

# Deep Learning-based rPPG Models towards Automotive Applications: A Benchmark Study

Tayssir Bouraffa

Chalmers University of Technology

Department of Computer Science and Engineering  
Gothenburg, Sweden

tayssir@chalmers.se

Dimitrios Koutsakis

Chalmers University of Technology

Gothenburg, Sweden

koutsakis.d@hotmail.com

Salvija Zelvyte

Chalmers University of Technology

Gothenburg, Sweden

salvija.zelvyte22@gmail.com

## Abstract

*Remote photoplethysmography (rPPG) has the potential to significantly enhance driver safety systems by enabling the detection of critical conditions, such as driver drowsiness and sudden illness, through non-invasive monitoring of cardio-respiratory functions. However, the dynamic environment within a vehicle, characterized by motion artifacts and varying illumination, presents unique challenges for accurate rPPG estimation. In this study, we conducted a comprehensive benchmark of various supervised and unsupervised rPPG algorithms using the MR-NIRP car dataset to assess their performance in automotive settings. Qualitative and quantitative experiments were performed to evaluate and compare several rPPG models designed in stable, noise-controlled environments, highlighting the impact of real-world conditions on model performance. Our findings highlight the promise of machine learning approaches, particularly neural network-based models, in overcoming these challenges and accurately estimating heart and respiration rates in real-world driving scenarios. This study underscores the potential for integrating rPPG-based monitoring systems into vehicles to enhance driver safety and well-being.*

## 1. Introduction

Traffic accidents claim over a million lives annually, with an additional 50 million people injured. Between 2000 and 2016, fatalities rose from 1.15 million to 1.35 million [1], emphasizing the urgent need for effective road safety measures. Driving requires continuous focus and situational

awareness, making Driver Monitoring Systems (DMS) essential for mitigating risks like fatigue, distractions, and sudden health emergencies.

Conventional methods for detecting sudden health emergencies rely on contact-based devices, such as radar-based systems [2], electrocardiography (ECG) or pulse oximeters [3]. While effective, these methods can be uncomfortable, impractical for long periods, or even distracting for drivers. High-frequency acoustic signals present an alternative but may cause discomfort to drivers and passengers, including pets and babies [2].

In-cabin cameras are increasingly being used in non-contact driver monitoring systems (DMS) to assess driver states such as drowsiness and distraction. These systems employ rPPG, a non-invasive technique that uses optical signals to detect subtle facial color changes caused by blood volume variations during cardiac cycles [4, 5]. rPPG facilitates the measurement of vital signs, including heart rate (HR) [6], heart rate variability (HRV) [7], and respiration rate (RR) [7], enabling DMS to detect distraction, fatigue, or sudden health issues that may affect driving safety.

Deep learning (DL)-based rPPG algorithms have demonstrated potential in controlled laboratory settings [8]. However, adapting them to vehicles introduces challenges like fluctuating illumination from external factors like streetlights, trees and buildings, and motion artifacts caused by road conditions, driver actions, and vibrations [5]. These challenges necessitate optimization for real-time performance and seamless integration into vehicles to ensure reliable health monitoring.

The primary motivation for this research lies in the significant potential of using cameras for remote vital sign

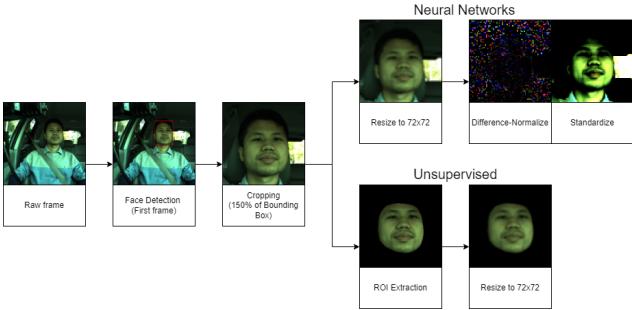


Figure 1. Preprocessing procedure for input videos used in both unsupervised methods and NN models.

measurement to improve DMS by eliminating the need for contact-based devices. This approach offers a more comfortable and seamless integration within vehicles, enhancing the convenience of health monitoring while enabling proactive identification of potential risks. Such advancements facilitate the implementation of preventive measures aimed at improving overall road safety. However, despite several years of research into rPPG for monitoring driver vital signs on the road, progress has been limited, particularly in addressing the unique challenges posed by the dynamic automotive environment.

This research aims to advance non-contact DMS by enhancing the rPPG-toolbox [8] to support automotive datasets and measure both HR and RR. We introduce a benchmarking framework for rPPG algorithms addressing challenges like vehicle motion and variable illumination. Using supervised and unsupervised neural network (NN) models, we estimate HR and RR from blood volume pulse signals and evaluate these models using five performance metrics. Cross-dataset testing across PURE [9] and UBFC-rPPG [10], SCAMPS [10] and MR-NIRP [5] ensures robustness and generalizability. Standardized pre-processing and post-processing workflows ensure consistent evaluations.

The remainder of this paper is structured as follows: Section 2 reviews rPPG algorithms and related research. Section 3 outlines the methodology, dataset, and comparison framework. Section 4 describes the experimental setup. Results and recommendations are presented in Section 5, followed by the conclusion in Section 6.

## 2. Related Work

In recent years, various rPPG algorithms have been developed to address challenges in HR and RR estimation. Traditional approaches, such as mathematical and signal processing techniques, have been widely explored. Nowara et al. proposed AutoSparsePPG, leveraging quasi-periodicity to enhance rPPG signal estimation with NIR recordings, addressing illumination changes and

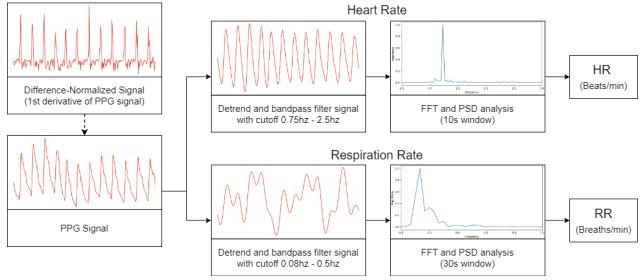


Figure 2. Postprocessing pipeline for the predicted and ground-truth PPG signals to extract the corresponding HR and RR.

head movement [5]. They also introduced MR-NIRP, the only publicly available dataset with synchronized NIR and RGB video recordings and ground-truth pulse oximeter measurements in vehicle environments. Hernandez-Ortega et al. improved HR estimation accuracy using NIR cameras with their Quality-Guided Spectrum Peak Screening method [11]. Gua et al. developed a non-intrusive HR and RR monitoring system using active infrared illumination and time-of-flight cameras to reduce motion artifacts and ambient light effects [12]. Xu et al. introduced Ivrr-PPG, which captures the nonlinear relationship between illumination changes and rPPG signals with NIR images [13].

