

# Feature-Disentangling RGB-NIR Fusion Network for Remote Driver Physiological Measurement

Tayssir Bouraffa\*      Ziyuan Wang\*  
Daniel Strüber  
Chalmers University of Technology  
Gothenburg, Sweden

tayssir.bouraffa@volvocars.com, wangzi@chalmers.se, danstru@chalmers.se

## Abstract

*Remote photoplethysmography (rPPG) is a crucial technique for non-contact heart rate (HR) estimation using facial videos, gaining significance in driver monitoring systems where contact-based measurements are impractical. Existing rPPG methods often rely on either RGB or NIR data, each susceptible to limitations under motion artifacts and varying illumination in real-world driving scenarios. To address these challenges, we introduce a novel RGB-NIR fusion model tailored for robust rPPG and HR estimation in dynamic vehicle environments. Our approach features two main contributions, an NIR-specific decoder that facilitates effective cross-modal knowledge transfer from RGB to NIR, enhancing model adaptability, and a dual autoencoder architecture for efficient feature disentanglement and reconstruction, mitigating noise from driver motion and changing lighting conditions. Comprehensive evaluations, including inter- and cross-dataset testing and ablation studies across various driving and garage conditions, demonstrate that our model achieves superior performance on the MR-NIRP car dataset, showcasing significant robustness in complex vehicular environments.*

## 1. Introduction

Each year, traffic accidents result in the loss of over 1 million lives globally, with an additional 50 million individuals sustaining injuries. Alarmingly, the number of traffic fatalities worldwide increased from 1.15 million in 2000 to 1.35 million in 2016 [36], highlighting an urgent need for improved road safety measures. Driving requires a high level of attention, as it demands both performance and awareness of external events. Consequently, enhancing driver monitoring systems to assess attention levels, addressing factors such as drowsiness, distraction, alcohol impairment,

and sudden illness, has become a top priority for automotive OEMs in efforts to reduce traffic fatalities.

Traditional approaches for monitoring sudden illness in drivers typically involve contact-based physiological measurement devices, such as pulse oximeters and electrocardiography (ECG) [32], or radar-based methods [26]. These techniques require direct contact between sensors and the driver’s skin, which can be uncomfortable and inconvenient. Additionally, these devices may distract drivers while they operate the vehicle. Although the use of high-frequency acoustic signals offers a novel alternative, it can cause discomfort for both drivers and sensitive passengers, such as infants and pets [26].

Non-contact driver monitoring systems (DMS) using on-board cameras are gaining popularity for detecting driver distraction and drowsiness. These cameras can also perform physiological measurements, such as remote photoplethysmography (rPPG) [3] [25]. rPPG is a non-invasive technique that uses optical methods to capture the pulsatile signal by detecting subtle color changes in the facial skin caused by blood volume fluctuations during each cardiac cycle. This enables the monitoring of vital signs, including heart rate (HR) [29], respiration rate (RR) [30], heart rate variability (HRV) [30], and blood pressure [2]. By continuously tracking these physiological parameters, the DMS can detect signs of distraction, drowsiness, and other physiological changes indicative of sudden illness, which could significantly affect driving performance.

Deep learning (DL)-based rPPG algorithms have demonstrated promising results in detecting vital signs within controlled laboratory environments, suggesting potential applications in driving scenarios [20]. However, unique challenges emerge in vehicles due to their dynamic nature. One major source of noise is the uneven distribution of ambient light, caused by environmental factors such as buildings, trees, and streetlights, which results in fluctuating facial illumination. Additionally, motion artifacts, arising from

---

\*Equal contribution.

vehicle movement, driver actions, and engine vibrations, further complicate accurate rPPG estimation [25]. These factors significantly impact the precision of rPPG measurements. Moreover, deploying DMS brings additional challenges, including model implementation and maintaining real-time performance. Overcoming these obstacles is crucial to ensuring driver safety, especially in cases where sudden illness could impair driving capabilities.

DL models for rPPG and HR estimation primarily rely on RGB facial videos as their primary input source. While effective, these models often struggle under complex real-world vehicle conditions, such as head movements and varying lighting [13]. To address these challenges, researchers have explored NIR imaging [25] [11] [37], which provides clearer visuals in low-light conditions but lacks accuracy in well-lit environments due to blood’s light absorption properties. This has led to the exploration of RGB-NIR fusion methods that combine the strengths of both modalities to enhance rPPG reliability [27] [6]. By leveraging RGB and NIR data, these methods improve HR estimation accuracy and robustness, making them better suited for diverse, real-world applications. However, current models often rely on basic fusion strategies that fail to separate physiological signals from noise, such as motion artifacts and lighting changes, and underutilize the potential for cross-modal knowledge transfer between RGB and NIR data, resulting in suboptimal feature integration. Additionally, many models overlook representations like Spatio-Temporal Maps (STMaps), which capture comprehensive signal features and enhance robustness in dynamic environments. These limitations highlight the need for advanced architectures that can better handle noise, improve feature representation, and optimize the strengths of RGB and NIR data for reliable rPPG and HR estimation.

Unlike CVD [24], which disentangles physiology/noise *within* RGB by swapping features between two RGB clips, our network enforces *cross-spectral* constraints by decoding NIR from *RGB-encoded* features ( $D_{NIR}$ ) and imposing pseudo-feature consistency across  $RGB \leftrightarrow NIR$ , changing both the information path and the training objective rather than merely adding NIR as an input. We introduce an RGB-NIR fusion model with two architectural innovations, an NIR-specific decoder for cross-modal reconstruction and an NIR-specific autoencoder that extends the CVD consistency across spectra, together with a new RGB representation (CPG-STMap).

- We generalize CVD from intra-RGB to *cross-spectral* disentanglement by decoding NIR from *RGB-encoded* features ( $D_{NIR}$ ) and enforcing pseudo-feature consistency across  $RGB \leftrightarrow NIR$ .
- We explore STMap variants and introduce **CPG-STMap** (CHROM/POS/G), which enhances pulse-bearing structure for improved rPPG prediction under motion/lighting

changes.

- We extensively evaluated inter- and cross-dataset testing as well as ablation studies across various driving scenarios, showing consistent improvements over state-of-the-art methods in challenging vehicular conditions.

The rest of this paper is organized as follows. Section 2 reviews related work on rPPG and HR estimation, focusing on methods using RGB, NIR, and multi-modal data fusion. Section 3 introduces the proposed RGB-NIR fusion model, including its architecture and methodology. The experimental results are detailed in Section 4. Finally, Section 5 concludes the paper.

