the driver is considered to be in a drowsy or fatigued state [71].

3.3.2 Yawn Detection. Frequent yawning is a strong indicator of fatigue. Prior work has employed face and mouth detection techniques based on the Viola-Jones algorithm [68]. They compute a histogram of pixel intensity in the driver's mouth region, which is then matched with a template to detect yawns [15]. While these techniques are fast, they perform poorly in the real-world due to the variation in illumination and head pose. We present an alternative approach to detect yawns based on the landmarks.

Mouth state: Eight landmarks represents the mouth region (Figure 3b). Inspired by EAR, we propose a metric called Mouth Aspect Ratio (MAR) that computes the ratio between the height and width of the mouth:

$$MAR = \frac{||p_{68} - p_{62}|| + ||p_{67} - p_{63}|| + ||p_{66} - p_{64}||}{3 \times ||p_{65} - p_{61}||} \tag{2}$$

where p_61, ..., p_68 are the landmarks corresponding to the upper and the lower lip. The value of MAR is close to zero when the mouth is closed and a much larger non-zero value when it is open.

`InSight`**-specific extension for detecting Yawns:** A yawn is detected when the mouth is open continuously for a prolonged period (say 0.5-2 seconds [31]). Like EAR, the MAR metric depends on the person (e.g., because of variation in how wide the mouth is opened). Therefore, to detect yawns, we monitor the value of MAR and use the same moving average method as used for detecting blinks to also detect yawns. A yawn is detected when there is a positive peak in the moving average of MAR values as shown in Figure 3b.

3.4 Estimating Driver's Gaze

Estimating the gaze direction of the driver can provide essential information about the attentiveness of the driver. Specifically, mirror scanning helps drivers maintain situational awareness of their surroundings and we can estimate if the driver is actively scanning the mirrors of the car while driving or is distracted. We formulate the problem of estimating the gaze direction of the driver as a head-pose estimation and classification problem.

We use an iterative method based on a Levenberg-Marquardt optimization provided in OpenCV [56] to solve the Perspective-n-Point problem [47] from the provided 3D-2D point correspondences to determine the head-pose of the driver. The method finds a pose that minimizes re-projection error, which is the sum of squared distances between the observed projections (facial landmarks) on the 2D image and the projected 3D points corresponding to these facial landmarks. Therefore, the output of this method is the yaw and pitch angles corresponding to the head pose of the driver.

`InSight`-**specific extension for Mirror Scan Detection:** Although the driver's gaze is not restricted in an uncontrolled setting, there are three frequently (re-)occurring states: (1) Left-mirror scan, (2) Right-mirror scan, and (3) Straight gaze (driver focusing on the road straight ahead). This is evident in Figure 4 which shows the gaze angle distribution for the first few minutes of the drive. There are three high density regions corresponding to the three aforementioned recurring states. Empirically, the high-density region closest to the mean yaw value corresponds to "Straight". To delineate "Left" and "Right" more clearly, we first remove the "Straight" region by removing the frames with $|y(t) - \mu_y| \leq \sigma_y$, where $y(t)$ is the gaze angle, μ_y is the mean yaw and σ_y the standard deviation. We then apply a moving average technique (similar to blinks and yawns), where frames with yaw values greater than three standard deviations is classified as either Left or Right. Thus eliminating any fixed thresholds to classify the mirror scan states.

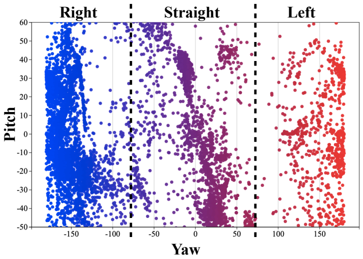


Fig. 4. Gaze distribution.

4 MONITOR THE STATE OF DRIVER IN LOW-LIGHT USING SPECIALIZED CAMERAS

In this section, we present existing approaches to monitor the state of the driver in low-light conditions using specialized cameras, *viz.,* IR and FLIR cameras.

4.1 Infrared Sensitive (IR) Cameras

Most high-end vehicles with ADAS support employ an IR camera for monitoring the driver in low-light conditions. IR cameras are equipped with an array of IR LEDs (940nm). Typical IR sensitive cameras does not include an IR filter, thus letting more IR light enabling clear vision of the subject under low-light conditions. Figure 5 shows an IR dashcamera and some sample images obtained using the setup in the vehicle.

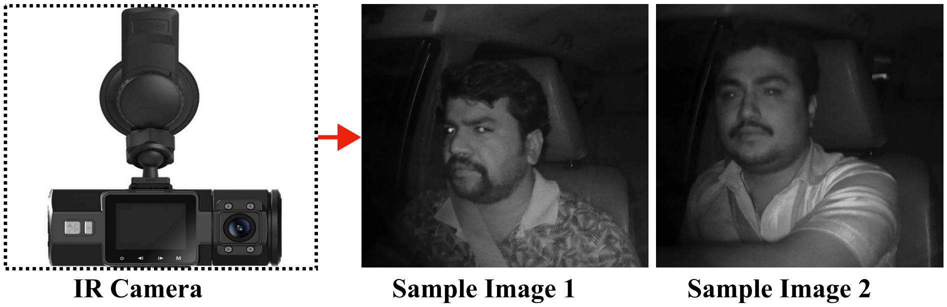


Fig. 5. IR camera setup along with few sample images.

Typical, IR camera imagery looks like black and white images due to the monochrome filter. Further, the existing computer vision algorithms for face detection and facial landmarks are trained on both RGB, and black and white images making it to work directly on the IR imagery. Previous work on using cameras with IR illuminators [17] for monitoring in low light has shown the efficacy of IR cameras. These systems perform very well even in complete darkness and their accuracy is significantly higher (>90%) as compared to other methods for monitoring in low light environments.

Despite their superiority in terms of performance, IR cameras are not a cost-effective choice besides the fact that this involves installation of an additional hardware setup in the vehicle specifically for monitoring driver behaviour in low light. *Hence, in the remaining of the paper, we use IR camera imagery for deriving only ground truth labels for other approaches such as FLIR thermal and NIR imagery.*

4.2 Forward Looking Infrared (FLIR) Thermal Cameras

A natural alternative to RGB images is to consider thermal imaging, especially since our subject of interest is a human face, with a distinctive heat signature. The key observation for selecting thermal image is that, it is largely invariant across good and poor lighting conditions and captures all the face attributes accurately.

Working. Heat radiations are emitted by the human body, however, since these radiations have a much longer wavelengths and very low frequencies compared to the visible spectrum, the human eyes cannot see these far-infrared radiations. Thermal cameras, on the other hand, are designed to capture these radiations and generate a heat map or a thermal image. This thermal image doesn't depend on the lighting conditions and we leverage this fact by making use of a thermal imaging camera to monitor drivers in low light. The major advantage of using thermal cameras to monitor drivers is that these are non-intrusive, passive devices which mean that there is no active emission of radiations and the camera only captures the heat radiated by a human body to generate thermal images.

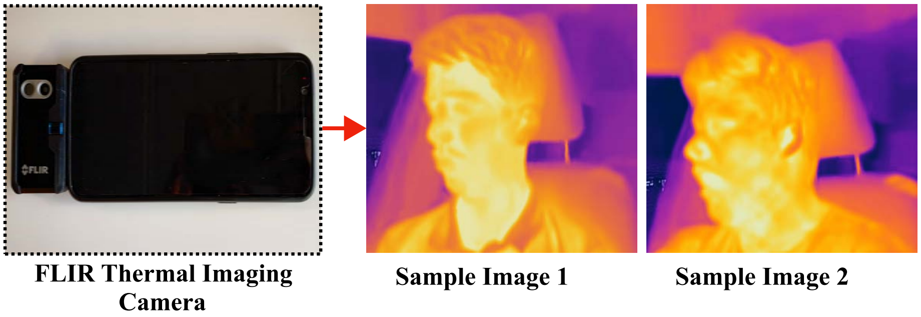


Fig. 6. FLIR thermal camera setup attached to the smartphone along with few sample images.

Setup. To obtain thermal images we use the FLIR One Pro Thermal Camera [66], which can be attached to the smartphone as shown in Figure 6. All the recorded data is stored on the smartphone itself. The FLIR One Pro is also equipped with a visible camera therefore allowing one to record both the thermal and RGB images at the same time. Figure 6 shows sample images obtained using an FLIR (Forward-Looking Infrared) camera. However, such cameras are costly (> $300) and the thermal imagery obtained by these cameras are of low resolution, leading to inaccuracies in detection of various facial features, specifically eye related features.

In Section 8 we present detailed evaluation of FLIR thermal imagery for monitoring the state of the driver in low-light conditions.

5 MONITOR THE STATE OF DRIVER IN LOW-LIGHT USING IMAGE SYNTHESIS TECHNIQUES

Since installation of specialized camera is intrusive and expensive, we present a cost-effective image synthesis approach that uses only the smartphone camera. Before describing the synthesis approach, we also applied few traditional image enhancement techniques to improve the quality of low-light images as described next.

5.1 Image Enhancement Using Exposure Compensation

Low illuminance leads to the scene being underexposed, which is the root of the problem in low-light RGB imagery.

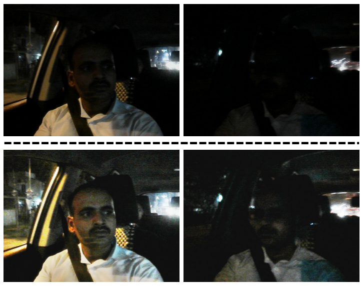


Fig. 7. Images in the bottom row are obtained using exposure compensation with top row images as input.

One way of addressing this is to employ exposure compensation to improve visibility [1, 22, 77]. Exposure compensation applied on low-light RGB images slightly improves the information level of low-light RGB images. To quantify the improvement, we conducted an experiment with 1000 low-light RGB images, which were processed with exposure compensation. The processed images were then fed to the standard face detectors, where faces were detected in only 98 out of 1000 low-light RGB images (<1%). This is mainly due to the noise introduced by adjusting exposure, especially, if there is any light source (e.g., reflected light on driver's face) it leads to overexposure and noise. Figure 7 shows few samples of low-light RGB images (top row) and the corresponding images with exposure compensation (bottom row). We can see that exposure compensated images look slightly better than the original images, but still not sufficient to detect face and landmarks accurately. Thus eliminating the usage of any image enhancement techniques.

5.2 Custom Model to Derive Landmarks Directly from Low-light RGB Images

Before we turn to image synthesis approach, we present another approach where we develop a custom model, which takes input as low-light RGB images and outputs face detection and facial landmark directly on a low-light RGB image. In order to develop such a model, it requires significant amount of training data, i.e., 1000s of low-light RGB images labelled with face detection and facial landmarks. We are not aware of any public dataset with 68 facial landmarks annotated on low-light RGB images; indeed, it is challenging to annotate low-light images due to lack of information available on RGB images.