Here's a more concise version of the text that reduces it by one line while retaining the essential details:

Recently, supervised ML has gained prominence, with DL and Convolutional Neural Networks (CNNs) achieving state-of-the-art results in HR detection. Mitsubishi demonstrated the effectiveness of a time-series U-net with GRU for accurate HR detection using the MR-NIRP dataset [14]. Toyota showcased contrastive learning models with unlabeled video data for HR detection [15]. Huang et al. introduced a transformer-based algorithm using RGB images and a custom dataset addressing motion and illumination artifacts [16]. Chiu et al. developed a CNN-based model for rPPG in driving scenarios, tackling head movement and illumination changes with an encoder-decoder for NIR images and the CHROM model for RGB images [17]. Wu et al. used a k-nearest neighbor classifier on frequency domain features from RGB videos for HR monitoring in vehicles, later extending it to outdoor driving data to improve HR estimation in outdoor conditions [18, 19]. Park et al. proposed a self-supervised transformer-based framework for rPPG estimation, combining RGB and NIR videos to extract spatiotemporal features and enhance HR estimation with contrastive learning [20].

Most rPPG algorithms focus on HR detection in vehicle environments, with limited research on RR monitoring, particularly using DL models. Many rely on custom datasets or subsets of the MR-NIRP dataset, often excluding the large-motion subset or focusing only on still garage conditions.

Additionally, these datasets and codes are often not publicly accessible. Wang et al. conducted a benchmark study with the MR-NIRP car dataset, comparing two unsupervised physiological-based methods and two modified PhysNet DL models [21]. Results showed physiological models often outperformed DL models in challenging scenarios, but the study’s limited exploration of DL methods left other algorithms unassessed. To date, no comprehensive benchmark has compared DL-based rPPG methods in automotive environments, a gap our research aims to address.

### 3. Methodology

To advance research in camera-based physiological measurements and improve experimental reproducibility in vehicular environments, we present a comprehensive benchmark study using the MR-NIRP car dataset. This study evaluates the performance of eleven rPPG models, including six conventional unsupervised methods and five supervised NN models, focusing on HR and RR estimation from BVP signals captured by camera devices. Examples of predicted HR and RR signals for PhysNet and PhysFormer are illustrated in Figure 3. Our analysis covers data pre-processing, model training, post-processing, and evaluation, with a detailed comparison of CNN and transformer-based architectures to assess their effectiveness in addressing the challenges of vehicular contexts.

**Unsupervised Algorithms:** Several traditional methods have been proposed to extract rPPG signals from red-green-blue (RGB) images. Blind source separation (BSS) methods, such as **POS**: Plane-Orthogonal-to-the-Skin (POS) method calculates a projection plane orthogonal to the skin tone, relying on optical and physiological principles [22]. **ICA**: Poh et al. introduced a method for extracting BVP by leveraging ICA to divide temporal RGB color signals into independent or uncorrelated sources [7]. Similarly, **CHROM**: chrominance-based rPPG (CHROM), introduced by de Haan et al., estimates the BVP through a linear combination of the chrominance signals obtained from the RGB data [23]. **LGI**: Local Group Invariance (LGI) is a feature representation technique developed to remain invariant to motion by applying differentiable local transformations, as introduced by Pilz et al [24]. The **GREEN**: GREEN method leverages the fact that the green channel, due to its mid-spectrum wavelength, exhibits a higher absorption contrast with blood compared to the blue and red channels, making it more responsive to variations in blood volume beneath the skin [25]. Lastly, the **PBV**: Blood Volume Pulse Vector (PBV) method, also proposed by de Haan et al., captures blood pulsations in the RGB channels through a distinct signature, enabling accurate isolation of the pulsatile component of the blood volume signal [25].

**Supervised NN Algorithms:** Most current supervised NN models for rPPG rely on a CNN-based architecture.

These models typically incorporate attention mechanisms to detect and extract regions of interest (ROI), which are subsequently processed through a series of convolutional and fully connected layers to extract the BVP. **DeepPhys**: Chen et al. developed DeepPhys, an end-to-end model for video-based, non-contact measurement of HR and RR. This model utilizes a two-branch 2D convolutional attention network, which combines a motion representation derived from a skin reflection model with an attention mechanism utilizing appearance information [26]. **EfficientPhys-C**: Liu et al. also introduced EfficientPhys-C, a 2D-CNN with a TSM designed for real-time on-device computations. Unlike TS-CAN and DeepPhys, EfficientPhys-C features a single-branch architecture that includes a normalization module and a self-attention mechanism, specifically targeting skin pixels relevant to the PPG signal [27]. **TS-CAN**: Liu et al. introduced TS-CAN, a 2D temporal shift convolutional attention network that incorporates a Temporal Shift Module (TSM) to capture and process spatiotemporal information, with dual branches for motion modeling and spatial feature extraction, operating concurrently with a gated attention mechanism [28]. **PhysNet**: Yu et al. proposed PhysNet, an end-to-end spatio-temporal network that employs a 3D-CNN to capture semantic rPPG features across both spatial and temporal dimensions, improving the learning of robust contextual information and minimizing temporal variations [29]. **PhysFormer**: PhysFormer adopts a video transformer-based architecture, dynamically integrating local and global spatio-temporal elements, with an emphasis on extracting long-term global features to improve rPPG signal representation [30].

