PhysDrive: A Multimodal Remote Physiological Measurement Dataset for In-vehicle Driver Monitoring

Jiyao Wang[1*], **Xiao Yang**[1*], **Qingyong Hu**[2*], **Jiankai Tang**[3], **Can Liu**[4], **Dengbo He**[1†], **Yuntao Wang**[3], **Yingcong Chen**[1], **Kaishun Wu**[1]
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]The Hong Kong University of Science and Technology, [3]Tsinghua University
[4]Sichuan Agricultural University
∗ Equal Contribution † Corresponding Author

Abstract

Robust and unobtrusive in-vehicle physiological monitoring is crucial for ensuring driving safety and user experience. While remote physiological measurement (RPM) offers a promising non-invasive solution, its translation to real-world driving scenarios is critically constrained by the scarcity of comprehensive datasets. Existing resources are often limited in scale, modality diversity, the breadth of biometric annotations, and the range of captured conditions, thereby omitting inherent real-world challenges in driving. Here, we present PhysDrive, the first large-scale multimodal dataset for contactless in-vehicle physiological sensing with dedicated consideration on various modality settings and driving factors. PhysDrive collects data from 48 drivers, including synchronized RGB, near-infrared camera, and raw mmWave radar data, accompanied with six synchronized ground truths (ECG, BVP, Respiration, HR, RR, and SpO2). It covers a wide spectrum of naturalistic driving conditions, including driver motions, dynamic natural light, vehicle types, and road conditions. We extensively evaluate both signal-processing and deep-learning methods on PhysDrive, establishing a comprehensive benchmark across all modalities, and release full open-source code with compatibility for mainstream public toolboxes. We envision PhysDrive will serve as a foundational resource and accelerate research on multimodal driver monitoring and smart-cockpit systems.

1 Introduction

As intelligent transportation moves toward human–machine co-driving, effective driver monitoring becomes essential to ensure safety and enable timely driver intervention [1, 2]. Such monitoring increasingly depends on understanding the driver's internal state, where physiological signals offer an objective and rich source of information for tasks like driver state monitoring [3], health assessment [4], and in-vehicle interaction [5]. However, traditional contact-based acquisition methods for these signals, such as electrocardiography (ECG) and respiratory belts, are often intrusive and costly, potentially distracting drivers and hindering user acceptance [6]. Remote physiological measurement (RPM) emerges as a compelling alternative, enabling the noninvasive, convenient, and simultaneous acquisition of multiple biological signals (e.g., heart rate (HR), respiration rate (RR)) without physical contact [7]. Compared to contact-based approaches, RPM's noninvasive and convenient nature [8, 9] facilitates its seamless integration into in-vehicle systems, thereby minimizing driver disruption and broadening its applicability within smart vehicles [3].

Contactless in-vehicle physiological monitoring primarily utilizes vision-based approaches and radio frequency (RF) sensing [7]. Vision-based methods include cost-effective RGB cameras for color information-based remote photoplethysmography (rPPG) and near-infrared (NIR) cameras for more

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

stable imaging than RGB under dynamic in-vehicle lighting. Millimeter-wave (mmWave) radar, as a typical RF technology, leverages its short wavelength to detect minute cardiorespiratory chest displacements at the millimeter level, offering robustness to illumination and enhanced privacy. While each modality has individual strengths, they also face distinct challenges: RGB is sensitive to light variations [10], NIR can have lower signal-to-noise ratios for some rPPG estimations [11, 12], and mmWave systems can be influenced by vehicle vibrations and incur higher costs [13, 14].

Considering the variety of driving scenarios, it is essential to collaboratively analyze the three modalities towards a practical solution. Despite the growing interest in these technologies, the field is significantly constrained by a lack of comprehensive public benchmark datasets. As shown in Table 1, existing datasets often focus on single-modality like RGB data [15, 16, 17, 18] or provide limited coverage of NIR [19] or coarse-grained measurement for in-vehicle RF sensing [20]. Besides, they are typically gathered in controlled indoor environments, which lack diversity in real-world settings.

In this paper, we propose PhysDrive, a multimodal non-intrusive dataset to facilitate the algorithm development for contactless driving physiological sensing. PhysDrive contains data from 48 drivers with over 1500k synchronized frames in total four conditions, from three contactless sensing modalities: RGB camera, NIR camera, and mmWave radar as well as six contact ground truths: ECG, blood volume pulse (BVP), respiration signals (RESP), HR, RR, and blood oxygen saturation (SpO2). To the best of our knowledge, PhysDrive is the first dataset that comprises all the modalities across real-world driving settings. The contributions and features of PhysDrive are as follows.

Diverse Sensing Modalities. PhysDrive aims to provide a comprehensive dataset considering the various applicability of existing sensors. It contains three typical contactless modalities in vision and RF sensors, with six contact ground truths as labeling across ECG, BVP, RESP, HR, RR, and SpO2.

Versatile Sensing under Real World Settings. PhysDrive features practical data collection from 48 subjects under various real-world driving settings, such as different illuminations, motions, and road conditions. This opens up new research possibilities for establishing open evaluation benchmarks for in-vehicle RPM and unexplored problems, e.g., generalization in contactless multi-modal sensing.