To overcome this difficulty, we employ a three-step process to generate labelled low-light RGB training data. We first employ an FLIR camera to capture thermal images in tandem with low-light RGB images. We then use the thermal landmark model to estimate the facial landmarks in the thermal image. Finally, we map these landmarks onto the low-light RGB image at the corresponding pixel positions (since the FLIR and RGB images are aligned). Thus, we are able to generate training data comprising low-light RGB images annotated with 68 facial landmarks.

With the generated training data, we train a low-light landmark model using the Dlib library, for which the input is a set of low-light RGB images, each annotated with 68 facial landmarks. We evaluated this low-light model on a test set of low-light RGB images and found that it detected landmarks only in 47.7% of the frames, with a landmark localization error, i.e., Normalized Mean Square Error (NMSE) of 0.21. NMSE quantifies the misalignment in the landmarks obtained on the low-light RGB and FLIR thermal images (i.e., the ground truth). The landmark localization error (NMSE) is larger than that considered acceptable, i.e., 0.1-0.15. The reason for this poor accuracy is that the low-light RGB image often contains large regions of dark pixels bearing little information, so annotating these with landmarks still does not enable a landmark model to learn effectively.

5.3 Synthesising Thermal Images from RGB Images.

As described in Section 4, a downside of using specialized cameras such as IR or FLIR cameras is the cost associated with the hardware. Since an IR or thermal image is agnostic to lighting changes and is easy to capture, we argue that this is an ideal intermediate representation. Furthermore, recent advancements in computer vision and deep learning has enabled realistic image-to-image translation. To this end, we present a technique that employs Generative Adversarial Networks (GANs) [33] (see below for elaboration) to synthesize a thermal image using just an RGB image captured using the smartphone camera. This paves the way for the detectors to run on RGB images from a smartphone obtained in low-light conditions, without requiring an FLIR thermal camera or other additional sensor attachment. We employ a state of the art framework called pix-to-pix [40] to translate an input RGB image to a synthesised thermal image. Note that, the same approach can be utilized to synthesize a IR image from a low-light RGB image. In this paper we restrict image synthesis to only thermal images.

Intuition. Before getting into the details, we briefly present the intuition behind this approach and why it works. In a typical low-light setting, the driver's face is partially illuminated, with the rest being in a shadow. Even with a partial view, say just the position of one eye or the nose or the chin, a human observer can, nevertheless, form a rather accurate picture of the head pose or whether the driver is yawning. What aids in this process is interpolation and extrapolation based on facial symmetry and other such high-level notions.

Computer vision community has shown that GANs are a powerful learning framework that can perform such interpolation effectively [40, 83]. Hence, we use a GAN to synthesize accurate thermal images based just on low-light RGB images. Once this is done, a relatively simple landmark model such as Dlib can function effectively. We believe that this split of functionality between a powerful GAN and the relatively simple landmark is appropriate from the viewpoint of the overall system. Specifically, it is relatively easy to obtain training data for the former (using an FLIR camera to capture thermal and RGB images in tandem) but much harder for the latter (it would require significant manual effort to annotate low-light RGB imagery as mentioned earlier). Finally, the structure provided by having the synthesized thermal image as an intermediate step would help improve accuracy, specifically in the context of landmark detection. Furthermore, it would likely be advantageous to have an informative intermediate representation for low-light RGB imagery, that is then amenable to various applications such as face recognition and pose estimation, which is beyond the scope of this paper.

Working. A GAN consists of two adversarial models — a *generator* and a *discriminator*. The generator (G) applies a transformation to the input image to generate an output image. The discriminator (D) outputs a single scalar value representing the probability that the output image came from the (real) training data rather than the (synthetic) generated data. The idea of a GAN is that the generator and discriminator networks would "compete" with each other, to reach an equilibrium. If the discriminator is able to tell that the generator's output is synthetic, that would result in a poor adversarial loss for the generator. This, in turn, would spur the generator to steadily improve over time and synthesize ever more "realistic" images, so much so that eventually the discriminator is unable to tell that these are synthetic.

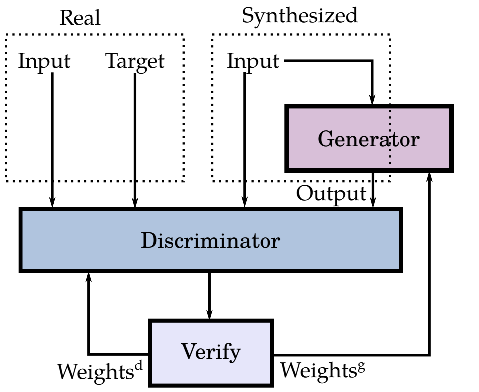


Fig. 8. GAN to synthesize thermal image. Input: low-light RGB, Target: FLIR thermal, Output: Synthesized thermal.

Training. Figure 8 shows the training process for the GAN. The generator and discriminator are trained using the input-target pairs i.e., RGB-FLIR thermal images. First, the generator generates an output image (synthetic thermal image) for the given input (real RGB image). The discriminator, which is trained with pairs of input images and the corresponding target thermal images, determines the probability that the generated output image is real, i.e., it corresponds to the input RGB image. The weights of the discriminator network are then adjusted based on the classification error. The weights of the generator network are also adjusted based on the output of the discriminator. This way the network calculates the gradients through the discriminator, and consequently, the generator learns the correct mapping function through gradient descent. Thus, as the discriminator gets better at telling between real and synthetic thermal images, so does the generator in synthesizing more realistic thermal images.

Testing. The trained network can now synthesize a thermal image that looks similar to the target thermal image based on a low-light input RGB image. We have also implemented this network on to a smartphone app to synthesize thermal images (Section 10.1). In Section 8 we show the efficacy of the synthesized images towards fatigue and distraction detection.

6 MONITORING THE STATE OF THE DRIVER IN LOW LIGHT USING SMARTPHONE CAMERAS AND NEAR-INFRARED LEDS

We now present an alternative low-cost approach to image synthesis that relies on a Near-IR (NIR) LED attached to the smartphone.

6.1 Working of NIR LED

Majority of the cameras are designed to capture an image of what humans can see and hence a good quality camera would only detect light in the visible spectrum, i.e., between 400nm to 700nm and block out other infrared light. However, camera sensors based on silicon including CCDs and CMOS sensors used in smartphones have sensitivies extending into near-infrared ranging from 700nm to 1000nm. Thus, camera manufacturers introduce an IR filter to block out this IR light. Furthermore, developing an IR filter that cuts off sharply at one particular wavelength is challenging and expensive. For instance, even in a very good quality high-end camera such as DSLRs the IR filter does not block 100% of the IR light [39].

On the other hand, smartphone cameras are usually mass-produced cheaply and are not as good as a high-end digital camera. Hence, the majority of smartphone cameras that are commercially available have a much thinner film/filter to block out infrared light. The lack of good filter is one of the reason that phone photographs do not look as good as when taken on a proper digital camera. This however, provides us with an opportunity to use our smartphone cameras to "see" in near-infrared light. Thus, all most all smartphone cameras that are commercially available let near-IR light pass through [69]. To this end, we conducted an experiment with 8 smartphone models, namely, OnePlus 3T, OnePlus 5T, OnePlus 7T, Nexus 5X, Moto G4, iPhone 10, Nokia 8, and Lenovo Zuk Z2, ranging from 150 USD to 700 USD. We found in all the smartphones we were able to accurately capture the NIR imagery with an LED operating at 810nm wavelength. Thus, standard smartphone cameras can capture NIR imagery without additional hardware.

We leverage this key insight to design our setup, which includes an NIR LED attached to the smartphone, where the NIR LED acts as an light source to illuminate the driver's face, which is not visible to the human eyes.

Choice of NIR LEDs. The choice of NIR LED depends on two factors: (i) the wavelength and (ii) the intensity. Typical NIR LEDs wavelength starts from 750 nm to 1000 nm. Figure 9 shows images captured by the smartphone when NIR LEDs emitting light at 750nm, 780nm, 810nm, and 850nm wavelengths was used, respectively. The closer the wavelength of the NIR LED is to visible range, the more the red color is visible in the emitted light, making the setup intrusive as it is easily perceived by the user. In our setup, we selected 810nm wavelength NIR LED due to its non-intrusiveness and it is not perceived by human eyes.

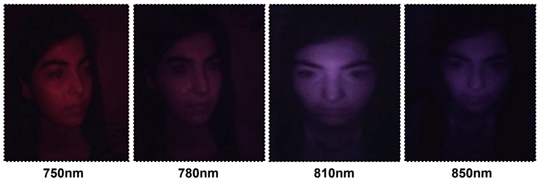


Fig. 9. NIR imagery with different Near-IR LED wavelengths.

The second factor affecting the choice of NIR LED is the intensity of the LED. Intensity describes the luminous flux per unit solid angle in a given direction from a point source. Typical NIR LEDs can be rated from 1W to 5W. The higher the intensity, the more area the NIR light illuminates (thus increasing the illuminated distance), and also resulting in higher power consumption. Hence it is important to select the correct intensity level to ensure both the driver's face is properly illuminated and the power consumption is kept low.

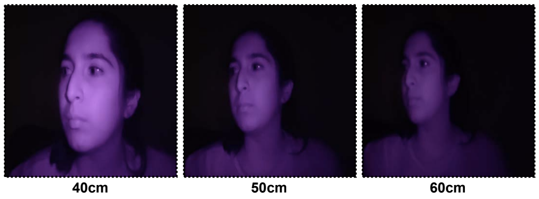


Fig. 10. NIR imagery captured at different distances.

Figure 10 shows images captured by the smartphone at distance 40cm, 50cm and 60cm for NIR LED with 810 nm wavelength. As we can see the shorter the distance, the more clear is the captured NIR images. Given the smartphone will be mounted on the wind-shield of the vehicle, the distance between the LED to the driver's face is typically around 40cm-60cm. After extensive experimentation we selected an NIR LED emitting light at 810 nm [12] with 1.8W intensity, which is not visible to human eyes but can be captured by the smartphone camera effectively. This also adheres to the safety guidelines for usage of NIR LEDs as described in Section 10.3.

6.2 InSight NIR LED setup

Typical IR cameras come with an array of IR LEDs, making the setup expensive. In InSight, we use just two 810nm NIR LEDs connected in parallel with a smartphone. To ensure optimal working of the LED, it needs to be powered using a constant current source. The current regulator ensures the same fixed current flows through the LED, thus regulating the power drawn by the LED (otherwise leads to burning the LED) and also ensures no fluctuations in the NIR light emitted by the LED. To this end, we created a small circuitry where NIR LED is powered by a constant current regulator and a Li-ion battery. Figure 11 shows the NIR LED setup, which comprises of a NIR LED emitting light in the range of 810 nm, a current regulator supporting 300mA-500mA of current and a 3.7V Li-ion battery.