**Datasets:** In our experiments, we evaluated the proposed method using the MR-NIRP car dataset [5], PURE [9] and UBFC-rPPG [10] and SCAMPS [10] using a five-fold subject-exclusive cross-validation technique. Few studies have captured rPPG measurements in car environment using camera sensors, with the MR-NIRP car dataset being the only publicly available dataset recorded in such settings [5], while PURE and UBFC were recorded in controlled indoor settings and SCAMPS is a synthetic dataset. The MR-NIRP car dataset includes recordings from 18 healthy participants, comprising 16 males and 2 females, aged between 25 and 60 years, with a variety of facial features and skin tones. The recordings were made using RGB and NIR (10-bit raw) cameras at a resolution of 640x640 pixels and a frame rate of 30 Hz. Ground-truth PPG signals were collected utilizing a finger pulse oximeter, recorded at 60 Hz. To ensure safety and signal quality, all subjects were seated in the passenger seat during the recordings [5]. The MR-NIRP car dataset videos were recorded in two main settings: inside a garage with the engine running and while driving in urban environments, which included stops, accelerations, and turns. For the driving scenario, the recordings captured a range of nat-

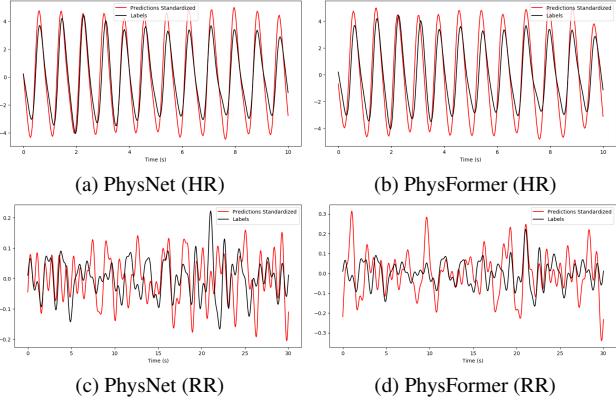


Figure 3. Example plots of the predicted HR and RR signals against the ground-truth signal (black) for PhysNet and PhysFormer models. These examples correspond to the "subject18\_garage\_still\_940" recording of the MR-NIRP car dataset.

ural lighting conditions, such as daytime in both sunny and overcast weather, as well as nighttime. The dataset is further categorized into three motion conditions: still, where participants remained seated with minimal head movement; small motion, where participants made natural head movements and engaged in conversation; and large motion, involving more pronounced and abrupt head movements.

## 4. Experiments

**Experimental Setup:** For both HR and RR estimation, we conducted quantitative and qualitative evaluations using five metrics: MAE, RMSE, MAPE,  $\rho$ , and SNR. Further details on the selected metrics and model training are provided in Appendix C and D. All implementations were conducted using PyTorch on a single Nvidia Tesla V100 GPU with 32GB of memory. Our implementation has been integrated with the open-source rPPG-toolbox [8] and is available at the following [GitHub repository](#).

**Pre-processing:** A standardized pre-processing procedure is adopted to ensure uniform data preparation across the evaluated algorithms, as shown in Figure 1. This procedure ensures that all rPPG algorithms are compared fairly, with only slight modifications to meet the specific input needs of each method.

Given the head movements that occur during driving, accurate face detection and tracking are essential for reliable rPPG estimation. For the supervised NN methods, face detection was performed on the first frame of each video using RetinaFace method, a strong single-stage face detection framework that utilizes multi-task learning for precise pixel-level face localization [31]. Once the face was detected, it was cropped into a rectangular region scaled to 150% of the bounding box to ensure that the entire face was captured even with slight movements. The cropped frames

were subsequently resized to 72x72 pixels for most NN models. However, for the Physformer model, the frames were adjusted to 128x128 pixels to meet the input requirements of the original model [30]. Additionally, the ground-truth PPG signal labels were downsampled from 60 Hz to 30 Hz to synchronize with the video frame rate.

For the unsupervised methods, MediaPipe FaceMesh framework was utilized to identify facial landmarks on each image frame [32]. The entire face was designated as the ROI, using a total of 478 landmark indices. This framework enabled the extraction of the ROI by defining a convex hull polygon around these landmarks. For the NN methods, all algorithms included a trainable attention mechanism, which autonomously learned to extract the appropriate ROIs. To further reduce dependence on overall frame brightness and the subject's skin tone, difference-normalized frames were provided as inputs for the NN models [28]. This process involved calculating the difference between every two consecutive frames and labels and normalizing these differences by their standard deviation. EfficientPhys-C did not require this step, as its already includes a normalization module that computes frame differences. Instead, standardized frames were used as input, and the labels were difference-normalized. For DeepPhys and TS-CAN, which have two branches, difference-normalized frames were fed into the motion branch, while standardized frames were applied to the appearance branch.

**Post-processing:** The post-processing procedure adopts several standardized steps to facilitate accurate and reliable performance evaluation, as shown in Figure 2. Initially, the predicted rPPG signal was uniformly processed across all algorithms to extract HR and RR. When the signal is difference-normalized, it reflects the first derivative of the PPG signal. To reconstruct the original rPPG signal, the cumulative sum of its values is calculated.

Moreover, to eliminate gradual changes or drifts in the signal that occur over a longer period and are not related to physiological measurements, the predicted and ground-truth signals were adjusted utilizing a fixed  $\lambda$  value set at 100. Subsequently, a second-order Butterworth bandpass filter was employed to eliminate noise. The selection of filter frequencies was tailored to the specific physiological measurements. For HR extraction, the filter's upper and lower cutoff frequencies were set to 2.5 Hz and 0.75 Hz, respectively, corresponding to a typical pulse range of 45 to 150 beats per minute (Beats/min) in healthy adults. For RR extraction, the cutoff frequencies were set to 0.5 Hz and 0.08 Hz, encompassing the typical adult RR range of 5 to 30 breaths per minute (Breaths/min). To achieve the desired cutoff frequencies and prevent aliasing, the actual frequencies used in the filter were doubled, following the Nyquist–Shannon sampling theorem [28] [33]. The filtered signals were subsequently converted into the frequency do-

Table 1. Qualitative HR Performance Evaluation for the Unsupervised Methods on the MR-NIRP Subsets.