## 2. Related Work

Several unsupervised and supervised methods for rPPG extraction from RGB images have been proposed for indoor settings. Traditional unsupervised approaches include: ICA for separating RGB signals into independent sources [31], POS, which projects a plane orthogonal to the skin tone for BVP estimation [35], CHROM, a chrominance-based method for extracting BVP through a combination of RGB signals [9], and GREEN, which leverages the green channel’s superior blood absorption contrast for improved signal extraction [34]. Techniques such as LGI, which ensures motion-invariant feature representation [28], and PBV, which isolates blood pulsations in RGB signals for accurate BVP estimation [8], have also been explored.

Supervised DL models have advanced rPPG extraction by incorporating CNNs with attention mechanisms for ROI and BVP extraction. Notable examples include: DeepPhys, a two-branch 2D attention network designed for HR estimation [5], TS-CAN, which uses a Temporal Shift Module for spatiotemporal information modeling [16], and EfficientPhys-C, a real-time single-branch CNN that integrates normalization and self-attention for efficient HR estimation [19]. RhythmNet employs a CNN-RNN hybrid for enhanced feature learning [23], while BigSmall utilizes dual-branch architectures to effectively model spatial-temporal features [22]. CVD strategy enhances the disentanglement of physiological signals from noise by using paired inputs and autoencoder-based RGB encoders for better signal isolation [24].

In vehicular environments, methods have been tailored to overcome specific challenges. Hernandez-Ortega et al. proposed Quality-Guided Spectrum Peak Screening for accurate HR estimation using NIR images [11]. Nowara et al. introduced AutoSparsePPG, enhancing rPPG signal estimation from narrow-band NIR recordings by addressing illumination changes and motion, and provided the MR-NIRP car dataset, a unique publicly available dataset featuring synchronized NIR/RGB videos and pulse oximeter data in vehicle setting [25]. Xu et al.’s Ivrr-PPG addressed illumination challenges by using NIR cameras for clearer rPPG

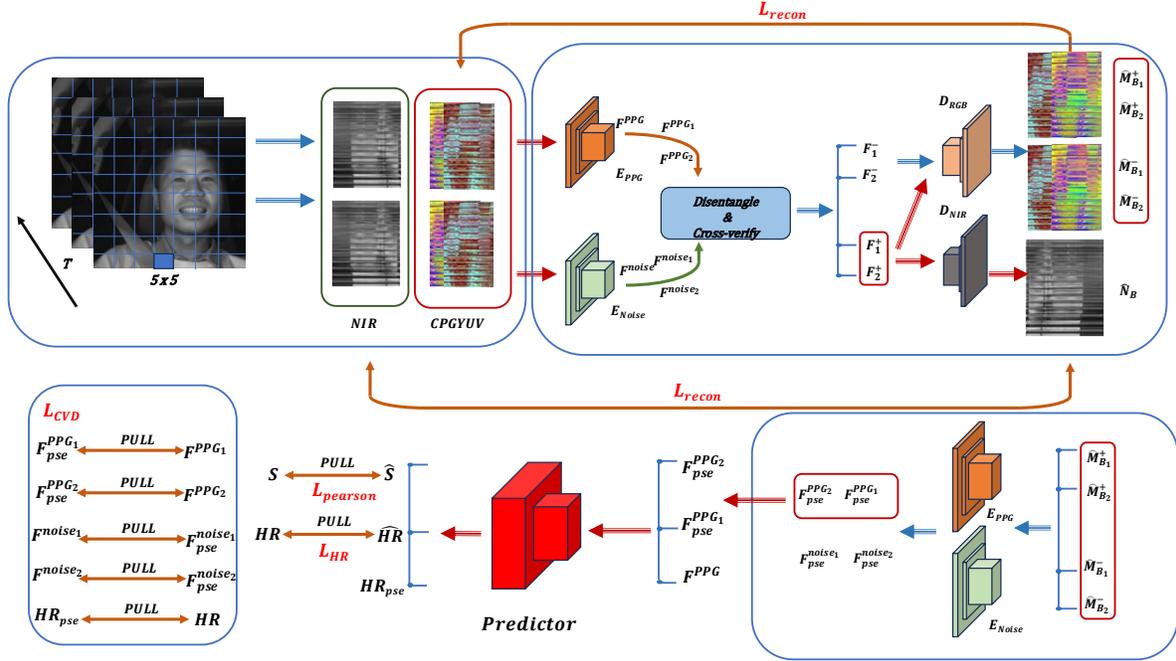


Figure 1. An overview of the feature disentangling RGB-NIR fusion network for robust rPPG and HR estimation.

signal capture [37], Guo et al. employed active infrared illumination and time-of-flight depth information to mitigate the impact of ambient light and motion artifacts [12].

Recent DL advancements have further pushed the boundaries of HR detection. Mitsubishi and Toyota have demonstrated the effectiveness of time-series U-net with GRU [7] and contrastive learning on unlabeled video data [14] using the MR-NIRP dataset. Huang et al. proposed a transformer-based model leveraging RGB data for handling motion and illumination artifacts [13]. Park et al. introduced a self-supervised transformer framework combining RGB and NIR for spatiotemporal feature extraction, utilizing contrastive learning for enhanced HR estimation [27]. Chiu et al. developed a CNN-based approach that applies an encoder-decoder structure to NIR and the CHROM method to RGB, effectively improving HR measurements under challenging conditions, such as head movement and changing illumination [6]. Unlike existing methods, the proposed RGB-NIR fusion model incorporates an NIR-specific decoder to enable cross-modal knowledge transfer and an NIR-focused autoencoder for effective feature disentanglement. This approach captures detailed spatio-temporal information for improved rPPG and HR estimation, ensuring robustness in diverse real-world driving conditions.

### 3. Methodology

In this section, we present the proposed feature disentangling RGB-NIR fusion model for HR and rPPG estimation.

As depicted in Figure 1, we detail the STMap representations and explain the cross-verification disentangling strategy employing two decoders. Lastly, we outline the modality fusion approach and the rPPG estimation process.

#### 3.1. Spatio-Temporal Representation

The Spatio-Temporal representation, known as STMap, was introduced to capture physiological signal information from the human face [23]. The STMap compresses video clips into a  $C$ -channel map derived from multiple facial regions of interest (ROIs) and it has gained traction in recent years for remote rPPG prediction [24] [21]. As a pre-processing step, the open-source face detector RetinaFace [10] was applied to the first frame of each video for precise face localization. Once detected, the face region was cropped using a bounding box and resized to  $72 \times 72$  pixels. Each frame was then divided into  $5 \times 5$  patches, resulting in 196 patches per frame, facilitating the identification of representative ROIs for physiological measurement.

