DepthPhys: Near-Infrared Remote Photoplethysmography in Driver Monitoring Systems

Timothy John Alder
u7287129

October 25, 2024

A THESIS SUBMITTED IN
PARTIAL FULFILMENT OF THE
REQUIREMENTS FOR THE DEGREE

Bachelor of Engineering
(Research and Development) (Honours)

College of Engineering, Computing and Cybernetics
Australian National University

Australian National University

seeingmachines

Project Area: **Biometrics, Deep Learning**
Industry Sponsor: **Seeing Machines**
Project Supervisor: **Dr. Andrei Komar**
Second Examiner: **Prof. Stephen Gould**

Abstract

DepthPhys: Near-Infrared Remote Photoplethysmography in Driver Monitoring Systems

by

Timothy John Alder

Subtle variations in the body can be used to detect underlying physiological signals from camera video data. Existing research in this space is largely focused on extracting physiological signals from traditional two-dimensional planar video data. This project explores how the addition of depth data may be used to augment the results of these approaches. In particular, a novel optical pathway is proposed for camera-based physiological sensing and is validated on a new dataset consisting of seven hours of three-dimensional volumetric near-infrared monochromatic video data. Both two-dimensional and three-dimensional signals are extracted from this dataset and the results of either approach are compared using state-of-the-art camera-based physiological sensing neural methods. These findings provide valuable and novel insight into the utility of three-dimensional signals for camera-based physiological sensing, as well as the capability and limitations of existing camera-based physiological sensing solutions when applied in a driver monitoring system application context.

Foreword

This work would not have been possible without the help and support of my friends, family, colleagues, mentors, and supervisor. In particular, I would like to thank Ricky, Cameron, Abraham, and John for their support in delivering the data collection campaign that made this project possible. The amount of engineering hours, resources, and technical skills required were well beyond what I would be capable of delivering alone and the final product is a testament to their expertise. I would also like to thank Tom for his role in championing this project at Seeing Machines, Pratyush for his guidance on the technical aspects of this project, Andrei for introducing me to the logical framework approach of thinking that allowed me to stay organised over the last year, Lachlan for the initial inspiration that birthed the core idea behind this project, and Charles for his continual support of me undertaking research roles throughout my internship at Seeing Machines.

Most of all, I am very thankful to Seeing Machines for giving me the opportunity to pursue this project. Few companies would put such faith in an undergraduate and provide them with such resources to pursue their vision. This project is a testament to the innovation that underpins Seeing Machines.

Contents

1 Introduction

In 2022, land transport accidents were one of the top three leading causes of death in Australia for persons 1-44 years of age [1]. 20%-30% of these accidents are attributed to fatigue [2] and an additional 16% are attributed to distraction [3]. Seeing Machines (SM) is a Tier Two automotive company focused on creating computer vision-based active driver monitoring systems (DMS) with the mission of eliminating transportation fatalities and injuries caused by distraction and fatigue.

Figure 1: *A typical Seeing Machines DMS system. Two external NIR illumination pods are paired with a camera body.*

Inclusion of physiological signals is a crucial step to the development of more capable DMS. For example; [4], [5], [6] demonstrate there is a 7.5 to 10 beats-per-minute (BPM) difference in heart rate between drowsy and alert states. Furthermore, [7] found a strong correlation between heart rate variability (HRV) and cognitive load, while [8] found that HRV can be used for intoxication detection. Lastly, numerous studies have demonstrated the capacity of camera-based heart rate sensing—specifically, the remote photoplethysmography (rPPG) signal—for spoof detection [9], [10], [11]. Accordingly, inclusion of physiological signals is expected to foster significant improvements and advances in existing DMS technology.

This project seeks to evaluate the capabilities of near-infrared (NIR) coherent structured light emitters for use in rPPG. While previous research has primarily focused on traditional incoherent illumination sources, the narrow linewidth of coherent sources provides a significant advantage in ability to reject interfering external light. Furthermore, structured light emitters also offer the ability to deterministically derive depth data [12], which could be valuable for detecting respiration rate. Moreover, structured patterns have previously proven to be effective for decoding the ballistocardiogram (BCG) signal [13]. Finally, coherent structured light emitters have been proven to be capable for material detection [14] which could help guide models to the most suitable regions of an image (i.e., skin pixels) for obtaining the rPPG signal.

In particular, a dot projector coherent illumination source delivers all these benefits while still retaining the underlying flood illumination profile found in traditional sources—albeit at a lower intensity. This project successfully extracts and applies these signals to detect the rPPG physiological signal in the context of DMS applications, demonstrating their value for training rPPG deep learning models.

Broadly summarising, the project scope covers (1) the development of a novel system for rPPG data collection; (2) collection of 7 hours of high bit-depth video training data including synchronised physiological signals for truthing; (3) benchmarking this new dataset against existing state-of-the-art rPPG deep learning algorithms [15], [16], [17], [18]; (4) development of novel deep learning algorithms using structured coherent illumination source signals; and (5) validation of these models using the newly compiled dataset. The findings of this project provide valuable insights into the capacity of existing Seeing Machines DMS for rPPG detection, and will guide the future direction of Seeing Machines rPPG research.

2 Background

2.1 Hemoglobin and the Cardiac Conduction Cycle

The action of the heart beating is referred to as the cardiac cycle. The cardiac cycle is comprised of two periods. In the first period, known as *diastole*, blood is received into both ventricles through their corresponding atrias. Secondly, during *systole*, the ventricles contract, pumping two distinct blood supplies—one to the lungs and one to the rest of the body [19]. Of particular interest to this project is the blood pumped to the lungs, as this is where oxygen is absorped by the protein hemoglobin (Hb).

Hb is responsible for the binding of inhaled oxygen to red blood cells and the subsequent transport of oxygen throughout the body. Hb-bound oxygen is transported throughout the body via the vascular system, and unloaded in preparation for delivery to tissues in need. This cycle, called the cardiac conduction cycle, repeats in perpetuity.

2.2 Photoplethysmography

As early as 1942 [20], it has been shown that Hb is primarily responsible for the absorption spectra of blood in the visible light region. The absorption spectrum of oxygenated and deoxygenated Hb are significantly different (**Figure 2**). It is this difference that forms the basis of photoplethysmography (PPG). The goal of rPPG, also known as imaging photoplethysmography (iPPG), is to detect this difference via remote sensing means (i.e., a camera), and thus, detect a subjects heart beat.

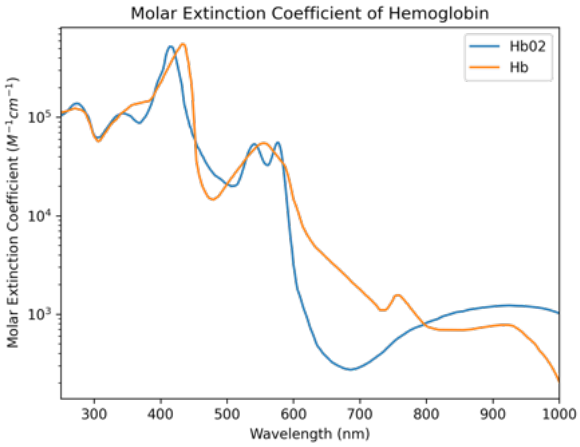

Figure 2: *Molar extinction coefficient for oxygenated (HbO_2) and deoxygenated (Hb) hemoglobin [21].*

In the NIR region, the absorption of hemoglobin is significantly less intense and suspended blood constituents play a more appreciable role in the absorption spectrum of blood. When compared to visible light, this makes NIR rPPG significantly more difficult, generally requiring more capable hardware (e.g., higher bit-depth). Conversely ambient sunlight rejection is significantly better in the NIR region (**Figure 3**), resulting in algorithms that are more robust to varying ambient illumination conditions. Furthermore, NIR light is invisible to the human eye making it a valid, non-invasive solution that can be used at night.

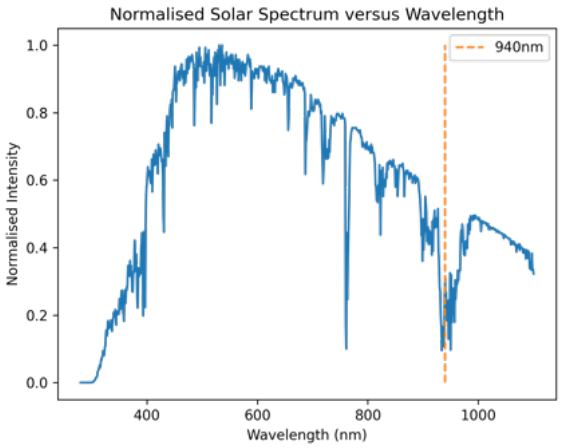

Figure 3: *Normalised solar emission spectrum. A significant decrease in relative intensity is observed at wavelength 940nm due to absorption by H_2O in the atmosphere [22].*

2.3 Blood Volume Pulse

The blood volume pulse (BVP) is closely related to photoplethysmography (PPG), and the two terms are often used interchangeably in the rPPG research field. BVP refers to the rhythmic changes in blood volume that occur with each heartbeat. PPG technology measures these changes in blood volume by taking advantage of the different light absorption characteristics of oxygenated and deoxygenated Hb. Because both BVP and PPG are derived from the same physiological phenomenon—the pulsatile flow of blood—they are directly correlated and represent the same underlying signal.

2.4 Electrocardiogram

The beating of the heart is regulated through the release of electrical impulses by the cardiac conduction system. These signals can be directly measured from the surface of the skin using electrodes. This process is referred to as electrocardiography.

The electrocardiogram (ECG) signal is not detectable through remote sensing means. Despite this, it is an important signal for the accurate validation of metrics like HRV. This is because the peaks of the ECG signal are significantly more localised than the peaks of the PPG signals. Furthermore, unlike other heart rate measurement techniques, ECG is resilient to noise factors like motion and heterogeneous illumination.

2.5 Ballistocardiogram

Ballistocardiography seeks to measure the reaction (displacement, velocity or acceleration) of the whole body resulting from ballistic forces generated by the heart. More specifically, the downward pumping of blood through the descending aorta produces an upward reactive recoil of the body. This recoil can be measured to determine the mass movements of circulating blood.

The displacement corresponding to the ballistocardiogram (BCG) signal ranges from 0.1 mm to 0.2 mm [23]. Accordingly, it is extremely difficult to reliably measure the BCG signal using a camera in a practical application context. Nonetheless, existing research has demonstrated it is possible to remotely measure the BCG signal using a camera in a controlled settings [13], [23]. This signal is denoted as the imaging ballistocardiogram (iBCG). Accordingly, it is

important to acknowledge the existence of BCG as a viable signal that may be used to boost the performance of existing rPPG algorithms.

2.6 Respiration

While not directly correlated with heart rate, respiration significantly influences cardiovascular dynamics through respiratory sinus arrhythmia, where heart rate varies with the breathing cycle. This makes respiratory information valuable in physiological monitoring.

Respiratory waveforms can be detected using piezoelectric sensors on chest straps or derived from ECG signals, where the movement of electrodes relative to the heart provides measurable variations. Alternatively, the respiratory rate can be estimated from PPG data, though this method is generally less precise due to the indirect nature of the signal.

Capturing respiratory information from camera-based systems is challenging, but recent methods have had great success with multi-modal deep-learning approaches that jointly learn to predict both the rPPG and respiration signals [15], [16], [18], [24]. Such models provide a comprehensive assessment of vital signs while retaining the computational complexity of a single modality model. Furthermore, the physiological link between the two signals allows these multi-modal models to leverage their correlation for improved performance.

2.7 Camera-based Vital Measurement

Remote camera-based vital sensing seeks to harness subtle variations in the human body to detect a subjects physiological signals. These variations may be any one/combination of the rPPG, iBVP, iBCG, and respiratory waveform signals.

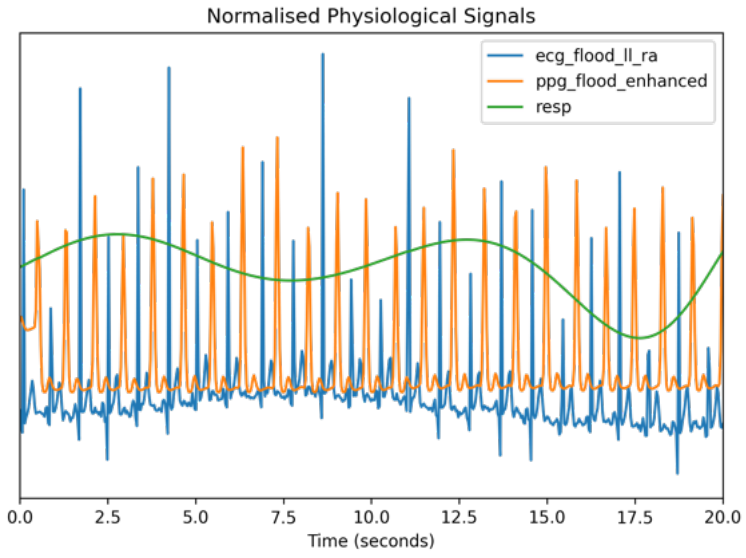

Figure 4: *The ECG, PPG and respiratory physiological signals*

Traditional signal-processing video-based physiological sensing frameworks begin with face lanbdmark detection, followed by segmentation, colour space transform, and filtering of the desired physiological signal [25], [26], [27], [28]. Such pipelines, shown in **Figure 5**, are difficult to implement, computationally expensive, and rely on handcrafted feature extractors that generalise poorly.

State-of-the-art methods have shifted to deep-learning based approaches [15], [16], [17], [18]. These models are more robust to heterogeneous illumination and other noise factors like subject movement; outperforming complex, handcrafted feature extractors. Furthermore, these deep

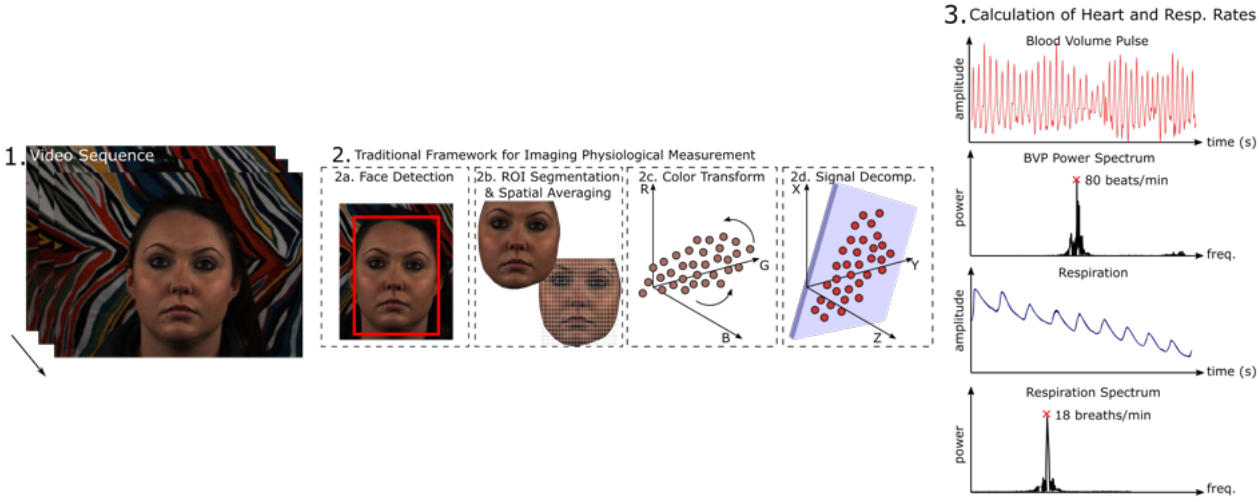

Figure 5: *Traditional unsupervised camera-based physiological sensing frameworks involve complicated, hand-crafted preprocessing and feature extraction steps* [15].

learning-based approaches have the capacity to be entirely end-to-end solutions, thereby saving significant compute resources and running in real-time on embedded targets.

2.8 Structured Coherent Light Illumination

As shown in **Figure** 1, a traditional DMS system typically uses one or more infrared light sources. NIR illumination at 850nm or 940nm is preferred due to reduced sunlight interference at these wavelengths. Conventionally, light emitting diodes (LEDs) are the chosen NIR light source for these systems due to their low-cost, high-efficiency, and ubiquity.

Recently, vertical-cavity surface-emitting lasers (VCSELs) have gained increasing popularity. Unlike traditional edge-emitting lasers, VCSELs are easier to mass-produce and are now used in devices like smartphones, time-of-flight sensors, and optical communications. While VCSELs remain more expensive than LEDs, they are seen as a promising alternative due to their potential for new applications that require coherent light sources.

The intensity profile of an individual laser emitter is far too small in diameter to evenly illuminate an entire scene. To overcome this limitation, it is common for hundreds of laser emitters to be combined to form an array that is passed through a diffusor and a diverging lens in a configuration referred to as a flood illuminator. At present, to ensure a stable phase relationship is maintained for imaging application scenarios, the number of emitters in a flood illuminator is restricted by manufacturing limitations to order of magnitude 10^2. VCSELs have greater efficiency, higher optical power output, narrower bandwidth, less latency, greater temperature stability and greater customisability than LEDs. For example, it is common for an aspherical lens to be used in conjunction with a beam splitter to create a structured light emission pattern. This configuration is referred to as a dot project (**Figure** 6).

Existing research has shown that dot pattern projector beam profiles can be used for material classification [14]. This classification signal could, in theory, guide a physiological sensing model to focus on regions of an image where the physiological signals are encoded (i.e., skin), while avoiding other areas. Additionally, dot pattern projectors can be paired with a camera to mimic a stereo setup and derive depth data through a process called depth from structured light [12]. Furthermore, no mass-manufactured dot projector is exhibits a perfectly discontinuous emission profile. Accordingly, a dot pattern project illumination profile can be thought of as a flood emitter operating in series with a structured light emitter.Thus, it should be possible to

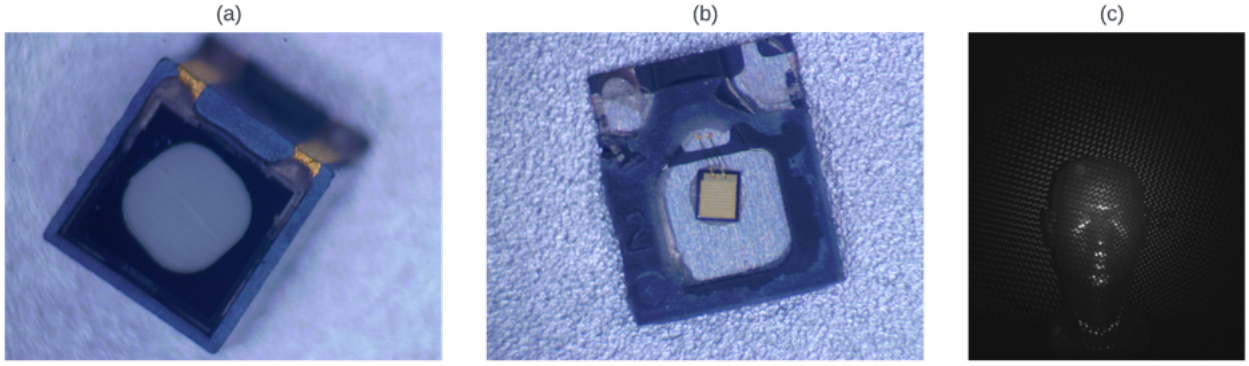

Figure 6: **(a)** *A dot pattern projector.* **(b)** *The same dot pattern projector with the diffusive cover removed.* **(c)** *An example of a scene illuminated by an NIR dot pattern projector. Notice that the light is focused in discontinuous regions throughout the scene.*

recover the flood emission rPPG signal from a scene illuminated by a dot pattern project, plus these additional depth and material classification signals. This concept is the core motivation of this project.

3 Literature Review

3.1 Camera-based Vital Measurement

3.1.1 Unsupervised Methods

Filtered Average Pixel Intensity. [29]

Perhaps the most rudimentary method of deriving the rPPG signal is to perform a series of filtering operations on the average pixel intensity of a static region of skin pixels. This is the approach utilised by Lindvqist and Lindelow [29], where a series of pre-filtering operations are performed to isolate the periodic rPPG signal from the raw pixel intensity values (**Figure 7**). Welch's method is then used to produce a heart rate estimate based on the filtered pixel intensities.

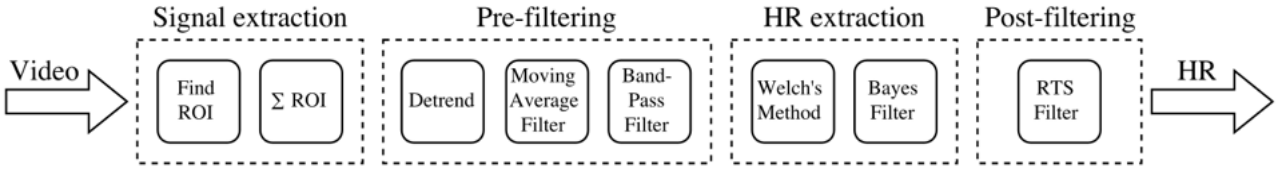

Figure 7: *The processing pipeline for camera-based rPPG proposed by [29].*

Firstly, a set of face landmarks are extracted from a face tracking algorithm. The authors use Smart Eye Pro, a proprietary headtracker developed by DMS company SmartEye[1]. Using the landmarks, regions-of-interest are defined encompassing the face, cheek, forehead and nose. For each frame, the average pixel intensity of each region is calculated. A high pass finite impulse response (FIR) filter with a cut-off frequency of 0.38 Hz is applied to the average intensity values to remove trends such as small head movements and lighting changes. A moving average filter with a temporal window duration of 0.125 seconds is then applied to remove high frequency dynamic noise. Finally, a Hamming window bandpassed FIR filter with cut-off frequencies [0.67, 4] Hz is applied to remove frequencies outside of the feasible range of heart rate values. After pre-filtering, the heart rate is calculated over a 15-second sliding window using Welch's method. A 50% overlap is used for each segment. A Bayes filter is then applied to select the most probable heart rate peak in the power spectral density domain.

The proposed method is capable of isolating the rPPG signal in controlled scenarios. To verify this, the authors collect a stationary dataset of ten subjects with 60-seconds of video per subject. They also collect a motion robustness dataset consisting of four subjects who perform scaling and translational movements in four different velocities with a duration of 30 seconds per movement, per velocity. Subjects are recorded in 850 nm light for all recordings.

The authors record a MAE of 0.02 BPM in predicted heart rate on the static dataset. For the motion robustness dataset, the authors find their method performs acceptably for velocities below 6 cm/s (< 8 BPM MAE). However, for any larger velocities performance rapidly degrades. This result is not suprising considering the methodology lacks any consideration of the impact of motion on the specular reflectance component of each skin pixel, as well as its impact on the variation in relative intensity of the source illumination source. Regardless, these findings demonstrate that the rPPG signal can effectively be derived from a monochromatic NIR system with an incoherent flood illumination source. Notably, the pre-filtering steps of this approach were implemented at the inception of this project to verify the existence of the rPPG signal in frames illuminated by a structured coherent illuminator (**Appendix A**).