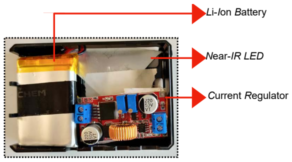


Fig. 11. NIR LED and its associated components [back view].

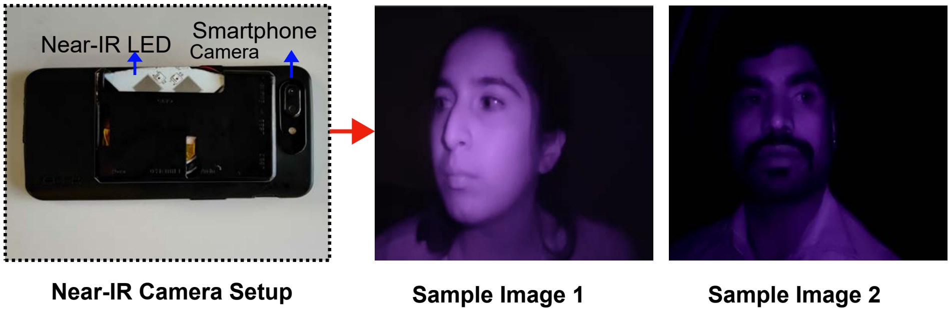


Fig. 12. NIR LED setup [front view] attached to the smartphone along with few sample images.

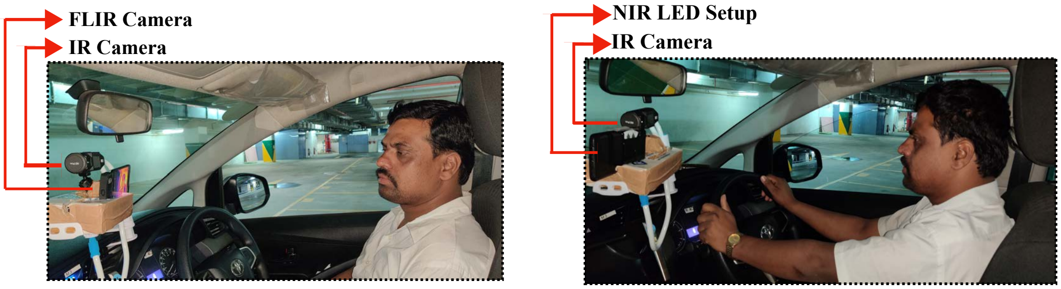
Figure 12 shows the NIR LED circuitry attached to the smartphone along with the sample images captured by the smartphone. We can clearly see the quality of NIR imagery is significantly better than raw low-light RGB images. Thus, without additional camera and with minimal electronics and NIR LEDs, we can capture high quality videos in low-light. This setup is non-intrusive (driver does not notice the near-IR light) and also low-cost. In Section 8 we show the evaluation of NIR imagery towards fatigue and distraction detection.

7 TRAINING FACE DETECTION AND FACIAL LANDMARKS MODEL FOR THERMAL, SYNTHESIZED THERMAL AND NIR IMAGERY

As described earlier, the state of the art face detectors and facial landmark extractors work only with good lighting RGB imagery or IR imagery. These models are not designed to work on NIR or thermal images as these images are significantly different leading to a different set of feature maps as compared to features obtained using standard RGB imagery. In this section we present mechanisms to train face detectors and landmark models that work with thermal and NIR imagery. We now present an automated approach to derive ground truth labels on NIR and thermal imagery, which can be used to train the detectors accurately.

7.1 Deriving Ground Truth Labels in Low-light Conditions

Since the existing face detector and facial landmark predictor models are able to get accurate face and facial landmark coordinates on video frames recorded with IR cameras in low-light conditions, we decide to use labels from IR imagery as the ground truth labels. In order use labels from IR imagery as ground truth, we need to translate the labels from IR imagery captured from dash camera shown in Figure 5 to smartphone camera imagery with NIR LED. To this end, during training data collection we mount both the IR camera and the smartphone camera with NIR LED on the dashboard of the vehicle and then we transfer the labels from the image plane of one camera (IR) to the other (smartphone). Figures 13 shows the mounting of IR-FLIR setup and IR-NIR setup to collect data and associated ground truth information.



(a)

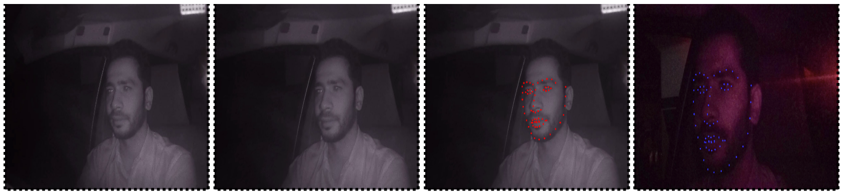
(b)

Fig. 13. Setup for ground truth data (a) IR and FLIR camera setup, and (b) IR and NIR camera setup.

We formulate this as a perspective transformation problem which is solved using Homography in computer vision [49, 67]. A Homography is a transformation (a 3×3 matrix) that maps the points in one image to the corresponding points in the other image. If we have points (x_i, y_i) in the IR image, the corresponding point (x_j, y_j) in the image captured from the smartphone can be obtained as:

$$s \begin{bmatrix} x_j \\ y_j \\ 1 \end{bmatrix} = H \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} \; and \; H = \begin{bmatrix} h_{00} & h_{01} & h_{02} \\ h_{10} & h_{11} & h_{12} \\ h_{20} & h_{21} & h_{22} \end{bmatrix} \tag{3}$$

To calculate the homography between two images, at-least 4 point correspondences are required between images captured from the IR camera and the smartphone camera. Therefore, we record a short video in good lighting condition before we begin to collect the real driving dataset in low light. In this video, the facial landmarks are recorded simultaneously by the IR camera and the smartphone camera in good lighting, leading to 68 point correspondences between IR and smartphone camera. This correspondences are then used to create a homography matrix. Note that, this is a one-time calibration to ensure we align the IR and smartphone imagery accurately.



IR image

Undistorted IR Image

Landmarks on Undistorted IR Image

Transformed Landmarks on NIR

Fig. 14. Distorted and undistorted IR images along with label translated NIR image.

The calculated homography matrix can now be used to map the coordinates obtained in IR images to the images captured by the smartphone camera in low light. In addition, since the IR dashboard camera that we used is a wide-angle fish-eye camera, we also perform an additional step of calibrating the camera and removing distortion [81] from the IR images before obtaining a translation of the 68 landmarks. Figure 14 shows distorted and undistorted IR image along with the translated landmark labels to smartphone camera images with NIR LED.

We can clearly see the accurate translations of landmarks from IR image to smartphone camera image with NIR LED. *Thus, we use this setup to derive ground truth labels for both NIR and thermal imagery.*

7.2 Models for Thermal and Synthesized Thermal Imagery

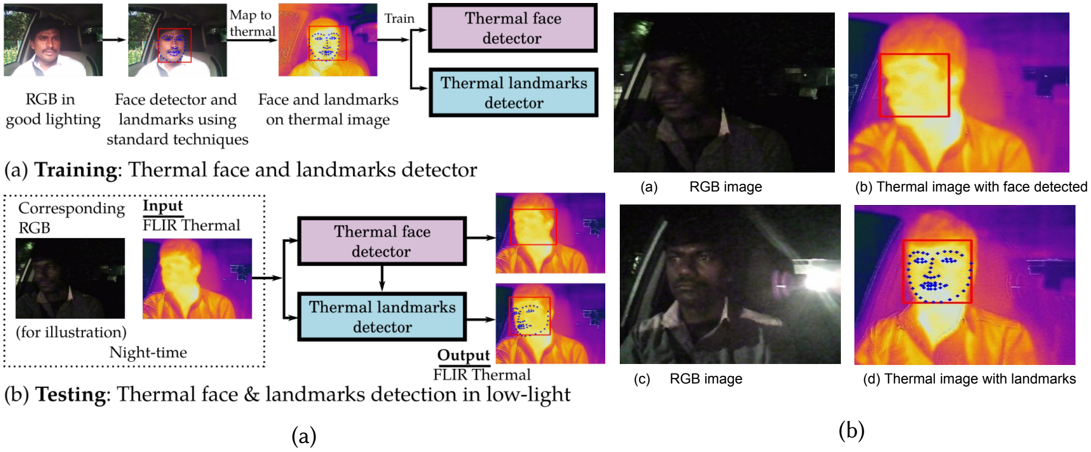

Map to thermal
Train
Thermal face detector
Thermal landmarks detector
RGB in good lighting
Face detector and landmarks using standard techniques
Face and landmarks on thermal image
(a) **Training**: Thermal face and landmarks detector
Corresponding RGB
Input
FLIR Thermal
(for illustration)
Night-time
Thermal face detector
Thermal landmarks detector
Output
FLIR Thermal
(b) **Testing**: Thermal face & landmarks detection in low-light
(a)
(a) RGB image
(b) Thermal image with face detected
(c) RGB image
(d) Thermal image with landmarks
(b)

Fig. 15. (a) Facial landmarks on FLIR thermal images, and (b) Face detection and facial landmarks on FLIR thermal image.

We now explain how to collect training and test data for developing models for thermal images. A key insight we leverage is that the thermal image of a face is largely unchanged across good lighting (e.g., daytime) and poor lighting (e.g., nighttime) conditions. Therefore, we can use thermal images, along with the corresponding RGB images, gathered during daytime for training and then apply the resulting model to thermal images from nighttime for detection. Indeed, our FLIR cameras [9] allow images to be captured in tandem (i.e., simultaneously and spatially aligned) using the thermal and RGB sensors.

Training. To develop a robust thermal face detector and landmark predictor, we first collect RGB and thermal images in tandem in daytime conditions across multiple drivers. A standard face detector and landmark predictor [27] as described in Section 3 is used to detect faces and corresponding landmarks in the daytime RGB image. This information is then transferred on to the corresponding FLIR thermal image at the same pixel positions since both the RGB and the thermal images are aligned. The FLIR thermal images, with face detection and facial landmarks transferred from RGB images, serve as our training data.

We train a face detector and landmark predictor with the thermal images. The trained thermal face detector and landmark predictor together are used to detect faces and landmarks on test thermal images obtained in low-light conditions. Figure 15a depicts both the train and test process. Figure 15b shows the low-light RGB image and its corresponding thermal image from FLIR camera along with the estimated face and facial landmarks using the trained models.

Also, we use the same trained thermal face detector and landmark predictor for synthesized thermal images as both the images appear the same. In Section 8 we present detailed evaluation of the face detectors and facial landmarks for both thermal and synthesized thermal images with labels from IR data as ground truth.

7.3 Models for NIR Imagery Captured from the Smartphone

NIR imagery captured using smartphones is significantly different than images captured using IR cameras. Hence, existing models for face detection and landmarks trained on RGB/IR images will not work on NIR imagery.

NIR to Grayscale Conversion.