To extract the physiological information from face videos, a video clip  $V \in \mathbb{R}^{T \times C \times H \times W}$  is transformed into a spatio-temporal map  $ST_t \in \mathbb{R}^{N \times T \times C}$ , where  $T$  represents the number of frames,  $C$  refers to the number of channels ( $C = 3$ , including R, G, B),  $H$  and  $W$  are height and width, and  $N$  denotes the number of ROIs. Like the approach in [21], we divide  $ST_t$  into non-overlapping slices  $ST_i$  of length  $T_0$ , where  $i \in [1, \lfloor \frac{T}{T_0} \rfloor]$ . Using interpolation, each slice  $ST_i$  is then resized to  $ST_i \in \mathbb{R}^{T_0 \times T_0 \times C}$ .

Following Liu et al. [21], we leveraged the CHROM

[9] and POS [35] methods to construct CHROM-STMap and POS-STMap, aimed at mitigating the effects of lighting variations and ambient noise. However, both CHROM-STMap and POS-STMap use identical  $R$ ,  $G$ ,  $U$  and  $V$  channels, resulting in redundancy when combined, as observed in [21]. To address this, we introduce a new STMap, termed CPG-STMap, retaining only the  $G$  channel due to its higher correlation with PPG signals compared to other channels. Additionally, given the effectiveness of the YUV color space for representing physiological information on the face [24], we transformed the RGB color space to YUV (Y, U, and V channels), constructing the YUV-STMap. For near-infrared images, we compute the NIR-STMap directly from the NIR stream, and represent it as  $(\mathcal{N}, \mathcal{N}, \text{and } \mathcal{N})$ , three identical channels, to match three-channel backbones. Thus, the five types of STMap are expressed as follows:

$$\begin{aligned} \text{CHROM-STMap} &= C(\mathcal{C}(\mathcal{R}, \mathcal{G}, \mathcal{B}), \mathcal{R}, \mathcal{G}), \\ \text{POS-STMap} &= C(\mathcal{P}(\mathcal{R}, \mathcal{G}, \mathcal{B}), \mathcal{R}, \mathcal{G}), \\ \text{CPG-STMap} &= C(\mathcal{C}(\mathcal{R}, \mathcal{G}, \mathcal{B}), \mathcal{P}(\mathcal{R}, \mathcal{G}, \mathcal{B}), \mathcal{G}), \quad (1) \\ \text{YUV-STMap} &= C(\mathcal{Y}, \mathcal{U}, \mathcal{V}), \\ \text{NIR-STMap} &= C(\mathcal{N}, \mathcal{N}, \mathcal{N}), \end{aligned}$$

where  $\mathcal{C}$  and  $\mathcal{P}$  denote the CHROM and POS algorithms, respectively,  $C(\cdot)$  indicates the concatenation operation.

### 3.2. Two-decoder Cross-Verified Model

The generated STMaps encode physiological information but also capture noise from factors like motion artifacts and lighting variations, which can interfere with accurate physiological signal prediction. To isolate physiological signals from non-physiological artifacts, we apply a cross-verified feature disentangling approach during model training [24]. This approach utilizes an auto-encoder with dual encoders,  $E_{PPG}$  and  $E_{Noise}$ . One is dedicated to learning physiological features and the other is to identifying non-physiological features. Pairwise STMaps,  $ST_i$  and  $ST_i^{NIR}$ , serve as input and reconstructed target, and the model is trained with constructed STMaps, where features are cross-decoded by swapping between different original STMaps.

A new separate decoder specifically for the NIR data is added. This NIR decoder is placed after the RGB encoder. The idea is that the RGB encoder processes the RGB STMap, representing spatial-temporal information from color channels, and the NIR decoder uses this encoded information to reconstruct an NIR-STMap. The purpose of this setup is to train the RGB encoder to capture invariant features useful for NIR-STMap prediction, enabling *knowledge transfer* from RGB to NIR [1]. This approach allows the encoder, trained on RGB, to learn features relevant across both modalities, effectively transitioning from a single modality (RGB) to multi-modalities (RGB + NIR) by leveraging shared features. A visualization example of

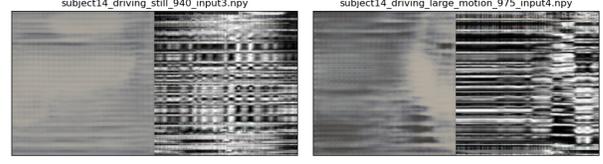


Figure 2. Visualization of reconstructed NIR-STMap (left) and ground truth map (right) examples from the MR-NIRP car dataset.

the reconstructed NIR-STMap is illustrated in Figure 2 for subjects 14 in driving scenarios with still and large motion conditions.

To predict the average heart rate  $HR_i$  and rPPG signal  $S_i$  over  $T_0$  time steps, by passing each input  $ST_i$  through the two encoders, we obtain separate feature representations for physiological and noise features:  $\mathcal{F}^{PPG_i} = E_{PPG}(ST_i)$  captures the rPPG-related features, while  $\mathcal{F}^{noise_i} = E_{Noise}(ST_i)$  isolates noise components. The  $ST_i$  input face videos  $\mathcal{B}$  are randomly split into two groups,  $\mathcal{B}_1$  and  $\mathcal{B}_2$ , each containing corresponding rPPG and noise features:  $\mathcal{F}^{PPG_1}$ ,  $\mathcal{F}^{noise_1}$  and  $\mathcal{F}^{PPG_2}$ ,  $\mathcal{F}^{noise_2}$ . These features are then combined to create true feature pairs of original STMap,  $\mathcal{F}^+$ , and manually constructed pairs with mismatched noise,  $\mathcal{F}^-$  presented, respectively:

$$\begin{aligned} \mathcal{F}_1^+ &= \mathcal{F}^{PPG_1} + \mathcal{F}^{noise_1}, \mathcal{F}_2^+ = \mathcal{F}^{PPG_2} + \mathcal{F}^{noise_2}, \\ \mathcal{F}_1^- &= \mathcal{F}^{PPG_1} + \mathcal{F}^{noise_2}, \mathcal{F}_2^- = \mathcal{F}^{PPG_2} + \mathcal{F}^{noise_1}. \end{aligned}$$

The RGB decoder reconstructs these features into maps  $\hat{M}_{\mathcal{B}_1}^+$ ,  $\hat{M}_{\mathcal{B}_2}^+$ ,  $\hat{M}_{\mathcal{B}_1}^-$ , and  $\hat{M}_{\mathcal{B}_2}^-$ . Additionally, the true features  $\mathcal{F}^+$  are sent to the NIR decoder to generate NIR-STMaps  $\hat{N}_{\mathcal{B}_1}$  and  $\hat{N}_{\mathcal{B}_2}$ , validating that the RGB encoder captures invariant features across RGB and NIR modalities for robust multi-modal representation.