Unsupervised Methods - HR - MR-NIRP																
Method:	Driving Large Motion					Driving Small Motion					Driving Still					
	MAPE↓	RMSE↓	MAE↓	SNR↑	$\rho\uparrow$	MAPE↓	RMSE↓	MAE↓	SNR↑	$\rho\uparrow$	MAPE↓	RMSE↓	MAE↓	SNR↑	$\rho\uparrow$	
ICA	14.77	15.42	11.14	-10.96	0.23	14.43	14.92	10.58	-10.11	0.22	13.34	14.07	9.46	-7.60	0.30	
POS	<b>13.01</b>	<b>14.22</b>	<b>9.16</b>	<b>-8.12</b>	<b>0.34</b>	<b>12.26</b>	<b>13.87</b>	<b>8.54</b>	<b>-6.96</b>	<b>0.33</b>	<b>12.37</b>	<b>13.97</b>	<b>8.14</b>	<b>-3.92</b>	<b>0.40</b>	
CHROM	15.36	16.19	10.90	-9.75	0.21	14.01	15.22	9.78	-8.07	0.27	12.68	14.68	8.38	-4.72	0.35	
GREEN	19.87	18.67	14.71	-14.16	0.01	17.81	17.31	13.27	-13.74	0.11	16.80	16.61	11.91	-10.86	0.17	
LGI	16.32	17.16	12.23	-10.37	0.09	14.36	15.41	10.68	-8.84	0.26	12.51	14.08	8.74	-5.32	0.34	
PBV	15.49	16.44	11.74	-10.66	0.11	16.19	16.88	12.01	-10.18	0.13	14.06	15.56	9.86	-6.55	0.23	
Unsupervised Methods - HR - MR-NIRP																
Method:	Garage Large Motion					Garage Small Motion					Garage Still					
	MAPE↓	RMSE↓	MAE↓	SNR↑	$\rho\uparrow$	MAPE↓	RMSE↓	MAE↓	SNR↑	$\rho\uparrow$	MAPE↓	RMSE↓	MAE↓	SNR↑	$\rho\uparrow$	
ICA	13.82	16.18	10.75	-9.22	0.13	8.57	11.93	6.19	-3.64	0.44	2.50	5.23	1.64	5.54	0.89	
POS	9.00	12.56	<b>5.97</b>	<b>-4.23</b>	0.53	4.43	7.46	2.97	<b>-0.24</b>	0.75	3.04	7.29	1.73	6.11	0.80	
CHROM	11.06	13.46	7.82	-6.98	0.48	8.22	11.50	5.53	-3.67	0.51	3.53	8.07	1.99	3.50	0.76	
GREEN	18.16	18.10	13.66	-13.69	0.10	17.73	17.43	12.85	-12.16	0.24	9.99	12.83	6.71	-5.72	0.42	
LGI	<b>8.93</b>	<b>11.14</b>	6.37	-5.83	<b>0.61</b>	<b>3.98</b>	<b>6.44</b>	<b>2.90</b>	-0.95	<b>0.82</b>	<b>1.30</b>	<b>2.50</b>	<b>0.91</b>	<b>6.19</b>	<b>0.98</b>	
PBV	10.89	12.89	8.04	-6.50	0.52	7.75	10.17	5.69	-5.37	0.66	7.87	12.22	5.41	-1.07	0.51	

Table 2. Overall HR Performance Evaluation for the Unsupervised Methods on the MR-NIRP Dataset.

Method:	MAPE↓	RMSE↓	MAE↓	SNR↑	$\rho\uparrow$
ICA	13.14	14.34	9.08	-6.60	0.32
POS	12.30	14.26	<b>7.74</b>	<b>-3.71</b>	0.39
CHROM	14.05	15.60	8.94	-5.41	0.31
GREEN	17.71	17.19	12.44	-11.59	0.16
LGI	<b>11.21</b>	<b>13.33</b>	7.87	-4.81	<b>0.44</b>
PBV	13.95	15.45	9.72	-7.13	0.28

main using Fast Fourier Transform (FFT). Power Spectral Density (PSD) analysis was then applied to detect the frequency with the highest power, corresponding to the predicted vital sign. Specifically, HR was extracted in 10-second non-overlapping windows, while RR was extracted in 30-second non-overlapping windows [8] [28].

## 5. Results

We compared the performance of six unsupervised methods: ICA, POS, CHROM, GREEN, LGI, and PBV, and five NN-based methods: DeepPhys, TS-CAN, EfficientPhys-C, PhysNet, and PhysFormer. All the rPPG algorithms were evaluated on the same MR-NIRP car dataset using 5-fold cross-validation, where some cases are excluded. The same pre-processing, post-processing, and time window parameters were applied for HR and RR estimation. More details about the dataset folders and the excluded cases can be found in Appendix A and B.

### 5.1. Heart Rate estimation

In this section, we provide the performance results for HR estimation utilizing both unsupervised methods and supervised NN models. Quantitative evaluation was conducted using the overall dataset, while qualitative evaluation was performed using different subsets, including settings such as the garage, where the illumination is stable,

and driving, which presents more challenging conditions with variable illumination. These two settings are further divided into three different motion conditions (still, small, and large motion).

#### 5.1.1 Unsupervised methods

**Quantitative evaluation:** The table 2 presents the overall HR performance of various unsupervised methods evaluated on the MR-NIRP dataset. LGI emerges as the best overall performer, achieving the lowest RMSE and MAPE, along with the highest  $\rho$ , indicating its superior accuracy and strong linear correlation between predicted and ground-truth HR. POS also demonstrates commendable performance, particularly excelling in MAE and SNR, which reflects its lower average error and better signal quality relative to noise. In contrast, GREEN consistently underperforms across all metrics, showing the highest MAE, RMSE, and MAPE, the lowest  $\rho$ , and the poorest SNR. These results suggest that while LGI and POS are effective for HR estimation in automotive environments, GREEN struggles significantly, highlighting the variability in the robustness and accuracy of different unsupervised rPPG methods.

**Qualitative evaluation:** The table 1 illustrates the HR performance for various unsupervised methods evaluated under different automotive conditions in the MR-NIRP dataset. POS consistently demonstrates superior performance in driving conditions, particularly excelling in large and small motion scenarios with the lowest MAE, RMSE, MAPE, and highest  $\rho$  and SNR. LGI, on the other hand, excels in garage settings, particularly when the subject is still, showcasing the lowest errors and highest correlation and SNR. In contrast, GREEN consistently underperforms across all conditions, indicating its limited effectiveness in both dynamic and stable environments. PBV, ICA, and CHROM show moderate performance, generally falling between the extremes of POS/LGI and GREEN.

Table 3. Qualitative HR Performance Evaluation for the Supervised NN Methods on the MR-NIRP Subsets.