[1]SmartEye is a direct competitor to Seeing Machines. The contents of this summary are entirely my own and it should not be assumed that they reflect the views of Seeing Machines.

LGI. [28]

Traditional blind source separation (BSS) rPPG methods struggle with signal degradation caused by head movement, facial expressions, and changing light conditions. Local group invariance (LGI) [28] is a novel unsupervised approach to heart rate estimation from face videos by incorporating prior knowledge about the invariance of facial transformations. This approach introduces features that are invariant to these disturbances, allowing the algorithm to isolate the rPPG signal from challenging scenarios. By applying group theory, the authors transform the blood volume signal into a more concentrated feature space, leading to improved heart rate predictions under real-world conditions. This method outperforms existing algorithms in dynamic scenarios, highlighting its potential for robust physiological monitoring.

Consider a pixel $\mathbf{p} \in \mathbb{R}^3 = \{R, G, B\}$. For a skin pixel, the expectation value can be written as dependent on a skin operator, s

$$\mathbb{E}[\{\mathbf{p}|s(\mathbf{p})\}] \tag{1}$$

The authors assume the time-dependent spatial expectation is drawn by a normal distribution

$$\mathbf{x}(\mathbf{t}) = \int_0^\infty \mathbb{E}[\{\mathbf{p}|s(\mathbf{p})\}]dt \sim \mathcal{N}(\mu, \sigma^2) \tag{2}$$

The authors go on to define a set of differentiable local transforms for local groups of $\mathbf{x}(\mathbf{t})$ that can be enforced by minimising a regulariser. By minimising the regulariser, the distribution of the transformed features becomes more compact, reducing the impact of variations caused by noise or motion. This transformation helps isolate the relevant signal related to physiological changes, such as blood volume variations, by ensuring that the features remain robust against disturbances.

To assess the performance of their approach, the authors collect a dataset of 25 subjects who are recorded while stationary, moving, on an exercise bike, and talking. Each session lasts 1-minute, except the exercise bike session which lasts 5-minutes. Subjects are recorded using an RGB webcam at 60 frames-per-second. A face detector is used to locate regions of interest. Skin segmentation is then performed by thresholding a blue and red chromatic difference. For the set of obtained pixels, the expectation value is computed and stored as an $\mathbb{R}^3$ time vector. The rPPG signal is derived from the raw data using independent component analysis (ICA) [30], spatial subspace rotation (SSR) [31], plane-orthogonal-to-skin (POS) [32], and LGI [28]. The output signal of each algorithm is bandpassed filtered about [0, 2.0] Hz, with the exception of the exercise bike session where the upper threshold is increased to 2.5 Hz.

Table 1: *Comparison of LGI BPM MAE for different behaviour configurations [28].*

Session	**ICA** [30] MAE↓	**SSR** [31] MAE↓	**POS** [32] MAE↓	**LGI** [28] MAE↓
Resting	1.4	2.0	2.1	3.3
Rotation	10.8	7.6	5.3	2.9
Gym	16.6	18.6	23.1	13.1
Talk	21.1	15.4	12.5	4.3

In all dynamic configurations, LGI outperforms the other algorithms, with the exception of the static configuration where LGI underperforms all other algorithms. The authors later clarify this can be attributed to an estimation bias of 4 BPM, obtained from a Bland-Altman plot of the estimation performance across the entire dataset.

CHROM. [25]

Chrominance (CHROM)-based rPPG [25] is one of the most widely-cited rPPG techniques. CHROM is a unsupervised rPPG method that uses a colour difference (chrominance) between RGB pixels to remove specular reflections. CHROM significantly outperformed pre-existing BSS techniques that assumed the component with the strongest periodicity corresponds to the rPPG signal. Consequently, periodic motion as in a fitness setting, resulted in significant performance degradation. CHROM sought to address this deficiencies by using a colour difference to remove specular reflections and then incorporating a skin-tone normalisation term that is initially estimated in a data-driven manner, then dynamically updated to correct for improper estimation.

Consider the intensity of a pixel in an image i in colour channel $C \in \{R, G, B\}$. The intensity of the pixel can be written as a modulation of the light source intensity, I_{C_i}, where $\rho_{C_{dc}}$ is a diffuse reflection component, s_i is a specular reflection component, and ρ_{C_i} is the zero-mean time-varying fraction caused by the pulsation of blood.

$$C_i = I_{C_i}(\rho_{C_{dc}} + \rho_{C_i} + s_i) \tag{1}$$

By normalising the colour channel of a pixel by its average value of a temporal window greater than one pulse duration, a pulse signal independent of I_{C_i} is found. Assuming that intensity modulations caused by movement of the skin are equal for all channels, taking the ratio of a channel where the pulse is strong (e.g., the green channel), and dividing by a channel where the pulse is weak (e.g., the red channel) undoes the impact of such motions. The remaining issue is the specular reflection term. Assuming a white light source (meaning the specular reflection affects all pixels equally), by taking the colour difference signals—i.e., chrominance—the specular reflection is undone. CHROM combines all of the above techniques, defining a set of orthogonal chrominance signals, X and Y, then taking their ratio to get the pulse signal, S:

$$X_i = R_i - G_i \quad Y_i = 0.5R_i + 0.5G_i - B_i \tag{2}$$

$$S = \frac{X_n}{Y_n} - 1 \tag{3}$$

Where n indicates the signals have being normalised by their average values. Lastly, to enable correct functioning with coloured light sources, the authors propose a skin-tone standardisation. After a large-scale data collection, the authors find a standardisation term that, when combined with normalisation using each colour channels mean, gives an algorithm that can work correctly regardless of the colour of illuminant.

$$R_{s_i} = 0.7682\frac{R_{C_i}}{\mu(R_{C_i})} \quad G_{s_i} = 0.5121\frac{G_{C_i}}{\mu(G_{C_i})} \quad B_{s_i} = 0.3841\frac{B_{C_i}}{\mu(B_{C_i})} \tag{4}$$

$$S = \frac{X_s}{Y_s} - 1 \tag{5}$$

with

$$X_s = \frac{R_s - G_s}{0.7682 - 09.5121} = 3R_n - 2G_n \tag{6}$$

$$Y_s = \frac{R_s + G_s - 2B_s}{0.7682 + 0.5121 - 0.7682} = 1.5R_n + G_n - 1.5B_n \tag{7}$$

Finally, to correct for misestimation of the skin-tone standardisation coefficients, the authors propose X_f and Y_f which are bandpassed filtered versions of X_s and Y_s. By taking the ratio of standard deviations of the bandpassed filtered signals, the in-band disturbances of the output signal are minimised.

$$S = X_f - \alpha Y_f \tag{8}$$

with

$$\alpha = \frac{\sigma(X_f)}{\sigma(Y_f)} \tag{9}$$

In a dataset of 117 stationary subjects , CHROM achieves a RMSE of 0.5 BPM, while ICA and PCA achieve an RMSE of 1.1 and 0.9 BPM, respectively. When considering a population exercising on an exercise bike and an optimisation interval of 1.6 seconds, CHROM is within 3 BPM of the measured pulse rate of 98% of samples. Comparatively, ICA and PCA achieve around 79%, and require a significantly longer optimisation interval of 25 seconds. Finally, for subjects recorded performing a stepping sequence, CHROM corresponds to the actual pulse rates 48% of the time, while ICA and PCA do not exceed 11% regardless of optimisation interval. In summary, CHROM was a significant milestone in the rPPG research space towards robust performance in uncontrolled scenarios. CHROM remains widely-used as a benchmark for comparison of new methods.

SparsePPG. [26]

SparsePPG is an unsupervised rPPG technique developed for DMS. The authors highlight that the main challenges in applying rPPG in DMS are subject motion and heterogeneous illumination. They propose that the issue of varying illumination can be effectively mitigated by using a camera with a narrow bandpass filter centered at 940 nm. By adaptively selecting optimal facial regions, the authors propose a methodology for denoising the estimated rPPG signals that leverages the principles that the pulsatile signal should be sparse in the frequency domain and exhibit low-rank characteristics across different facial regions. A high-level summary of the SparsePPG methodology is shown in **Figure 8**.

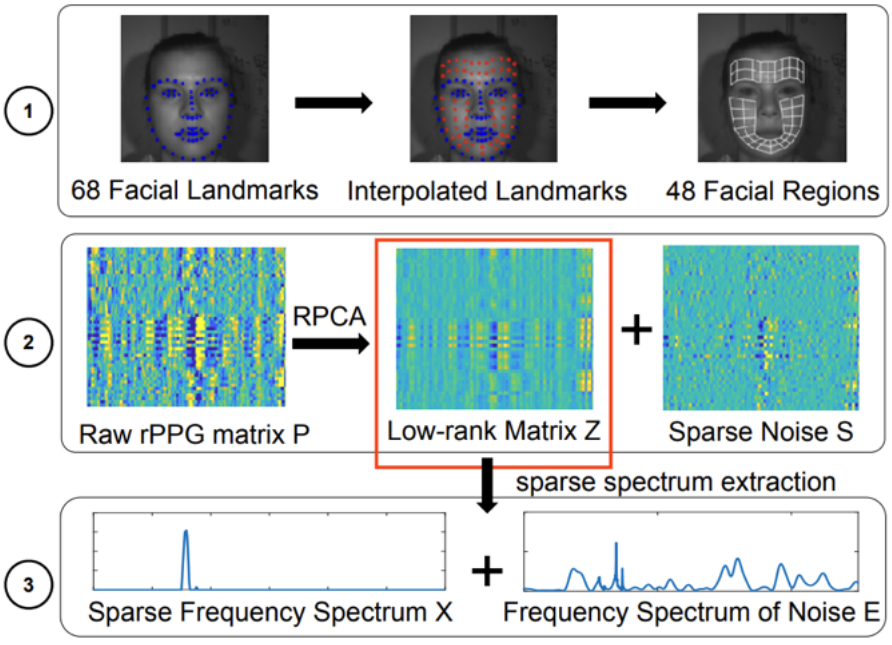

Figure 8: *Overview of the SparsePPG methodology [26].*

Consider N facial regions focused around the forehead, cheek, and chin area. For every facial region, $j \in \{1, ..., N\}$, the measured mean pixel intensity, $p_j(t)$, is recorded across T video frames. The measured signal is modelled as:

$$p_j(t) = h_j(t) * y_j(t) + n_j(t) \tag{1}$$

Where $y_j(t)$ is the pulsatile signal at region j, $h_j(t)$ is the channel impulse response and $n_j(t)$ is the channel noise. Rewriting the pulsatile signal in frequency form where $x_j \in \mathbb{C}^T$ denotes the sparse frequency spectrum of the heartbeat signal $\boldsymbol{y_j} \in \mathbb{R}^T$:

$$\boldsymbol{p_j} = \boldsymbol{h_j} * \boldsymbol{\mathcal{F}^{-1}} \boldsymbol{x_j} + \boldsymbol{n_j} \tag{2}$$

The authors go on to exclude the impact of the channel impulse response function considering only:

$$\boldsymbol{p_j} = \boldsymbol{\mathcal{F}^{-1}} \boldsymbol{x_j} + \boldsymbol{n_j} \tag{3}$$

SparsePPG is unique in its decision to model the rPPG signal as a multi-channel wireless communications problem. Through this formulation, the authors formulate an optimisation problem and use robust PCA to recover a denoised version of the rPPG signal.

To evaluate their model, the authors collect a dataset of 12 subjects recorded under static illumination with natural motion using both RGB and NIR cameras. On the NIR footage, SparsePPG achieves a RMSE of 13.6 BPM with a percentage of time that the heart rate error is less than 6 BPM (PTE6) of 79.7%. Comparatively, on the RGB footage CHROM achieves an RMSE of 13.6 BPM with PTE6 85.9%. Notably, SparsePPG outperforms CHROM on the RGB footage achieving an RMSE of 9.6 BPM with a PTE6 of 88.5%.

The pre-processing pipeline for SparsePPG requires complex facetracking, segmentation and averaging steps that introduce undesirable dependencies and complexity. For example, the authors are unable to robustly apply their algorithm in the case of rapidly varying ambient light due to poor face tracking results. Despite this, their findings suggest the the use of narrowband NIR illumination is a valid pathway to robust rPPG in a DMS application context.

3.1.2 Neural Methods

DeepPhys. [15]

In 2016, Chen and McDuff introduced DeepPhys, a convolutional neural network (CNN) architecture for rPPG. DeepPhys is a seminal model in the rPPG research space, having been the first deep-learning approach to significantly outperform unsupervised methods. Chen and McDuff were motivated by the idea of an end-to-end model that reads motion information from video frames and jointly learns a spatial mask to detect regions-of-interest and recover the rPPG and respiration signals. At time of inception, this was a unique idea as existing rPPG algorithms were solely focused on recovering the rPPG signal embedded in temporal variations of pixel intensity [29], [28], [25], [26].

The basic principle of DeepPhys is to use sequential difference frames to generate a motion representation that can better capture physiological motions under heterogeneous illumination. This motion representation is jointly learned alongside an auxiliary appearance representation that acquires attention from the human appearance. Both branches use a VGG-style backbone [33]. Notably, this approach reverses the traditional use of attention, which typically derives attention from motion to guide the learning of appearance representation. The output of the appearance branch is a set of masks that are convolved with the motion branch to assist the motion learning. The attention masks help the network minimize the negative effects introduced by motion and lighting noise.

Accordingly, the inputs to DeepPhys are raw video frames (appearance branch inputs) coupled with normalised difference frames (motion branch inputs). These frames are downsampled

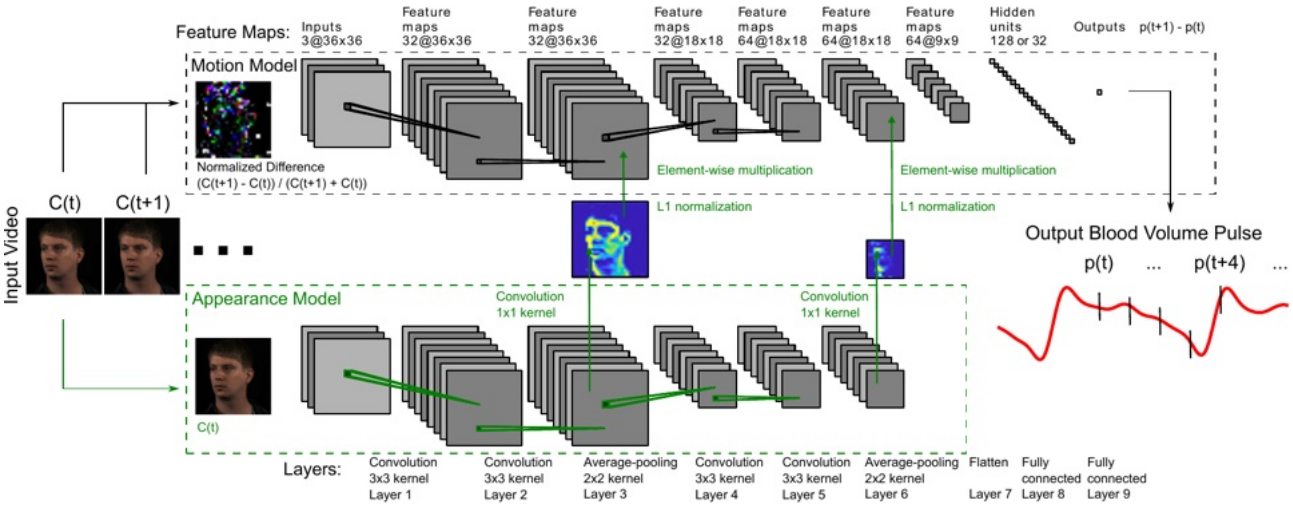

Figure 9: *DeepPhys architecture summary [15].*

to L pixels by L pixels using bicubic interpolation. The authors found $L = 36$ to be a good trade-off between suppressing camera noise while retaining spatial information.

The training signal for DeepPhys is the derivative of the PPG signal. The objective of the model is to predict this differentiated PPG signal using the input data. As is standard with papers in the physiological sensing space, heart rate is computed by taking the dominant frequency component of the physiological signal. Comparatively, the use of an mean-squared error loss function means the model is trained to minimise the temporal error. Although these generally have high correlation, a small temporal error does not guarantee a small frequency error. To counteract this, the authors use an ensemble learning approach where models were trained for an extra 16 epochs after convergence. These models were applied to the training data and the model with the smallest frequency error was chosen.

On all tasks of the AFRL [34] and MAHNOB-HCI [35] datasets, DeepPhys was able to significantly outperform traditional handcrafted signal-processing approaches like ICA [30], [32], and PBV [36] (an extension of CHROM). For example, on the rapid random head movement task of the AFRL dataset, DeepPhys outperformed ICA and PBV by more than ten times (13.4 and 10.3 versus 1.78 BPM MAE). Furthermore, the authors demonstrate their model is capable of running on monchromatic NIR videos, achieving a MAE of 0.55 BPM on the MAHNOB-HCI dataset.

MTTS-CAN. [16]

MTTS-CAN was the first application of a temporal-modeling to the rPPG space. Temporal shift modules (TSMs) allow CNNs to capture complex temporal dynamics with the computational efficiency of 2D convolutions [37]. This is of particular relevance to rPPG, where high frame-rates (at least 100 Hz) are required for the precise measurement of arterial fibrillation [38], hypertension [39], and heart rate variability [40]. These were the principles that guided Liu, Fromm, Patel, and McDuff in the construction of MTTS-CAN, a novel multi-task temporal shift convolutional attention network that runs at 150 frames-per-second (FPS) while achieving state-of-the-art performance on benchmark datasets.

The optical model proposed by the authors extends upon that which is outlined in [15] by leveraging the interdependence of the respiration and pulse signals to define a new rPPG signal that is a combination of iBCG, rPPG, and respiration. Since the rPPG and respiration signal are closely intertwined, the authors argue a temporal multi-task learning approach seems optimal as it could leverage the cross-correlation of the two signals.

The architecture of MTTS-CAN is identical to DeepPhys [15], with the addition of temporal-shift modules (TSMs) before each conv2D block in the motion branch. This allows the model

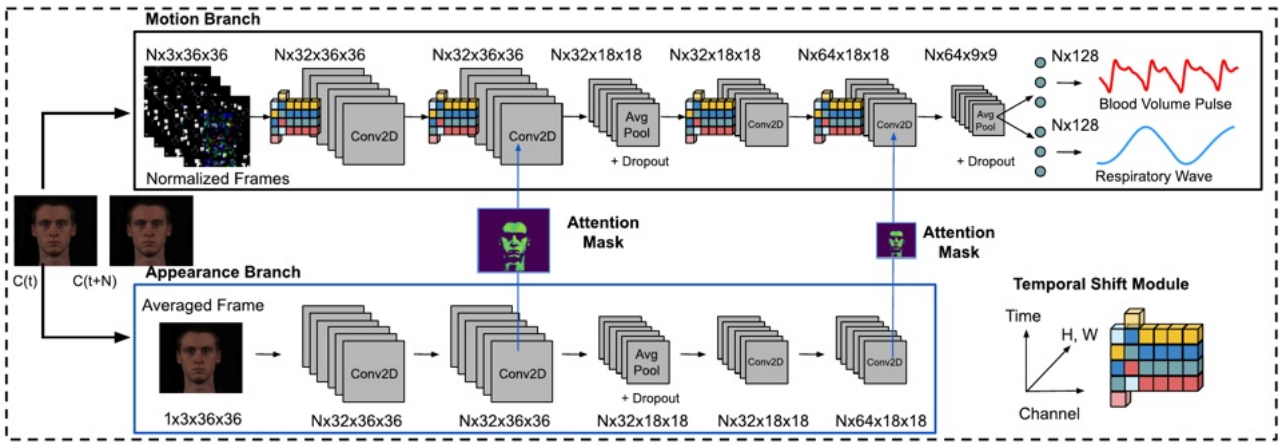

Figure 10: *MTTS-CAN architecture summary [16].*

to efficiently capture complex temporal dynamics that extend beyond those within a single set of consecutive frames—i.e., difference frames. Unlike DeepPhys, the input to the appearance branch is a single averaged frame. This is done so that the attention mask (1) need only be calculated once, (2) captures most of the pixels that contain skin, and (3) reduces camera quantisation error.

Through these changes, MTTS-CAN operates at 150 FPS, while achieving state-of-the-art prediction performance. In particular, the authors find that the addition of TSMs leads to greater resilience to noise (e.g., motion, varying ambient illumination, etc.). Through a systematic evaluation on videos with varying velocities of rotational head motion, the addition of TSMs is shown to perform strongly on tasks with greater velocity head motion. Moreover, although tensor shifting provides important temporal information, it also introduces extra noise. Despite this, the authors find that the attention module is effective at separating the desired signal from the added noise.

EfficientPhys. [17]

EfficientPhys is a lightweight architecture for camera-based vitals measurement with an emphasis on low-latency performance. Prior neural rPPG methods rely on hand-crafted pre-processing steps like normalised difference frames [15], [16]; and traditional signal-processing methods rely on facial landmark detection, segmentation, colour space transforms, and pixel averaging. These pre-processing steps are computationally expensive and prevent the network from learning these features in a data-driven manner. Furthermore, most of these steps are non-trivial to implement and optimise for real-time performance on an embedded target but privacy preservation is crucial for camera-based physiological sensing meaning cloud-based processing is strongly undesirable.

EfficientPhys seeks to address these deficiencies by building upon pre-existing research in the rPPG deep learning space (primarily DeepPhys [15] and MTTS-CAN [16]) while seeking to provide a truly end-to-end solution that requires no pre-processing steps and can be practically run in real-time on a mobile device while maintaining state-of-the-art performance on heterogeneous scenes.

The authors propose two variants of EfficientPhys: one with a convolutional architecture and another with a transformer architecture. The input to either variant is unprocessed video frames without accurate face cropping. Before the stem layer of each model, a normalisation module is inserted that performs equivalent pre-processing to what is described for DeepPhys [15]. The primary difference between the two is that the batchnorm implementation allows the model to learn the optimal normalisation parameters that can magnify the contained physiological signals in comparison to a bespoke hand-crafted normalisation operation.