The model’s predictor,  $\mathcal{P}$ , is designed to integrate both the average HR and rPPG signals, offering complementary supervision during training to facilitate the learning of more robust features. It takes as input  $\mathcal{F}^{PPG_i}$ , the features extracted by the rPPG encoders, and is optimized using  $L_{pred}$ . The predictor outputs both the predicted rPPG signal and the average HR.

**Reconstruction Loss:** To ensure that the decoder accurately reconstructs the STMap, both physiological and non-physiological features encoded from the same RGB-STMap, as well as physiological features encoded from the NIR-STMap, are used to reconstruct the original STMaps. The reconstruction loss is formulated as follows:

$$L_{recon} = \sum_{i=1}^2 |ST_{\mathcal{B}_i} - \hat{M}_{\mathcal{B}_i}^+| + \sum_{i=1}^2 |ST_{\mathcal{B}_i}^{NIR} - \hat{N}_{\mathcal{B}_i}| \quad (2)$$

where  $ST_{\mathcal{B}_i}$  and  $ST_{\mathcal{B}_i}^{NIR}$  denotes the inputs of RGB-STMap and NIR-STMap in  $\mathcal{B}_i$ , respectively.

**Prediction Loss:** This loss function consists of two components: one for the average HR branch and one for the rPPG branch. For the average HR, the  $L_1$  loss function is used to minimize the difference between the predicted average HR ( $\hat{H}R_i$ ) and the target average HR ( $HR_i$ ), formulated as follows:

$$L_{HR} = \frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} |\hat{H}R_i - HR_i| \quad (3)$$

For the prediction of rPPG signals, the negative Pearson correlation is applied to assess the alignment between the predicted and target rPPG signals.

$$\begin{aligned} L_{\text{Pearson}} &= \frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \left( 1 - \frac{\text{Cov}(\hat{S}_i, S_i)}{\sqrt{\text{Cov}(\hat{S}_i, \hat{S}_i) \cdot \text{Cov}(S_i, S_i)}} \right) \\ &= \frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \left( 1 - \frac{T_0 \sum_{t=1}^{T_0} \hat{S}_{i,t} S_{i,t} - \sum_{t=1}^{T_0} \hat{S}_{i,t} \sum_{t=1}^{T_0} S_{i,t}}{\sqrt{T_0} \sigma(S_i) \sigma(\hat{S}_i)} \right) \end{aligned} \quad (4)$$

where  $\sigma$  denotes the standard deviation, and  $\hat{S}_{i,t}$  represents the  $i^{\text{th}}$  predicted signal at time  $t$ .

**Frequency Domain Loss:** Given an HR range, the average HR can be estimated from rPPG branch features by computing the power spectral density ( $PSD$ ) of the predicted rPPG signals. The frequency with the highest power in the  $PSD$  represents the HR. This estimate is then optimized using cross-entropy loss  $CE$ , formulated as follows:

$$L_{CE} = CE(PSD(\hat{S}_i), HR_i) \quad (5)$$

where  $PSD(\hat{S}_i) \in \mathbb{R}^{\text{len}(HR^{\text{BPM}})}$  and its  $j^{\text{th}}$  entry of  $PSD(\hat{S}_i)$  is  $\sum_{t=1}^{T_0} |\hat{S}_{i,t} e^{i2\pi \frac{t}{T_0} \frac{\text{HR}_j^{\text{BPM}}}{60}}|^2$ ,  $\text{HR}_j^{\text{BPM}}$  refers to the  $j^{\text{th}}$  element in  $\text{HR}^{\text{BPM}}$ .

Therefore, the final prediction loss is expressed as a combination of these terms:

$$L_{\text{pred}} = L_{HR} + \lambda_{\text{Pearson}} L_{\text{Pearson}} + \lambda_{CE} L_{CE} \quad (6)$$

where  $\lambda_{\text{Pearson}}$  and  $\lambda_{CE} \in (0, 1)$  are two hyperparameters.

**Cross-verified Feature Disentangling (CVD) Loss:** To enhance the disentanglement between rPPG and noise representations, as described in [24],  $\hat{M}_{\mathcal{B}_1}^-$  and  $\hat{M}_{\mathcal{B}_2}^-$  are further processed through two separate encoders  $E_{\text{PPG}}$  and  $E_{\text{Noise}}$ , producing the pseudo features  $\mathcal{F}_{\text{pse}}^{\text{PPG}1}$ ,  $\mathcal{F}_{\text{pse}}^{\text{noise}2}$ ,  $\mathcal{F}_{\text{pse}}^{\text{PPG}2}$ ,  $\mathcal{F}_{\text{pse}}^{\text{noise}1}$ . Considering both physiological and non-physiological components, along with the predicted heart rates  $HR_1^{\text{pse}}$  and  $HR_2^{\text{pse}}$ , the  $CVD$  loss function is defined

Table 1. HR estimation results of the proposed method compared to several state-of-the-art unsupervised and supervised methods. All methods were retrained from scratch on the MR-NIRP car dataset under the same subject-exclusive 5-fold splits.

Method	MAE↓	RMSE↓	MAPE↓	$\rho$ ↑	SNR↑	Latency↑
<b>Unsupervised</b>						
ICA [31]	12.82	18.81	16.21	0.09	-10.91	—
POS [35]	10.64	16.55	14.22	0.25	-8.59	6,767
CHROM [9]	11.15	17.31	15.01	0.22	-9.00	10,957
GREEN [34]	14.31	20.05	17.94	0.06	-12.37	46,766
LGI [28]	12.55	18.99	15.68	0.11	-10.13	36,101
PBV [8]	13.77	19.56	17.50	0.08	-11.99	33,240
<b>Supervised</b>						
DeepPhys [5]	7.92	13.67	10.63	0.31	-8.52	16,833
EfficientPhys [19]	8.13	13.67	11.06	0.32	-4.25	46,991
TS-CAN [16]	8.08	13.72	11.02	0.30	-4.20	27,766
BigSmall [22]	16.70	19.12	22.22	0.01	—	17,557
RhythmNet [23]	9.17	11.10	12.56	0.07	—	<b>120,057</b>
Baseline (RGB+YUV) [24]	8.93	11.80	12.31	0.28	-4.90	31,311
PhysFormer [38]	<b>6.98</b>	11.65	<b>9.65</b>	0.32	<b>-2.41</b>	10,072
MetaPhys [17]	7.57	12.96	10.80	0.34	-4.42	35,937
$\mathcal{F}_{\text{REC}}$ (CPG+YUV+NIR)	7.51	<b>10.45</b>	10.35	<b>0.42</b>	-2.60	22,445