NN Models - HR - MR-NIRP															
NN Model:	Driving Large Motion					Driving Small Motion					Driving Still				
	MAPE↓	RMSE↓	MAE↓	SNR↑	ρ↑	MAPE↓	RMSE↓	MAE↓	SNR↑	ρ↑	MAPE↓	RMSE↓	MAE↓	SNR↑	ρ↑
DeepPhys	15.71	16.49	11.79	-10.18	0.08	13.00	14.56	9.95	-8.02	0.21	9.74	12.06	7.00	-3.47	0.32
TS-CAN	15.58	15.98	11.73	-10.45	0.18	11.81	13.48	8.83	-8.09	0.31	9.91	11.73	7.09	-4.20	0.36
EfficientPhys-C	14.78	15.48	11.07	-10.50	0.20	12.67	14.41	9.64	-8.28	0.24	10.42	12.52	7.44	-3.98	0.31
PhysNet	<b>9.20</b>	<b>11.55</b>	<b>7.04</b>	<b>-4.21</b>	<b>0.35</b>	<b>7.56</b>	<b>10.47</b>	<b>5.80</b>	<b>-1.27</b>	<b>0.44</b>	<b>6.11</b>	<b>9.00</b>	<b>4.37</b>	<b>2.02</b>	<b>0.51</b>
PhysFormer	12.08	12.97	9.17	-7.22	0.19	10.16	12.28	7.88	-3.39	0.26	8.57	10.81	6.35	-0.10	0.37
Garage Large Motion															
NN Model:	Garage Large Motion					Garage Small Motion					Garage Still				
	MAPE↓	RMSE↓	MAE↓	SNR↑	ρ↑	MAPE↓	RMSE↓	MAE↓	SNR↑	ρ↑	MAPE↓	RMSE↓	MAE↓	SNR↑	ρ↑
DeepPhys	13.31	14.65	10.47	-8.72	0.24	6.65	9.46	4.99	-3.36	0.51	1.65	3.80	1.14	6.60	0.87
TS-CAN	10.87	12.90	8.45	-7.13	0.31	4.58	7.66	3.32	-1.97	0.63	1.13	2.32	0.81	7.38	0.88
EfficientPhys-C	10.62	12.70	8.25	-7.64	0.31	4.98	8.27	3.66	-1.86	0.61	1.41	3.08	0.96	7.12	0.87
PhysNet	<b>6.53</b>	<b>10.01</b>	<b>5.43</b>	<b>-0.17</b>	<b>0.52</b>	<b>2.46</b>	<b>5.37</b>	<b>1.96</b>	<b>5.54</b>	<b>0.78</b>	<b>0.52</b>	<b>1.32</b>	<b>0.35</b>	<b>11.78</b>	<b>0.98</b>
PhysFormer	9.10	11.28	7.07	-2.45	0.35	5.00	7.96	4.11	2.92	0.54	3.75	6.72	3.19	9.05	0.73

Table 4. Overall HR Performance Evaluation for the Supervised NN Methods on the MR-NIRP Dataset.

NN Models - HR					
NN Model:	MAPE↓	RMSE↓	MAE↓	SNR↑	ρ↑
DeepPhys	9.95	12.69	7.45	-4.37	0.31
TS-CAN	9.17	11.76	6.78	-4.20	0.40
EfficientPhys-C	9.46	12.20	7.01	-4.25	0.36
PhysNet	<b>5.67</b>	<b>9.02</b>	<b>4.28</b>	<b>2.02</b>	<b>0.55</b>
PhysFormer	8.33	11.01	6.42	-0.39	0.38

## 5.2. Neural Network models

**Quantitative evaluation:** The table 4 presents the HR performance for various supervised NN models evaluated on the MR-NIRP dataset. PhysNet stands out as the best overall performer, demonstrating the lowest errors and highest correlation and signal quality. TS-CAN and PhysFormer also show strong performance with relatively low errors and good correlation, making them reliable for HR estimation. However, DeepPhys and EfficientPhys-C exhibit moderate performance, with higher errors and lower correlation compared to PhysNet. These evaluations highlight PhysNet as the most effective supervised NN model for HR estimation on the MR-NIRP car dataset.

**Qualitative evaluation:** The table 3 illustrates the HR performance for various supervised NN methods evaluated under different automotive conditions in the MR-NIRP dataset. PhysNet consistently outperforms others across all conditions, achieving the lowest errors and highest correlations and SNR values. PhysFormer also shows strong performance, particularly in driving conditions and in garage settings with large motion, achieving relatively low errors and high correlations. TS-CAN performs well in stable conditions like garage small motion and still, but is less effective in more dynamic scenarios. DeepPhys and EfficientPhys-C exhibit moderate performance overall, showing some strengths in garage small motion and still conditions, but with higher errors and lower correlations compared to PhysNet and PhysFormer.

## 5.3. Respiration Rate estimation

This section outlines the performance results for RR estimation using both unsupervised methods and supervised NN models. Similar to the HR analysis, we conducted both quantitative and qualitative evaluations using the MR-NIRP car subsets.

### 5.3.1 Unsupervised methods

**Quantitative evaluation:** The table 6 presents the RR performance for various unsupervised methods evaluated on the MR-NIRP dataset. ICA consistently outperforms others across all metrics, achieving the lowest errors and highest correlations. POS and CHROM demonstrate moderate performance, with POS slightly outperforming CHROM, both methods show relatively lower errors and better correlation. In contrast, GREEN, LGI, and PBV consistently underperform, showing higher errors and lower correlations and signal quality. This indicates that these methods are less effective in predicting RR accurately.

**Qualitative evaluation:** The table 5 shows the RR performance for various unsupervised methods evaluated under different conditions on the MR-NIRP subsets. ICA outperforms others across almost all conditions in both driving and garage settings, achieving the lowest errors and highest SNR values, although its correlation  $\rho$  is often negative or weakly positive. POS demonstrates moderate performance with competitive errors  $\rho$  values, though it falls short of ICA's performance. CHROM consistently shows higher errors and lower correlations, indicating poorer performance compared to ICA and POS. In contrast, GREEN, LGI, and PBV underperform across all conditions, with significantly higher errors, lower correlations, and poor SNR values.

### 5.3.2 Neural Network models

**Quantitative evaluation:** The table 8 presents the RR performance for various supervised NN models evaluated on the MR-NIRP dataset. PhysNet stands out as the best

Table 5. Qualitative RR Performance Evaluation for the Unsupervised Methods on the MR-NIRP Subsets.