Since standard detectors convert RGB/IR imagery to grayscale before detection, one approach is to convert NIR images to grayscale and verify the detection accuracy. However, the features extracted for an NIR image converted to grayscale is significantly different than the features obtained from RGB/IR images used to train the model. This is due to the low contrast and significant change in lighting conditions in the low-light NIR imagery. We also performed several image enhancements to improve the contrast, but still the standard detectors performed poorly. To validate this we conducted an experiment with grayscale images as described next.

Figure 16 shows the original image (RGB, IR, NIR), corresponding grayscale converted image and detection on grayscale with standard detector.

(i) Good lighting RGB image converted to grayscale: We fed 100 RGB to grayscale-converted images to a standard face detector (DLIB face detector [43]). On all 100 images the standard detector was able to detect faces. A sample image is shown in the first row of the Figure 16.

(ii) IR image converted to grayscale: Similar to RGB scenario, the standard detector detected all the faces in 100 IR to grayscale converted images. This is mainly due to the fact that these detectors were trained on RGB and IR images, and both have good contrast variations making the detection accurate. A sample IR image and its corresponding grayscale image is shown in second row of Figure 16.

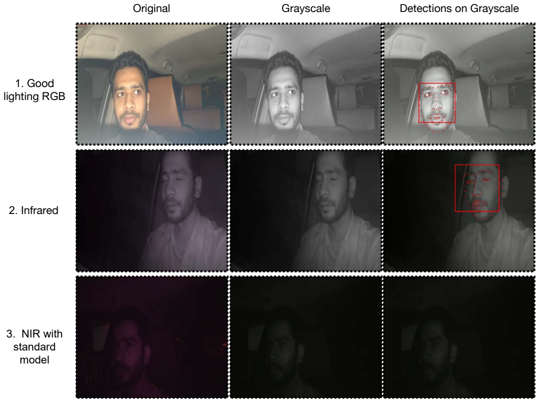


Fig. 16. Detection on grayscale images using standard models.

(iii) NIR image converted to grayscale: We applied the same technique as before where we fed 100 NIR to grayscale converted images and in that faces were detected on only 12 images using the standard detector. This is mainly due to the significant change in features extracted due to lighting and contrast variations. Last row of Figure 16 shows the NIR image, its corresponding grayscale image and grayscale image with no detection using standard detector. We can see that this grayscale image has much lower contrast than the grayscale image obtained using RGB or IR imagery.

In a typical setting, RGB to grayscale conversion assigns higher weightage to the green channel. However, in extremely low light conditions, where NIR LEDs are the only source of illumination, the information captured by the red channel is more compared to the green and blue channels. Hence assigning more weightage to the red channel during conversion might lead to improved low-light image. However, on real-roads there is typically a much wider variation of light conditions and it is the green channel that still captures most of the ambient light. Hence, assigning higher weightage to one channel does not lead to significant improvement in NIR image processing. Thus necessitating development of a new model as described next.

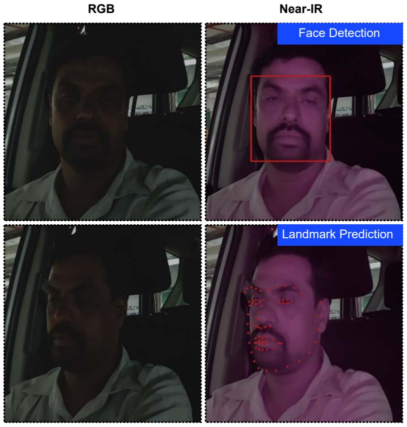


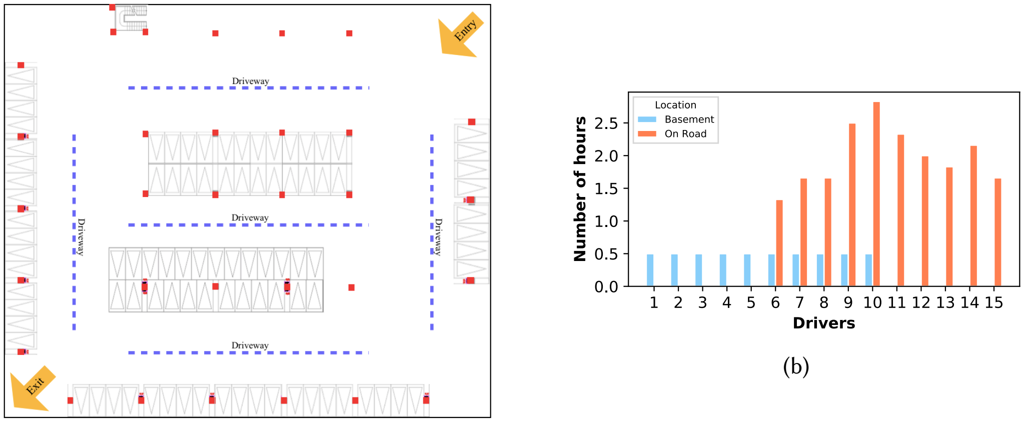
Fig. 17. Face detection and facial landmarks on NIR image from smartphone.

NIR model Training. To train the face detector and facial landmark models for NIR images captured by a smartphone, we used the setup shown in Figure 13b. The setup includes an NIR LED attachment to the smartphone camera along with the IR camera. The labels from the IR camera is then translated to the NIR images as described in Section 7.1. Both the cameras recorded the driver's activities simultaneously at a resolution of 1080p and a frame rate of 30fps. Thus, using the techniques mentioned earlier, we now generate a dataset of labelled (face detection and landmarks) NIR imagery, which is used to train the state of the art models (described in Section 3).

Figure 17 shows the RGB image in low-light condition and the NIR image captured by the smartphone along with face bounding box and facial landmarks using the trained model. In the next section we present detailed evaluation of the face detectors and facial landmarks for NIR images with labels from IR data as ground truth.

8 EXPERIMENTAL SETUP AND RESULTS

We present detailed evaluation of `InSight` and report accuracy of the trained detectors on the two proposed approaches (image synthesis and NIR LED) for monitoring the state of the driver in low-light conditions.



(a)

(b)

Fig. 18. (a) Basement floor plan, and (b) Data distribution per driver in basement and on-roads.

8.1 Setup and Data Collection

Our setup includes an Android smartphone (OnePlus 5T running Android 8) that is mounted just below the centre mirror. We gather imagery from the smartphone camera (driver-facing) at 1080p and 30 fps. For thermal imagery, we attach a FLIR camera [10] to the smartphone and similarly, for NIR imagery we attach the proposed NIR LED setup. Further, for ground truth information we use the labels from the IR imagery obtained using IR dashcamera [13]. Thus our setup includes a IR camera (used for ground truth) and a NIR LED setup or FLIR setup as shown in Figure 13a and 13b. We collected low-light imagery in two environments, namely,

(i) Basement data collection: We collected 5 hours of IR, NIR and FLIR thermal imagery from 10 different drivers in the basement. Each driver was asked to drive for 30 minutes duration in the basement of a building, which replicated the low-light conditions. Typically in the basement the illuminance was around 30-40lx near the driver's face indicating low-light conditions. The reason for collecting this data in a basement was to get wide-variety of data in a safe environment. The basement parking area is of 3740 m^2 with length of 68 meters and width of 55 meters, and Figure 18a shows the floor map of the basement with red icons indicating the pillars present in the basement and triangles indicating the parking spots.

(ii) On-road data collection: We also collected 20 hours of data on real roads across 10 drivers in the nighttime for robust evaluation of the proposed techniques. The vehicles in our deployment, is part of an office fleet, which shuttles employees from office to their house. The on-road data was collected between 8 PM to 4 AM across all drivers, which represents low-light conditions. Since the data collection was on real roads, it captures realistic traffic conditions (such as relatively high traffic during 8PM to midnight, sparse traffic conditions after midnight to 4AM) along with natural low-lighting variations.

Overall, we collected 25 hours of data from 15 drivers, out of which 5 hours of data was collected in the basement with 10 drivers (D_1-D_{10}) and 20 hours of data on real roads from 10 drivers (D_5-D_{15}). Note that, for 5 drivers (D_5-D_{10}) we have data both in basement and on real roads. Figure 18b shows the distribution of data collected per driver in hours in basement and on real roads.

8.2 Model Training and Evaluation

We now discuss the details of the data used to train the face detection and landmark models, along with an extensive evaluation on the trained model.

8.2.1 Metrics. We first define the metrics used to evaluate the performance of the models.

Metrics - Face detection. We use the Precision (P), Recall (R), and F1 score to evaluate the accuracy of face detection and is defined as:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{4}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{5}$$

$$F_1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{6}$$

Metrics - Facial landmarks. To evaluate the accuracy of the predicted landmarks we compare the misalignment between the landmarks obtained on the IR imagery and FLIR thermal images (or synthesized thermal or NIR images) using Normalized Mean Square Error (NMSE). NMSE is the average point-to-point Euclidean error normalized by the inter-ocular distance. Inter-ocular distance is measured as the Euclidean distance between the outer corners of the eyes, which ensures the accuracy measure is independent of the face size. NMSE is a standard metric to compare the accuracy of landmarks and an error under 0.15 is considered as acceptable [20].

8.2.2 Model Evaluation. We train a single face detector and facial landmark model for all drivers. To this end, we aggregated data from all the drivers and split it into train and test data. Further, we used data only from the basement for training and testing. We conducted thorough evaluation to determine the efficacy of the models as described next.

(i) 10-fold cross validation (10-fold CV): We sub-sampled 1000 images for each of the 10 drivers in the basement dataset for training the face detection and landmark models. In each iteration the data is split equally into 10 parts, with 9 parts used for training and the remaining 1 part used for testing. This technique shows the capability of the model performance on unseen data.

(ii) Leave one out cross validation (LOOCV): In this case, among the 10 driver's data, we excluded one driver data and used the remaining 9 driver data (1000 images per driver) for training the model and the excluded driver data (1000 images) is used for testing. We repeated this for each of the 10 drivers and then reported the average results across all the iterations. This shows generalizability of the trained model on a new driver data, which was previously unseen.

(iii) Custom validation: In this case, for training, we used a subset of data from 9 drivers (1000 per driver) and for testing we used unseen new data from the 9 drivers who are part of the training (2000 images per driver) along with a new driver data (2000 images). This method verifies the model efficacy on both unseen data of the driver used in training and unseen data from a new driver.

We apply the above model evaluations for both face detection and facial landmarks with NIR and FLIR thermal images as described next.

Table 1. Face Detection Evaluation.

Imagery	Evaluation	Precision	Recall	F_1-Score
FLIR	10-fold	99.7	70.5	77.9
	LOOCV	99.8	82.2	86.9
	Custom	99.8	47.4	64.3
NIR	10-fold	99.0	80.0	88.4
	LOOCV	98.9	78.7	87.6
	Custom	98.3	88.2	92.9

Table 2. Facial Landmark Prediction Evaluation.