In designing the convolutional network architecture, the authors effectively flattened the

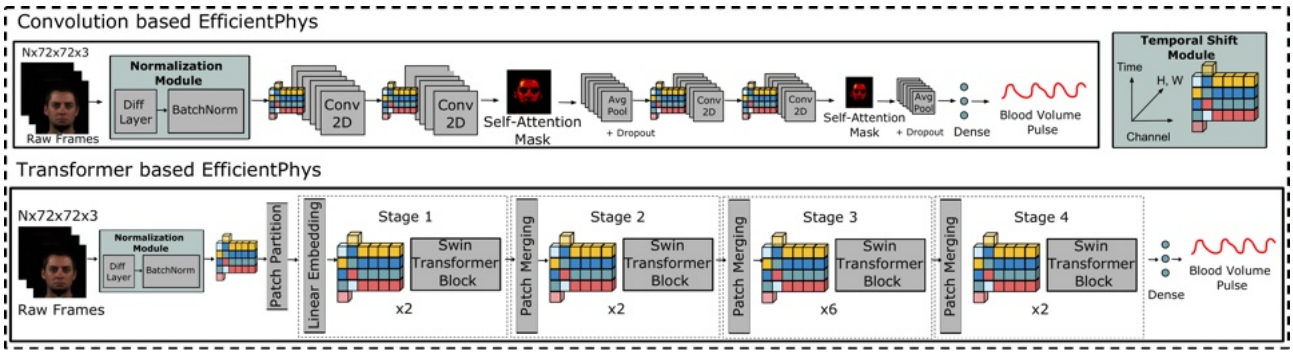

Figure 11: *EfficientPhys architecture summary [17].*

two branches of MTTS-CAN into a single branch end-to-end network. This is done under the assumption that the attention masks can be learned through self-attention modules, rather than requiring a separate appearance branch. They are softmax attention layers with 1D convolutions followed by a sigmoid activation function. Then, normalisation is applied to remove the outlier values in the attention mask, and the final normalised attention mask is element-wise multiplied with the output from the tensor-shifted convolution.

Comparatively, the transformer-based architecture replaces the conv2D blocks with 2D Swin transformer blocks [41]. Inspired by the idea of shifting spatial window partitions, the authors add TSMs before each Swin transformer block to facilitate information exchange across the temporal axis. The TSM modules give the architecture the ability to perform efficient spatial-temporal modeling and attention by combining shifting window partitions spatially and shifting frames temporally. Notably, the TSMs do not introduce any learnable parameters.

The authors found the convolutional-based architecture to significantly outperform the transformer-based architecture in both accuracy and latency. Performance further degraded when the transformer architecture was shrunk to use a similar number of parameters to the convolutional architecture. These results indicate shallow transformer architectures struggle to model subtle changes of skin pixels. However, the authors do acknowledge transformers typically require larger datasets to achieve state-of-the-art results while there is a limited quantity of data available in the rPPG research space.

BigSmall. [24]

BigSmall is a multi-task neural model for disparate spatial and temporal physiological measurements. BigSmall produces predictions for PPG, respiration, and facial activation units. The primary inspiration behind BigSmall was that a multi-task model could leverage cross correlation between these three signals to produce more robust and reliable predictions. Ultimately, the authors found this to be false, however, they were able to produced a multi-task model that could efficiently produce predictions for all three signals with the same computational cost as a single-task model.

BigSmall is comprised of a Big branch that takes high-resolution inputs for deriving spatial texture features, and a Small branch with downsampled inputs that compresses noise from spatial features while retaining temporal dynamics. Both the Big and Small branches have a VGG-style CNN backbone [33] similar to DeepPhys [15], however, the small branch includes the addition of wrapping temporal-shift modules (WTSMs). These modules are an extension of traditional TSMs [37], where shifted-out folds are wrapped to fill previously zero-padded folds. Thus, WTSM can leverage the temporal benefits of TSM without increasing the proportion of zeroed features—even with a small N. This is particularly important for facial AU classification where N must be small for optimal performance [42]. Furthermore, the temporal information added by WTSM is not dependent on time-series order meaning it helps convolutions learn a time-invariant mapping. As a result, shared features between non-adjacent frames do not

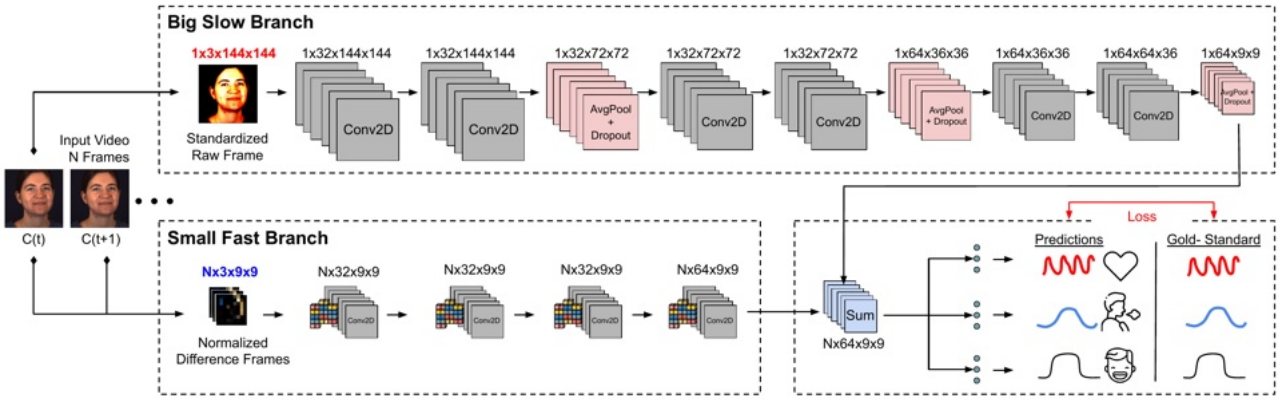

Figure 12: *BigSmall architecture summary [24].*

disturb the temporal representation, and the architecture learns a representation that best balances the information across all frames.

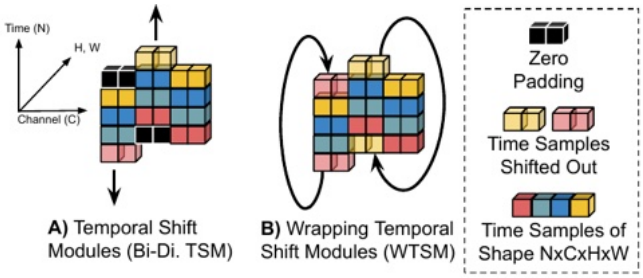

Figure 13: *A comparison of TSM and WTSM. For modeling time variant signals with small window sizes, the authors found that WTSMs provide superior performance [24].*

Notably, the authors show that training a model on a single modality and then fine-tuning the learned embedding on other modalities resulted in significantly worse performance than task-optimised models. This findings suggests that, despite all signals occurring in the face, each physiological signal manifests in a distinctly unique way. BigSmall was able to match state-of-the-art rPPG performance, while simultaneously outputting respiration and facial activation unit expressions for the same computational cost as a single modality network. These results highlight the capacity for models to jointly learn physiological signals.

3.1.3 rPPG-Toolbox.

rPPG-Toolbox [43] is an open-source repository designed to standardise the training, testing, and comparison of vision-based physiological sensing models. Existing papers in the rPPG research space have custom pre-processing and post-processing pipelines that makes meaningful performance comparisons difficult. For example, different researchers compute the ground-truth heart rate signal using their own methods (e.g., frequency-based, peak-based, window sizes, etc.). This leads to different labels, causing a fundamental issue when it comes to benchmarking performance. Furthermore, researchers often benchmark their designs against so-called "baseline" unsupervised algorithms (e.g., ICA [30], CHROM [25], POS [32]) but it is an inefficient use of researchers time to repeatedly re-implement these methods.

The authors of rPPG-Toolbox seek to address these issues by releasing an end-to-end toolbox for camera-based physiological measurement. rPPG-Toolbox includes implementations of all of the aforementioned unsupervised and neural camera-based physiological sensing methods except SparsePPG [26]. It also includes support for six public datasets, pre-processing code to format the datasets for neural models, implementations of additional neural and unsupervised

learning methods, evaluation and inference pipelines for supervised and unsupervised methods, and advanced neural training and inference such as weakly supervised psuedo labels, motion augmentation, and multi-task learning.

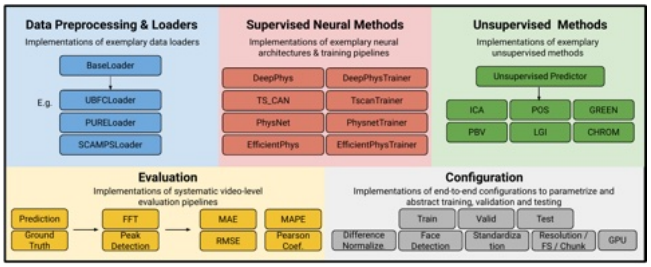

Figure 14: *An overview of the rPPG Toolbox-codebase [43].*

In releasing the toolbox, the authors publish clear and reproducible benchmarks for each dataset. They also release pre-trained models to allow researchers to perform model inference. A detailed discussion of the metrics output by rPPG-Toolbox and their calculations is provided in **Appendix B**. rPPG-Toolbox sets the foundation for rigorous and informative comparison of existing and novel camera-based physiological sensing algorithms. The core training pipeline from which the results of this project were generated was built atop of the rPPG-Toolbox repository.

4 Dataset

4.1 SM-EOD

Datasets in the rPPG research field are limited, primarily due to privacy concerns related to their collection and distribution. Commonly used datasets include AFRL [34], UBFC-rPPG [44], PURE [45], UBFC-PHYS [46], MMPD [47], SCAMPS [48], VIPL-HR [49], and MAHNOB-HCI [35]. However, accessing these datasets can be challenging. For instance, AFRL lacks clear instructions on how researchers can obtain access, while datasets like MMPD require a formal request for use. Importantly, all of these datasets prohibit commercial use, which directly impacts this project as it is sponsored by a commercial entity and its findings are the intellectual property of Seeing Machines. Consequently, using these datasets would not be suitable for this work.

As a result, it was necessary to create a new dataset consisting of synchronised camera and physiological data. While the data collection process consumed a significant duration of the projects timeline, this challenge also became a key advantage. The newly assembled dataset ensured that the models were evaluated on an optical setup that closely mirrors the technology Seeing Machines offers to its customers, providing greater relevance and alignment with commercial applications.

The dataset was defined through two documents: a configuration document and a protocol document. The configuration document details the physical configuration of the data collection such as cameras, sensors, and software. The protocol document details unique activities that are specific to this data collection and are likely to change for future data collection activities.

4.1.1 Configuration

When designing the data collection study configuration, it was important the methodology sought to replicate existing Seeing Machines systems. This was a critical detail in ensuring the projects results accurately reflected the performance attainable using existing Seeing Machines systems.

Hardware.

Seeing Machines typical system is a wide-field of view camera system that operates using 940 nm NIR light. Paradoxically, the rPPG signal is significantly degraded in the NIR spectrum—about ten times weaker than in the green spectrum of visible light [26]. Despite this, the NIR spectrum offers superior sunlight rejection performance and can be utilised non-invasively at night time. Accordingly, it was a fixed requirement that this data collection utilise cameras equipped with NIR illumination sources, and that an NIR narrow bandpass filter was fitted within the optical pathway.

DMS are typically packaged in the instrument cluster (IC), center console (CC), headliner, or rear-view mirror (RVM). Recently, RVM DMS solutions have become increasingly popular as they offer the capability to perform both DMS and occupant monitoring from a single package. The primary business of Seeing Machines is IC DMS and RVM occupant monitoring systems (OMS). Accordingly, the dataset was collected from two viewpoints: a narrow-field of view camera positioned at the IC and a wide field-of-view camera located at the RVM position.

Seeing Machines also uses a proprietary automatic exposure controller (AEC) to ensure consistent subject appearance under heteregeneous external illumination. In theory, such a controller could significantly degrade rPPG performance as its objective would be to undo the variations in pixel intensity caused by the rPPG signal. After some testing, it was verified that the AEC controller was not sensitive to the rPPG signal though it was still expected that

the AEC controller would degrade rPPG performance. Accordingly, an additional camera was added to the instrument cluster position that did not have an AEC (i.e., was operating with static exposure and gain). This camera provided data most representative of existing public rPPG datasets.

A summary of these three cameras is provided in **Table 2**. Each camera was hand-focused to a target effective working distance of 750mm using Imatest software [50].

Table 2: *Summary of the three cameras used in the project's data collection campaign.*

	Camera 1	**Camera 2**	**Camera 3**
Location	IC	RVM	IC
HFOV (°)	35	160	45
VFOV (°)	25	130	35
Resolution (MP)	5		2.4
Bit depth (bits)	10		12
Frame-rate (fps)	60		
Illumination	4x 940nm LED + 940nm dot-pattern projector		N/A
AEC	✓		✗
3D capable	✓		✗
Bandpass filter (nm)	940 ± 50		
Effective working distance (mm)	750		
MTF50 (lp/mm)	65	77	

A proprietary Seeing Machines camera controller was used to pulse the illumination sources and control the triggering of the camera exposures. The illumination-exposure timing diagram is shown in **Figure 15**.

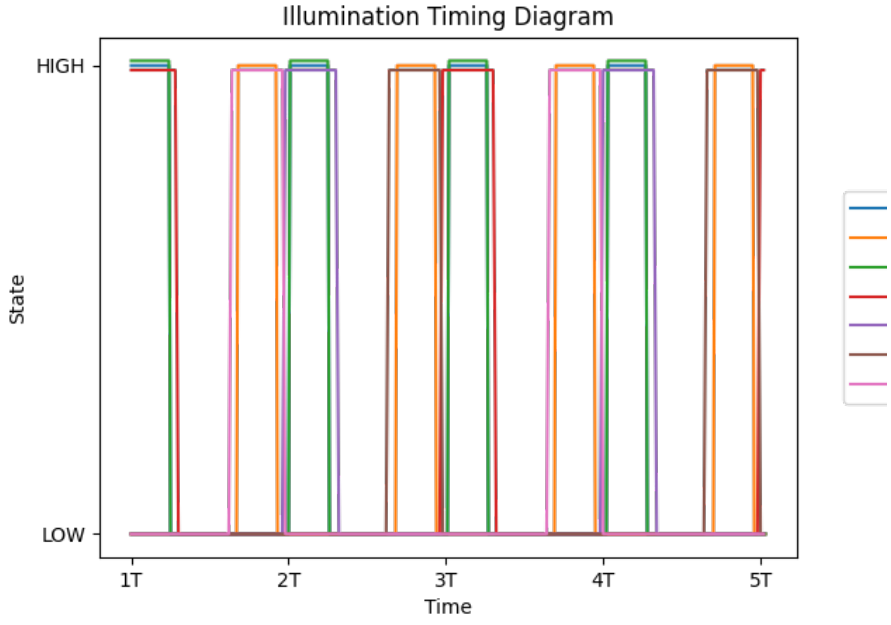

Figure 15: *Illumination-exposure timing diagram. The two IC cameras (camera 1 and camera 3) were exposed synchronously, while the RVM (camera 2) was exposed independently.*

For monitoring of the physiological signals, a Shimmer 3 GSR+ and ExG unit were used with sampling rate set to 1024 Hz. They were controlled using the Shimmer LabVIEW API [51] and synchronised with the recorded video data in post-processing.

A single National Instruments PXIe-1092 was used to record the synchronised data. Recordings were performed inside one of Seeing Machines indoor data collection simulators (**Figure 16**). The simulator consists of 3 Corsair Xeneon Flex 45WQHD420 OLED screens, a MOZO R9 Base, and a MOZA CRP pedal set all mounted on TrakRacer TR160 racing simulator base. The simulator was operated by a custom desktop with a single NVIDIA GeForce 4090 RTX.

Figure 16: *The Seeing Machines indoor data collection simulator used for the data collection campaign. Subjects physiological signals are recorded through the use of ECG electrodes and a finger clip PPG sensor. The physiological sensing data is synchronised with three camera streams: two located on the IC and one located in the RVM position.*

With a total data rate exceeding 1.4 GBps, implementation of this configuration presented a significant engineering challenge. This would not have been possible without the Seeing Machines Test Engineering and Data Engineering Teams and significant praise goes to them for their efforts in delivering this system.

Software.

Recording and signal synchronisation was performed using a custom LabVIEW application (**Figure 17**). This application provided the Data Engineering Team with a live view of the three camera streams as well as the synchronised physiological data waveform. This was important as the Data Engineer was required to continuously monitor the PPG waveform and verify proper contact of the sensor was maintained for the duration of each recording.

The downside of this application was that illumination controller metadata could not be obtained as the illumination control was performed by an independent system that was configured in series. This meant it was not possible to know the gain and exposure parameters for each frame, nor was it possible to encode the illumination channel of a frame (i.e., flood or dot). Instead, the channel information had to be manually initialised and than encoded in post-processing.

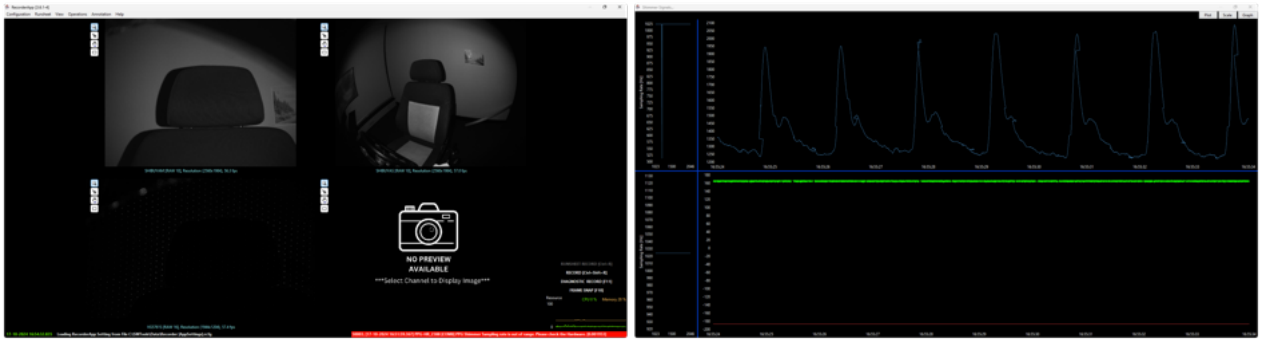

Figure 17: *The DataRecorder App software. A synchronised live view of the three camera streams and physiological signal waveforms is provided so that the Data Engineering Team can monitor the collected data in real-time.*

4.1.2 Protocol

In collecting this dataset, one of the key objectives was to compile a dataset that would be of significant utility to future physiological sensing endeavours at Seeing Machines. Ideally, the protocol should be comprehensive and consider a variety of behaviours that satisfy the needs of this project while considering behaviours that might be desirable for future projects.

Traditional public rPPG datasets focus on motion and/or heterogeneous external illumination noise factors. These datasets fail to capture the diversity of noise factors encountered in a DMS application context, which include: motion, heterogeneous external illumination, makeup, facial hair, skin pigment, masks, eyewear, headwear, intoxication, medication (e.g., stimulants), pacemakers, and more. Accordingly, a protocol was devised that sought to capture a more comprehensive variety of noise factors.

The other objective was to compile a dataset that covered behaviours similar to those in public datasets to allow for meaningful comparison of the projects results.

Typical Subject.

Inspired by the AFRL dataset [34], a variety of tasks with varying degrees of movement were included in the protocol. For a typical subject, three unique guided recordings were specified:

- **iGR1:** Subject fixes their gaze on a target and maintains static head position. Five gaze targets were defined: the instrument cluster, the center console, on-road, the left-side mirror, and the right-side mirror.
- **iGR2:** Subject continuously moves their head position for 2 minutes in a random pattern.
- **iGR3:** Subject repeatedly leans forwards and backwards for 1 minutes.

Lastly, data representative of a typical DMS scene was captured. Namely, a naturalistic recording configuration (**iNR1**) was defined that required the subject to drive for 10 minutes uninterrupted along the Tokyo Shuto Expressway using the simulator rig. Behaviour during this configuration is entirely naturalistic, meaning each video may contain any combination of static, moving, and talking behaviour.

The data collection procedure for a typical subject is shown in **Figure 18**

Makeup Subject.

Makeup products like foundation and contour have a significant impact on the transmission and reflectance of skin [52]. Existing research has shown the reduced transmissivity of cosmetically-enhanced skin degrades the signal-to-noise ratio of the rPPG signal. Although the impact of makeup on rPPG signal strength is more severe in the visible spectrum, a reduction of 54.5% to 61.4% is still observed in the NIR spectrum [53].

Thus, it was important our dataset include data with varying applications of makeup.

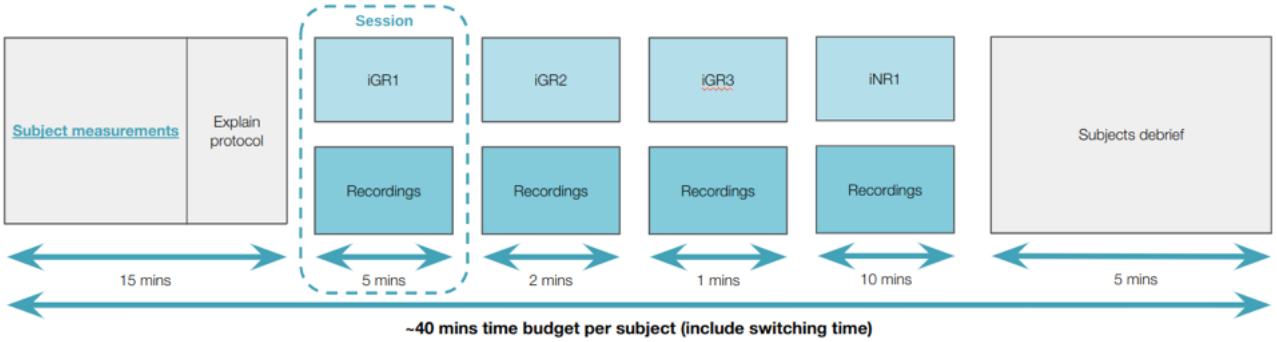

Figure 18: *Data collection procedure for a typical subject.*

More specifically, a protocol was defined requiring a subject to cover half their face (split vertically down the center) with makeup. The subject was then instructed to stare directly at the instrument cluster for a two minute recording (**iGR4**). The intention was not to use this data for training, but as a source of investigation in supplementary evaluation. This analysis is shown in **Section 7.2.4**.

Exercised Subject.

The distribution of average resting heart rate amongst the general population is fairly narrow. As shown in the results of this paper, on the MR-NIRP rPPG public dataset, it is possible to score a MAE of less than 10 BPM by guessing the mean heart rate of the entire dataset. Accordingly, for our dataset, we deemed it important to include a noise factor where a subject is required to elevate their heart rate to 70%-80% of their maximum heart rate (determined using the Karvonen formula) before commencing recording. For exercised subjects, three additional recording configurations are specified:

- **iGR5-1:** Subject elevates their heart rate to zone three before commencing recording. Subject then repeats iGR1.
- **iGR5-2:** Subject elevates their heart rate to zone three before commencing recording. Subject then repeats iGR2.
- **iNR2:** Subject elevates their heart rate to zone three before commencing recording. Subject then repeats iNR1.

Importantly, a baseline set of guided recordings was taken for exercised subjects. The typical recording session for an exercised subject is shown in **Figure 19**.