Extensive Benchmarks. To demonstrate the utility of PhysDrive, we have extensively implemented and evaluated the performance of mainstream baseline models across all factors. We also provide open-source resources[1] to facilitate future research, including raw and preprocessed data, code for benchmarking setup, and tutorials for use with the public RPM toolbox.

2 Related Works

2.1 Remote Physiological Measurement

Effective driving monitoring systems are crucial for improving driver safety and well-being, with a growing demand for unobtrusive, contactless solutions [21, 22]. Among these, RGB cameras have attracted considerable attention, largely due to their ubiquity in vehicles. They mainly leverage the periodic color modulations in skin pixels induced by blood flow changes to perform remote photoplethysmography (rPPG) for estimating vital signs such as heart rate and respiratory rate [23, 24, 25]. While promising, the efficacy of RGB-based rPPG is notably susceptible to fluctuations in ambient lighting conditions [10]. To address this illumination challenge, near-infrared (NIR) cameras with active NIR illumination have been explored in the invisible light spectrum and share resilience to variations in visible light [26, 27, 28]. However, due to the reduced sensitivity to blood oxygenation changes, it has low signal-to-noise ratio (SNR), which leads to higher requirements for algorithm development [29]. Concurrently, RF-based solutions, such as mmWave radar, have been investigated from a different modality perspective. These systems typically transmit electromagnetic waves and analyze the reflected signals to capture mechanical movements associated with physiological processes, including chest vibrations from respiration and heartbeats. Thus, they inherently overcome challenges related to dynamic lighting conditions and can better preserve users' privacy compared to video collections. However, radar modules generally entail higher costs than camera systems [7] and are susceptible to interference from vehicle vibrations during driving and diverse reflection paths from different vehicle interior structures [13, 14]. The distinct advantages and limitations of each sensing modality underscore the need for comprehensive, multi-modal datasets from realistic

[1] https://github.com/WJULYW/PhysDrive-Dataset

Table 1: Comparison of existing public remote physiological measurement datasets.

Dataset	Subjects	Modalities	Physiological Signals	Environment	Experiment Conditions
COHFACE [30]	40	RGB	BVP, RR	Indoor	Illumination; Motion
MAHNOB-HCI [31]	27	RGB	ECG, RR	Indoor	Illumination; Emotion
PURE [17]	10	RGB	BVP, HR, SpO2	Indoor	Motion
UBFC-rPPG [18]	43	RGB	BVP, HR, SpO2	Indoor	Motion
VIPL-HR [19]	107	RGB, NIR	BVP, HR, SpO2	Indoor	Illumination; Skin color
MMSE-HR [32]	140	RGB	HR, RR, Blood pressure	Indoor	Skin color; Expression; Emotion
HCW [25]	48	RGB	BVP, RESP, HR, RR	Indoor	Emotion
ECG-Fitness [33]	17	RGB	ECG, HR	Indoor	Special state; Motion
MMPD [15]	33	RGB	BVP, HR	Indoor	Motion; Skin color; Illumination
SUMS [16]	10	RGB	BVP, HR, RR, SpO2	Indoor	Motion
MMDE [34]	64	RGB	BVP, HR	Indoor	Illumination; Motion
EquiPleth [35]	91	RGB, mmWave	BVP	Indoor	Body posture
4TU.ResearchD [36]	10	mmWave	ECG	Indoor	Angle; Special state
MR-NIRP [37]	18	RGB, NIR	BVP	Driving	Illumination
Wu et al. [38]	14	RGB	HR	Driving	Emotion; Illumination; Motion
PhysDrive	48	RGB, NIR, mmWave	ECG, RESP, BVP, HR, RR, SpO2	Driving	Illumination; Motion; Road condition; Car type

in-vehicle conditions. Our dataset fills this gap with synchronized modalities collected under diverse real-world scenarios to facilitate future research..

2.2 Multi-modal RPM Dataset

Currently, there are a number of datasets available for vision-based RPM; however, as indicated in Table 1, most of these datasets offer only a single modality (i.e., RGB video).The factors typically included in existing datasets focus on indoor motions, lighting conditions, and skin color. Some datasets also consider changes in human emotions [31, 25, 32] or various recording devices, such as webcams and mobile phone cameras [19, 15]. The algorithms developed and tested using these datasets are primarily suited for indoor monitoring scenarios instead of real driving scenarios. Only two datasets, MR-NIRP [37] and Wu et al. [38], concentrate on driving scenarios. Among these, MR-NIRP [37] is the only multimodal dataset specific to driving. Unfortunately, both datasets share common shortcomings: they are small in scale, provide only a single physiological label (either BVP or HR), and do not account for varying driving conditions or perceptual modalities.

Furthermore, there is a limited choice of raw mmWave data available. EquiPlet [35] and 4TU.ResearchD [36] are the only two public datasets that include mmWave data, with EquiPlet also offering RGB video. However, these datasets are focused on indoor environments and do not address in-vehicle monitoring. Since data collected indoors cannot effectively simulate the complex interactions of road surface feedback and signal reflections that occur during driving, such datasets are not suitable for developing and evaluating in-vehicle monitoring models with mmWave. Most previous mmWave-based in-vehicle monitoring systems [39, 40] have relied on private data, resulting in a lack of a public and unified evaluation benchmark.