Imagery	Evaluation	NMSE
FLIR	10-fold	0.08
	LOOCV	0.10
	Custom	0.13
NIR	10-fold	0.03
	LOOCV	0.07
	Custom	0.08

8.3 Face Detection and Facial Landmark Results on NIR and FLIR Imagery.

8.3.1 Face Detection Results. Table 1 shows the precision, recall and F_1-Score of face detection models with NIR and FLIR thermal images for 10-fold, LOOCV and custom training scenarios, respectively. In general. we see that precision for both FLIR thermal and NIR models are over 98%, thanks to very few false positives in both the cases. However, FLIR thermal model fails to detect faces in many instances in low-light conditions leading to a poor recall and hence a low F_1-Score. Upon inspection, we reason that this is because thermal images are sensitive to body temperature which varies from person to person and also depends on the temperature inside/outside of the vehicle. Since it is hard to capture all these variations in a training set, the trained model is not be able to detect faces accurately across all frames. On the contrary, NIR model is not dependent on such intrinsic features and thereby, more robust in detecting faces under low-light conditions.

From Table 1, we can also see that NIR model performance with 10-fold performs similar to LOOCV, with 10-fold models having slightly better accuracy. This is mainly because the 10-fold trained model includes data from all the drivers during training and testing is done only on new unseen data from the same drivers included in the training. Whereas, in LOOCV the trained model does not have any data of the test driver. In general, both 10-fold and LOOCV models have high accuracy showcasing the efficacy of the trained models on unseen data.

8.3.2 Facial Landmark Results. Deriving accurate facial landmarks is key towards detecting fatigue and distraction. We now present the results of facial landmark models on FLIR thermal and NIR data. As mentioned previously, we use NMSE metric which compares the misalignment of facial landmarks from FLIR thermal or NIR images to ground truth IR images.

Table 2 shows the NMSE for facial landmarks models with NIR and FLIR thermal imagery for 10-fold, LOOCV and custom training scenarios, respectively. We can clearly see that both FLIR thermal and NIR models have NMSE under the acceptable limits (i.e., under 0.15 [20]), with NIR model performing significantly better than FLIR thermal model. Upon manual inspection, we found that landmarks on FLIR images are erroneous especially around the eye region. This is because of uniformity in temperature surrounding eye region (see Figure 15b), leading to smooth eye patches and erroneous landmarks. Further, the NIR facial landmarks model performs accurately for both unseen data from same drivers and unseen data from new drivers.

8.4 Face Detection and Landmark Results on Synthesized Thermal Images

In order to train a GAN network, we need large amount of labelled data, specifically in this case, pairs of thermal and low-light RGB image. To this end, we created a dataset of 13000 thermal and low-light RGB image pairs from 10 drivers. Further, unlike labelling low-light RGB images, which is non-trivial, collecting pairs of thermal and low-light RGB imagery to train the network is relatively simple using the specialized FLIR camera and smartphone setup in tandem.

We used 8000 images from 8 drivers (1000 images per driver) for training the GAN network. The trained GAN model can now be used to generate synthesized thermal images using only low-light RGB image as input. In order to evaluate the performance of the synthesized thermal image on face detection and landmark models, we created a test dataset of 5000 images of which, 2000 images are from 2 new unseen drivers and 3000 unseen images from the 8 drivers included in the training. We used the thermal models trained previously to detect faces and facial landmarks on the synthesized thermal images. On our test data, the face detection **F_1-Score was 81%**. Further, we compare the misalignment of landmarks by computing NMSE. We see that even with large yaw values accurate landmarks were predicted on synthesized thermal images. **NMSE of 0.11** was found across all yaw values, which is under the acceptable error. Thus showing the effectiveness of the proposed thermal image synthesis approach to extract reliable landmarks in low-lighting.

Impact of Illuminance on synthesized thermal. We collected RGB and FLIR thermal images at various illuminance values. Figure 19 shows the RGB, synthesized thermal, and FLIR thermal images for varying illuminance values. When it is pitch dark, RGB images have little information and hence the network fails to synthesize a proper thermal image. However, when the illuminance is higher, which is typically the case during driving due to street lights, opposing vehicle headlights, etc., (e.g., it is rarely below 10 lx in our drives) the network is able to synthesize a realistic thermal image. The ability to accurately synthesize a thermal image from partially illuminated RGB images is mainly due to the powerful learning capability of GANs with respect to facial symmetry and interpolation based on the training data. To analyze the trade-off between variation in illuminance and accuracy of landmarks predicted, we collected images at 5 illuminance levels: <10 lx, 15-20 lx, 40-60 lx, 100-115 lx and 150-180 lx (as shown in Figure 19). The corresponding NMSE between Synthesized thermal and FLIR thermal are NA, 0.196, 0.121, 0.117, 0.09, respectively. As the illuminance improves, the NMSE decreases, indicating improved accuracy for the synthesized thermal images.

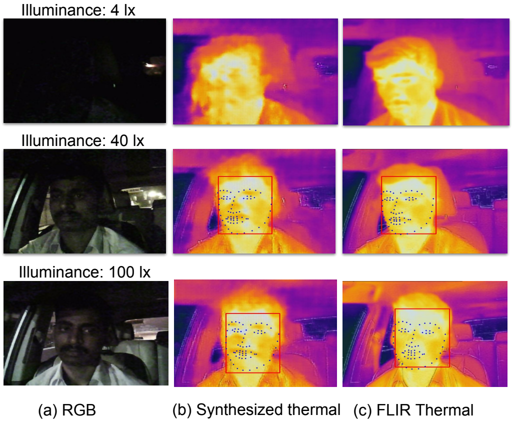


Fig. 19. Input RGB, synthesized thermal, and FLIR thermal images, with varying illuminance.

9 DRIVER FATIGUE AND DISTRACTION MONITORING

In the previous section we showed the efficacy of the trained models for FLIR thermal, synthesized thermal and NIR images. We now present detailed experimental results for driver fatigue and distraction monitoring using data from both basement and on-real roads.

Table 3. Number of groundtruth events.

	FLIR thermal	Synthesised thermal	Near IR
Blink	200	373	150
Yawn	160	120	180
Gaze	2000	5000	8000

9.1 Fatigue Monitoring Using Blink and Yawn Detection

As mentioned earlier, in this paper we detect driver fatigue by monitoring eye blinks and yawns. Table 3 shows the number of blink, yawn and gaze events labelled using IR imagery on FLIR thermal, synthesized thermal and NIR imagery in low-light conditions for basement data. Furthermore, we manually verified all the labels of the detected events across all types of imagery.

Table 4. Blink and yawn detection using FLIR and NIR - Moving average

Model	Event	Metric		
		Precision	Recall	F1-Score
FLIR thermal	Blink	93.6	50.2	65.4
	Yawn	96.2	47.2	63.3
Synthesized thermal	Blink	NA	NA	NA
	Yawn	84.0	78.0	80.8
NIR	Blink	87.32	83.22	**85.22**
	Yawn	82.56	86.46	**84.46**

Table 5. NIR LED - Fixed threshold

Event	Metric		
	Precision	Recall	F1-Score
Blink	74.38	90.0	81.44
Yawn	74.76	87.91	80.80

9.1.1 Basement Data Evaluation. Table 4 shows the blink and yawn detection accuracy on the FLIR thermal, synthesized thermal and NIR test images using the custom model described in the previous section. We detect eye blinks and yawns based on the moving average of EAR and MAR metrics defined in Section 3.3. We restrict evaluation of only yawn events from the synthesized images as it is very hard to detect the eye patch region in a synthesized image due to the noise and low-resolution. Note that, we obtain ground truth for these events using the IR camera imagery.

Detecting blinks and yawns relies heavily upon the accuracy of the landmarks. Since NIR models detect precise facial landmarks, we can clearly see from Table 4 eye blink and yawn detection accuracy of NIR models outperform FLIR thermal and synthesized thermal models. NIR models have around 85% accuracy in detecting both eye blinks and yawns as compared to 64% using FLIR models. Furthermore, several research efforts show that during good lighting conditions the eye blink and yawn accuracy is around 91% [29, 53]. Thus our NIR low-light models are as powerful as the good-lighting state of the art models. Figure 20 shows a sample FLIR thermal, synthesized thermal and NIR images with yawn detection in our dataset.

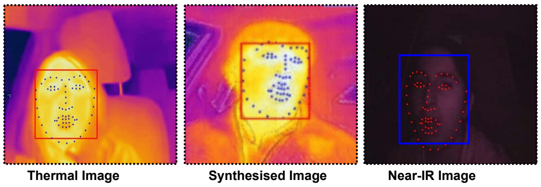
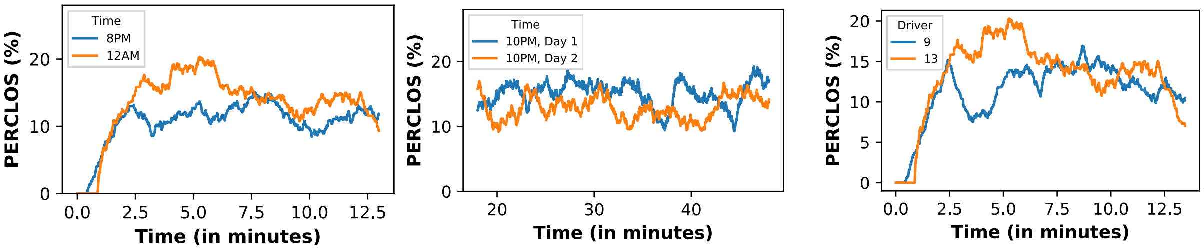


Fig. 20. Yawn detection on FLIR thermal, synthesized thermal and NIR images.

In addition, we also compare our method based on moving averages of EAR and MAR to detect blink and yawns with a naive approach of using fixed thresholds. The fixed thresholds are obtained during a calibration phase where the driver looks in different directions at the beginning of the driving session and the average values of EAR and MAR are recorded. These values are then used to detect eye blinks and yawns accordingly. Table 5 shows the blink and yawn detection accuracy on NIR images. As expected, using fixed thresholds yields lower F_1-accuracy due to a higher number of false positives. This is due to the fact that the EAR/MAR thresholds for a person varies over time even in the same drive, due to change in facial expressions, gaze distribution, and seating position.



(a) Same driver, different hours, same day. (b) Same driver, same hour, different days. (c) Different drivers, same hours, same day.

Fig. 21. PERCLOS analysis on data collected from real roads across drivers.

9.1.2 On-road Data Evaluation. As described in Section 8.1, we collected around 20 hours of data across 10 drivers on real roads between 8PM to 4AM using the NIR setup (shown in Figure 12). Since this was on real roads, we could not include the IR camera for ground truth as it is bulky and can obstruct driver's view. Hence, we now show analysis on how eye closure rate (PERCLOS) can be used to monitor driver fatigue.

PERCLOS indicates the percentage of time eyes are closed in a minute and several studies indicate when the value of PERCLOS exceeds 30%, the driver is considered to be in a drowsy or fatigued state [58]. In the 20 hours of data collected in low-light conditions, average PERCLOS of all the drivers was below 22% indicating drivers were not drowsy. We now present detailed PERCLOS analysis on few drives from our on road data.