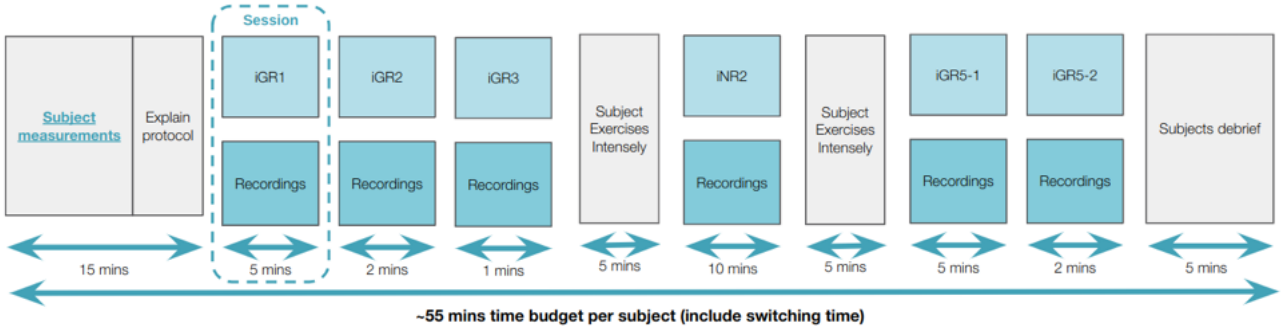

Figure 19: *Data collection procedure for an exercised subject.*

Likewise, the intention was not to use this data for training, but as a source of investigation in supplementary evaluation. This analysis is shown in **Section 7.2.3**.

Intoxicated Subject.

Existing DMS systems have demonstrated significant efficacy in detecting fatigue and reducing the number of overall fatigue-induced accidents [54]. The next technological revolution

in the DMS space will be the capability to reliably detect and prevent distraction and intoxication related accidents. Existing research has sought to use heart rate statistics like HRV to infer cognitive load [7] and intoxication [8].

Unfortunately, very little synchronised video and subject physiological signal data exists where subjects are also intoxicated. To combat this deficiency, one of the major goals of this project was to produce such data that would be useful for other endeavours at Seeing Machines while still being valuable training data for camera-based rPPG.

Accordingly, an additional final naturalistic configuration, **iNR3**, was defined in which subjects were required to reach a blood alcohol concentration >0.08 (verified using a calibrated breathalyser) before performing 10-minutes of naturalistic driving. Importantly, all intoxicated subjects were required to perform a baseline set of typical subject recordings. The recording procedure for an intoxicated subject is shown in **Figure 20**.

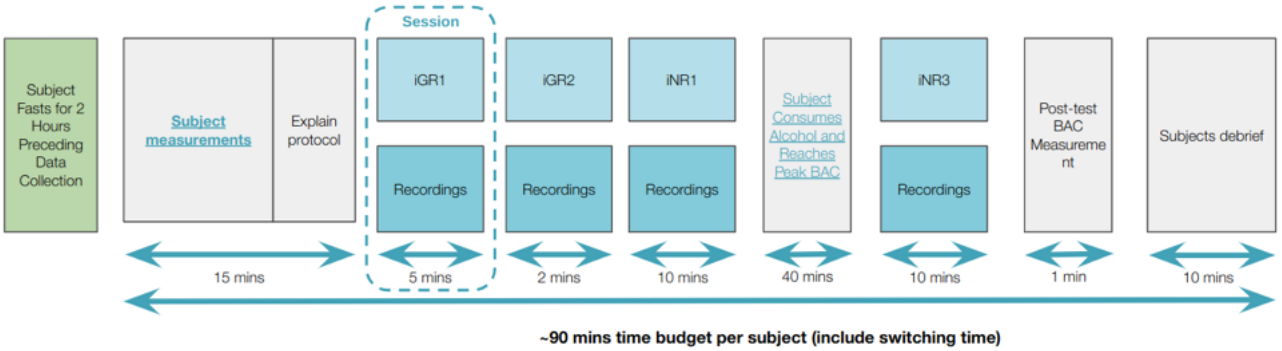

Figure 20: *Data collection procedure for an intoxicated subject.*

Mannequin Subject.

Lastly, existing research has used the rPPG signal as a means of spoof detection [9], [10], [11]. Reliable spoof detection has been achieved using deterministic unsupervised rPPG techniques, however, neural networks are susceptible to hallucinatory output signals if they have overfit to their training dataset. Accordingly, a single 10-minute recording was performed with a mannequin head for which the ground-truth PPG signal is all-zeros. This data was not used for training and was instead separated for supplementary evaluation. This analysis is shown in **Section 7.2.5**.

4.1.3 Post-processing

EOD Conversion.

For each recording, the output of the data collection system was a (1) set of PNG images labelled by frame number, (2) a CSV file relating each frame number to its associated Shimmer physiological sensing data and timestamp, (3) a folder of calibration images from which the camera extrinsics could be derived, and (4) a CSV file containing a raw dump of the 1024 Hz Shimmer data. The objective of the post-processing was to convert this data into Seeing Machines proprietary video format (called EOD) where the physiological data and camera intrinsincs & extrinsics were embedded in each frame.

A series of Python scripts were written to accomplish this. An existing internal company script was used to derive the camera extrinsics from the calibration images and output a CSV file. Unfortunately, there was no fixed relationship between the raw data's channel frame number and the active illumination channel (i.e., dot or flood illuminated frame). A new script was developed that displayed the first frame of each recording and allowed the viewer to use key-bindings to manually label the frame as a dot or flood frame. From there on, the remaining frame numbers could be classified as dot or flood based on their parity. Finally, another Python

script was written that compiled all of this data into a single viewable EOD. The conversion workflow for a single subject is shown in **Figure 21**.

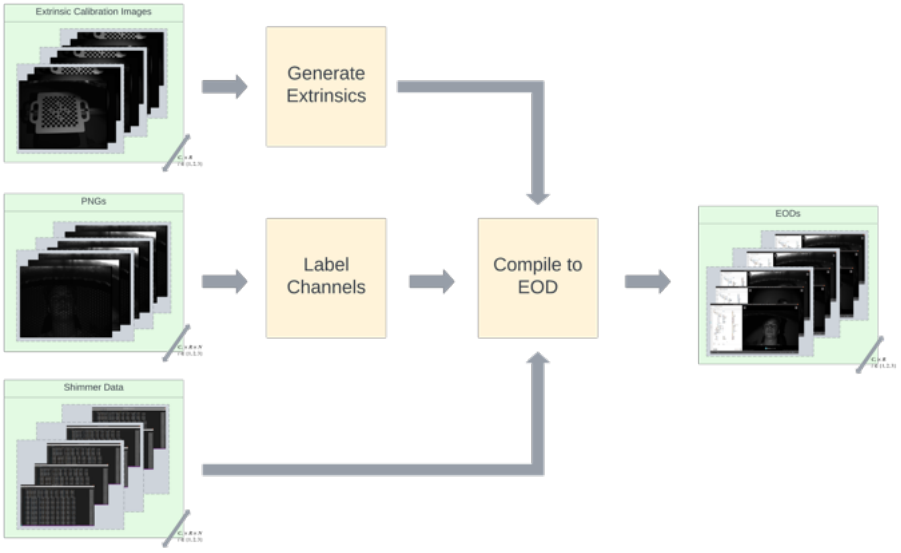

Figure 21: *Programmatic block diagram of the conversion workflow for a single subject.*

Feature Extraction.

Following conversion to EOD, the data was uploaded to an Amazon s3 bucket. At this point, the total dataset size exceeded 40 TB, meaning extraction of the desired training features posed a significant logistical issue. Another Python script was written to accomplish this. The script took an EOD as input and extracted a set of training feature, than saved them in a HDF5 file. It was not possible to read most EODs into memory (a single 5MP 10-minute EOD totalled $>$ 330 GB of data), so a windowing approach was used where an EOD was read-in, processed and saved in 10-second intervals. This also provided an opportunity for judgement of the recorded ground-truth PPG signal quality. To judge the signal quality, the HeartPy Python package [55] was used to and derive an enhanced version of the raw PPG signal (**Figure 22**) and compute the RMSSD[2] of each window. If the RMSSD was greater than 130 ms, the window was shifted 1-second forward. This process was repeated until a good window was found, at which point the data was processed and saved to the HDF5 file, and the window was then shifted by a full 10-seconds. This Python script was deployed with AWS Batch Jobs, allowing the entire dataset to be processed in parallel. This meant the total time to process and extract the desired training features from the dataset **was reduced to just 3 hours**. A programmatic block diagram of the feature extraction workflow for a single EOD is shown in **Figure 23**.

A visualisation of the extracted training signals for a 3D camera (i.e., the two 5 MP cameras) is shown in **Figure 24**. The sparse depth maps were obtained through an existing Seeing Machines technology that derives depth from calibrated cameras equipped with structured light emitters. The dense depth maps were obtained from the sparse depth maps using nearest neighbour interpolation followed by a median filtering operation. For the 2.4 MP camera, 3D training features could not be derived as this camera was not equipped with its own illumination (it was exposed in synchronisation with the IC 5 MP camera). A visualisation of the extracted training features is shown in **Figure 25**.

[2]The root-mean square of successive differences between consecutive RR intervals. The typical range of RMSSD values is 25 ms to 80 ms [56]. A RMSSD significantly higher than this provides a good indication the PPG signal is corrupted with noise artefacts.

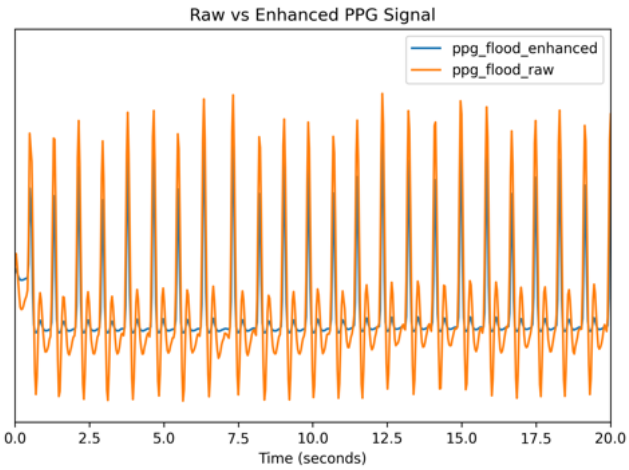

Figure 22: *The raw versus enhanced PPG signals. Enhanced PPG signals were derived using two iterations of the HeartPy enhance function [55]. The enhance algorithm normalises the PPG signal amplitude, then increases R-peak amplitude relative to the rest of the signal. It only uses linear transformations, meaning absolute peak positions are not disturbed.*

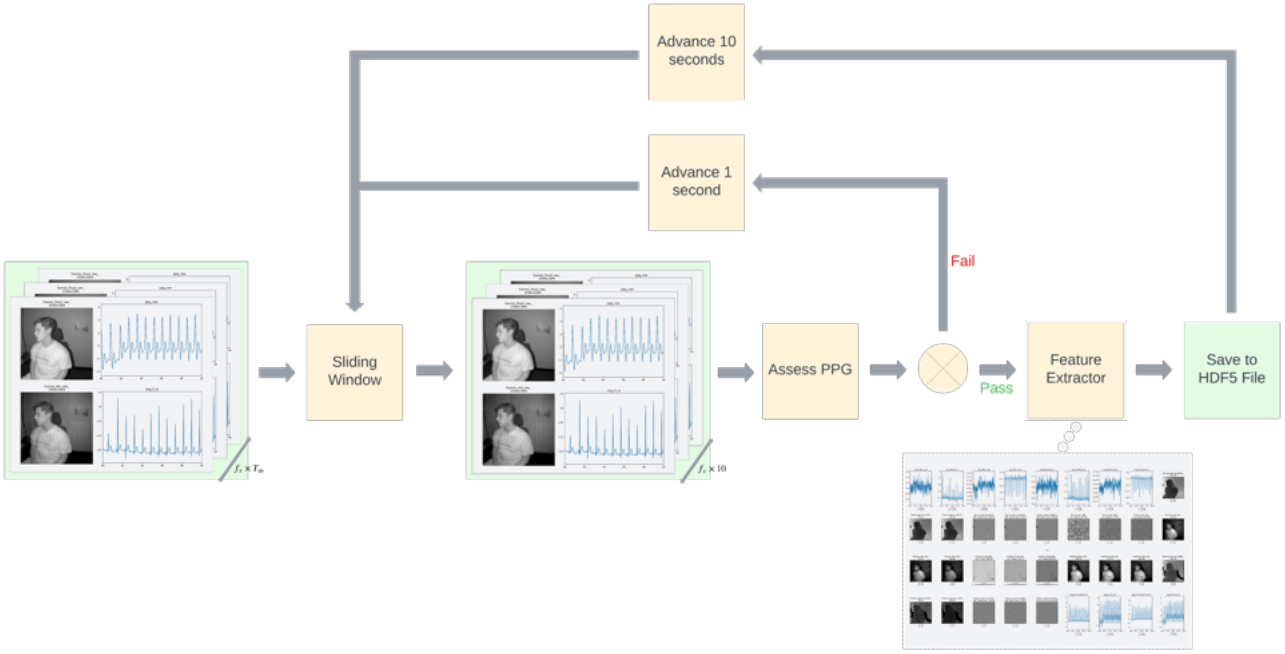

Figure 23: *Programmatic block diagram of the feature extraction workflow for a single EOD.*

4.1.4 Bugs, Blockers and other Disruptions

Data collection was the longest and most perilous phase of this entire project, with a total duration of 6-months from proposal to upload of the final converted recordings to s3. In this section, disruptions encountered throughout the duration of the data collection process are discussed.

Simulator sickness.

20% to 30% of candidates from the general population may experience significant levels of simulator sickness. Symptoms range from mild discomfort through to severe nausea. Some subjects are immediately overcome with sickness, while a smaller number are unaware of it building up until they suddenly vomit. Simulator sickness impacts the subjects natural tendency to become drowsy. Often they will become increasingly agitated instead of drowsy.

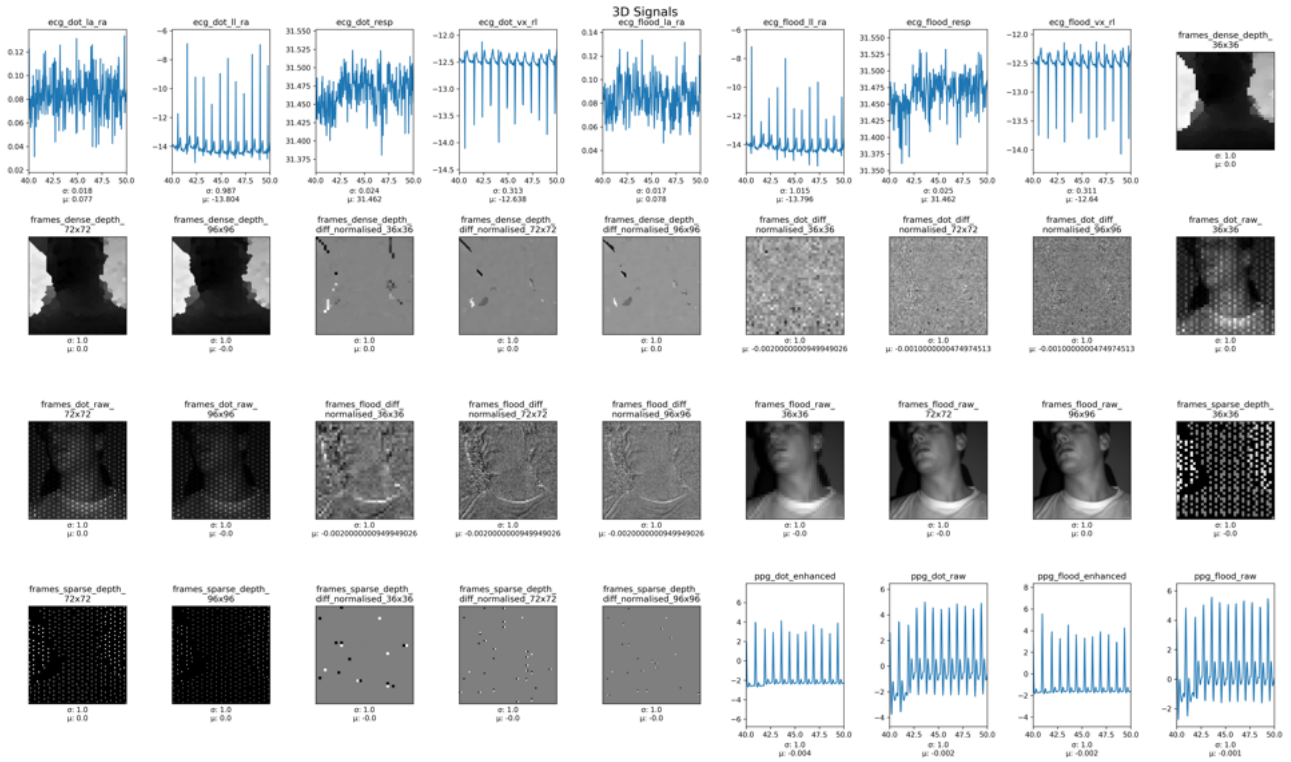

Figure 24: *The extracted 3D training features for the IC 5MP camera.*

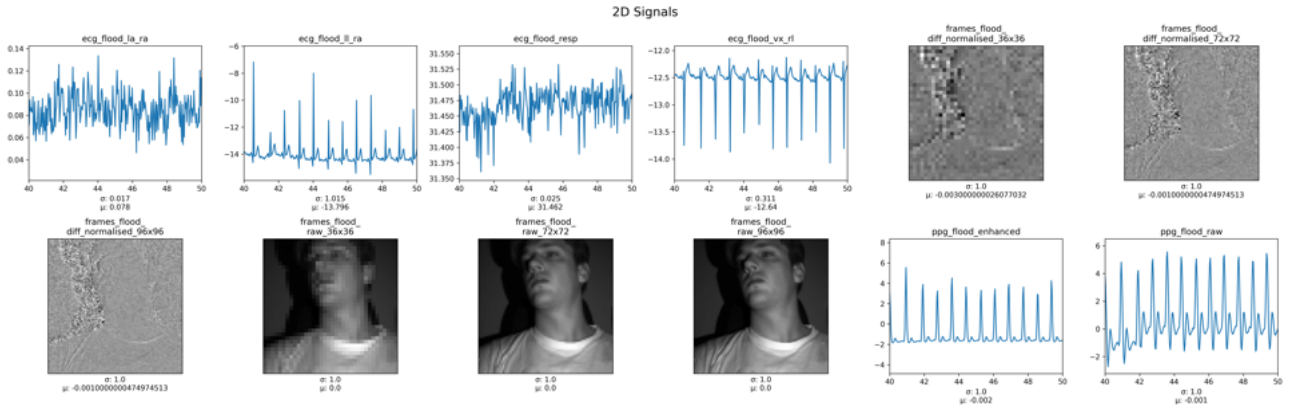

Figure 25: *The extracted 2D training features for the IC 2.4 MP camera.*

Subjects who experience significant simulator sickness were deemed unsuitable for this data collection. Unfortunately, there are no absolute rules to establishing criteria for selecting subjects who will not get sick. A history of playing 3D computer games is the best indicator a subject will not experience significant simulator sickness, but, unfortunately only a small part of the population is in this category. This made planning difficult, as it was often only discovered at their orientation session and therefore a higher number of candidates was required to allow for potential variability in subject demographics and noise factors.

Motion artefacts.

Recording of the ground truth PPG signal is extremely sensitive to changes in the optical pathway. Movement of the subject displaces the PPG sensor and causes motion artefacts that corrupt the recorded ground truth signal. These motion artefacts (**Figure 26**) are significantly larger than the physiological signals. Notably, there's also a settling time following any motion induced artifacts before the sensor output returns to steady state. This means that even a few motion events could corrupt large portions of recordings.

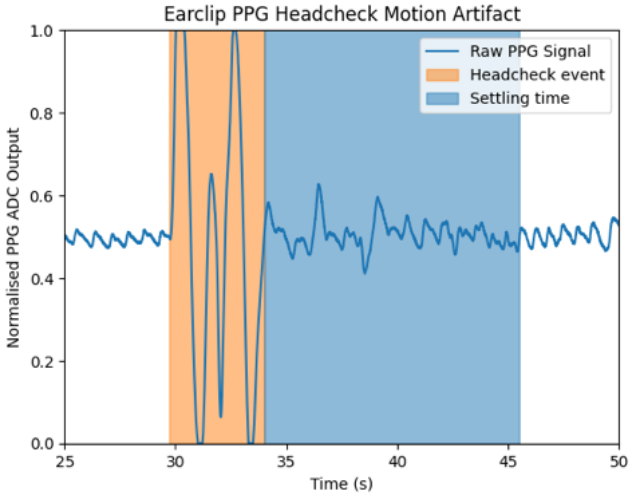

Figure 26: *Example of an earclip PPG sensor signal motion artefact caused by a headcheck event.*

Paradoxically, it was important that uncontaminated PPG signals were recorded for model training purposes. Thus, it was important that the PPG sensor was properly fixed to the subject, and movement of sensor was minimised during recordings.

It was initially intentioned to use an earclip PPG sensor, however, after some experimentation it was determined that a hook-and-loop fastener finger PPG sensor was significantly more robust to motion artefacts.

Bugged software.

The large data rate of the proposed system presented a significant engineering challenge. Work on the data collection system began on April 8th, one month after the project officially commenced. Three months later, on July 12th, the data collection software was officially released. However, following pilot testing, a bug was uncovered in the software that required a re-release that delayed collection until July 31st. After a second pilot test, the system was deemed operational and the data collection campaign was run from the 19th of August until the 23rd of September. This included conversion of the data to EODs and uploading to s3. A timeline of the data collection process and all associated subtasks is shown in **Appendix C**.

Corrupted data.

Unfortunately, 11.14% of the collected data was corrupted due to loss of the Bluetooth connection to the Shimmer physiological sensors. In some cases, it was possible to recover lost data using the raw 1024 z dump of the Shimmer data, however, in many cases the data was not recoverable. Disconnection of the Shimmer sensors was irregular and unpredictable. For some subjects, no data was lost while for other subjects a significant portion of their data was lost. Most of the lost data corresponded to naturalistic recordings. A summary of the lost data (relative to total recording duration) for each subject is provided in **Appendix D**.

4.2 MR-NIRP Car Dataset

The MERL-Rice Near-Infrared Pulse (MR-NIRP) Car Dataset [57], is designed for camera-based vital signs estimation using NIR imaging during driving. It is the only publicly available dataset in the rPPG research space that does not disallow commercial use, and it can be freely downloaded from Google Drive. The dataset includes face video recordings collected simultaneously in both broadband RGB and narrow-band NIR, with pulse oximeter data providing ground-truth vital signs. It contains data from 18 subjects, including both male and female

participants with diverse skin tones, recorded in various driving conditions to analyse the impact of motion and ambient light on rPPG in a practical scenario. The data is intended for researchers studying the challenges of estimating vital signs from facial videos under realistic driving conditions.

Videos in the dataset were recorded using a Point Grey Grasshopper NIR camera with different 10 nm passband bandpass filters (940 nm and 975 nm) and a FLIR Grasshopper RGB camera, along with Bosch EX12LED illuminators to provide consistent NIR illumination.

Each subject's data contains recordings from ten different driving experiments. These recordings are categorised into three main subdirectories: NIR images, pulse oximeter data, and RGB images, which were all synchronised during data collection. The dataset captures data in both static and dynamic scenarios. For this project, to maintain consistency with the SM-EOD dataset, only those experiments recorded using a 940 nm imaging system were considered. This meant four recordings were available per subject, corresponding to conditions "still" and "small motion", with environments "garage" and "driving". Considering only these recordings, the total dataset duration was three hours.