Unsupervised Methods - RR - MR-NIRP															
Method:	Driving Large Motion					Driving Small Motion					Driving Still				
	MAPE↓	RMSE↓	MAE↓	SNR↑	$\rho \uparrow$	MAPE↓	RMSE↓	MAE↓	SNR↑	$\rho \uparrow$	MAPE↓	RMSE↓	MAE↓	SNR↑	$\rho \uparrow$
ICA	<b>31.19</b>	<b>6.29</b>	<b>3.20</b>	<b>20.13</b>	-0.08	<b>27.50</b>	<b>5.05</b>	<b>2.59</b>	<b>21.16</b>	<b>0.16</b>	<b>22.90</b>	<b>4.20</b>	<b>2.08</b>	<b>22.24</b>	<b>0.07</b>
POS	50.07	8.15	4.24	16.34	<b>0.05</b>	59.57	8.21	4.33	16.81	0.12	65.51	9.04	4.79	16.60	-0.01
CHROM	85.41	10.80	6.45	12.67	-0.03	86.46	10.15	5.86	13.54	0.07	109.32	11.39	7.25	12.78	-0.12
GREEN	215.38	15.85	13.58	1.40	-0.09	218.05	15.81	13.72	0.43	-0.04	209.30	15.60	13.20	2.84	-0.01
LGI	210.18	15.55	13.37	0.51	-0.05	203.44	15.30	13.05	0.07	-0.04	204.57	15.10	12.92	1.96	0.00
PBV	209.13	15.30	13.13	1.13	-0.04	223.91	16.07	14.08	0.31	-0.05	191.03	14.73	12.13	2.57	0.06
Unsupervised Methods - RR - Garage															
Method:	Garage Large Motion					Garage Small Motion					Garage Still				
	MAPE↓	RMSE↓	MAE↓	SNR↑	$\rho \uparrow$	MAPE↓	RMSE↓	MAE↓	SNR↑	$\rho \uparrow$	MAPE↓	RMSE↓	MAE↓	SNR↑	$\rho \uparrow$
ICA	<b>28.47</b>	<b>5.50</b>	2.94	<b>21.63</b>	-0.19	<b>27.22</b>	<b>4.96</b>	<b>2.54</b>	<b>21.41</b>	<b>0.01</b>	<b>25.41</b>	<b>3.43</b>	<b>1.89</b>	<b>23.33</b>	-0.05
POS	30.04	5.99	<b>3.05</b>	19.19	-0.03	30.90	5.92	2.79	20.99	-0.04	30.58	4.28	2.22	22.37	<b>-0.02</b>
CHROM	70.10	9.30	5.44	15.97	<b>0.11</b>	39.45	6.96	3.39	18.15	-0.01	42.93	5.72	3.05	19.42	-0.03
GREEN	219.71	16.00	14.06	1.59	-0.05	141.86	12.80	9.37	8.30	-0.13	92.18	9.95	6.04	13.26	-0.05
LGI	191.88	14.88	12.55	3.71	-0.05	111.93	11.28	7.57	9.95	0.01	61.15	7.48	3.98	17.35	-0.05
PBV	206.93	15.72	13.60	2.04	-0.23	132.74	12.07	8.65	10.69	-0.07	77.84	8.92	5.01	16.49	-0.09

Table 6. Overall RR Performance Evaluation for the Unsupervised Methods on the MR-NIRP Dataset.

Unsupervised Methods - RR					
Method:	MAPE↓	RMSE↓	MAE↓	SNR↑	$\rho \uparrow$
ICA	<b>26.45</b>	<b>4.86</b>	<b>2.45</b>	<b>21.69</b>	<b>0.04</b>
POS	53.16	7.88	4.03	18.02	0.02
CHROM	77.65	9.60	5.50	14.90	-0.01
GREEN	185.25	14.62	11.81	4.26	-0.04
LGI	168.96	13.80	10.87	4.82	-0.01
PBV	176.63	14.15	11.26	4.97	-0.03

overall performer, achieving the lowest MAE, RMSE, and MAPE values, indicating the smallest errors, as well as the highest  $\rho$  and SNR, suggesting the strongest correlation and best signal quality. PhysFormer also shows strong performance with relatively low errors and decent correlation and signal quality, making it a reliable model for RR estimation. TS-CAN and EfficientPhys-C demonstrate moderate performance, with lower errors compared to DeepPhys but falling short of PhysNet and PhysFormer. They exhibit reasonable accuracy and signal quality. DeepPhys, however, is the least effective model, with the highest errors and lowest correlation and signal quality, suggesting major potential for progress.

**Qualitative evaluation:** The table 7 outlines the RR performance for various supervised NN models evaluated under different conditions on the MR-NIRP subsets. PhysNet consistently outperforms others across all conditions, achieving the lowest MAE, RMSE, and MAPE values, along with relatively high  $\rho$  and SNR values, indicating its outperformance in accuracy and signal quality in both dynamic and stable environments. PhysFormer also demonstrates strong performance, particularly in garage settings, with relatively low errors and decent  $\rho$  and SNR values, making it a reliable model for RR estimation in various conditions. TS-CAN and EfficientPhys-C exhibit moderate per-

formance, performing reasonably well in stable settings like garage still motion but showing degraded performance in more dynamic scenarios such as driving with large motion. DeepPhys generally exhibits the highest errors and lowest correlation and signal quality across most conditions, exposing considerable gaps for improvement.

#### 5.4. Cross-Dataset testing

We evaluated NN models trained on PURE, UBFC-RPPG, and SCAMPS and tested on MR-NIRP, revealing significant performance declines compared to models trained and tested on MR-NIRP for HR estimation, as illustrated in Figure 4. Models like DeepPhys, TS-CAN, and EfficientPhys-C showed moderate adaptability in controlled environments but struggled with dynamic lighting and motion, reflecting the limitations of indoor datasets in capturing vehicular complexities. Surprisingly, PhysNet and PhysFormer, which excelled in MR-NIRP-specific training, performed poorly across datasets, likely due to overfitting to their training environments. These findings highlight the need for more diverse and representative datasets to improve model generalization in automotive settings.

#### 5.5. Discussion and Future Work

In our experiments, we evaluated unsupervised and supervised NN models for HR and RR estimation in a vehicular setting using RGB videos. The results, shown in Tables 1-4 and Figure 3 in Appendix, indicate that NN models performed well in the dynamic vehicle environment, generally surpassing conventional unsupervised methods in HR estimation under both ideal and challenging conditions. This superior performance is attributed to NN models' ability to learn complex patterns and adapt to varying conditions, crucial in a moving vehicle.

PhysNet showed exceptional accuracy across all conditions, with notable resistance to motion artifacts. Although all models experienced some decline under varying lighting

Table 7. Qualitative RR Performance Evaluation for the Supervised NN Methods on the MR-NIRP Subsets.