Scenario 1- Same driver at different hours on the same day: Figure 21a shows the PERCLOS values over time for driver D_{11} at two different drives taken at 8PM and 12AM on the same day. Each drive was an hour long and for our analysis, we show PERCLOS values for a specific 12 minute window. The blue and orange lines indicate the PERCLOS from 8PM and 12AM drives, respectively. We can see that in general, the PERCLOS values are below 30% indicating the driver is not drowsy. However, we can observe that in the 12AM drive the PERCLOS values are significantly higher than 8PM indicating the driver is getting fatigued. Thus, we can now alert the driver when his/her PERCLOS values varies significantly compared to other drives and also when PERCLOS exceeds 30%.

Scenario 2- Same driver at same hours on different days: Figure 21b shows the PERCLOS values over time for driver D_{12} at 10PM drives on two different days. We can see that on both the days the PERCLOS values of the driver was pretty similar and below 30%. We can now use such analysis to determine if the PERCLOS values of a driver vary significantly compared to rest of the days/drives and alert accordingly.

Scenario 3- Different drivers at same hours on the same day: Figure 21c shows the PERCLOS values over time for two drivers D_9 and D_{13} at 8PM drives on the same day. We can see that for driver D_{13} PERCLOS values are higher than driver D_9 in the early parts of the drive, however, as the drive progresses, the PERCLOS values of driver D_{13} reduce indicating driver is being attentive. PERCLOS analysis at different time periods and against different drivers can be effective to determine if the drivers are drowsy and alert them accordingly.

Since we did not collect ground truth in real road data collection using IR camera, we manually analysed a subset of 10 hour data to label yawn events. We identified 15 yawns from our dataset and NIR models were able to accurately identify all the yawns with very few false positives (2) and false negatives (1). The false positives were mainly due to the large variation in MAR values due to combination of turning back and talking to the passenger, which resulted in false detection of yawn.

9.2 Distraction Monitoring Using Gaze information

We now present results on mirror scanning event detection using the gaze information. Mirror scanning is a key factor to determine the attentiveness of the driver.

Table 6. Three-way gaze classification accuracy.

Monitoring Method	Gaze Direction		
	Left	Straight	Right
FLIR thermal	**98.96**	91.03	**98.89**
Synthesized thermal	74.32	82.50	78.67
NIR	96.31	**98.41**	86.82

9.2.1 *Basement Data Evaluation.*

We first show the accuracy of detecting mirror scans (driver looking at the left mirror, right mirror and straight) for the data collected in the basement. Table 3 shows the number of mirror scan events recorded for FLIR, synthesized and NIR imagery. Note that, we obtain ground truth for these events using the IR camera imagery. Table 6 shows the gaze direction accuracy corresponding to the three states, i.e., left, straight, right across FLIR thermal, synthesized thermal and NIR test data.

FLIR thermal models have an overall accuracy of 96% for left, straight and right classifications slightly outperforming NIR models, which has an overall accuracy of 93.8%. Furthermore, synthesized images perform poorly compared to FLIR and NIR, as the GAN model does not generalize well when the yaw angles are large and hence it has higher accuracy (82.5%) for straight direction as compared to left and right directions. Figure 22 shows sample images with the left, right gaze classifications on FLIR thermal, synthesized thermal and NIR images.

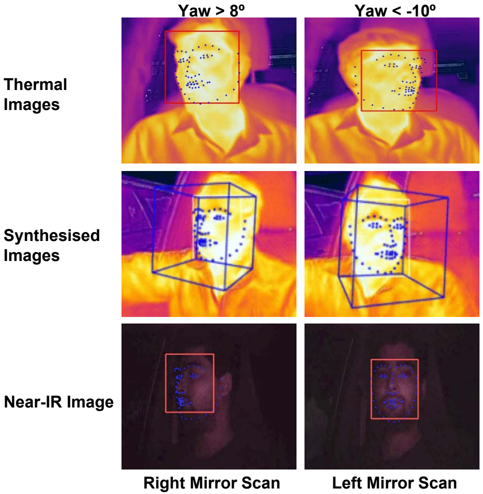


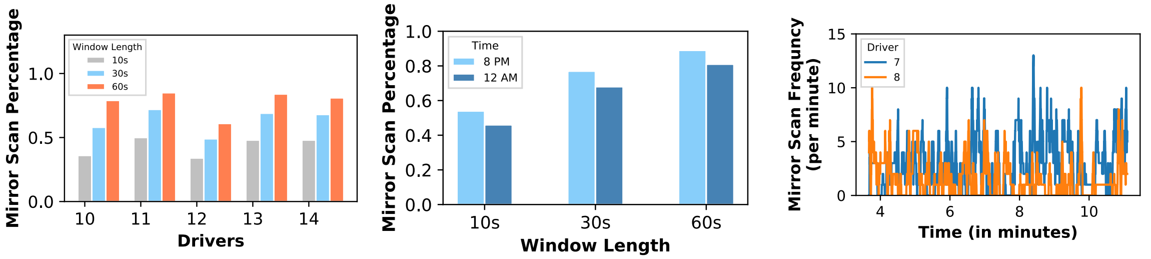
Fig. 22. Mirror scan detection in FLIR, synthesized thermal and NIR images.

9.2.2 On-road Data Evaluation. We now show analysis of mirror scanning behavior across drivers using our on-road data. Various studies have reported that drivers should scan both left and right mirrors every 7-12 seconds to be situationally aware [2, 50].

Adherence to mirror scanning. We classify a driver to be situationally aware, if the driver is scanning his/her mirrors at least once every 10s continuously during the drive. Specifically, we compute the % of time the driver adheres to 10s mirror scan rule. For instance, a fully situational-aware driver will have 100% mirror scan value, where he scans the mirror once every 10 seconds through out the drive. Figure 23a shows the mirror scan % for 10s, 30s and 60s window for 5 drivers from our on-road data. As we can see when the mirror scan rule is 10s, the mirror scan % during the drive for all drivers is around 40%, indicating only 40% of the time during the drive, the drivers adhere to 10s rule of checking mirrors. Further, if the 10s rule is relaxed to 30s or 60s, the mirror scan adherence increases to 63% and 78%, respectively. Thus, we can say in our dataset drivers scan at least one of the mirrors within a minute.

Figure 23b shows the mirror scan % for driver D_9 from a drive at 8PM and 12AM on the same day. We can clearly see that the adherence to mirror scan rule drops by 10% for 12AM drive as compared to 8PM drive for 10s, 30s and 60s window intervals. This shows that driver at 12AM drive is not being fully situationally aware as compared to 8PM drive. We can now perform such analysis to provide actionable feedback and improve road safety in low-light conditions.

Mirror scan frequency. We also compared mirror scan frequency per minute to understand driver's mirror scanning behavior over time. Figure 23c shows the mirror scan frequency per minute for two drivers, D_7 and D_8 at 10PM. In general, D_8 mirror scan frequency is much lower compared to D_7 indicating the driver D_8 is not situationally aware. During this data collection we ensured both the drivers were driving in similar traffic conditions. Furthermore, mirror scan frequency varies abruptly over time across both the drivers, indicating a poor driving behavior. We can use mirror scan frequency to effectively determine driver attentiveness and accordingly provide actionable feedback.



(a) Mirror scan adherence across drivers.

(b) Mirror scan adherence at 8PM and 12AM.

(c) Mirror scan frequency of two drivers.

Fig. 23. Mirror scan analysis on data collected from real roads across drivers.

10 PERFORMANCE BENCHMARKS AND DISCUSSIONS

In this section, we first present details of implementing `InSight` on smartphones. We then compare the proposed synthesized thermal and NIR imagery approaches and finally, provide discussion on the generalizability of the proposed approach, trade-offs and few limitations.

Table 7. Time take and energy consumption for different models.

Models	Time Taken in ms	Energy in J
Face detection and landmarks (CPU)	58	0.12
Gaze estimation (CPU)	45	0.08
GAN (CPU)	600	5.36
GAN (GPU)	140	0.42

10.1 Implementation on Smartphones

We have implemented all the four detectors, *viz.,* face detector, facial landmark model, blink and yawn detector and gaze classification on the smartphone. We used OnePlus 5T smartphone to benchmark these detectors, which has Kryo 280 4 CPU cores and a Adreno 540 GPU core. The entire pipeline is written in native C++ code based on OpenCV, and custom libraries interfaced with Java using JNI wrappers. Since facial landmarks form the basis of InSight's detectors, the main thread of InSight is tasked with detecting face and facial landmarks for incoming video frames. A set of parallel threads are then spawned to use the landmarks for tasks such as fatigue and distraction.

Table 7 shows the time taken and energy consumed for each of the models. The face and facial landmark detection takes 58 ms per frame and an additional 10 ms for tracking facial landmarks in subsequent frames. Furthermore, the fatigue monitoring component that computes EAR and MAR to determine eye closure and yawns takes about 5 ms. The gaze estimation takes 45 ms per frame. Overall, InSight takes 80-100 ms for running all the detectors, supporting up to 10 fps. This can be further improved by pairing down the number of landmarks as described in [54].

We also benchmark the energy consumption in Joules (J) for the above models. We employed the techniques described in [74] to measure energy consumption. Specifically, we use snapdragon profiler to measure the average power consumed by the device. Before benchmarking, we measure the idle power consumption by the phone (P_{idle}) and we then run our workload and measure P_{active}. Thus, power consumed per frame is P_{frame} = (P_{active} - P_{idle}), further, P_{frame} is multiplied by the time taken for processing per frame , i.e., T_{frame} to obtain energy consumption in J. Thus, energy consumed per frame is derived using, E_{frame} = P_{frame} x T_{frame} and Table 7 shows the energy consumed per frame for each of the model.

Finally, we employ the techniques proposed in the literature to make use of the low-end GPU to run the GAN network [18, 46, 55]. The GAN network for synthesizing a thermal image from a low-light RGB takes 600 ms on the CPU but only 140 ms on the GPU. With the additional 45-65 ms for the remainder of the processing on the GPU added in, the overall time per frame in low-light conditions would be about 200 ms, yielding an acceptable frame rate of 5 fps even under such challenging conditions. Thus, by utilizing both the CPU and GPU resources available on the smartphone, InSight can perform effective driver state monitoring in low-lighting conditions.

10.2 Comparison of the Proposed Approaches

We now summarize the applicability of the proposed based on the following key requirements:

Accurate. Our exhaustive experimental evaluation shows that results from NIR imagery outperforms synthesized thermal images across all detectors (face detection, facial landmarks, fatigue and distraction). Even more, NIR LED based setup performs significantly better than a specialized camera such as thermal cameras and has comparable accuracy to the IR cameras (which was used as the ground truth device in this paper). On the other hand, synthesized thermal images generally outperform the thermal cameras, mainly due to the power of GANs to extrapolate/interpolate information. Since, we focus on just the driver's face this is very confined problem for GANs. One key challenge today with synthesized thermal is that it relies purely on the training data and hence it is hard to generalize.