as follows:

$$\begin{aligned} L_{\text{feat}} &= \lambda_{\text{feat}} \left( \sum_{i=1}^2 |\mathcal{F}_{\text{pse}}^{\text{PPG}i} - \mathcal{F}^{\text{PPG}i}| \right. \\ &\quad \left. + \sum_{i=1}^2 |\mathcal{F}_{\text{pse}}^{\text{noise}i} - \mathcal{F}^{\text{noise}i}| \right) \end{aligned} \quad (7)$$

$$L_{HR_{\text{pse}}} = \sum_{i=1}^2 |HR_i^{\text{pse}} - HR_{\mathcal{B}_i}| \quad (8)$$

$$L_{CVD} = L_{\text{feat}} + L_{HR_{\text{pse}}} \quad (9)$$

where  $\mathcal{F}_{\text{pse}}^{\text{PPG}i}$  and  $\mathcal{F}_{\text{pse}}^{\text{noise}i}$  denote the pseudo-rPPG and pseudo-noise features of the RGB-STMap, and  $HR_{\mathcal{B}_i}$  denotes the target average HR in  $\mathcal{B}_i$ . Additionally, the true pseudo features  $\mathcal{F}_i^+$  are fed to the model's predictor  $\mathcal{P}$  to estimate rPPG and HR. The total loss of the model is then the sum of these three components:

$$L = \lambda_{\text{recon}} L_{\text{recon}} + L_{\text{pred}} + L_{CVD} \quad (10)$$

### 3.3. Modality Fusion Comparison

To argue the effect for reconstruction of NIR-STMap, we compare the model using modality fusion, by incorporating a separate encoder and decoder for the NIR-STMap, treating it as an additional input. This setup involves an auto-encoder structure for NIR-STMap, including two encoders,  $E_{\text{PPG}}^{\text{NIR}}$  and  $E_{\text{noise}}^{\text{NIR}}$ , and a decoder  $D^{\text{NIR}}$ , which extract rPPG and noise features  $\mathcal{F}_{\text{NIR}}^{\text{PPG}i}$  and  $\mathcal{F}_{\text{NIR}}^{\text{noise}i}$ , respectively. The same disentangled recipe is used for NIR-STMap. Then, fusion occurs by combining the rPPG features from the RGB-STMaps and NIR-STMaps ( $\mathcal{F}^{\text{PPG}i}$  and  $\mathcal{F}_{\text{NIR}}^{\text{PPG}i}$ ) into a unified representation,  $\mathcal{F}_{\text{FUSED}}^{\text{PPG}i}$ , which is used for

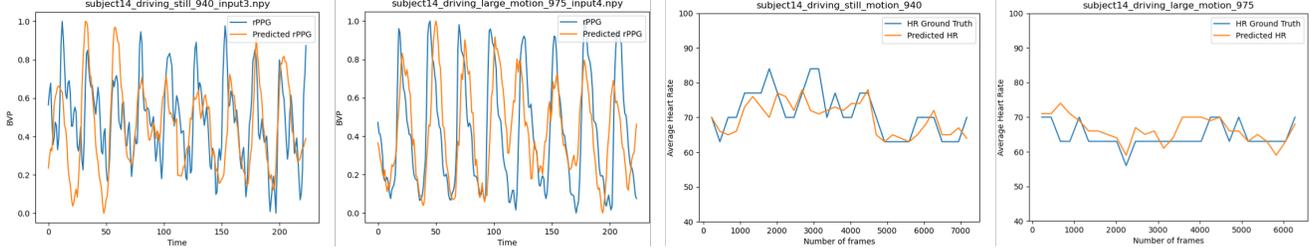


Figure 3. Examples of rPPG and HR predictions for driving scenarios under varying conditions.

Table 2. Case-specific results of the ablation study on MR-NIRP

Model	Driving Large Motion					Driving Small Motion					Driving Still				
	MAE↓	RMSE↓	MAPE↓	$\rho$ ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	$\rho$ ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	$\rho$ ↑	SNR↑
NIR	10.85	13.09	14.99	0.13	-7.13	10.19	12.55	14.02	0.06	-6.45	10.44	12.89	14.85	0.13	-6.99
RGB+YUV	11.03	13.93	14.98	0.05	-6.15	10.47	13.09	14.10	0.08	-6.19	9.89	12.83	13.89	0.17	-5.89
CPG+YUV	10.54	13.27	14.45	0.10	-6.02	8.92	11.40	12.00	0.24	-4.95	8.42	11.13	11.95	0.27	-4.05
$\mathcal{F}_{REC}$ (CPG+YUV+NIR)	<b>9.12</b>	<b>11.92</b>	<b>12.45</b>	<b>0.27</b>	<b>-5.85</b>	<b>8.15</b>	<b>10.88</b>	<b>10.86</b>	<b>0.34</b>	<b>-4.13</b>	<b>7.80</b>	<b>10.62</b>	<b>11.04</b>	<b>0.31</b>	<b>-3.74</b>
Model	Garage Large Motion					Garage Small Motion					Garage Still				
	MAE↓	RMSE↓	MAPE↓	$\rho$ ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	$\rho$ ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	$\rho$ ↑	SNR↑
NIR	9.87	12.34	13.21	0.24	-6.75	9.87	12.27	13.79	0.05	-6.57	8.18	10.74	11.47	0.32	-3.93
RGB+YUV	10.69	13.49	14.46	-0.01	-5.97	8.16	11.54	11.79	0.29	-3.86	5.61	8.03	8.00	0.74	-0.80
CPG+YUV	10.35	12.74	14.17	0.09	<b>-5.64</b>	6.91	9.08	9.49	0.52	-3.80	5.15	7.34	7.29	<b>0.79</b>	0.09
$\mathcal{F}_{REC}$ (CPG+YUV+NIR)	<b>9.09</b>	<b>11.57</b>	<b>12.32</b>	<b>0.21</b>	-5.65	<b>6.05</b>	<b>8.38</b>	<b>8.31</b>	<b>0.56</b>	<b>-0.67</b>	<b>4.88</b>	<b>7.32</b>	<b>6.96</b>	0.78	<b>1.43</b>

predicting average HR and rPPG signals. This unified representation is formed using either addition  $\mathcal{F}_{ADD}$  or concatenation  $\mathcal{F}_{CONC}$  of the two modalities to compare our reconstructed method  $\mathcal{F}_{REC}$  in the previous section.