For this project, the MR-NIRP dataset was preprocessed using the AWS Batches framework described above and used for both independent training and cross-dataset validation. A visualisation of the extracted training features is shown in **Figure 27**. Note that the dark frames included in the MR-NIRP dataset were not used for training.

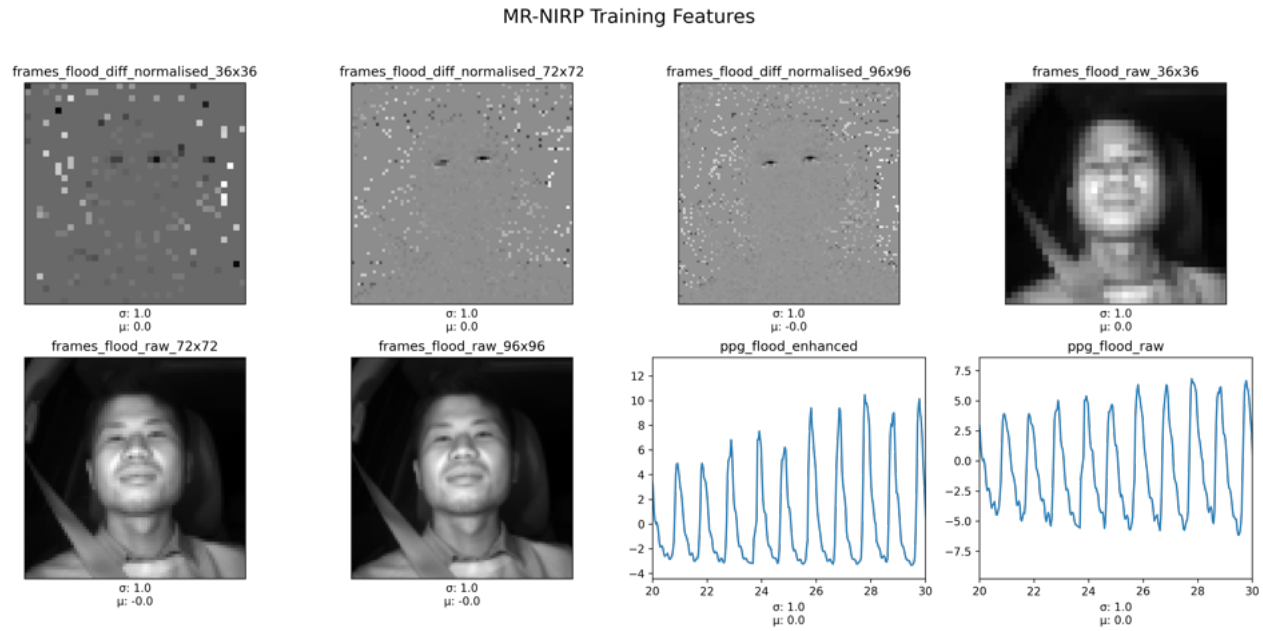

Figure 27: *Training features extracted from each recording in the MR-NIRP dataset using the AWS Batches framework outlined in* ***Section 4.1.3****.*

5 Optical Theory

For a monochromatic flood frame, skin pixels can be modelled using Schafer's Dichromatic Reflectance Model as defined in [15]. Consider the intensity of the *kth* skin pixel, denoted $C_k(t)$.

$$C_k(t) = I_f(t) \cdot (v_s(t) + v_d(t)) + v_n(t) \tag{1}$$

The luminance intensity of the flood source, $I_f(t)$, is modulated by a time-varying specular $v_s(t)$ and diffuse $v_d(t)$ reflection component; while $v_n(t)$ denotes the quantisation noise of the image sensor. The luminance intensity, reflection, and noise terms are time varying because they change with the distance between the light source, skin tissue, and camera.

The luminance intensity and modulating reflection components can be decomposed into a stationary and a time dependent part through a linear transformation [32]. In the case of the diffuse reflection, a stationary reflection strength, d_0, is compiled with the PPG signal, $p(t)$, which is scaled by the skin tissues hemoglobin and melanin absorption.

$$v_d(t) = d_0 + u_p \cdot p(t) \tag{2}$$

The specular reflection consists of a stationary component, s_0, and a varying component, $\Phi(m(t), p(t))$ where $m(t)$ denotes all non-physiological variations such as flickering of the light source and movement of the skin pixel.

$$v_s(t) = s_0 + \Phi(m(t), p(t)) \tag{3}$$

$$I_f(t) = I_0 \cdot (1 + \Psi(m(t), p(t)) \tag{4}$$

$I_0 \cdot \Psi((m(t), p(t))$ is the intensity variation observed by the camera. Decoupling the PPG signal from the non-physiological motions requires modelling of complex non-linear dynamics. This is why unsupervised methods typically struggle to isolate the rPPG signal in scenarios with motion. In this regard, a neural method is expected to outperform an unsupervised extractor.

Combining the specular and diffuse components into a single component representing the stationary skin reflection gives

$$c_0 = s_0 + d_0 \tag{5}$$

Substituting **Eq. 2**, **Eq. 3**, **Eq. 4**, and **Eq. 5** into **Eq. 1** yields

$$C_k(t) = I_0 \cdot (1 + \Psi(m(t), p(t))) \cdot (c_0 + \Phi(m(t), p(t)) + p(t)) + v_n(t) \tag{6}$$

The time-varying components are significantly smaller than the stationary components, allowing for the approximation

$$C_k(t) \approx I_0 \cdot c_0 + I_0 \cdot c_0 \cdot \Psi(m(t), p(t)) + I_0 \cdot \Phi(m(t), p(t)) + I_0 \cdot p(t) + v_n(t) \tag{7}$$

Eq. 7 is the core representation derived by [15] from which the DeepPhys architecture is formulated. To extract $p(t)$, the raw inputs are first downsampled to remove $v_n(t)$. Let $C_l(t)$ denote the new pixel index for each downsampled frame. The stationary skin reflection components can then be removed by taking the first-order derivative with respect to time.

$$C_l'(t) \approx I_0 \cdot c_0 \cdot (\frac{\partial \Psi}{\partial m} m'(t) + \frac{\partial \Psi}{\partial p} p'(t)) + I_0 \cdot (\frac{\partial \Phi}{\partial m} m'(t) + \frac{\partial \Psi}{\partial p} p'(t)) + I_0 \cdot p'(t) \tag{8}$$

Finally, by normalising the above with the temporal mean of $C_l(t)$, the stationary luminance intensity term is removed resulting in a representation that is invariant to the video recording setup.

$$\frac{C_l'(t)}{\bar{C_l}(t)} \approx 1 \cdot c_0 \cdot (\frac{\partial \Psi}{\partial m} m'(t) + \frac{\partial \Psi}{\partial p} p'(t)) + \frac{1}{c_0} \cdot (\frac{\partial \Phi}{\partial m} m'(t) + \frac{\partial \Psi}{\partial p} p'(t)) + \frac{1}{c_0} \cdot p'(t) \tag{9}$$

Importantly, $\bar{C_l}(t)$ should be computed per pixel over a short time window to minimise the impact of external noise factors. In particular, computing over consecutive frames yields

$$D_l(t) = \frac{C_l'(t)}{\bar{C_l}(t)} \sim \frac{C_l(t) + \Delta(t) - C_l(t)}{C_l(t + \Delta(t) + C_l(t)} \tag{10}$$

Accordingly, difference normalised frames provide a valid source for detection of the rPPG signal.

Comparatively, a dot frame can be thought of as a flood frame crested with localised dot beam profiles components. Majority of the pixels are solely influenced by the flood component $I_f(t)$, while those pixels on which dots are projected are also influenced by a dot component $I_d(t)$.

$$C_k(t) = I_f(t) \cdot (v_s(t) + v_d(t)) + I_d(t) \cdot (v_s(t) + v_d(t)) + v_n(t) \tag{11}$$

Through the same mathematics, we arrive at

$$\begin{aligned} \frac{C_l'(t)}{\bar{C_l}(t)} \approx\ & 1 \cdot c_{f_0} \cdot (\frac{\partial \Psi}{\partial m} m'(t) + \frac{\partial \Psi}{\partial p} p'(t)) + \frac{1}{c_{f_0}} \cdot (\frac{\partial \Phi}{\partial m} m'(t) + \frac{\partial \Psi}{\partial p} p'(t)) + \frac{1}{c_{f_0}} \cdot p'(t) \\ & + 1 \cdot c_{d_0} \cdot (\frac{\partial \Psi}{\partial m} m'(t) + \frac{\partial \Psi}{\partial p} p'(t)) + \frac{1}{c_{d_0}} \cdot (\frac{\partial \Phi}{\partial m} m'(t) + \frac{\partial \Psi}{\partial p} p'(t)) + \frac{1}{c_{d_0}} \cdot p'(t) \end{aligned} \tag{12}$$

Importantly, the temporal impact on the dot component of translational motion parallel to the optical axis of the camera causes the centroid of each dot to change. This is what enables the measurement of depth from structured light sources [12]. Furthermore, the stationary component of the luminance intensity for localised dots (i.e., the beam profile) is dependent on the material of the incident surface. This is what enables material detection using the beam profiles [14]. Again, computing over difference frames provides a valid source for detection of the rPPG signal

$$D_l(t) = \frac{C_l'(t)}{\bar{C_l}(t)} \sim \frac{C_l(t) + \Delta(t) - C_l(t)}{C_l(t + \Delta(t) + C_l(t)} \tag{13}$$

where spatial downsampling for dot frames should be carefully considered so that the beam profile information of each dot is not lost.

6 Method

6.1 Architecture Instantiation

The collated dataset was benchmarked using existing state-of-the-art neural rPPG methods as well as novel methods. The instantiation for each of these methods is described below. The training, testing, and evaluation pipeline for this project was adapted from rPPG-Toolbox [43].

LGI Instantiation. Unsupervised rPPG methods have historically been used to establish baseline performance on physiological sensing datasets. In line with this, for this project, LGI [28] serves as a benchmark for comparing the performance of neural methods on the compiled dataset.

While most unsupervised methods require RGB video data, LGI is one of the few that can be applied to monochromatic frames. The input to LGI was standardised raw flood frames, without face cropping.

DeepPhys Instantiation. The DeepPhys model [15] was trained using the implementation from rPPG-Toolbox with inputs being difference-normalised and standardised raw flood frames. For the DeepPhysFormer model, the same inputs were used, but the loss function was replaced by the PhysFormer loss function from [18]. The PhysFormer authors posit that temporal domain supervision signals like MSE provide signal trend-level constraints which are straightforward and easy for model convergence, but then overfit. Comparatively, frequency domain constraints enforce the model to learn periodic features within the target frequency bands, which are hard to converge to due to the rPPG-irrelevant noise. To overcome this, the authors propose a dynamic supervision approach that gradually enlarges the frequency constraints, alleviating the overfitting issue and benefiting the intrinsic rPPG-aware feature learning gradually. The PhysFormer dynamic loss is formulated as:

$$\begin{aligned} \mathcal{L}_{overall} &= \alpha \cdot \mathcal{L}_{time} + \beta \cdot (\mathcal{L}_{CE} + \mathcal{L}_{LD}), \\ \beta &= \beta_0 \cdot (\eta^{\frac{Epoch_{current}-1}{Epoch_{total}}}), \end{aligned} \tag{1}$$

where hyperparameters α, β_0, and η equal 0.1, 1.0, and 5.0; respectively. The temporal domain loss $\mathcal{L}_{time}$ is a linear combination of MSE, L1, cross-entropy and Negative Pearson loss. The frequency domain loss is a linear combination of the cross entropy loss $\mathcal{L}_{CE}$ between the predicted and target heart rate frequency, and the label distribution loss $\mathcal{L}_{LD}$ which is formulated using the Kullback-Leibler divergence between predicted and target power spectrum distribution.

Similarly, the DeepPhys-ConvNeXt model used the same inputs but replaced the VGG-style backbone [33] with a ConvNeXt-style backbone [58] to explore potential performance improvements using a modern architecture. A architectural diagram of the DeepPhys-ConvNeXt model is shown in **Figure 28**.

TS-CAN Instantiation. The TS-CAN model [16] was trained using difference-normalised and standardised raw flood frames as inputs. Note that TS-CAN is a variant of MTTS-CAN that produces predictions for only the rPPG signal, versus both rPPG and respiration waveform predictions.

EfficientPhys Instantiation. The EfficientPhys model [17] used standardised raw frames as inputs.

PhysFormer Instantiation. The PhysFormer model [18] was also trained with standardised raw frames as inputs.

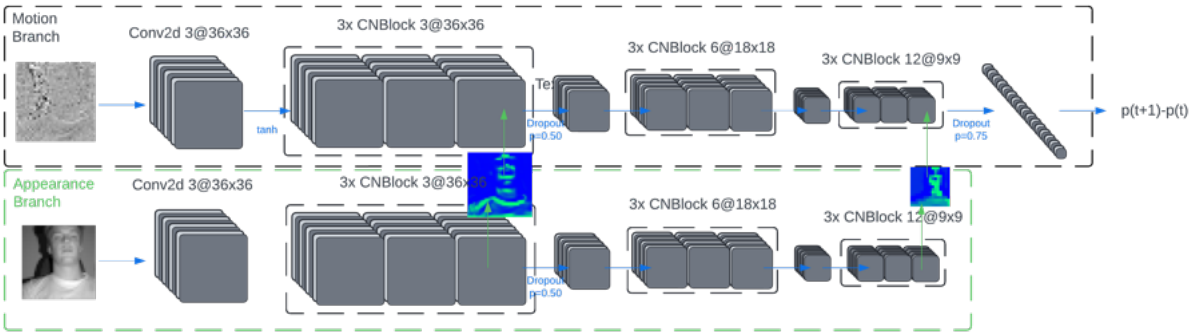

Figure 28: *Architectural diagram of the DeepPhys-ConvNeXt model.*

Dot Variants. For the *Dot* variants of each model, the raw flood and difference frames were replaced with dot frames.

FloodDotDepth Variants.

For the *FloodDotDepth* variants of each model; the inputs are flood, dot and dense depth frames; as well as their difference counterparts where applicable. Because depth calibrations were only possible for the 5 MP cameras, these models were evaluated on the 5 MP IC camera. Notably, an AEC was running on this camera meaning these models were exposed to an additional noise factor.

6.2 Experimental Details.

Each model was trained on the SM-EOD and MR-NIRP datasets. Cross-dataset validation was performed, where models were either trained and tested on the same dataset or trained on one dataset and tested on the other. Importantly, to our knowledge, this paper marks the first time these neural methods have been applied to the MR-NIRP dataset.

For all experiments, models were trained for 30 epochs. In each case, the epoch with the best validation score was used for testing. An ablation analysis was performed where models were trained with Negative Pearson [59], mean-squared error (MSE), and L1 loss. The exception is PhysFormer models, which were trained using the PhysFormer loss function.

On the SM-EOD dataset, models were trained with learning rates (LRs) 9e−4 and 9e−5. On the MR-NIRP dataset, models were trained with LRs 9e−3 and 9e−4. All models are trained using a 1cycle LR policy [60], with the exception of the PhysFormer models which did not use a LR policy. All models were trained using an AdamW optimiser [61]. For all models, a five-fold analysis is performed. In the case that a model is trained and tested on the same dataset, the average performance across the five-folds is recorded. When trained on one dataset and tested on another, the model with the best performance was selected, as the folds were applied only to the train and validation data, while the test dataset remained intact.

All models were trained using inputs of resolution 36x36, with the exception of PhysFormer which was trained using 96x96 resolution inputs that were zero-padded to 128x128 to match the input size expected by the architecture. Although high bit-depth data was recorded, the results of this project were generated using solely 8-bit input data. For all training and test runs, the makeup **iGR4** and exercised **iGR5** configurations were excluded unless otherwise stated. Unless otherwise stated, all results from the SM-EOD dataset were generated using the 2.4 MP IC camera.

6.3 Metrics

The ground-truth labels and model predictions were detrended and filtered using a first-order Butterworth bandpass filter with cutoffs at [0.75, 2.5] Hz (corresponding to [45, 150] BPM). Heart rate was measured in the frequency domain, with 10-second buffers of prediction/label data processed for each heart rate measurement. The heart rate for each filtered window was determined by the frequency with the peak spectral power.

For each method, the mean absolute error (MAE) and mean absolute percentage error (MAPE) between predictions and ground truth were evaluated.

MAE. MAE is the average of the absolute differences between predicted and measured heart rates, indicating how many BPM the model over or underestimates on average. For a given recording, each measurement (denoted i) is taken from a 10-second sliding window with no overlap (i.e., stride 10-seconds).

$$MAE(BPM) = \frac{\sum_{i=0}^{N} |(A_i - B_i)|}{N}$$

MAPE. The mean absolute percentage error (MAPE) is defined as the average percentage error between predictions and measurements. For a given recording, each measurement (denoted i) is taken from a 10-second sliding window with no overlap (i.e., stride 10-seconds).

$$MAPE(BPM) = \sum_{i=0}^{N} \frac{|(A_i - B_i)|}{B_i} \times 100$$

7 Results and Discussion

7.1 SM-EOD and MR-NIRP

All Behaviours.

Table 3: *Results for models trained and tested on the corrupted SM-EOD dataset and the MR-NIRP dataset. The corrupted dataset cells are shaded in red.*

		Test Set			
		SM-EOD		MR-NIRP [57]	
Method	**Train Set**	MAE↓	MAPE↓	MAE↓	MAPE↓
LGI [28]	N/A	25.05	22.27	14.56	22.87
Guess Mean HR	N/A	8.85	12.30	9.19	13.27
DeepPhys [15]	SM-EOD	13.99	17.17	10.77	14.94
	MR-NIRP [57]	7.11	9.00	3.75	4.70
TS-CAN [16]	SM-EOD	14.37	18.07	12.01	16.91
	MR-NIRP [57]	5.62	7.01	3.92	4.83
EfficientPhys [17]	SM-EOD	14.60	17.98	11.55	16.28
	MR-NIRP [57]	8.11	10.34	3.56	4.91
PhysFormer [18]	SM-EOD	30.62	44.31	8.95	13.55
	MR-NIRP [57]	10.60	13.64	5.95	7.10
DeepPhys-ConvNeXt	SM-EOD	13.28	16.46	8.31	10.63
	MR-NIRP [57]	8.31	10.63	8.40	10.47
DeepPhysFormer	SM-EOD	12.69	15.91	13.25	18.34
	MR-NIRP [57]	6.50	7.96	4.72	5.73
DeepPhys-Dot	SM-EOD	14.98	19.54		
TSCAN-Dot	SM-EOD	15.10	19.68		
EfficientPhys-Dot	SM-EOD	15.88	20.73		
PhysFormer-Dot	SM-EOD	37.82	54.39		
DeepPhysFormer-Dot	SM-EOD	17.18	21.57		

Results on the SM-EOD dataset were consistently poor. After some investigation, it was uncovered that the software developed to record the synchronised video and physiological signal data suffered a bug that meant **none of the ground-truth data was reliably synchronised**. In short, the capture software assumed that each sample of the physiological data was exactly $T = \frac{1}{1024}$ seconds apart, but the physiological sensors transmitted over a BlueTooth channel meaning packet loss was inevitable. A single dropped packet was enough to compromise the synchronisation of the recording and without knowledge of when packets were dropped (which was not recorded in the data dump), it was not possible to properly synchronise the ground-truth data. This finding was very disappointing as it compromised the results of this project, but it does explain the discrepancy in performance observed between the SM-EOD and MR-NIRP [57] datasets. Results for models trained and tested on the MR-NIRP dataset and the corrupted SM-EOD dataset are shown in **Table 3**.

Guessing the mean heart rate of either dataset yields better performance than any of the models trained on SM-EOD dataset. Comparatively, performance on the MR-NIRP dataset is consistently good, with errors as low as 3.56 BPM MAE when considering all behavioural and environmental settings. Even when testing on the corrupted ground-truth SM-EOD data, models trained on the MR-NIRP dataset still achieve respectable results. This is because the

time-varying phase shift of the SM-EOD ground-truth signal compromises the utility of the dataset for training, but does not severely compromise validation capacity—particularly the beginning of each recording before any substantial packet loss is experienced. An example of the predictions for the TS-CAN model trained on the MR-NIRP dataset and tested on the corrupted SM-EOD dataset are shown in **Figure 29**. Even though the instantaneous magnitude of the predicted rPPG and corrupted PPG labels are weakly correlated, the frequency content of the two signals is similar (i.e., they have the same number of peaks for a given window. Accordingly, the corrupted data can still be used for validation.

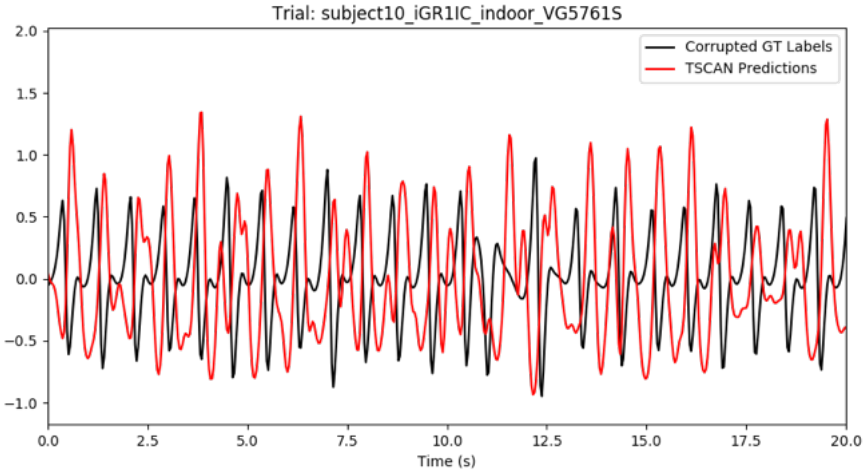

Figure 29: *Plot of the TSCAN predictions and corresponding corrupted ground-truth signal for an **iGR1** static behaviour recording. A time-varying phase shift is observed in the corrupted ground-truth data due to unaccounted packet loss.*

Behaviour Specific Results.

Table 4: *Behaviour-specific results for models trained and tested on the MR-NIRP dataset.*

		Test Set							
		MR-NIRP [57]							
		Behaviour							
		garage still		garage small motion		driving still		driving small motion	
Method	**Train Set**	MAE↓	MAPE↓	MAE↓	MAPE↓	MAE↓	MAPE↓	MAE↓	MAPE↓
DeepPhys [15]	MR-NIRP [57]	0.09	0.10	1.94	2.63	2.01	2.73	8.33	10.13
TS-CAN [16]	MR-NIRP [57]	0.35	0.47	1.49	1.89	0.36	0.43	7.63	9.36
EfficientPhys [17]	MR-NIRP [57]	0.32	0.47	4.35	5.96	1.61	2.27	6.64	8.91
PhysFormer [18]	MR-NIRP [57]	2.55	2.60	10.37	11.32	4.56	5.64	7.27	8.95
DeepPhys-ConvNeXt	MR-NIRP [57]	0.09	0.10	1.49	1.91	1.48	1.92	10.17	12.65
DeepPhysFormer	MR-NIRP [57]	0.40	0.54	1.93	2.63	3.19	3.98	10.04	11.97

Due to the corrupted ground-truth signals in the SM-EOD dataset, analysis in this section only examines the MR-NIRP dataset. In **Table 4**, behaviour specific results are shown for models trained and tested on the MR-NIRP dataset. Best results are obtained for the indoor static behavioural configuration. This is not unexpected, as prior research has shown that motion and heterogeneous illumination are the two biggest noise factors for rPPG sensing [26], [27], [25], [28], [15].