Non-intrusive. Both synthesized thermal and NIR LED based approaches are non-intrusive. Furthermore, both the approaches can leverage the existing smartphone systems to monitor the state of the driver without any additional specialized hardware.

Low-cost. In case of synthesized thermal images, there is a need for specialized hardware atleast during the training phase to collect paired FLIR thermal and RGB images. Thus to achieve a robust system, one may need multiple FLIR cameras resulting in slightly expensive solution as each FLIR camera costs around 300$. On the other hand, NIR LED based setup relies on just a few NIR LEDs and minimal additional electronics. As we showed in this paper, the setup can be easily attached to the existing phones and the entire unit would cost <10$. Thus enabling support to monitor driver state at large scale.

In conclusion, for large scale deployments NIR based setup is preferable due to its robustness, accuracy and low-cost setup. However, with the advancements in deep learning synthesizing thermal images from RGB images is a promising avenue, especially given it can work with just a standard RGB camera (which is the holy grail for monitoring in low-light conditions), thus eliminating any additional attachment to the smartphone.

10.3 Discussions

We provide some discussion on the generalizability of the proposed approach, few limitations and directions for our future work.

(i) NIR LED selection: As we discussed in Section 6.1, while majority of the smartphones have an IR filter, they do not completely block out infrared light. We leverage this insight to use NIR LEDs in our setup. However the choice of NIR LED wavelength is very important for its effective operation. Even-though NIR LEDs can range from 700-1000nm, using NIR LEDs in the range of 810-850nm is preferable based on our extensive experimentation as, (i) they do not emit too much red light, which might be intrusive to the driver and (ii) most smartphone cameras can capture NIR light in this range effectively.

(ii) Safety implications on usage of NIR LEDs: IEC-62471-1 [38] standard is followed for IR LEDs in lamp applications, which define the limits and other risks when IR LEDs are used. Specifically, the standard mention NO hazard or risk for long time duration, when the emitted radiant intensity of LED is within 100 W/m^2 at a fixed distance of d=200mm [39]. The NIR LED [12] (OSRAM SFH 4787S) used in this work has radiant intensity of 25 W/m^2 at d=200mm (derived from the datasheet [12]), which is well within the acceptable limits of 100 W/m^2. Furthermore, given that the driver is seated beyond 200mm from the smartphone with NIR LED in a typical vehicle setting (typically around 40-60cm) and the emitted intensity is significantly lower than the limits, there is none or negligible impact on human eyes.

(iii) Image synthesis GAN training and generalizability: Although image synthesis approach is cost-effective, as the system scales it is imperative to collect large amount of data in order to train a GAN. This would require deploying multiple FLIR cameras resulting in slightly expensive solution as each FLIR camera costs around 300$. Furthermore, the trained model works well for a specific setting on which it is trained and needs to be re-trained based on the camera mounting and the target geography (as the person's face attributes vary from one geography to another). One approach to make the model applicable widely is to train with much diverse data and as mentioned earlier since the GAN requires just pairs of thermal and RGB images it is easier to collect one, but is time consuming. Another approach to overcome this relies on advancements in computer vision and deep learning algorithms, where recent works have used unsupervised cycle GAN [16, 84]. Cycle GANs are similar to the GANs used in this paper for image to image translation, but with a key distinction that cycle GANs no more require a paired set of low-light RGB and thermal images to synthesize a thermal image. This reduces the burden in collecting the paired image data for training GANs, but has limitations in synthesizing accurate images.

(iv) Improving facial landmarks with GANs: In this work we showed the power of GANs to synthesize a thermal image using just a low-light RGB image. The synthesized image leverages the learning from the paired images to derive an intermediate representation, which is useful for low-light conditions. Furthermore, recent works [28] show that GANs can be used to create different image styles, i.e., light, dark, grayscale, etc., which can be used to train a robust facial landmarks. This coupled with the proposed thermal image synthesis can pave the way towards robust facial landmark detection in low-light or nighttime conditions.

(v) Accuracy of fatigue and distraction models: The proposed models for face detection, landmarks, driver fatigue and gaze extends state of the art techniques and has an overall accuracy around 85-90% in detecting events. However, to deploy a system that alert drivers in real-time based on fatigue/distracted events in real-world, this accuracy needs to be further improved. The current choice of detectors is dictated by the resource constraints in a smartphone, as the goal in this paper is to perform on-device driver monitoring. Depending on the network connectivity and bandwidth, one can train a much larger and accurate deep neural network in the cloud. However, this requires sending video snippets to the cloud for processing and thus necessitates development of new edge and cloud architectures for video processing.

(vi) Large scale evaluation: In this paper, we have taken the first step towards developing a low-cost end-to-end system for monitoring the state of the driver in low-light. Compared to prior works we perform extensive evaluation of the proposed approaches from data collected with 15 drivers in real world conditions (basement area and on real-roads). While we show excellent model performance with unseen data, new drivers and in extreme lighting conditions, the number of drivers in the study is still limited. Hence, as part of our future work we plan to deploy and test the proposed NIR setup with larger number of drivers in more diverse scenarios.

11 CONCLUSION

InSight is a smartphone-based system for monitoring driver fatigue and distraction in low-light conditions. In this paper, we presented two novel and practical solutions *viz.,* image synthesis and NIR LED setup, that enables the InSight to be accurate, low-cost and non-intrusive. We demonstrated the efficacy of both the approaches using data collected from controlled basement drives and real-world on-road drives.

ACKNOWLEDGMENTS

We owe a debt of gratitude to the many drivers, and other members of staff at Microsoft Research India for their patience as we iterated through many versions of our system and during data collection.

REFERENCES

[1] 2009. Exposure compensation. https://en.wikipedia.org/wiki/Exposure_compensation.

[2] 2009. Mirror scan duration. https://www.thesafedriver.ca/2009/10/08/how-often-should-you-check-your-mirror/.

[3] 2017. Honda CR-V SUV. https://venturebeat.com/2017/03/09/this-small-suv-knows-when-you-get-sleepy-and-can-wake-you-up/

[4] 2017. Nighttime Car Accident Statistics. https://seriousaccidents.com/legal-advice/top-causes-of-car-accidents/nighttime-driving/

[5] 2017. Receive Warnings About Your Level of Alertness While Driving With Honda's Driver Attention Monitor. http://www.hiltonheadhonda.com/blog/how-does-the-honda-driver-attention-monitor-work/

[6] 2018. Night Sight: Seeing in the Dark on Pixel Phones . https://ai.googleblog.com/2018/11/night-sight-seeing-in-dark-on-pixel.html.

[7] 2018. WHO Global Health Observatory (GHO) data . http://www.who.int/gho/road_safety/mortality/en/.

[8] 2018. WHO Road traffic injuries . http://www.who.int/mediacentre/factsheets/fs358/en/.

[9] 2019. FLIR one for Android. http://www.flir.com/flirone/pro/

[10] 2019. *Flir one for android*. https://www.flir.com/flir-one/

[11] 2019. Infrared cameras? https://en.wikipedia.org/wiki/Thermographic_camera

[12] 2019. SFH 4787S- OSRAM 810nm IR LED. https://www.mouser.com/ProductDetail/OSRAM-Opto-Semiconductors/SFH-4787S

[13] 2019. VANTRUE: High-end Dash Cam for Your Drive. https://www.vantrue.net/

[14] 2019. What is an Infrared Illuminator? https://www.flir.in/discover/ots/what-is-an-infrared-illuminator/

[15] Shabnam Abtahi, Mona Omidyeganeh, Shervin Shirmohammadi, and Behnoosh Hariri. 2014. YawDD: A yawning detection dataset. In *Proceedings of the 5th ACM Multimedia Systems Conference*. 24–28.

[16] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. 2018. Recycle-gan: Unsupervised video retargeting. In *Proceedings of the European conference on computer vision (ECCV)*. 119–135.

[17] Luis Miguel Bergasa, Jesus Nuevo, Miguel A Sotelo, Rafael Barea, and María Elena Lopez. 2006. Real-time system for monitoring driver vigilance. *IEEE Transactions on Intelligent Transportation Systems* 7, 1 (2006), 63–77.

[18] Ravi Bhandari, Akshay Uttama Nambi, Venkata N Padmanabhan, and Bhaskaran Raman. 2018. DeepLane: camera-assisted GPS for driving lane detection. In *Proceedings of the 5th Conference on Systems for Built Environments*. 73–82.

[19] Adrian Bulat and Georgios Tzimiropoulos. 2017. How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks). In *International Conference on Computer Vision*.

[20] Oya Çeliktutan, Sezer Ulukaya, and Bülent Sankur. 2013. A comparative study of face landmarking techniques. *EURASIP Journal on Image and Video Processing* 2013, 1 (07 Mar 2013), 13. https://doi.org/10.1186/1687-5281-2013-13

[21] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. 2018. Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3291–3300.

[22] Chen-Jui Chung, Wei-Yao Chou, and Chia-Wen Lin. 2013. Under-exposed image enhancement using exposure compensation. In *2013 13th International Conference on ITS Telecommunications (ITST)*. IEEE, 204–209.

[23] Reinier C Coetzer and Gerhard P Hancke. 2011. Eye detection for a real-time vehicle driver fatigue monitoring system. In *2011 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 66–71.

[24] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. 2001. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence* 23, 6 (2001), 681–685.

[25] David Cristinacce and Timothy F Cootes. 2006. Feature detection and tracking with constrained local models.. In *Bmvc*, Vol. 1. Citeseer, 3.

[26] C. Cudalbu, B. Anastasiu, R. Radu, R. Cruceanu, E. Schmidt, and E. Barth. 2005. Driver monitoring with a single high-speed camera and IR illumination. In *International Symposium on Signals, Circuits and Systems, 2005. ISSCS 2005.*, Vol. 1. 219–222 Vol. 1.

[27] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection.

[28] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. 2018. Style Aggregated Network for Facial Landmark Detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 379–388.

[29] Isha Dua, Akshay Nambi, C. V. Jawahar, and Venkat Padmanabhan. 2019. AutoRate: How attentive is the driver?. In *The 14th IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE.

[30] Tiziana D'Orazio, Marco Leo, Cataldo Guaragnella, and Arcangelo Distante. 2007. A visual approach for driver inattention detection. *Pattern recognition* 40, 8 (2007), 2341–2355.

[31] Andrew C Gallup, Allyson M Church, and Anthony J Pelegrino. 2016. Yawn duration predicts brain weight and cortical neuron number in mammals. *Biology letters* 12, 10 (2016), 20160545.

[32] Miguel García-García, Alice Caplier, and Michèle Rombaut. 2018. Sleep deprivation detection for real-time driver monitoring using deep learning. In *International Conference Image Analysis and Recognition*. Springer, 435–442.