NN Models - RR - MR-NIRP															
NN Model:	Driving Large Motion					Driving Small Motion					Driving Still				
	MAPE↓	RMSE↓	MAE↓	SNR↑	ρ↑	MAPE↓	RMSE↓	MAE↓	SNR↑	ρ↑	MAPE↓	RMSE↓	MAE↓	SNR↑	ρ↑
DeepPhys	166.99	13.56	10.48	4.42	0.14	191.16	14.58	12.30	3.72	0.05	186.00	13.96	11.42	4.80	0.04
TS-CAN	198.33	14.18	12.16	5.87	-0.01	202.08	15.01	12.94	3.31	-0.17	181.08	14.31	11.39	7.58	-0.13
EfficientPhys-C	211.18	15.12	13.18	4.64	0.08	187.24	14.29	11.85	5.05	0.01	171.87	13.66	10.78	6.43	-0.16
PhysNet	<b>44.99</b>	<b>7.71</b>	<b>4.09</b>	<b>14.08</b>	<b>0.03</b>	<b>54.83</b>	<b>8.11</b>	<b>4.42</b>	<b>13.93</b>	<b>-0.03</b>	<b>36.39</b>	<b>5.90</b>	<b>2.95</b>	<b>16.80</b>	<b>0.40</b>
PhysFormer	73.63	8.77	5.36	12.99	-0.01	60.39	8.28	4.93	13.41	-0.02	55.40	7.59	4.22	15.92	-0.04
NN Models - RR - Garage															
NN Model:	Garage Large Motion					Garage Small Motion					Garage Still				
	MAPE↓	RMSE↓	MAE↓	SNR↑	ρ↑	MAPE↓	RMSE↓	MAE↓	SNR↑	ρ↑	MAPE↓	RMSE↓	MAE↓	SNR↑	ρ↑
DeepPhys	211.11	16.35	13.77	6.06	-0.49	194.79	14.09	11.50	5.57	-0.09	79.03	8.30	5.20	12.50	0.23
TS-CAN	115.00	10.61	8.06	9.27	0.00	202.99	15.02	12.30	5.86	0.04	59.31	7.05	3.66	15.74	0.04
EfficientPhys-C	167.22	14.22	10.99	4.79	-0.33	163.82	13.13	9.89	5.89	0.05	77.15	9.35	4.61	14.49	-0.05
PhysNet	<b>24.04</b>	<b>3.45</b>	<b>1.79</b>	<b>17.45</b>	<b>0.07</b>	<b>34.58</b>	<b>4.03</b>	<b>2.35</b>	<b>19.25</b>	<b>0.19</b>	<b>18.77</b>	<b>2.88</b>	<b>1.25</b>	<b>22.50</b>	<b>0.40</b>
PhysFormer	68.49	8.10	4.81	15.27	-0.08	45.92	6.39	3.43	18.43	-0.11	20.72	2.92	1.44	21.43	0.39

Table 8. Overall RR performance metrics of all the supervised NN models on the MR-NIRP dataset.

NN Models - RR					
NN Model:	MAPE↓	RMSE↓	MAE↓	SNR↑	ρ↑
DeepPhys	175.18	13.70	10.95	5.23	0.07
TS-CAN	164.64	13.40	10.37	6.11	0.04
EfficientPhys-C	166.78	13.52	10.51	5.81	0.03
PhysNet	<b>39.39</b>	<b>6.36</b>	<b>3.12</b>	<b>16.84</b>	<b>0.15</b>
PhysFormer	54.09	7.49	4.10	15.95	-0.01

conditions, PhysNet remained the most resilient, maintaining higher accuracy. PhysFormer also performed well, accurately capturing the scale of the ground-truth data, likely due to its transformer-based architecture and specialized loss function. While this feature isn't critical for HR or RR estimations, it may benefit other rPPG signal processing applications.

For RR estimation, results indicated room for improvement. Despite producing signals with stronger correlations to the ground truth than unsupervised methods, even the best NN models, PhysNet and PhysFormer, did not surpass unsupervised methods in error metrics, as illustrated in Tables 5-8, and in Figure 4 in Appendix, ICA outperforms with RMSE the NN models in terms of RMSE under driving conditions. PhysNet was the most robust in a garage setting but struggled with lighting variations. The challenges of RR detection, due to the narrow and low-frequency range of the respiration signal, contributed to generally low performance across all models, with low SNR making it difficult to separate the signal from noise. Future research should address the limitations of the MR-NIRP car dataset by including more diverse in-vehicle data and incorporating advanced data augmentation and meta-learning techniques. Expanding the scope to include NIR recordings alongside RGB could offer a more comprehensive benchmark, as both methods have unique strengths in detecting vital signs. Regular updates to the benchmark are also essential to incorporate the latest NN-based advancements, ensuring it remains

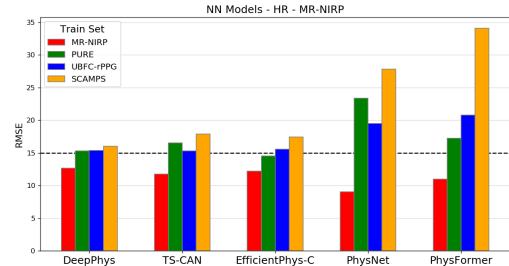


Figure 4. Cross-dataset RMSE results for HR estimation, comparing NN models trained on PURE, UBFC, SCAMPS and evaluated on MR-NIRP car dataset.

a relevant and reliable resource for the community.

## 6. Conclusion

This paper presents a comprehensive benchmark study on non-contact HR and RR monitoring in automotive settings using the MR-NIRP car dataset. Through extensive experimental evaluation, we compared the performance of various supervised and unsupervised rPPG algorithms. The results demonstrate that supervised NN-based algorithms generally outperform their unsupervised counterparts in both HR and RR estimation. However, the findings also indicate that there is room for further improvement, particularly in scenarios involving significant motion and the challenges posed by real-world driving environments. These insights underscore the potential for advancing rPPG technologies to enhance the reliability of non-invasive vital sign monitoring in vehicles.

## Acknowledgments

This research is supported by the SAFER funded pre-study "ViDCoM" and by the research project 'SUNRISE', which has received funding from the European Union's Horizon 2020 Research & Innovation Actions under grant agreement No. 101069573.