For the loss functions, similar to the RGB-STMaps, reconstruction loss is computed for the NIR-STMap using the decoder  $D^{NIR}$ , with outputs  $\hat{N}_{B_1}^-$  and  $\hat{N}_{B_2}^-$ . The NIR-specific CVD loss is also calculated by comparing pseudo-rPPG and pseudo-noise features with the true NIR features:

$$\begin{aligned}
 L_{feat}^{NIR} = & \lambda_{feat} \left( \sum_{i=1}^2 |\mathcal{F}_{pseNIR}^{PPG_i} - \mathcal{F}_{NIR}^{PPG_i}| \right. \\
 & \left. + \sum_{i=1}^2 |\mathcal{F}_{pseNIR}^{noise_i} - \mathcal{F}_{NIR}^{noise_i}| \right), \quad (11)
 \end{aligned}$$

where  $\mathcal{F}_{pseNIR}^{PPG_i}$  and  $\mathcal{F}_{pseNIR}^{noise_i}$  denotes the pseudo-rPPG and pseudo-noise features of NIR-STMap in group  $i$ .

The total CVD loss is then changed as:

$$L_{CVD} = L_{feat} + L_{feat}^{NIR} + L_{HR_{pse}} \quad (12)$$

## 4. Experiments and Discussion

### 4.1. Experimental Setup

**Datasets:** In our experiments, we evaluated the proposed method using the MR-NIRP car dataset [25], PURE[33] and UBFC[4] using a five-fold subject-exclusive cross-validation technique. The MR-NIRP is the only publicly

accessible dataset specific to vehicular environments, while PURE and UBFC were recorded in controlled indoor settings. Our analysis covered scenarios with still, small, and large motions, conducted in both garage and driving contexts, details are provided in the Appendix A. All implementations were conducted using PyTorch on Nvidia A100 GPU. To train and evaluate the performance of various existing neural network models and unsupervised methods on the MR-NIRP, we rebuilt and extend the rPPG-Toolbox [18], an open-source repository for training and benchmarking NN-based and unsupervised rPPG algorithms.

**Training Setup:** To effectively assess the models' performance in vehicular environments, the MR-NIRP car dataset was used for both training and testing. All methods were *retrained from scratch on MR-NIRP car dataset to guarantee a direct, fair comparison across models*. To prevent overfitting and ensure unbiased predictions, distinct train-validation-test splits were implemented. Since the MR-NIRP dataset lacks predefined folds for training and testing, we use five-fold subject-exclusive cross-validation technique outlined by Gideon et al. [14]. As detailed in Table 3, the dataset was divided into 5 folds based on subject ID, with a different test set held out for each fold. All models were trained and evaluated on each fold and the average of the performance across all folds was reported.

**Model Training:** For all experiments, the average heart rate was computed over  $T_0 = 224$  frames, and extracted from PPG signals using Fast Fourier Transform (FFT). The spatio-temporal map included 196 ROIs. Hyperparameters

Table 3. Subject IDs allocated to training, validation, and test sets.

	Train Set	Validation Set	Test Set
<b>Fold 1</b>	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11	13, 14, 15	16, 17, 18, 19
<b>Fold 2</b>	5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16	17, 18, 19	1, 2, 3, 4
<b>Fold 3</b>	1, 9, 10, 11, 13, 14, 15, 16, 17, 18, 19	2, 3, 4	5, 6, 7, 8
<b>Fold 4</b>	1, 2, 3, 4, 5, 13, 14, 15, 16, 17, 18, 19	6, 7, 8	9, 10, 11
<b>Fold 5</b>	1, 2, 3, 4, 5, 6, 7, 8, 16, 17, 18, 19	9, 10, 11	13, 14, 15

were set as follows:  $\lambda_{\text{pearson}} = 20$ ,  $\lambda_{\text{CE}} = 10$ ,  $\lambda_{\text{recon}} = 50$ , and  $\lambda_{\text{feat}} = 10$ . The Adam optimizer [15] with an initial learning rate of  $5 \times 10^{-4}$  was adopted in the proposed model. The learning rate was scheduled with a multi-step decay strategy, reducing by a factor of 0.5 at specific milestones (epochs 15 and 25) to ensure stable convergence. The maximum number of training epochs was set to 70 for the MR-NIRP dataset, with early stopping applied after 10 epochs of no improvement. Our implementation and all the training configurations of the models are shared as open source and is available at [GitHub repository](#).

## 4.2. Intra-database Testing

In our experiments, we evaluated our method and compared it with several state-of-the-art methods.

Table 1. presents a comprehensive evaluation of our method against both unsupervised and supervised approaches for average HR estimation on the MR-NIRP car dataset. We compare performance across key metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), Pearson Correlation Coefficient ( $\rho$ ), Signal-to-Noise Ratio (SNR), and Frames Per Second (FPS), more details about the selected metrics are provided in Appendix C.

Our method attains the best RMSE (10.45 bpm) and the highest  $\rho=0.42$ , while maintaining competitive MAE (7.51 bpm) and MAPE (10.35). Compared to the strongest unsupervised baseline, POS [35] (MAE 10.64, RMSE 16.55), our errors are substantially lower. Among supervised baselines, DeepPhys [5] and EfficientPhys [19] have higher MAEs (7.92 and 8.13, respectively). PhysFormer [38] achieves a lower MAE/MAPE (6.98/9.65) but exhibits a higher RMSE (11.65) and a lower correlation ( $\rho=0.32$ ) than ours, indicating that our predictions are more consistent and better aligned with ground truth across clips/subjects. In SNR, PhysFormer is strongest ( $-2.41$  dB) and our model is comparable ( $-2.60$  dB).