[33] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[34] Samuel W Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. 2016. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (TOG)* 35, 6 (2016), 1–12.

[35] Hua Huang, Hongkai Chen, and Shan Lin. 2019. MagTrack: Enabling Safe Driving Monitoring with Wearable Magnetics. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. 326–339.

[36] Shih-Chia Huang, Fan-Chieh Cheng, and Yi-Sheng Chiu. 2012. Efficient contrast enhancement using adaptive gamma correction with weighting distribution. *IEEE transactions on image processing* 22, 3 (2012), 1032–1041.

[37] Haidi Ibrahim and Nicholas Sia Pik Kong. 2007. Brightness preserving dynamic histogram equalization for image contrast enhancement. *IEEE Transactions on Consumer Electronics* 53, 4 (2007), 1752–1758.

[38] IEC IEC. 2006. 62471: 2006/CIE S 009/E: 2002 Photobiological safety of lamps and lamp systems.

[39] IR Illumination and Eye Safety. [n.d.]. https://medium.com/@alex.kilpatrick/ir-illumination-and-eye-safety-f0804673ca7

[40] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125–1134.

[41] Qiang Ji, Zhiwei Zhu, and Peilin Lan. 2004. Real-time nonintrusive monitoring and prediction of driver fatigue. *IEEE transactions on vehicular technology* 53, 4 (2004), 1052–1068.

[42] Sang-Joong Jung, Heung-Sub Shin, and Wan-Young Chung. 2014. Driver fatigue and drowsiness monitoring system with embedded electrocardiogram sensor on steering wheel. *IET Intelligent Transport Systems* 8, 1 (2014), 43–50.

[43] Vahid Kazemi and Josephine Sullivan. 2014. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1867–1874.

[44] Marcin Kopaczka, Justus Schock, and Dorit Merhof. 2019. Super-realtime facial landmark detection and shape fitting by deep regression of shape model parameters. *arXiv preprint arXiv:1902.03459* (2019).

[45] Neslihan Kose, Okan Kopuklu, Alexander Unnervik, and Gerhard Rigoll. 2019. Real-Time Driver State Monitoring Using a CNN Based Spatio-Temporal Approach. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 3236–3242.

[46] N. D. Lane and P. Warden. 2018. The Deep (Learning) Transformation of Mobile and Embedded Computing. *Computer* 51, 5 (May 2018), 12–16. https://doi.org/10.1109/MC.2018.2381129

[47] Shiqi Li, Chi Xu, and Ming Xie. 2012. A robust O (n) solution to the perspective-n-point problem. *IEEE transactions on pattern analysis and machine intelligence* 34, 7 (2012), 1444–1450.

[48] Stan Z Li, RuFeng Chu, Meng Ao, Lun Zhang, and Ran He. 2006. Highly accurate and fast face recognition using near infrared images. In *International Conference on Biometrics*. Springer, 151–158.

[49] Charles Loop and Zhengyou Zhang. 1999. Computing rectifying homographies for stereo vision. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, Vol. 1. IEEE, 125–131.

[50] Zhenji Lu, Xander Coster, and Joost de Winter. 2017. How much time do drivers need to obtain situation awareness? A laboratory-based study of automated driving. *Applied ergonomics* 60 (2017), 293–304.

[51] Krzysztof Małecki, Paweł Forczmański, Adam Nowosielski, Anton Smoliński, and Daniel Ozga. 2019. A new benchmark collection for driver fatigue research based on thermal, depth map and visible light imagery. In *International Conference on Computer Recognition Systems*. Springer, 295–304.

[52] Bappaditya Mandal, Liyuan Li, Gang Sam Wang, and Jie Lin. 2016. Towards detection of bus driver fatigue based on robust visual analysis of eye state. *IEEE Transactions on Intelligent Transportation Systems* 18, 3 (2016), 545–557.

[53] Akshay Uttama Nambi, Shruthi Bannur, Ishit Mehta, Harshvardhan Kalra, Aditya Virmani, Venkata N. Padmanabhan, Ravi Bhandari, and Bhaskaran Raman. 2018. HAMS: Driver and Driving Monitoring Using a Smartphone. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking (MobiCom '18)*. ACM, New York, NY, USA, 840–842. https://doi.org/10.1145/3241539.3267723

[54] Akshay Uttama Nambi, Ishit Mehta, Anurag Ghosh, Vijay Lingam, and Venkata N Padmanabhan. 2019. ALT: towards automating driver license testing using smartphones. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*. 29–42.

[55] Akshay Uttama Nambi, Adtiya Virmani, and Venkata N Padmanabhan. 2018. FarSight: A Smartphone-based Vehicle Ranging System. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–22.

[56] OpenCV. 2014. Camera calibration and 3d reconstruction. *URL https://docs. opencv. org/2.4/modules/calib3d/doc/camera_calibration_and_3d_reconstruction. html* (2014).

[57] Wu Qing, Sun BingXi, Xie Bin, and Zhao Junjie. 2010. A perclos-based driver fatigue recognition application for smart vehicle space. In *2010 Third International Symposium on Information Processing*. IEEE, 437–441.

[58] W. Qing, S. BingXi, X. Bin, and Z. Junjie. 2010. A PERCLOS-Based Driver Fatigue Recognition Application for Smart Vehicle Space. In *2010 Third International Symposium on Information Processing*. 437–441. https://doi.org/10.1109/ISIP.2010.116

[59] Shanto Rahman, Md Mostafijur Rahman, Mohammad Abdullah-Al-Wadud, Golam Dastegir Al-Quaderi, and Mohammad Shoyaib. 2016. An adaptive gamma correction for image enhancement. *EURASIP Journal on Image and Video Processing* 2016, 1 (2016), 1–13.

[60] Antoni Rogalski, Piotr Martyniuk, and Małgorzata Kopytko. 2016. Challenges of small-pixel infrared detectors: a review. *Reports on Progress in Physics* 79, 4 (2016), 046501.

[61] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 2013. 300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge. *2013 IEEE International Conference on Computer Vision Workshops* (2013), 397–403.

[62] Arun Sahayadhas, Kenneth Sundaraj, and M Murugappan. 2012. Detecting driver drowsiness based on sensors-a review. (2012).

[63] Gulbadan Sikander and Shahzad Anwar. 2018. Driver fatigue detection systems: A review. *IEEE Transactions on Intelligent Transportation Systems* 20, 6 (2018), 2339–2352.

[64] Tereza Soukupová and Jan Cech. 2016. Eye blink detection using facial landmarks. In *21st Computer Vision Winter Workshop, Rimske Toplice, Slovenia*.

[65] Tereza Soukupová and Jan Cech. 2016. Real-Time Eye Blink Detection using Facial Landmarks.

[66] FLIR Systems. 2019. Flir one for android. https://www.flir.com/flir-one/

[67] Etienne Vincent and Robert Laganiére. 2001. Detecting planar homographies in an image pair. In *ISPA 2001. Proceedings of the 2nd International Symposium on Image and Signal Processing and Analysis. In conjunction with 23rd International Conference on Information Technology Interfaces (IEEE Cat.* IEEE, 182–187.

[68] Paul Viola and Michael Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, Vol. 1. IEEE, I–I.

[69] Michael Vollmer, Klaus-Peter Möllmann, and Joseph A Shaw. 2015. The optics and physics of near infrared imaging. In *Education and Training in Optics and Photonics*. Optical Society of America, TPE09.

[70] Yu Wang, Qian Chen, and Baeomin Zhang. 1999. Image enhancement based on equal area dualistic sub-image histogram equalization method. *IEEE Transactions on Consumer Electronics* 45, 1 (1999), 68–75.

[71] Walter W Wierwille, SS Wreggit, CL Kirn, LA Ellsworth, and RJ Fairbanks. 1994. *Research on vehicle-based driver status/performance monitoring; development, validation, and refinement of algorithms for detection of driver drowsiness. final report.* Technical Report.

[72] Hong Wu, Michael J Hayes, Albert Weiss, and Qi Hu. 2001. An evaluation of the Standardized Precipitation Index, the China-Z Index and the statistical Z-Score. *International Journal of Climatology: A Journal of the Royal Meteorological Society* 21, 6 (2001), 745–758.

[73] Yue Wu and Qiang Ji. 2019. Facial landmark detection: A literature survey. *International Journal of Computer Vision* 127, 2 (2019), 115–142.

[74] Chunwei Xia, Jiacheng Zhao, Huimin Cui, Xiaobing Feng, and Jingling Xue. 2019. DNNTune: Automatic Benchmarking DNN Models for Mobile-cloud Computing. *ACM Transactions on Architecture and Code Optimization (TACO)* 16, 4 (2019), 1–26.

[75] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2016. WIDER FACE: A Face Detection Benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[76] Shuochao Yao, Yiran Zhao, Aston Zhang, Shaohan Hu, Huajie Shao, Chao Zhang, Lu Su, and Tarek Abdelzaher. 2018. Deep Learning for the Internet of Things. *Computer* 51, 5 (2018), 32–41.

[77] Zhenqiang Ying, Ge Li, Yurui Ren, Ronggang Wang, and Wenmin Wang. 2017. A new image contrast enhancement algorithm using exposure fusion framework. In *International Conference on Computer Analysis of Images and Patterns*. Springer, 36–46.

[78] Chuang-Wen You, Nicholas D. Lane, Fanglin Chen, Rui Wang, Zhenyu Chen, Thomas J. Bao, Martha Montes de Oca, Yuting Cheng, Mu Lin, Lorenzo Torresani, and Andrew T. Campbell. 2013. CarSafe App: Alerting Drowsy and Distracted Drivers using Dual Cameras on Smartphones. In *ACM MobiSys*.

[79] Chuang-Wen You, Nicholas D Lane, Fanglin Chen, Rui Wang, Zhenyu Chen, Thomas J Bao, Martha Montes-de Oca, Yuting Cheng, Mu Lin, Lorenzo Torresani, et al. 2013. Carsafe app: Alerting drowsy and distracted drivers using dual cameras on smartphones. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*. 13–26.

[80] Amir Zadeh, Yao Chong Lim, Tadas Baltrusaitis, and Louis-Philippe Morency. 2017. Convolutional experts constrained local model for 3d facial landmark detection. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2519–2528.

[81] Zhengyou Zhang. 2000. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence* 22, 11 (2000), 1330–1334.

[82] Zutao Zhang and Jiashu Zhang. 2010. A new real-time eye tracking based on nonlinear unscented Kalman filter for monitoring driver fatigue. *Journal of Control Theory and Applications* 8, 2 (2010), 181–188.

[83] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.

[84] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*.

[85] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. 2016. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 146–155.

[86] Zhiwei Zhu and Qiang Ji. 2004. Real time and non-intrusive driver fatigue monitoring. In *Proceedings. The 7th International IEEE Conference on Intelligent Transportation Systems (IEEE Cat. No. 04TH8749)*. IEEE, 657–662.