Most models exhibit similar performance for the garage small motion and driving still recording configurations, with a small degradation in performance for the driving still configuration. However, [27] finds that the average movement of each subject in the MR-NIRP dataset is

slightly larger in the driving still configuration versus the garage small motion configuration due to the influence of vehicle momentum on the passenger kinematics. Accordingly, these results seem to indicate that heterogeneous illumination is a significantly less impactful noise factor than motion for outdoor rPPG performance when using a narrow bandpass filter around 940 nm as employed within the MR-NIRP optical pathway. One caveat to this is that no measurements were taken of ambient illumination intensity for the MR-NIRP dataset, so it is not possible to quantitatively validate this conclusion.

A significant degradation in performance is observed for all models for the driving small motion configuration. This result suggests additional training data or architectural changes will be necessary for neural methods to robustly operate in a DMS rPPG application context. For example, incorporation of temporal filtering of model outputs could harness the property that heart rate is typically a wide sense stationary signal to produce more robust predictions in the face of motion. An observer (e.g., Kalman filter) seems like the most obvious solution here. Although this would likely significantly boost performance under normal circumstances, such approaches would introduce complexity in ensuring reliable prediction in the face of legitimate rapid increases in heart rate (e.g., myocardial infarction).

DeepPhys performs strongest in the simplest behavioural condition, but significantly degrades with the addition of motion. TS-CAN shows strong performance in the presence of motion (particularly the driving still configuration). This result is supported by [16], who found the addition of TSMs significantly improved performance on tasks with greater velocity head motion. EfficientPhys performs poorly with the addition of motion, however, demonstrates robustness to heterogeneous illumination. This results suggests the inclusion of the difference and batch norm layers in the model stem allows it to learn a normalisation more robust to heterogeneous illumination than the handcrafted normalisation employed by DeepPhys and TS-CAN.

The PhysFormer model consistently underperforms all models. This can be attributed to a lack of sufficient training data (PhysFormer is a large transformer architecture meaning it is expected to require more training data than the other CNN-based architectures). This analysis is supported by [17], who found transformer architectures were consistently outperformed by CNN equivalents for the same reasons. Another consideration is the ablation analysis used for the training regimes. The best performance for all CNN models was obtained from LR 9e−3, while PhysFormer was consistently better with LR 9e−4. Furthermore, the training curves for the PhysFormer models did not converge until epochs 25 to 28, while the CNN models all converged before epoch 20. It is expected that training PhysFormer with an even smaller LR and for more epochs could improve performance.

The DeepPhys-ConvNeXt model demonstrates equivalent performance to the DeepPhys model, with a small boost in performance for the garage small motion configurations. Notably, however, the use of ConvNeXt blocks means this model has significantly more parameters than the DeepPhys model (102.797 MFLOPs versus 21.7791 GFLOPS, respectively), resulting in a much slower inference time. Finally, the DeepPhysFormer model is consistently outperformed by the DeepPhys model, suggesting that the PhysFormer loss function [18] translates poorly to the DeepPhys architecture.

7.2 Pseudo-Labelled SM-EOD and MR-NIRP

Due to the corrupted ground-truth data, it was not possible to infer any meaningful results by training on the raw SM-EOD dataset. However, as demonstrated in **Table 3**, most models were able to achieve respectable performance on the MR-NIRP dataset. In particular, as shown in **Table 4**, most results on the static indoor condition of the MR-NIRP dataset are less than 0.5 BPM MAE. These results suggest that models trained on the MR-NIRP dataset could be used to generate pseudo-labels for the SM-EOD dataset—at the very least for static behavioural configurations. These results are the topic of this discussion.

The TSCAN model trained and tested on the entire MR-NIRP dataset achieving a MAE of 5.62 BPM across all conditions of the corrupted SM-EOD dataset was used to generate pseudo ground-truth PPG labels for the SM-EOD dataset. These predicted labels were detrended and bandpassed filtered before use. An example of the generated pseudo-labels and corrupted underlying ground-truth signal is shown in **Figure 29**. Because the performance of this model was significantly worse on moving configurations in the MR-NIRP dataset (**Table 4**), it was not appropriate to use such recordings. Accordingly, models were only trained on the static behaviours (**iGR1** recordings) of the SM-EOD dataset. These models were then tested on the static behaviours of the pseudo-labelled SM-EOD dataset and the entirety of the MR-NIRP dataset. Results are shown in **Table 5**. No results were generated for models trained on the MR-NIRP dataset and tested on the pseudo-labelled SM-EOD dataset as they would not provide any valuable insights.

All Behaviours.

In the case of the SM-EOD dataset, at best, these results indicate that models as good as the TSCAN model used to generate the pseudo-labels could be trained using the SM-EOD dataset with properly synchronised ground-truth. These results are not particularly interesting, with the exception of the results recorded for 3D signals which provide insight into the viability of dot and depth frames for rPPG. Of more interest is the results obtained by training on the pseudo-labelled SM-EOD dataset and tested on the MR-NIRP dataset which has proper ground-truth. These results provide insight into the capability of the developed data collection system and optical pathway for sensing of the rPPG signal assuming the ground-truth signal is properly synchronised.

Considering results obtained by training and testing on the pseudo-labelled dataset, the TS-CAN model was most effective at reproducing the pseudo-label signal. This result is not surprising considering the pseudo-labels were generated using a TS-CAN architecture model. In particular, the TS-CAN and PhysFormer models trained on the pseudo-labelled SM-EOD dataset are within 1 BPM MAE of their counterparts trained on the MR-NIRP dataset, with the PhysFormer model actually surpassing its counterpart. These results suggest that the use of pseudo-labels was a valid training strategy that has empowered these models to learn to detect the rPPG signal. All models are able to substantially surpass the performance of guessing the MR-NIRP dataset's mean heart rate.

The *Dot* and *FloodDotDepth* variants both demonstrate a capacity to learn the pseudo-labelled rPPG signal. It was not possible to test these models on the MR-NIRP dataset as there is no equivalent optical pathway in the MR-NIRP dataset. Despite this, the results do indicate that it is possible to isolate the rPPG signal from *Dot* frames using neural methods. No increases in performance were observed for the *FloodDotDepth* variants, contradicting the hypothesis of this project. Despite this, it is important to note that the addition of depth as an rPPG signal was expected to provide the most benefit in scenarios with movement, but it was not possible to examine such scenarios given that the pseudo PPG signal cannot be considered as a reliable ground truth when the subject is moving. Notably, results for the *Dot* variants were generated using the 5 MP IC camera. This was done as a significant degradation

Table 5: *Results for models trained and tested on the pseudo-labelled SM-EOD dataset and the MR-NIRP dataset. The SM-EOD test and training sets consist of only the* ***iGR1*** *behavioural recordings.*

		Test Set			
		SM-EOD (**iGR1** only)		MR-NIRP [57]	
Method	**Train Set**	MAE↓	MAPE↓	MAE↓	MAPE↓
LGI [28]	N/A	15.24	18.73	14.56	22.87
Guess Mean HR	N/A	9.10	13.04	9.19	13.27
DeepPhys [15]	SM-EOD (**iGR1** only)	0.64	0.97	7.38	9.88
	MR-NIRP [57]			3.75	4.70
TS-CAN [16]	SM-EOD (**iGR1** only)	0.31	0.48	5.23	7.38
	MR-NIRP [57]			3.92	4.83
EfficientPhys [17]	SM-EOD (**iGR1** only)	0.85	1.27	6.93	9.71
	MR-NIRP [57]			3.56	4.91
PhysFormer [18]	SM-EOD (**iGR1** only)	7.64	10.51	5.18	8.57
	MR-NIRP [57]			5.95	7.10
DeepPhys-ConvNeXt	SM-EOD (**iGR1** only)	0.66	0.99	9.53	12.75
	MR-NIRP [57]			4.25	5.31
DeepPhysFormer	SM-EOD (**iGR1** only)	1.01	1.47	8.40	11.46
	MR-NIRP [57]			4.72	5.73
DeepPhys-Dot	SM-EOD (**iGR1** only)	0.90	1.34		
TSCAN-Dot	SM-EOD (**iGR1** only)	0.94	1.40		
EfficientPhys-Dot	SM-EOD (**iGR1** only)	1.44	2.03		
PhysFormer-Dot	SM-EOD (**iGR1** only)				
DeepPhysFormer-Dot	SM-EOD (**iGR1** only)	1.16	1.71		
DeepPhys-FloodDotDepth	SM-EOD (**iGR1** only)	0.95	1.40		
DeepPhysFormer-FloodDotDepth	SM-EOD (**iGR1** only)	2.23	2.86		

in performance was observed for the 2.4 MP IC camera. This degradation is discussed in **Section 7.2.1**. Results for the *FloodDotDepth* variants were generated using the 5 MP IC camera as this was the only IC camera with depth capabilities.

Because the linewidth of the dot projector is an order of magnitude smaller than that of a flood emitter, it is reasonable to hypothesise that dot frames could outperform flood frames in conditions with heterogeneous illumination. Unfortunately, such conditions were not included within the SM-EOD dataset so it is not possible to quantitatively validate this hypothesis. Furthermore, it is also important to note that the results obtained in **Table 4** suggest that motion is a more significant noise factor for neural rPPG methods than heterogeneous illumination.

Behaviour Specific.

In **Table 6**, behaviour specific results are shown for models trained on the static recordings of the pseudo-labelled SM-EOD dataset and tested on the MR-NIRP dataset. A clear degradation is observed under the addition of movement—a similar but more severe degradation as is observed in **Table 4**. These results are supported by [15], who also showed that models trained on only static recordings generalise poorly to recordings containing motion. Similar to **Table 4**, these results indicate that motion is a more significant noise factor for rPPG than heterogeneous illumination. Interestingly, these results suggest that models trained with controlled illumination data generalise well to heterogeneous illumination scenarios. These results support the mathematics shown in **Section 5**, which posit that difference normalised frames are an effective method for removing the influence of controlled illumination as well as slow fading external illumination. Despite this, it is important to note that difference normalised

Table 6: *Behaviour-specific results for models tested on the MR-NIRP dataset.*

		Test Set							
		MR-NIRP [57]							
		Behaviour							
		garage still		garage small motion		driving still		driving small motion	
Method	**Train Set**	MAE↓	MAPE↓	MAE↓	MAPE↓	MAE↓	MAPE↓	MAE↓	MAPE↓
DeepPhys [15]	SM-EOD (**iGR1** only)	2.69	3.54	5.74	7.50	5.85	9.20	9.79	12.08
	MR-NIRP [57]	0.09	0.10	1.94	2.63	2.01	2.73	8.33	10.13
TS-CAN [16]	SM-EOD (**iGR1** only)	2.64	3.56	4.39	6.28	2.01	2.76	8.64	11.55
	MR-NIRP [57]	0.35	0.47	1.49	1.89	0.36	0.43	7.63	9.36
EfficientPhys [17]	SM-EOD (**iGR1** only)	4.20	6.71	3.98	6.07	5.18	7.19	11.74	15.16
	MR-NIRP [57]	0.32	0.47	4.35	5.96	1.61	2.27	6.64	8.91
PhysFormer [18]	SM-EOD (**iGR1** only)	0.48	0.63	10.06	15.88	3.37	5.90	6.20	9.83
	MR-NIRP [57]	2.55	2.60	10.37	11.32	4.56	5.64	7.27	8.95
DeepPhys-ConvNeXt	SM-EOD (**iGR1** only)	3.37	4.58	5.48	7.67	8.01	11.71	10.96	14.19
	MR-NIRP [57]	0.09	0.10	1.49	1.91	1.48	1.92	8.33	10.13
DeepPhysFormer	SM-EOD (**iGR1** only)	3.76	5.12	7.70	10.58	7.01	10.92	11.50	13.92
	MR-NIRP [57]	0.40	0.54	1.93	2.63	3.19	3.98	10.04	11.97

frames are not effective for removing fast fading external illumination noise factors (i.e., external illumination varying faster than the Nyquist rate of the difference frame frequency). As aforementioned, the MR-NIRP dataset does not contain measurements of external illumination power for each recording. However, these results, as well as those in **Table 4** suggest that the MR-NIRP dataset is unlikely to contain large quantities of driving data with fast fading external illumination. It is reasonable to assume the performance of all neural methods would degrade significantly for such data.

7.2.1 Higher Resolution Results

Using dot frames, the 2.4 MP IC camera at 36x36 resolution produced significantly worse results than the 5 MP IC camera at the same resolution on the pseudo-labelled SM-EOD dataset (e.g., 6.31 versus 0.64 BPM MAE for the DeepPhys model). This result raises questions regarding the difference between the two configurations. It was suspected that higher-resolution dot frames would yield better results because the high-frequency content of the dot frames creates artifacts in low-resolution images. This affects the 2.4 MP camera more due to its wider field-of-view, which more severely compresses the dot beam profiles when downsampled to 36x36 pixels as shown in **Figure 30**.

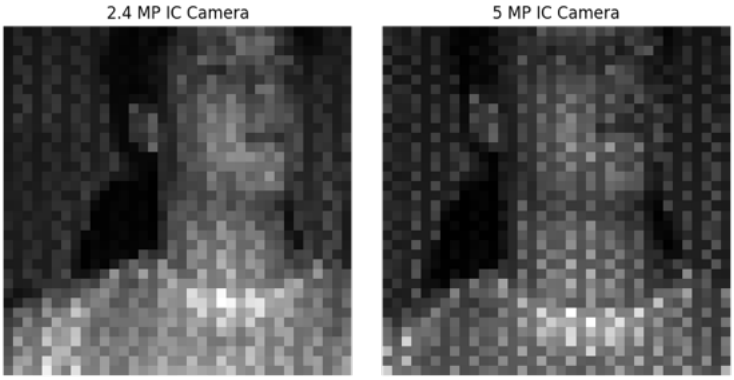

Figure 30: *Comparison of the downsampled dot frames for the 2.4 MP and 5 MP IC cameras. More high-frequency artefacts resultant of the downsampling are observed in the 2.4 MP IC camera due to it's wider field-of-view.*

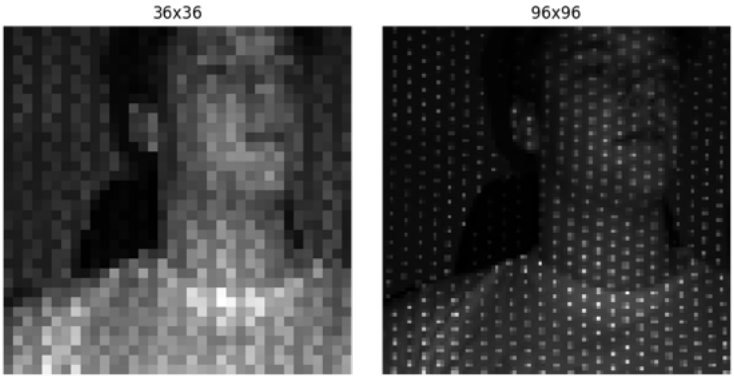

Figure 31: *Comparison of the downsampled dot frames for resolutions 36x36 and 96x96 pixels. At 36x36, the beam profiles of individual dots are no longer discernible.*

To test this hypothesis, models were trained on 96x96 pixel frames from each camera, with results shown in **Table 7**. Notably, the discrepancy persisted, thereby invalidating the hypothesis. The significant variation across models suggests that the dot frames for the 2.4 MP IC camera are highly sensitive to hyperparameter tuning (recall that these results are based on an ablation analysis where each cell represents the outcome of training with three loss functions, two LRs, and five folds; except for PhysFormer variants, which has a single loss function).

Furthermore, the performance of most models actually improves for the lower resolution dot frames. These results suggest that the underlying flood profile of the dot pattern projector is a more important signal for sensing the rPPG signal than the actual beam profiles of the dots. However, if we visualise the attention masks of the DeepPhys model for high-resolution inputs (**Figure 32**), we can see that the opposite is actually true—the model has learned an attention representation that actively focuses on the dot pixels! Thus, these results suggest that is is possible to isolate the rPPG signal solely based of the signals encoded in a dot pattern projector's beam profiles (i.e., intensity, dilation, centroid, etc.).

Table 7: *Results for the dot models trained and tested on the pseudo-labelled static behavioural conditions of the SM-EOD dataset for varying input image resolution.*

		Test Set							
		SM-EOD (**iGR1** only)							
		Camera							
		2.4 MP IC				5 MP IC			
		Resolution							
		36x36		96x96		36x36		96x96	
Method	**Train Set**	MAE↓	MAPE↓	MAE↓	MAPE↓	MAE↓	MAPE↓	MAE↓	MAPE↓
DeepPhys-Dot	SM-EOD (**iGR1** only)	6.31	9.25	7.39	11.33	0.90	1.34	0.79	1.18
TSCAN-Dot	SM-EOD (**iGR1** only)	2.22	2.86	1.21	2.00	0.94	1.40	1.25	1.71
EfficientPhys-Dot	SM-EOD (**iGR1** only)	4.05	4.95	3.36	4.27	1.44	2.03	1.61	2.14
PhysFormer-Dot	SM-EOD (**iGR1** only)	N/A		12.66	19.08	N/A		9.72	14.55

7.2.2 Rear-view Mirror Results

The RVM poses a significant challenge for rPPG signal monitoring due to several factors: (1) the subject occupies a much smaller portion of the frame, (2) only half of the subject's face is typically visible during normal driving conditions, and (3) the wide field-of-view lens distorts the subject's appearance when the RVM is positioned naturally. A comparison between a typical scene for an IC recording versus an RVM recording is shown in **Figure 33**. In this

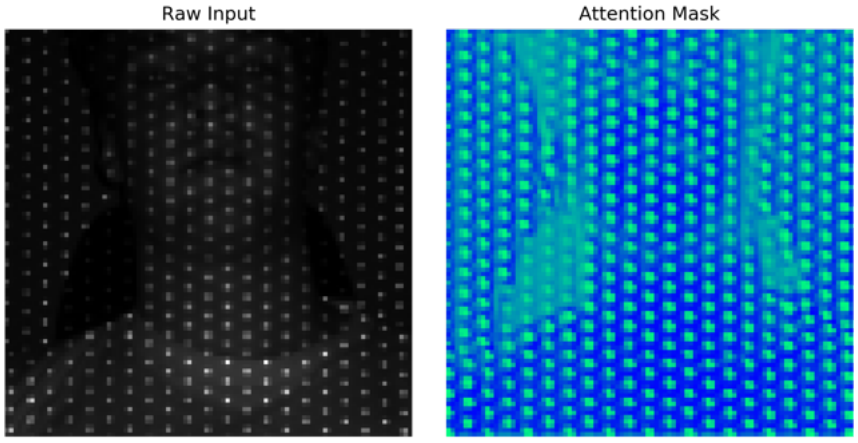

Figure 32: *Attention masks produced by the DeepPhys-Dot model for high resolution inputs. The attention mask demonstrates that the model has learned to isolate the rPPG signal using signals encoded in the beam profiles of the dot pattern projector.*

section, results are generated for the RVM camera position to quantify the impact of these challenges on rPPG performance. Results obtained training on the RVM camera configuration with pseudo-labelled ground truth PPG are shown in **Table 8**.

Figure 33: *Comparison of a 96x96 pixel flood image for an IC recording versus an RVM recording. The RVM configuration poses a significantly more challenging scene for reliable rPPG measurement.*

A small degradation in performance is observed for results obtained from the RVM position. Notably, the AEC was running on the RVM camera meaning that the observed degradation may even just be attributed to the AEC's impact on the intensity of the illumination source. These results were surprising considering the small portion of pixels occupied by the subject for the RVM configuration. It was suspected that the RVM configuration would benefit from higher resolution inputs, however, the results obtained in **Table 8** show this to be false. These results provide insight into the importance of spatial downsampling for removing image sensor noise, and suggest that further spatial downsampling may actually improve performance for the other camera configurations. This analysis is supported by [24], who are able to successfully isolate the rPPG signal after spatial downsampling down to only 9x9 pixels. Furthermore, these results suggest that rPPG may be possible in OMS systems.

Table 8: *Results for models trained and tested on the pseudo-labelled static behavioural conditions of the SM-EOD dataset for the 5MP RVM camera.*

		Test Set			
		SM-EOD (**iGR1** only)			
		Input Resolution			
		36x36		96x96	
Method	**Train Set**	MAE↓	MAPE↓	MAE↓	MAPE↓
DeepPhys [15]	SM-EOD (**iGR1** only)	0.69	1.03	0.77	1.13
TSCAN [16]	SM-EOD (**iGR1** only)	0.54	0.81	0.89	1.36
EfficientPhys [17]	SM-EOD (**iGR1** only)	0.75	1.09	1.19	1.85
PhysFormer [18]	SM-EOD (**iGR1** only)	N/A		10.21	16.53
DeepPhysFormer	SM-EOD (**iGR1** only)	0.79	1.14	1.24	1.92
DeepPhys-Dot	SM-EOD (**iGR1** only)	0.78	1.16	0.86	1.30
TSCAN-Dot	SM-EOD (**iGR1** only)	0.64	0.95	1.05	1.57
EfficientPhys-Dot	SM-EOD (**iGR1** only)	0.86	1.27	0.88	1.32
PhysFormer-Dot	SM-EOD (**iGR1** only)	N/A		12.22	18.04
DeepPhysFormer-Dot	SM-EOD (**iGR1** only)	0.79	1.15	1.24	1.92
DeepPhys-FloodDotDepth	SM-EOD (**iGR1** only)	0.69	1.05	0.78	1.17
DeepPhysFormer-FloodDotDepth	SM-EOD (**iGR1** only)	2.23	2.86	3.01	3.78

7.2.3 Exercised Subject.

Most public rPPG datasets fail to capture recordings with a wide range of heart rates. In a DMS application context, this is particularly limiting as one of the desired use cases would be the detection of abnormal events such as cardiac arrest and myocardial infarction. In this section, results are presented for the exercised condition of the SM-EOD dataset for models trained on the pseudo-labelled static conditions and tested using the corrupted ground-truth signal for labelling. In generating the post-processed waveforms for both the predicted and corrupted ground-truth signals, the cutoff frequency of the bandpass filter was raised to [0.75, 3.5] Hz. Results are shown in **Table 9**.