Speed is *model-only* throughput on precomputed STMaps (FPS; excludes detection/cropping/STMap construction). Under this metric, our model reaches 22,445 FPS; EfficientPhys and RhythmNet attain 46,991 and 120,057 FPS, respectively, while exhibiting higher HR error under the same evaluation. This reflects a standard accuracy–efficiency trade-off: our approach prioritizes robustness/accuracy for safety-critical applications while main-

Table 4. Cross-Dataset Results on the UBFC and PURE Datasets

Model	UBFC			PURE		
	MAE↓	MAPE↓	$\rho$ ↑	MAE↓	MAPE↓	$\rho$ ↑
PhysFormer [38]	42.62	40.67	0.04	21.04	24.58	-0.02
RhythmNet [23]	32.67	30.94	-0.02	18.28	25.75	0.05
Baseline (RGB+YUV) [24]	23.48	21.77	0.19	20.62	30.60	0.05
CHROM+YUV	30.68	28.70	-0.03	18.81	23.89	0.02
$\mathcal{F}_{\text{REC}}$ (CHROM+YUV+NIR)	<b>20.17</b>	<b>18.51</b>	0.23	15.34	21.65	<b>0.34</b>
CPG+YUV	20.73	25.97	<b>0.24</b>	18.10	24.08	0.13
$\mathcal{F}_{\text{REC}}$ (CPG+YUV+NIR)	20.97	19.17	0.23	15.35	21.96	0.33
POS+YUV	21.14	19.57	0.22	16.55	22.13	0.04
$\mathcal{F}_{\text{REC}}$ (POS+YUV+NIR)	21.95	20.10	0.08	<b>14.72</b>	<b>20.33</b>	0.25

taining competitive throughput.

Figure 3 further illustrates our model’s performance under still and large-motion driving scenarios. As shown for Subject 14, our predicted signals closely align with the ground truth, capturing peak locations and amplitude variations even under challenging motion conditions. These results, together with the table analysis, validate the effectiveness of our RGB-NIR fusion and reconstruction framework for real-time HR monitoring. They confirm that our method offers a more accurate and robust solution for physiological signal estimation in dynamic real-world environments.

## 4.3. Cross-database Testing

The cross-dataset evaluation, where models are trained on the MR-NIRP and tested on the UBFC and PURE, demonstrates the strong generalization ability of our method, as summarized in Table 4. Notably, across most of the configurations, the introduction of the NIR modality leads to consistent and significant performance gains.

On the UBFC, our model  $\mathcal{F}_{\text{REC}}$ , when applied to the CPG+YUV+NIR configuration, achieves an MAE of 20.97, a MAPE of 19.17, and a Pearson correlation coefficient ( $\rho$ ) of 0.23, clearly outperforming both the RGB+YUV baseline and RhythmNet. Similar improvements are observed for other traditional pipelines: CHROM+YUV+NIR reduces MAE from 30.68 to 20.17 and increases  $\rho$  from  $-0.03$  to 0.23. On the PURE, the improvements brought by NIR are even more pronounced. The CPG+YUV+NIR configuration with  $\mathcal{F}_{\text{REC}}$  achieves an MAE of 15.35 and  $\rho = 0.33$ . Particularly, the POS+YUV+NIR combination obtains the lowest MAE of 14.72 among all models, demonstrating the benefit of NIR in capturing robust physiological signals under diverse real-world conditions.

## 4.4. Ablation study

**Loss functions:** To evaluate the contribution of each loss function, we conduct an ablation study by removing one loss term at a time during training. Table 5 summarizes the results on the MR-NIRP dataset. Removing the cross-entropy loss ( $L_{\text{CE}}$ ) leads to a slight performance degradation. However, excluding the Pearson correlation loss

Table 5. Ablation studies on muted loss functions.

Ablated Loss	MAE↓	RMSE↓	MAPE↓	$\rho$ ↑	SNR↑
$L_{CE}$	7.70	10.50	10.67	0.43	-3.23
$L_{Pearson}$	8.88	11.06	12.11	0.044	-11.91
$L_{HR}$	68.15	68.73	99.63	-0.002	-9.13
None	7.51	10.45	10.35	0.42	-2.60

Table 6. Overall HR estimation results of the ablation study.

Model	MAE↓	RMSE↓	MAPE↓	$\rho$ ↑	SNR↑
NIR	10.02	12.62	13.96	0.15	-6.39
CHROM+YUV	8.73	11.37	11.46	0.12	-6.72
POS + YUV	8.28	11.00	11.33	0.34	-4.26
RGB + YUV	9.34	12.35	12.84	0.23	-4.90
CPG+YUV	8.32	11.12	11.51	0.36	-4.02
$\mathcal{F}_{REC}$ (CHROM+YUV+NIR)	8.16	11.21	11.36	0.34	-2.92
$\mathcal{F}_{REC}$ (POS+YUV+NIR)	8.19	11.28	11.30	0.32	-3.07
$\mathcal{F}_{ADD}$ (CPG+YUV+NIR)	7.75	10.92	10.75	0.40	<b>-2.51</b>
$\mathcal{F}_{CONC}$ (CPG+YUV+NIR)	7.63	10.60	10.49	<b>0.44</b>	-2.59
$\mathcal{F}_{REC}$ (CPG+YUV+NIR)	<b>7.51</b>	<b>10.45</b>	<b>10.35</b>	0.42	-2.60

( $L_{Pearson}$ ) or the heart rate regression loss ( $L_{HR}$ ) significantly deteriorates model’s accuracy and correlation with ground truth. In particular, omitting  $L_{HR}$  results in a drastic increase in error metrics, indicating its critical role in guiding precise heart rate estimation. The model with all loss components achieves the best performance, demonstrating the necessity of each loss term in optimizing the model.

**Modality Fusion and Feature Combination:** An ablation study on the MR-NIRP car dataset assessed the impact of different fusion strategies and feature combinations, evaluating overall outcomes and case-specific scenarios.

**Overall HR Estimation Results:** Table 6 summarizes the overall HR estimation performance across five different STMap configurations and feature combinations. Our results highlight that the combination of CPG, YUV, and NIR modalities significantly improves model performance across most metrics. Specifically, the CPG+YUV+NIR configuration with  $\mathcal{F}_{REC}$  achieved the best overall results in terms of MAE (7.51 bpm), RMSE (10.45 bpm), and MAPE (10.35), while also demonstrating a high  $\rho$  of 0.42. Although  $\mathcal{F}_{CONC}$  attained the highest correlation coefficient ( $\rho = 0.44$ ),  $\mathcal{F}_{REC}$  maintained a competitive balance across all other metrics. The SNR of -2.60 in  $\mathcal{F}_{REC}$  indicates effective noise handling, solidifying its robustness for HR estimation in challenging conditions.

Using NIR configuration alone yielded the highest error rates and the lowest correlation, underscoring the limitations of single-modality inputs in capturing comprehensive physiological signals. The results also reveal that while addition and concatenation of CPG+YUV+NIR modalities performed similarly in terms of SNR, the concatenation approach provided superior overall accuracy. These results highlight the robustness of the reconstruction-based fusion modalities in capturing comprehensive physiological sig-

nals, particularly in complex vehicular environments.