## References

- [1] WHO. Global status report on road safety. Available: <https://apps.who.int/iris/bitstream/handle/10665/277370/WHO-NMH-NVI-18.20-eng.pdf?ua=1>, 2018. 1
- [2] Paulson Eberechukwu Numan, Hyunwoo Park, Jaebok Lee, and Sunwoo Kim. Machine learning-based joint vital signs and occupancy detection with ir-uwb sensor. *IEEE Sensors Journal*, 23(7):7475–7482, 2023. 1
- [3] Michaela Sidikova, Radek Martinek, Aleksandra Kawala-Sterniuk, Martina Ldrova, Rene Jaros, Lukas Danys, and Petr Simonik. Vital sign monitoring in car seats based on electrocardiography, ballistocardiography and seismocardiography: A review. *Sensors*, 20(19), 2020. 1
- [4] Shahina Begum. Intelligent driver monitoring systems based on physiological sensor signals: A review. In *6th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013), The Hague, Netherlands*, pages 282–289, 2013. 1
- [5] E. M. Nowara, T. K. Marks, H. Mansour, and A. Veeraraghavan. Near-infrared imaging photoplethysmography during driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(4):3589–3600, 2020. 1, 2, 3
- [6] Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 18(10):10762–10774, 2010. 1
- [7] Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, 58(1):7–11, 2010. 1, 3
- [8] Xin Liu, Girish Narayanswamy, Akshay Paruchuri, Xiaoyu Zhang, Jiankai Tang, Yuzhe Zhang, Roni Sengupta, Shwetak Patel, Yuntao Wang, and Daniel McDuff. rppg-toolbox: Deep remote ppg toolbox. In *Advances in Neural Information Processing Systems, Vancouver, Canada*, volume 34, 2024. 1, 2, 4, 5
- [9] Ronny Stricker, Steffen Müller, and Horst-Michael Gross. Non-contact video-based pulse rate measurement on a mobile service robot. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 1056–1062. IEEE, 2014. 2, 3
- [10] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124:82–90, 2019. 2, 3
- [11] Zheng Gong, Xuezhi Yang, Rencheng Song, Xuesong Han, Chong Ren, Hailin Shi, Jianwei Niu, and Wei Li. Heart rate estimation in driver monitoring system using quality-guided spectrum peak screening. *IEEE Transactions on Instrumentation and Measurement*, DOI: 10.1109/TIM.2024.3352710, 73, 2024. 2
- [12] Kaiwen Guo, Tianqu Zhai, Elton Pashollari, Christopher J. Varlamos, Aymaan Ahmed, and Mohammed N. Islam. Contactless vital sign monitoring system for heart and respiratory rate measurements with motion compensation using a near-infrared time-of-flight camera. *Applied Sciences*, 11(22):10913, 2021. 2
- [13] Ming Xu, Guang Zeng, Yongjun Song, Yue Cao, Zeyi Liu, and Xiao He. Ivrr-ppg: An illumination variation robust remote-ppg algorithm for monitoring heart rate of drivers. *IEEE Transactions on Instrumentation and Measurement*, DOI: 10.1109/TIM.2023.3271760, 72, 2023. 2
- [14] Armand Comas, Tim K. Marks, Hassan Mansour, Suhas Lohit, Yechi Ma, and Xiaoming Liu. Turnip: Time-series u-net with recurrence for nir imaging ppg. In *IEEE International Conference on Image Processing (ICIP), Anchorage-Alaska*, 2021. 2
- [15] Gideon John and Simon Stent. The wayto myheart is through contrastive learning: Remote photoplethysmography from unlabelled video. In *Proceedings of the IEEE/CVF international conference on computer vision, Montreal, BC, Canada*, 2021. 2
- [16] Po-Wei Huang, Bing-Jhang Wu, and Bing-Fei Wu. A heart rate monitoring framework for real-world drivers using remote photoplethysmography. *IEEE journal of biomedical and health informatics*, 25(5):1397–1408, 2020. 2
- [17] Li-Wen Chiu, Yang-Ren Chou, Yi-Chiao Wu, and Bing-Fei Wu. Deep-learning based remote photoplethysmography measurement in driving scenarios with color and near-infrared images. *IEEE Transactions on Instrumentation and Measurement*, DOI: 10.1109/TIM.2023.3328703, 72, 2023. 2
- [18] Bing-Fei Wu, Yun-Wei Chu, Po-Wei Huang, Meng-Liang Chung, and Tzu-Min Lin. A motion robust remote-ppg approach to driver’s health state monitoring. *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, Revised Selected Papers*, Springer International Publishing, 13:463–476, 2017. 2
- [19] Bing-Fei Wu, Yun-Wei Chu, Po-Wei Huang, and Meng-Liang Chung. Neural network based luminance variation resistant remote-photoplethysmography for driver’s heart rate monitoring. *IEEE Access*, 7:57210–57225, 2019. 2
- [20] Bo-Kyeong Kim Park, Soyeon and Suh-Yeon Dong. Self-supervised rgb-nir fusion video vision transformer framework for rppg estimation. *IEEE Transactions on Instrumentation and Measurement*, DOI: 10.1109/TIM.2022.3217867, 71:1–10, 2022. 2
- [21] Zhiyu Wang, Xuezhi Yang, Hongzhou Lu, Caifeng Shan, and Wenjin Wang. Benchmark of physiological model based and deep learning based remote photoplethysmography in automotive applications. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece*, 2023. 3
- [22] Wenjin Wang, Brinker Albertus C. den, Stuijk Sander, and Gerard de Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, DOI: 10.1109/TBME.2016.2609282, 64(7):1479–1491, 2016. 3
- [23] Gerard De Haan and Jeanne Vincent. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013. 3

- [24] Pilz Christian S, Zaunseder Sebastian, Krajewski Jarek, and Blazek Vladimir. Local group invariance for heart rate estimation from face videos in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Salt Lake City, UT, USA, pages 1254–1262, 2018. 3
- [25] Wim Verkruyse, Svaasand Lars O., and Nelson J. Stuart. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008. 3
- [26] Weixuan Chen and McDuff Daniel. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the european conference on computer vision (ECCV)*, Munich, Germany, pages 349–365, 2018. 3
- [27] Xin Liu, Hill Brian, Jiang Ziheng, Patel Shwetak, and McDuff Daniel. Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, Waikoloa, Hawaii, pages 5008–5017, 2023. 3
- [28] Xin Liu, Fromm Josh, Patel Shwetak, and McDuff Daniel. Multi-task temporal shift attention networks for on-device contactless vitals measurement. In *Advances in Neural Information Processing Systems*, virtual, pages 19400–19411, 2020. 3, 4, 5
- [29] Zitong Yu, Li Xiaobai, and Zhao Guoying. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. *arXiv preprint arXiv:1905.02419*, 2019. 3
- [30] Zitong Yu, Shen Yuming, Shi Jingang, Zhao Hengshuang, Torr Philip HS, and Zhao Guoying. Physformer: Facial video-based physiological measurement with temporal difference transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, New Orleans, LA, USA, pages 4186–4196, 2022. 3, 4
- [31] Jiankang Deng, Guo Jia, Ververas Evangelos, Kotsia Irene, and Zafeiriou Stefanos. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Seattle, WA, USA, pages 5203–5212, 2022. 4
- [32] C Lugaresi, J Tang, H Nash, C McClanahan, E Ubweja, M Hays, F Zhang, CL Chang, MG Yong, J Lee, and WT Chang. Mediapipe: a framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, page 1906, 2019. 4
- [33] Claude Elwood Shannon. Communication in the presence of noise. *Journal of the Institute of Radio Engineers*, 37(1):10–21, 1949. 4