Table 9: *Behaviour-specific results for the exercised recording configuration. Models are trained on the static behaviours of the pseudo-labelled SM-EOD dataset and tested on the exercised behaviours of the SM-EOD dataset with corrupted ground-truth signals.*

		Test Set					
		SM-EOD					
		Behaviour					
		iGR5-1		iGR5-2		iGR5-3	
Method	**Train Set**	MAE↓	MAPE↓	MAE↓	MAPE↓	MAE↓	MAPE↓
DeepPhys [15]	SM-EOD (**iGR1** only)	68.03	55.48	47.07	44.91	34.80	32.59
TS-CAN [16]	SM-EOD (**iGR1** only)	70.40	57.42	41.13	39.25	35.60	33.33
EfficientPhys [17]	SM-EOD (**iGR1** only)	75.94	61.94	55.77	53.21	52.21	48.89
PhysFormer [18]	SM-EOD (**iGR1** only)	73.56	60.00	28.48	27.17	35.60	33.33

The ground-truth heart rate for the exercised configurations ranges from 100 to 120 BPM, but predictions of all models are consistently in the 60 to 80 BPM range. These results indicate that all models generalise poorly to scenes with higher frequency heart rates than are present in their training data. For all models, better performance is observed for the moving configura-

tions, however, considering the prediction waveforms (shown in **Figure 34**), it is clear that the noisier outputs associated with moving subjects push the peak power spectral density of the model predictions higher. As such, they do not actually reflect better prediction performance. These results suggest that a comprehensive set of training data with a wide range of heart rate values is required for neural rPPG methods to perform robustly in the face of abnormal heart rate events.

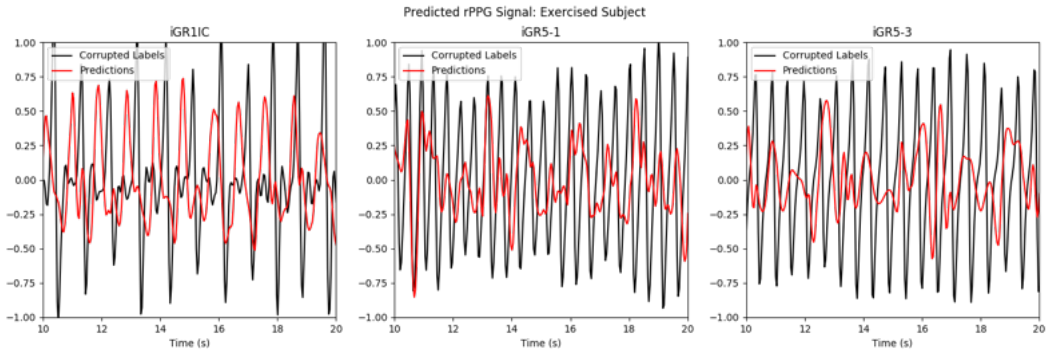

Figure 34: *Plot of the DeepPhys rPPG predictions for the exercised recording.*

7.2.4 Makeup Impact

The rPPG signal is around ten times weaker in the NIR spectrum than in visible light [26]. Prior research has demonstrated that the rPPG signal is 54.5% to 61.4% weaker in the presence of makeup [53]. Accordingly, makeup may pose a significant barrier to the application of rPPG in a DMS context.

In **Figure 35**, the attention masks of the DeepPhys model trained on the pseudo-labelled SM-EOD dataset are shown for the makeup subject, both with and without makeup. Note that, in the makeup case, only half of the subjects face is covered in makeup as described in **Section 4.1.2**. Suprisingly, there is no discernible difference between the attention masks, suggesting that the appearance branch does not learn a representation that priorities unoccluded skin pixels, but instead learns to isolate the broader face within the scene.

Figure 35: *Comparison of an attention mask generated by the DeepPhys appearance branch for a makeup recording versus a typical recording.*

In **Figure 36**, the predictions of the same model are shown for the makeup subject, both with and without makeup. For both recordings, the subject is static with gaze oriented at the IC camera. Note that the ground-truth signal is the corrupted PPG signal recorded using the finger clip sensor. Only the beginning of the recording is considered, as packet loss is minimal

meaning the correlation the prediction and ground-truth labels remains high. Suprisingly, there is not discernible difference in magnitude between the rPPG signal of the subject with and without makeup. This result may be explained by the fact that only half of the subjects face was covered in makeup. Despite this, these results suggest that neural methods are capable of making rPPG predictions in the presence of makeup, though a more comprehensive test dataset is needed to quantitatively validate this assertion.

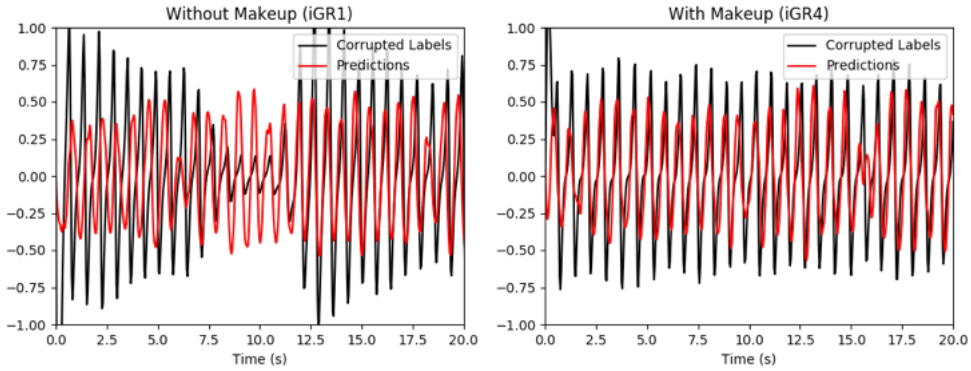

Figure 36: *Plot of the DeepPhys model rPPG predictions for the makeup recording.*

7.2.5 Spoof Performance

Prior research has demonstrated that the rPPG signal can be used for spoofing detection [9], [10], [11]. Spoof is a desired feature in DMS—primarily for DMS in the fleet, aviation, and mining industries where unauthorised operators present a significant safety hazard.

During the data collection for this project, a single 10-minute recording of a mannequin head was captured. In this section, predictions from the DeepPhys model, trained on the pseudo-labelled SM-EOD dataset, are evaluated to determine if existing neural rPPG methods can be directly applied to the spoof detection problem. A comparison between the model's predictions and the ground-truth PPG signal (or lack thereof) is shown in **Figure 37**.

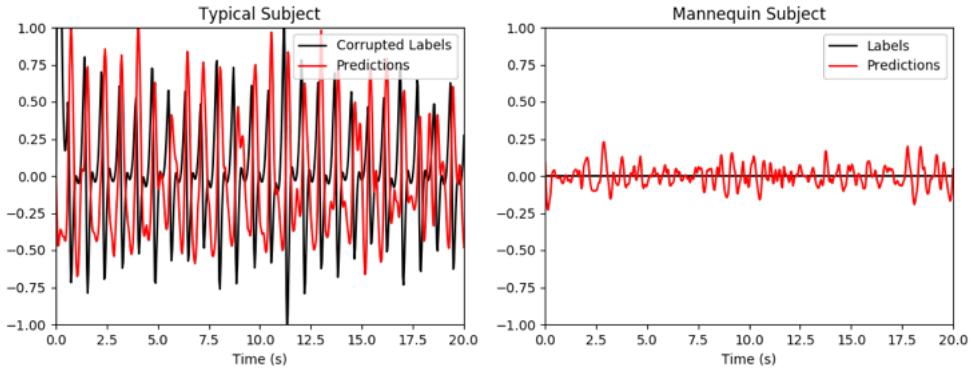

Figure 37: *Plot of the DeepPhys rPPG predictions for the spoof recording.*

Notably, the model's predictions are non-zero throughout the spoof recording but are capped around 0.2, about one third of the typical magnitude seen for a human subject. Despite this, it cannot be said that a binary threshold of the rPPG predictions would serve as an effective spoof detection method as the mannequin head is perfectly static for the duration of the recording. It is likely that the magnitude of predictions would change for a moving spoof target (e.g., a person wearing a mask). When considering the predicted waveform for the mannequin head, they bear little resemble to an rPPG waveform. These results suggest that it may be possible

to train a classifier network for spoof detection using the predicted rPPG signal as an input. However, a more comprehensive spoof dataset is required to properly validate this hypothesis.

Notably, the attention mask for the spoof recording bear no obvious differences to those obtained for a regular recording (**Figure 38**). These results suggest that the attention branch of the DeepPhys model learns features that isolate the structure of the human face, rather than isolating actual skin pixels. This conclusion is in support of the analysis presented in **Section 7.2.4**.

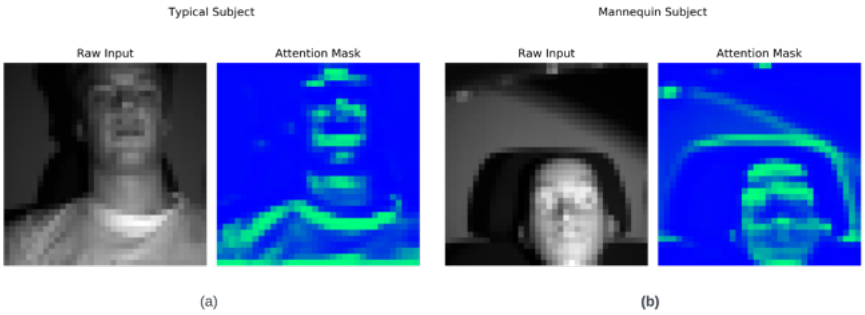

Figure 38: *Comparison of an attention mask generated by the DeepPhys appearance branch for a spoof recording versus a typical recording.*

8 Conclusion

This project laid the groundwork for rPPG research at Seeing Machines, developing both the infrastructure and methodologies necessary to advance this research topic towards a consumer-grade feature. A comprehensive data collection system was built to capture synchronised physiological and video data. Despite initial setbacks due to corrupted data, this issue has since been rectified and efforts are already underway to extend the protocol for synchronised outdoor data collection, incorporating ambient illumination criteria to enhance the robustness of the dataset. Furthermore, it was shown that although this corrupted data is of no utility for model training it may still be useful for validation purposes.

In addition to data collection, a streamlined pipeline was developed for the efficient extraction of heart rate training features from EODs using AWS Batches. A significant aspect of this project was the adaptation and extension of the rPPG-Toolbox respository, which was modified to support not only the new training feature extraction process but also offer enhanced functionalities such as ablation analysis, attention mask visualisation, metric compilation, neural pseudo-label generation, principal component analysis, ensemble learning, and more.

Through the MR-NIRP dataset, these preprocessing and training pipelines have been validated, confirming the effectiveness of the developed methods. Furthermore, testing on the MR-NIRP dataset using pseudo-labelled data verified the capability of Seeing Machines' camera systems to perform rPPG measurements.

Although the primary project objective of assessing the viability of three-dimensional signals for heart rate sensing could not be fully realised due to the corrupted ground-truth labels of the training data, preliminary analysis demonstrated that three-dimensional signals indeed offer a valid basis for rPPG signal extraction. In particular, this project demonstrates that the rPPG signal can effectively be decoded from the beam profile of structured light NIR illumination sources—an entirely novel finding that has not been examined by any prior research.

Furthermore, the findings of this project demonstrate that motion poses a more significant noise factor than heterogeneous illumination for rPPG in DMS equipped with a narrowband bandpass filter about 940 nm. The results also suggest that makeup is not a significant noise factor for NIR rPPG performance, that existing architectures generalise poorly to subjects with heart rates significantly outside of the range of their training data, and that spoof detection is feasible using existing state-of-the-art neural rPPG architectures.

Though the RVM camera configuration was initially expected to result in a significant reduction in rPPG performance, the results of this project show that this degradation is actually minimal and effective rPPG sensing is possible from the RVM camera configuration. These results also suggest that rPPG may be performed in OMS from a single camera configuration.

These initial findings establish a solid starting point for future research, with the GitHub repository that houses the programmatic infrastructure developed throughout this project and the new data collection protocols and systems setting the stage for further advancements in rPPG technology at Seeing Machines once a properly synchronised and sufficiently large dataset is acquired.

8.1 Privacy Concerns

Incorporation of physiological sensing has the potential to greatly enhance the capabilites of existing DMS. If done properly, this could result in a significant reduction in injuries and fatalities as a result of distraction, fatigue, and intoxication. However, it also raises serious privacy concerns, as these methods could be used to measure personal physiological information without the subject's knowledge.

To address these concerns, it is crucial to ensure transparency when using non-contact

physiological sensing methods. Just as with traditional contact sensors, individuals must be informed and provide explicit consent before their physiological data is measured or recorded. Furthermore, there should be no penalty for those who choose not to participate. The ability to measure physiological signals in more contexts does not inherently justify doing so without proper regulation. In fact, as the reach of these technologies expands, stricter regulations should be applied, not lessened.

To mitigate such risks, the findings and methodologies developed in this project are protected under Seeing Machines' strict data governance policies, ensuring that only authorised personnel have access to this sensitive data. Additionally, future deployments of this technology must comply with industry-standard data privacy laws and regulations, such as the Australian Privacy Act 1988, to ensure that users' rights are protected.

8.2 Future Works

This project explored the potential of using novel optical pathways and combined planar two-dimensional and volumetric three-dimensional signals to improve rPPG performance. While a comprehensive evaluation of this approach was limited due to data constraints, the groundwork was laid for investigating how integrating depth information could enhance physiological signal extraction. Moving forward, expanding the analysis to include more diverse architectures, proper fusion with depth estimation, and extending models to include respiration sensing will be provide greater insights into the value of depth data for three-dimensional sensing.

The training process for this project did not involve data augmentation, which is traditionally a straightforward way to enrich the training dataset. While augmenting physiological signals presents unique challenges, prior studies, such as the augmentation techniques proposed by [62], have demonstrated its feasibility. In particular, the addition of motion augmentation has been shown to be a powerful tool in boosting the performance of neural rPPG methods [63]. Integrating similar augmentation strategies could provide a significant boost to model performance and should be considered in future research.

Configurations for the outdoor dataset have been established to further validate the model under real-world conditions. These scenarios will include static subjects in a car park environment with varying levels and frequencies of ambient illumination. This campaign aims to provide deeper insights into each model's handling of heterogeneous illumination in conjunction with subject movement.

Despite the challenges posed by varying illumination conditions, with a sufficiently large dataset, it is anticipated that the model can be trained to manage these variations as long as they remain below the Nyquist rate of the camera's sampling frequency. To handle more extreme fluctuations, future models may need to incorporate predictive capabilities, generating extended forecasts of the PPG signal from shorter video segments. This approach would resemble an observer mechanism in control systems, enabling the algorithm to maintain robust predictions even when confronted with noisy image data.

Existing Seeing Machines OMS and DMS systems already include and budget for the compute time needed for face tracking. Future research should investigate whether cropping of the input data based on face landmarks can be used to boost performance. Furthermore, the results of this project found that the RVM camera configuration does not pose a significant challenge to existing rPPG neural methods. This suggests that rPPG sensing could be applied in OMS. Future datasets should include the addition of occupants in vehicle to assess this capability.

Finally, the findings of this project suggest more aggressive spatial downsampling may be beneficial for flood-based rPPG systems, and that higher resolution inputs could be beneficial for dot-based rPPG systems in obtaining a more rich set of features from the dot beam profiles. These should also be investigated by future research.

9 Reflection

The scope of this project was wide-reaching and of significant depth. As an academic, an engineer, an industry member, and an employee; this project has significantly furthered my development. I am extremely fortunate for the privileged opportunity I was provided in leading such a large project and I am grateful to Seeing Machines for investing in my professional development through their support of this project.

Within industry, this project has provided me with experience in the planning and management of long-term, novel, medium-scale research projects. My responsibility in planning and organising this project from feature definition to results familiarised myself with some of the commonly encountered pitfalls within industry. In particular, the data collection campaign of this project required me to manage the development of an entirely novel configuration and protocol whose realisation was a multi-disciplinary effort between Human Factors, Data Engineering, Test Engineering, Machine Learning, Advanced Software Engineering and Hardware Engineering Teams. The systems engineering framework taught by the Australian National University was essential in enabling me to effectively coordinate work across these teams. The opportunity to lead this multi-disciplinary effort from a technical level significantly expanded my leadership skills and I am confident I will be able to apply the lessons learned throughout this process to my future work. Furthermore, the delays experienced throughout the data collection campaign forced me to think creatively about how I could amend the project's timeline such that the critical path did not exceed the final end date.

Most of all, I am proud of the results I was able to salvage after it was discovered that the ground-truth physiological signals of the recorded dataset were corrupted. Before this finding, there was significant uncertainty as to why results training on the SM-EOD dataset were so poor. Despite spending a significant amount of time comparing the two datasets, I was unable to explain why such large discrepancies were observed. Although it was disappointing to uncover that the ground-truth data was corrupted, the use of pseudo-labels to obtain better results validated my training and testing methodologies, building my confidence in the training pipeline that I created. Furthermore, through the use of these pseudo-labels I was still able to answer all of the initial questions this project sought to address, albeit at a more constrained level.

This project required me to prepare several demonstrations for presentation both internally and externally. This process taught me to manage the expectations of all parties to ensure results were properly and fairly represented.

The results of this project will guide future research of this topic at Seeing Machines. To enable a smooth handover, all work conducted over the course of the project is stored in a GitHub repository that includes extensive documentation.

On a technical level, this project furthered my abilities with regards to data science, machine learning, computer vision, biometric sensing, three-dimensional sensing, software development, digital organisation and documentation, optics, systems engineering, and experimental design; amongst countless others. I look forward to continuing to develop each of these skills as well as new skills on future research endeavors.

To conclude, I would like to reiterate that this project was the culmination of several minds. Without Ricky, Abraham, Cameron, and John; no data would have been collected. Without Tom, this project would have been severely bottlenecked by a lack of internal resources. Without Andrei, the scope of this project would have been even more unrealistic with less results to show. These colleagues and friends have my everlasting thanks for the time that they committed to this project and the results are a testament to both their contributions as well as my own.

References

[1] Australian Institute of Health and Welfare, "Deaths in australia," 2024. Accessed: 2024-10-10.

[2] Australian Automobile Association, "Fatigued driving," 2024. Accessed: 2024-10-10.

[3] Australian Automobile Association, "Distracted driving," 2024. Accessed: 2024-10-10.

[4] M. Patel, S. Lal, D. Kavanagh, and P. Rossiter, "Applying neural network analysis on heart rate variability data to assess driver fatigue," *Expert Systems with Applications*, vol. 38, no. 6, pp. 7235–7242, 2011.

[5] G. Loudon, "The physiological response during divergent thinking," *Journal of Behavioral and Brain Science*, vol. 6, pp. 28–37, 01 2016.

[6] D. Kurian, P. L. Johnson Joseph, K. Radhakrishnan, and A. A. Balakrishnan, "Drowsiness detection using photoplethysmography signal," in *2014 Fourth International Conference on Advances in Computing and Communications*, pp. 73–76, 2014.

[7] S. Solhjoo, M. C. Haigney, E. McBee, *et al.*, "Heart rate and heart rate variability correlate with clinical reasoning performance and self-reported measures of cognitive load," *Scientific Reports*, vol. 9, p. 14668, 2019.

[8] J. Utama and Y. M. Aminudin, "Design of drunk detection device using non-linear analysis of the heart rate variability method," *AIP Conference Proceedings*, vol. 2510, p. 030027, 10 2023.

[9] S.-H. Kim, S.-M. Jeon, and E. C. Lee, "Face biometric spoof detection method using a remote photoplethysmography signal," *Sensors*, vol. 22, no. 8, 2022.

[10] F. S. Mousavi, A. Esmaeilzehi, and D. Hatzinakos, "Development of an efficient ecg and ppg signal processing-based spoof detection system using convolutional neural networks," in *2023 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, pp. 558–563, 2023.

[11] X. Jin, D. Ye, and C. Chen, "Countering spoof: Towards detecting deepfake with multidimensional biological signals," *Security and Communication Networks*, vol. 2021, no. 1, p. 6626974, 2021.

[12] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE Multimedia - IEEEMM*, vol. 19, pp. 4–10, 02 2012.

[13] J. W. Krug, F. Lüsebrink, O. Speck, and G. Rose, "Optical ballistocardiography for gating and patient monitoring during mri: an initial study," in *Computing in Cardiology 2014*, pp. 953–956, 2014.

[14] C. Lennartz *et al.*, "Beam profile analysis for 3d imaging and material detection whitepaper," 2020. Accessed: 2022-12-01.

[15] W. Chen and D. McDuff, "Deepphys: Video-based physiological measurement using convolutional attention networks," 2018.

[16] X. Liu, J. Fromm, S. Patel, and D. McDuff, "Multi-task temporal shift attention networks for on-device contactless vitals measurement," 2021.

[17] X. Liu, B. L. Hill, Z. Jiang, S. Patel, and D. McDuff, "Efficientphys: Enabling simple, fast and accurate camera-based vitals measurement," 2022.

[18] Z. Yu, Y. Shen, J. Shi, H. Zhao, P. Torr, and G. Zhao, "Physformer: Facial video-based physiological measurement with temporal difference transformer," 2022.

[19] L. National Heart and B. Insititute, "How the heart beats," *International Journal of Sample References*, 2022. Publisher: [United States Government].

[20] B. Horecker, "The absorption spectra of hemoglobin and its derivatives in the visible and near infra-red regions," *Journal of Biological Chemistry*, vol. 148, no. 1, pp. 173–183, 1943.

[21] S. Prahl, "Optical absorption of hemoglobin." [Online], 1999. Available at: `https://omlc.org/spectra/hemoglobin/`. [Accessed 16 May 2024].

[22] National Renewable Energy Laboratory, "Reference air mass 1.5 spectra," 2024. Accessed: 2024-10-10.

[23] D. Shao, F. Tsow, C. Liu, Y. Yang, and N. Tao, "Simultaneous monitoring of ballistocardiogram and photoplethysmogram using a camera," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 5, pp. 1003–1010, 2017.

[24] G. Narayanswamy, Y. Liu, Y. Yang, C. Ma, X. Liu, D. McDuff, and S. Patel, "Bigsmall: Efficient multi-task learning for disparate spatial and temporal physiological measurements," 2023.

[25] G. de Haan and V. Jeanne, "Robust pulse rate from chrominance-based rppg," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 10, pp. 2878–2886, 2013.

[26] E. M. Nowara, T. K. Marks, H. Mansour, and A. Veeraraghavan, "Sparseppg: Towards driver monitoring using camera-based vital signs estimation in near-infrared," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1353–135309, 2018.

[27] E. M. Nowara, T. K. Marks, H. Mansour, and A. Veeraraghavan, "Near-infrared imaging photoplethysmography during driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 4, pp. 3589–3600, 2022.

[28] C. Pilz, S. Zaunseder, J. Krajewski, and V. Blazek, "Local group invariance for heart rate estimation from face videos in the wild," 06 2018.

[29] A. Lindqvist and M. Lindelöw, "Remote heart rate extraction from near infrared videos: An approach to heart rate measurements for the smart eye head tracking system," master's thesis, Chalmers University of Technology, Gothenburg, Sweden, 2016. Supervisor: Björn Lindahl, Smart Eye AB; Academic Supervisors: Lars Hammarstrand and Karl Granström; Examiner: Tomas McKelvey.