Small gaps among  $\mathcal{F}_{ADD}/\mathcal{F}_{CONC}/\mathcal{F}_{REC}$  are expected since all heads read the same disentangled features, the primary gain stems from the cross-spectral constraint. We choose  $\mathcal{F}_{REC}$  for the best MAE/RMSE/MAPE trade-off with comparable SNR, while  $\mathcal{F}_{ADD}$  yields the highest SNR and  $\mathcal{F}_{CONC}$  the highest  $\rho$ .

**Case-specific HR Estimation Results:** To further assess model robustness under varying conditions, we evaluated different configurations across specific driving and garage scenarios, as shown in Table 2.

- **Driving Scenarios:** In high-motion driving scenarios, the  $\mathcal{F}_{REC}$  (CPG+YUV+NIR) model again showed superior performance, achieving the lowest MAE (9.12 bpm) and RMSE (11.92 bpm) in the large-motion case. This highlights the robustness of the fused configuration in handling dynamic environments. In small-motion and still driving cases, the fusion approach maintained strong results, with MAEs of 8.15 bpm and 7.80 bpm, respectively, significantly outperforming NIR-only, RGB+YUV and CPG+YUV combinations.

- **Garage Scenarios:** In garage environments, the  $\mathcal{F}_{REC}$  (CPG+YUV+NIR) configuration similarly demonstrated the lowest error rates across all motion levels, with an MAE of 9.09 bpm in large-motion and 4.88 bpm in still conditions. The high  $\rho$  and positive SNR in the garage still scenario further indicate this configuration’s ability to capture stable physiological signals in low-motion settings.

## 5. Conclusion

In this paper, we present a novel RGB-NIR fusion model for robust rPPG and HR estimation. The proposed RGB-NIR network incorporates two key innovations, an NIR-specific decoder that enables the RGB-trained encoder to learn features applicable across both modalities, and an NIR-specific autoencoder alongside an RGB-specific autoencoder for effective disentanglement and reconstruction of physiological features. By utilizing combinations of CPG-STMap representations as input and integrating the NIR-specific CVD loss, our model achieves superior performance across various scenarios in the MR-NIRP, demonstrating significant robustness in challenging vehicular environments. Future work will explore the robustness of RGB-NIR fusion methods within self-supervised learning models to leverage richer and more transferable features from unlabeled data.

## Acknowledgments

This research is supported by the SAFER funded pre-study “ViDCoM”, the Department funds of the CSE department at Chalmers University of Technology, and the University of Gothenburg, and by Vetenskapsrådet, grant 2021-04881\_VR.

## References

- [1] Sk Miraj Ahmed, Suhas Lohit, Kuan-Chuan Peng, Michael J Jones, and Amit K Roy-Chowdhury. Cross-modal knowledge transfer without task-relevant source data. In *European Conference on Computer Vision*, pages 111–127. Springer, 2022. 4
- [2] John Allen. Photoplethysmography and its application in clinical physiological measurement. *Physiological measurement*, 28(3):R1–R39, 2007. 1
- [3] 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
- [4] 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. 6
- [5] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the european conference on computer vision (ECCV)*, pages 349–365, 2018. 2, 5, 7
- [6] 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, 3
- [7] 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. 3
- [8] Gerard De Haan and Arno Van Leest. Improved motion robustness of remote-ppg by using the blood volume pulse signature. *Physiological measurement*, 35(9):1913, 2014. 2, 5
- [9] Gerard De Haan and Jeanne Vincent. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013. 2, 4, 5
- [10] 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. 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. 3
- [13] 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, 3
- [14] 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. 3, 6
- [15] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7
- [16] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems*, 33:19400–19411, 2020. 2, 5
- [17] Xin Liu, Ziheng Jiang, Josh Fromm, Xuhai Xu, Shwetak Patel, and Daniel McDuff. Metaphys: few-shot adaptation for non-contact physiological measurement. In *Proceedings of the Conference on Health, Inference, and Learning*, page 154–163, New York, NY, USA, 2021. Association for Computing Machinery. 5
- [18] Xin Liu, Girish Narayanswamy, Akshay Paruchuri, Xiaoyu Zhang, Jiankai Tang, Yuzhe Zhang, Yuntao Wang, Soumyadip Sengupta, Shwetak Patel, and Daniel McDuff. rppg-toolbox: Deep remote ppg toolbox. *arXiv preprint arXiv:2210.00716*, 2022. 6
- [19] Xin Liu, Brian Hill, Ziheng Jiang, Shwetak Patel, and Daniel McDuff. Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5008–5017, 2023. 2, 5, 7
- [20] 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*, 2024. 1
- [21] Xin Liu, Yuting Zhang, Zitong Yu, Hao Lu, Huanjing Yue, and Jingyu Yang. rppg-mae: Self-supervised pretraining with masked autoencoders for remote physiological measurements. *IEEE Transactions on Multimedia*, 2024. 3, 4
- [22] Girish Narayanswamy, Yujia Liu, Yuzhe Yang, Chengqian Ma, Xin Liu, Daniel McDuff, and Shwetak Patel. Bigsmall: Efficient multi-task learning for disparate spatial and temporal physiological measurements. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7914–7924, 2024. 2, 5
- [23] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Transactions on Image Processing*, 29:2409–2423, 2020. 2, 3, 5, 7
- [24] Xuesong Niu, Zitong Yu, Hu Han, Xiaobai Li, Shiguang Shan, and Guoying Zhao. Video-based remote physiological measurement via cross-verified feature disentangling. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 4, 5, 7
- [25] E. M. Nowara, T. K. Marks, H. Mansour, and A. Veeraghavan. Near-infrared imaging photoplethysmography during

- driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(4):3589–3600, 2020. 1, 2, 6
- [26] 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
- [27] 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, 3
- [28] Christian S Pilz, Sebastian Zaunseder, Jarek Krajewski, and Vladimir Blazek. Local group invariance for heart rate estimation from face videos in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1254–1262, 2018. 2, 5
- [29] 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
- [30] 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
- [31] 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. 2, 5
- [32] Michaela Sidikova, Radek Martinek, Aleksandra Kawala-Sterniuk, Martina Ladrova, 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
- [33] 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. 6
- [34] Wim Verkruyse, Svaasand Lars O., and Nelson J. Stuart. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008. 2, 5
- [35] Wenjin Wang, Albertus C. den Brinker, Sander Stuijk, and Gerard de Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2017. 2, 4, 5, 7
- [36] 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
- [37] 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, 3
- [38] 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 con-*
- ference on computer vision and pattern recognition, New Orleans, LA, USA*, pages 4186–4196, 2022. 5, 7