[30] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation.," *Opt. Express*, vol. 18, pp. 10762–10774, May 2010.

[31] W. Wang, S. Stuijk, and G. de Haan, "A novel algorithm for remote photoplethysmography: Spatial subspace rotation," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 9, pp. 1974–1984, 2016.

[32] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, "Algorithmic principles of remote ppg," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1479–1491, 2017.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[34] J. Estepp, E. Blackford, and C. Meier, "Recovering pulse rate during motion artifact with a multi-imager array for non-contact imaging photoplethysmography," *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, vol. 2014, pp. 1462–1469, 10 2014.

[35] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2012.

[36] G. Haan and A. Leest, "Improved motion robustness of remote-ppg by using the blood volume pulse signature," *Physiological measurement*, vol. 35, pp. 1913–1926, 08 2014.

[37] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," 2019.

[38] P. Chan, C. Wong, Y. C. Poh, L. Pun, W. W. Leung, Y. Wong, M. M. Wong, M. Poh, D. W. Chu, and C. Siu, "Diagnostic performance of a smartphone-based photoplethysmographic application for atrial fibrillation screening in a primary care setting," *Journal of the American Heart Association*, vol. 5, no. 7, p. e003428, 2016.

[39] M. Hosanee, G. Chan, K. Welykholowa, R. Cooper, P. A. Kyriacou, D. Zheng, J. Allen, D. Abbott, C. Menon, N. H. Lovell, N. Howard, W.-S. Chan, K. Lim, R. Fletcher, R. Ward, and M. Elgendi, "Cuffless single-site photoplethysmography for blood pressure monitoring," *Journal of Clinical Medicine*, vol. 9, no. 3, p. 723, 2020.

[40] D. McDuff, S. Gontarek, and R. W. Picard, "Remote detection of photoplethysmographic systolic and diastolic peaks using a digital camera," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 12, pp. 2948–2954, 2014.

[41] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021.

[42] Z. Shao, Z. Liu, J. Cai, and L. Ma, "JÂa-net: Joint facial action unit detection and face alignment via adaptive attention," *International Journal of Computer Vision*, vol. 129, p. 321–340, Sept. 2020.

[43] X. Liu, G. Narayanswamy, A. Paruchuri, X. Zhang, J. Tang, Y. Zhang, Y. Wang, S. Sengupta, S. Patel, and D. McDuff, "rppg-toolbox: Deep remote ppg toolbox," *arXiv preprint arXiv:2210.00716*, 2022.

[44] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois, "Unsupervised skin tissue segmentation for remote photoplethysmography," *Pattern Recognition Letters*, 2017.

[45] R. Stricker, S. Müller, and H.-M. Gross, "Non-contact video-based pulse rate measurement on a mobile service robot," in *Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication (Ro-Man 2014)*, (Edinburgh, Scotland, UK), pp. 1056–1062, IEEE, 2014.

[46] R. Meziati, Y. Benezeth, P. De Oliveira, J. Chappé, and F. Yang, "Ubfc-phys," 2021.

[47] J. Tang, K. Chen, Y. Wang, Y. Shi, S. Patel, D. McDuff, and X. Liu, "Mmpd: Multi-domain mobile video physiology dataset," 2023.

[48] D. McDuff, M. Wander, X. Liu, B. L. Hill, J. Hernandez, J. Lester, and T. Baltrusaitis, "Scamps: Synthetics for camera measurement of physiological signals," 2022.

[49] X. Niu, H. Han, S. Shan, and X. Chen, "Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video," 2018.

[50] I. LLC, "Imatest it, version 23.2." https://www.imatest.com/products/imatest-it/, 2024. Accessed: 2024-10-13.

[51] Shimmer Sensing, "Shimmer labview instrument driver," 2024. Accessed: 2024-10-13.

[52] K. Yoshida and N. Okiyama, "Estimation of reflectance, transmittance, and absorbance of cosmetic foundation layer on skin using translucency of skin," *Opt. Express*, vol. 29, pp. 40038–40050, Nov 2021.

[53] W. Wang and C. Shan, "Impact of makeup on remote-ppg monitoring," *Biomedical Physics & Engineering Express*, vol. 6, 03 2020.

[54] "Fleet — seeing machines," 2024. Accessed: 2024-10-20.

[55] Python Heart Rate Analysis Toolkit Developers, "Python heart rate analysis toolkit," 2024. Accessed: 2024-10-13.

[56] B. Tegegne, T. Man, A. Roon, H. Snieder, and H. Riese, "Reference values of heart rate variability from 10-second resting electrocardiograms: the lifelines cohort study," *European Journal of Preventive Cardiology*, vol. 27, p. 204748731987256, 09 2019.

[57] E. M. Nowara, T. K. Marks, H. Mansour, and A. Veeraraghavan, "Near-infrared imaging photoplethysmography during driving," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–12, 2020.

[58] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," 2022.

[59] Z. Yu, X. Li, and G. Zhao, "Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks," in *British Machine Vision Conference (BMVC)*, 2019.

[60] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," 2018.

[61] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019.

[62] E. M. Nowara and K. Anastasios, *Towards Robust Imaging Photoplethysmography in Unconstrained Settings*. PhD thesis, USA, 2021. AAI28735953.

[63] A. Paruchuri, X. Liu, Y. Pan, S. Patel, D. McDuff, and S. Sengupta, "Motion matters: Neural motion transfer for better camera physiological measurement," 2023.

[64] T. O'Brien, "Heart rate monitoring via remote photoplethysmography in real-time utilising visible light," June 2022.

A Preliminary Study

Background

It may be possible to detect heart rate using NIR light sources such as LEDs and dot pattern projectors. Existing research has already shown the capacity of NIR LEDs for heart rate sensing [26], [15]; and existing internal research has demonstrated the capability of 8-bit RGB-IR optical configurations for heart rate sensing [64].

However, the monochromatic nature of NIR light makes this problem significantly more difficult, as the pulsatile signal is ≈10 times stronger in the green spectrum of light than in NIR [26]. Thus, for NIR detection, higher resolution hardware will likely be necessary (e.g. 12-bit 5MP image sensor).

Before allocating internal resources into an extensive 12-unit research project, it was desired to perform a preliminary assessment of the viability of this project. Accordingly, a short, week-long feasibility study was conducted. This appendix summarises the findings of this feasibility study.

Experiment

Aim.

Can the intensity of a projected dot be used to measure heart rate by detecting variation of blood volume in the microvascular bed of tissue through the photoplethysmography (PPG) signal.

Pitfalls.

The pulsatile signal is ≈10 times stronger in the green spectrum of light than in NIR [26]. Thus, for NIR detection, higher resolution hardware will likely be necessary (e.g. 12-bit 5MP image sensor).

Equipment.

- Ximea MC050MG-SY-UB camera
- 5MP Sony Image Sensor
- 12-bit monochromatic stream with 500us exposure and 12dB analog gain at 30 f.p.s.
- Dot Pattern Projector pulsed producing a single 2A, 47us pulse for each exposure of the camera.
- Edmund Optics 17.5mm FL f/2.5 Blue M12 MicroVideo Lens

Setup.

- 155mm from subjects forehead to lens.
- Illuminator 300mm from subjects forehead.
- Subject maintains constant positioning through use of a chinstrap.
- Ground truth data recorded using pulse oximeter sampling at 100 Hz.

Methodology.

1. Equipment was set up as shown in **Figure A.1**. The camera was positioned such that the subjects forehead comprised majority of the image FOV.
2. A 30-second recording was taken. For each sample recorded from the camera and pulse oximeter. The Unix time was recorded. This was later used to synchronise the two data streams.

Post-processing

Several methods of heart rate detection were attempted. Firstly, the fast Fourier transform (FFT) was used to analyse the frequency content of the (1) peak intensity, (2) dot dilation, (3) image centroid, and (4) sum intensity of various dots in the recording. These did not yield and

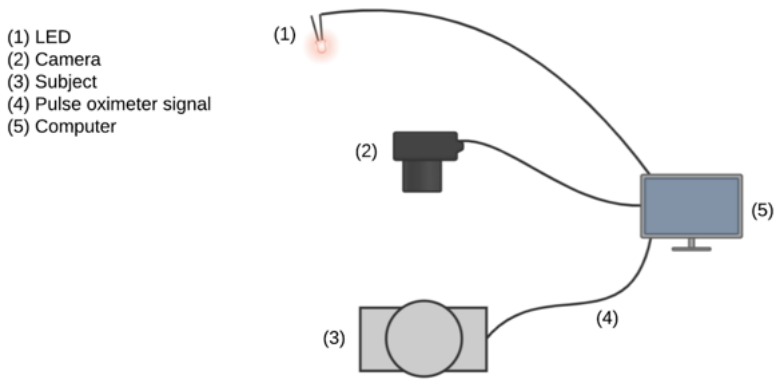

Figure A.1: *Experimental setup.*

immediately promising results. Notably, however, zero filtering of these signals was performed, so they cannot be entirely invalidated as viable methods.

Secondly, the methodology presented by [29] for NIR flood light detection was adapted. The method can be summarised as:

1. For a given region (i.e., dot), compute the sum of the pixel intensity values for each frame.
2. Apply a "detrending" finite impulse response (FIR) high-pass filter with a cutoff of 0.38 Hz to remove trends such as small head movements and small light changes.
3. Apply a moving average filter with a time window of 0.125 seconds to remove high-frequency dynamic noise.
4. Apply a Hamming window band-pass FIR filter with cutoff frequencies [0.67, 4] Hz, corresponding to [40, 240] BPM.
5. Compute the DFT using Welch's method with a 50% window overlap for each time segment.
6. For each analysed dot, sum the DFTs to produce a final, single DFT.

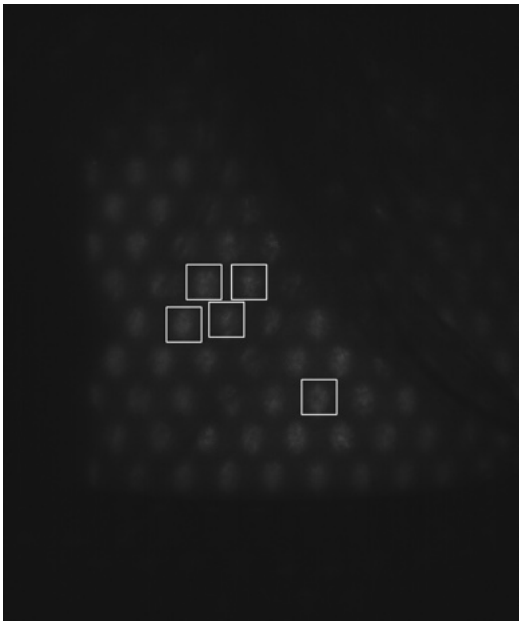

Figure A.2: *The regions-of-interest from which each dot signal was computed.*

This methodology yielded promising results, which are shown in **Figure A.3**.

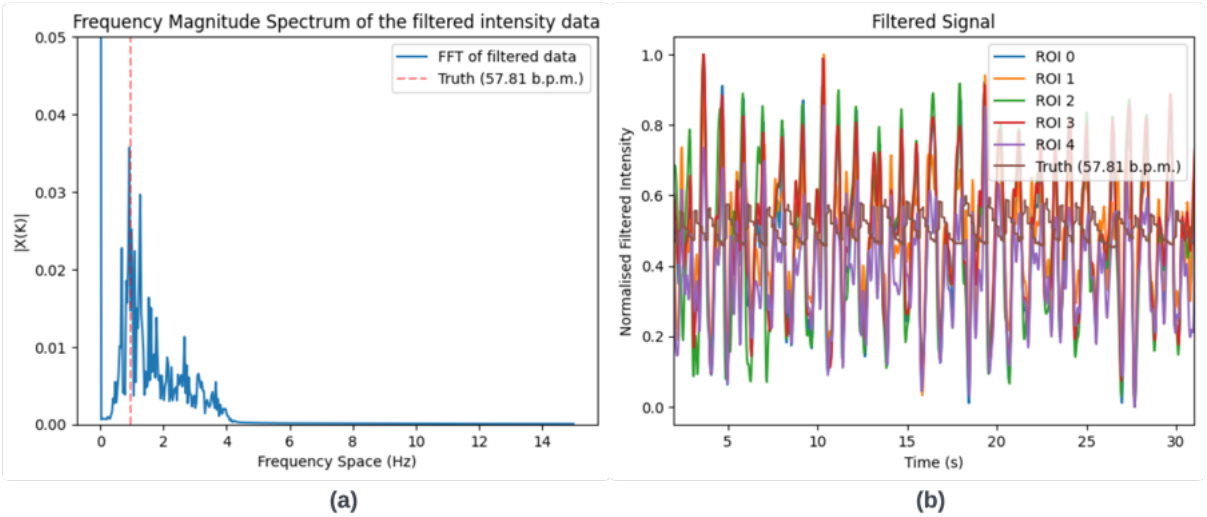

Figure A.3: **(a)** *The FFT of the filtered data plotted with the ground-truth HR data. A strong peak in the frequency magnitude is clearly visible corresponding to the ground-truth HR frequency.* **(b)** *The filtered signal plotted alongside the pulse oximeter truth. Each peak observed from the pulse oximeter corresponds to a trough in the filtered signal*

Results and Discussion

The FFT of the filtered signal is shown in **Figure A.3**, with the frequency corresponding to the ground-truth heart rate signal marked in orange. There is a clear peak in the filtered signals frequency content observed at the frequency of the ground-truth heart rate. Furthermore, by observing the filtered signal alongside the ground-truth PPG signal, there is a clearly discernible trough in the filtered signal with each peak of the truth signal.

Notably, [29] found the forehead to provide the worst detection accuracy out of all regions in the face. Thus, it is likely that significantly better performance would be observed using a different region of the face.

Furthermore, the ground truth frequency for heart rate was computed by counting the number of beats observed in a recording and dividing by the recording duration. This method is susceptible to irregularities in the heart rate rhythm.

B Metrics

Heart Rate.

Heart rate is measured in the frequency domain. A buffer of 10 seconds worth of PPG/rPPG data is processed for each measurement of heart rate. Firstly, a detrending filter is applied to the window. The detrending filter applies Whittaker smoothing where the amount of detrending is determined by the smoothing parameter, λ. The detrended signal is passed through a first-order Butterworth filter with cutoff frequencies [0.75, 2.5] Hz (corresponding to [45, 150] BPM). The heart rate for the window is then recorded as the frequency of the filtered window with peak spectral power.

MAE.

The mean absolute error (MAE) is the mean of the absolute value of differences between the predicted and measured heart rate. The MAE can be interpreted as, on average, how many BPM the model over or underestimates a windows heart rate.

$$MAE = \frac{\sum_{i=0}^{N} |(A_i - B_i)|}{N} \pm \frac{std(|(A_0 - B_0)| + ... + |(A_N - B_N)|)}{\sqrt{N}}$$

RMSE.

The root mean squared error (RMSE) is the square root of the mean of the square of differences between predicted and measured heart rate.

$$RMSE = \sqrt{\frac{\sum_{i=0}^{N} (A_i - B_i)^2}{N}} \pm \frac{std((A_0 - B_0)^2 + ... + (A_N - B_N)^2)}{\sqrt{N}}$$

RMSE is very similar to MAE, but penalises larger errors (i.e., outliers) more significantly. This makes RMSE a useful loss term when trying to heavily penalise large deviations from actual values. In comparison, MAE treats all errors equally. A large RMSE and a comparatively small MAE means there are many outliers.

MAPE.

Mean absolute percentage error (MAPE) is the average percentage error between predictions and measurements.

$$MAPE = \sum_{i=0}^{N} \frac{|(A_i - B_i)|}{B_i} \times 100 \pm \frac{std(\frac{|(A_0 - B_0)|}{B_0} + ... + \frac{|(A_N - B_N)|}{B_N})}{\sqrt{N}} \times 100$$

MAPE provides an intuitive measure of prediction accuracy by expressing errors as a percentage of the actual values. Unlike RMSE and MAE, which are absolute measures of error, MAPE normalizes the error, allowing for easier comparisons across different datasets or scales. While this makes MAPE a valuable metric in cases where understanding the relative error is important, it can also cause issues when the actual values are close to zero, leading to disproportionately large error values.

Pearson.

The Pearson correlation coefficient measures the linear relationship between predictions and measurements. The closer the absolute value of r is to 1, the stronger the linear relationship. The closer to 0, the weaker the relationship.

The Pearson correlation coefficient:

- Only measures linear relationships. Non-linear relationships might have a low r value even if they are strong.

- Is sensitive to outliers.
- Assumes homoscedasticity (the variability of prediction is constant across levels of measurement).

SNR.

For each window, the signal-to-noise ratio (SNR) is calculated as the ratio of the area under the curve of the frequency spectrum around the first and second harmonics of the ground truth heart rate frequency to the area under the curve of the remainder of the frequency spectrum, from 0.75Hz to 2.5Hz.

The total SNR is given by the mean of the SNR of each window.

C Dataset Collection Campaign

	Task Name	Start	Finish	Duration
1	Data Collection Campaign	11/04/24	23/09/24	116d
2	Prepare data collection protocol	11/04/24	14/06/24	45d
3	Investigate influence of AEC to heart rate detection	02/05/24	03/05/24	2d
4	Choose camera configurations	11/04/24	06/05/24	18d
5	Write protocol documentation	11/04/24	14/05/24	24d
6	Approve protocol with stakeholders	15/05/24	14/06/24	21d
7	Build data collection setup	11/04/24	21/08/24	93d
8	Choose lens	11/04/24	17/04/24	1w
9	Cameron complete synchronised 10bpp 60fps camera streaming	11/04/24	21/06/24	50d
10	Deliver lens to Seeing Machines	18/04/24	07/05/24	14d
11	Build cameras	06/05/24	24/05/24	15d
12	Test AICO lens	05/06/24	05/06/24	1d
13	Find lens for data collection	05/06/24	11/06/24	4d
14	John complete DataRecorder app	11/04/24	12/07/24	65d
15	Derisk use of Shimmer Consensys software	20/06/24	21/06/24	2d
16	Generate camera intrinsic calibrations	24/06/24	25/06/24	2d
19	Generate MagikEye camera calibrations	25/06/24	26/06/24	2d
21	Create PNG to EOD pipeline	16/05/24	21/08/24	68d
22	Add 16 bit images support to pyeod	01/08/24	21/08/24	15d
23	Write image to eod conversion script	16/05/24	17/05/24	2d
24	Pilot test	24/06/24	04/09/24	53d
25	Setup hardware	24/06/24	25/06/24	2d
26	John to debug and fix DataRecorder App issues	26/06/24	31/07/24	26d
27	Collect data	01/08/24	01/08/24	1d
28	Validate data	02/08/24	04/09/24	24d
29	Collect data with data collection team	19/08/24	23/09/24	26d
30	Collect 24 subjects data	19/08/24	02/09/24	11d
31	Assess ability to recover corrupted data	05/09/24	05/09/24	1d
32	DAQ to finish uploading data to s3	10/09/24	23/09/24	10d
33	Prepare preprocessed data	13/08/24	23/09/24	30d
34	Create preprocessing pipeline for collected data	13/08/24	10/09/24	21d
35	Format and process data for 24 collected subjects	18/09/24	23/09/24	4d

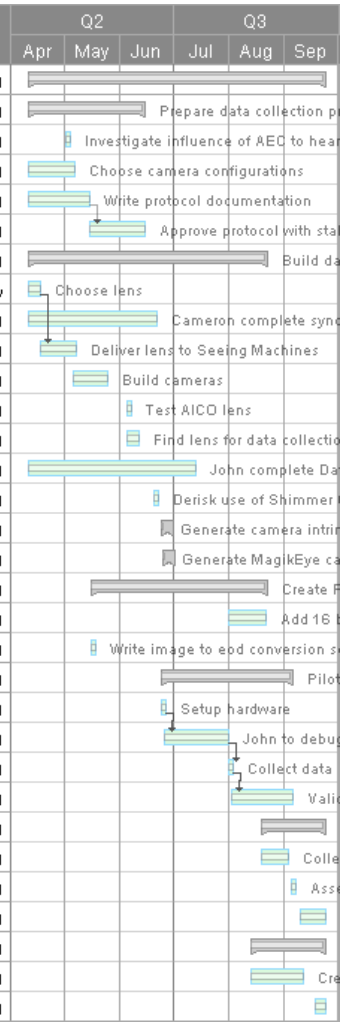

Figure C.1: *Timeline of the project's data collection campaign.*

D Corrupted Data Summary

	iGR1IC	iGR1RVM	iGR1CC	iGR1LSM	iGR1RSM	iGR3	iGR2	iNR1	iNR3	iNR2	iGR5-1	iGR5-2	iGR5-3	iGR4	iGR6
D01	0.0	0.9997	0.0	0.0	0.0	0.0	0.9999	0.0	0.0	-	-	-	-	-	-
D02	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	0.0	0.0	0.0	0.0	-	-
D03	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	-	-	-	-	-	-
D04	0.0	0.0	0.4683	0.0	0.0	0.0	0.0	0.5148	-	-	-	-	-	-	-
D05	0.0	0.0	0.0	0.0	0.0	0.0	0.6261	0.6096	-	-	-	-	-	-	-
D06	0.0	0.0	0.3519	0.0	0.0	0.0	0.0	0.0	-	-	-	-	-	-	-
D07	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	-	-	-	-	-	-	-
D08	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	-	-	-	-	-	-
D09	0.0	0.0	0.0	0.4930	0.0	0.9991	0.0	0.0	-	-	-	-	0.0	-	-
D10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	-	-	-	-	-
D11	0.5023	0.4688	0.4815	0.0	0.0	0.0	0.4594	0.0	0.0	-	-	-	-	-	-
D12	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5268	-	-	-	-	-	-	-
D13	0.0	0.0	0.0	0.0	0.0	-	0.0	0.0	-	-	-	-	-	-	0.0
D14	0.0	0.0	0.0	0.2074	0.0	-	0.0	0.0	0.0	-	-	-	-	-	0.0
D15	0.3245	0.5771	0.0	0.3198	0.3277	-	0.0	0.6446	-	-	-	-	-	-	0.0
D16	0.9994	0.0	0.4580	0.0	0.0	-	0.0	0.8203	-	-	-	-	-	-	0.0
D17	0.7866	0.2270	0.0	0.0	0.0	0.0	0.0	0.0	-	-	-	-	-	-	-
D18	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0941	-	-	-	-	-	-
D19	0.0	0.2593	0.0	0.0	0.0	0.0	0.0	0.0	0.9999	-	-	-	-	-	-
D20	0.0	0.0	0.0	0.2337	0.0	0.0	0.0	0.0	0.3876	-	-	-	-	-	-
D21	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	-	-	-	-	-
D22	-	-	-	-	-	-	-	0.0	-	-	-	-	-	-	-
D23	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	-	-	-	-	-
D24	0.0	0.2286	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	-	-	-	-	-
total	0.1089	0.1150	0.0733	0.0522	0.0137	0.0416	0.0869	0.1298	0.0617	0.0	0.0	0.0	0.0	0.0	0.0