For these reasons, we propose PhysDrive as the first public dataset that offers a comprehensive range of modal and physiological signal labels while encompassing the challenging factors found in real-world driving scenarios.

3 Dataset

3.1 Data Collection

As mentioned earlier, dynamic lighting conditions and driver movements during driving can significantly impact the extraction of physiological signals from video data. Additionally, movements and road conditions pose challenges to the mmWave sensing method. To further investigate these factors, we designed a real-world driving experiment.

Experiment Design. Our data collection experiment was conducted in Guangzhou City, Guangdong Province, China. As shown in the Figure 1, we incorporated lighting conditions and vehicle types as between-subject factors in the experimental design. For the lighting conditions, we considered

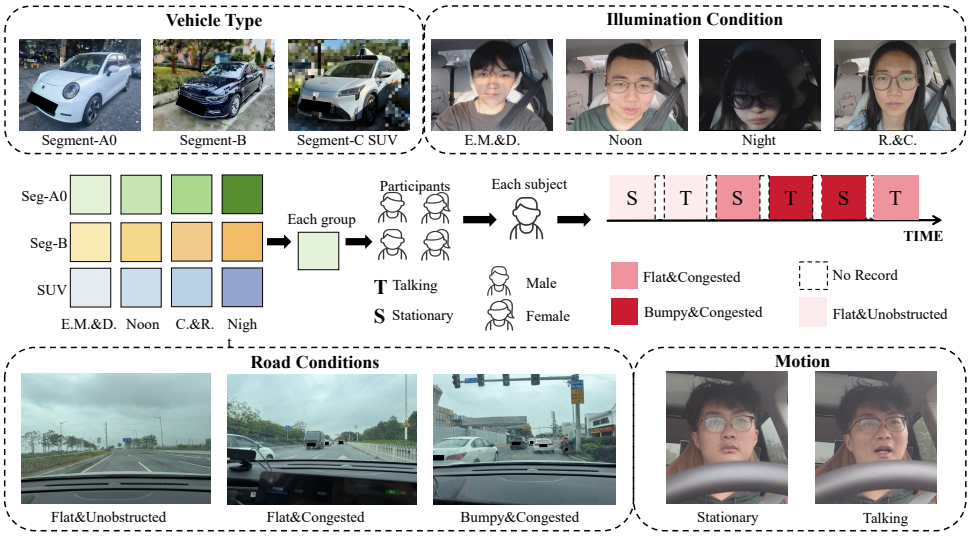

Figure 1: A visual illustration of our data collection experiment. Participants are divided into 12 groups by three types of vehicles and four types of illumination conditions. Each group consists of 4 subjects. Each subject data recordings of each participant are collected under three road conditions. In this figure, 'E.M.&D.' means 'early morning and dusk', and 'C.&R.' is 'cloudy and rainy day'.

changes in angle and brightness under natural light. Specifically, we compared stable lighting conditions at *Noon* (between 11 AM and 1 PM on a clear, cloudless day) to uneven light exposure and unstable angles experienced in the *Early Morning* (7 AM to 9 AM) or *Dusk* (5 PM to 7 PM) also on a clear day. Additionally, we accounted for two situations of varying light intensity: *Rainy/Cloudy Days* and *Nighttime* (after 8 PM). Regarding vehicle models, we considered the impact of different vehicle types on the mmWave scheme. We selected three common vehicle models according to the wheelbase classification method used by China: a *A0-segment vehicle*, a *B-segment vehicle*, and a *C-segment SUV*. Each driver participated in only one type of vehicle under one specific lighting condition, ensuring that all drivers were evenly distributed among the 12 groups formed by the 3 vehicle models and 4 lighting conditions.

Furthermore, we treated the driver's actions and road conditions as within-subject variables. Each driver was required to navigate through a total of two action conditions and three road conditions. For the action variables, we defined two states: 1. *Stationary State* - where the driver was instructed to avoid any distractions except for necessary head and hand movements for normal driving; and 2. *Speaking State* - where we engaged the driver in conversation, encouraging them to perform additional safe actions, such as looking around and relaxing their shoulders. For the road condition variables, we defined three progressively difficult scenarios: 1. *Flat and Unobstructed Road* - a newly constructed three-lane road with minimal traffic and no traffic lights; 2. *Flat but Congested Road* - a three-lane road that, despite being flat, experiences heavy traffic, causing frequent stops and starts, leading to additional body movements; and 3. *Bumpy and Congested Roads* - similar to the second condition but compounded by potholes, resulting in extra shaking. These factors together created six driving segments, each lasting approximately five minutes.

All drivers were required to manually drive all six segments sequentially, as shown in Figure 1. It is important to note that the segments were not continuous, and we did not collect data for non-target driving segments between them, ensuring that each driver was recorded for six driving segments, about 30 minutes with 16.5 km driving distance in total.

Participant. A total of 48 participants, aged from 18 to 41 (Mean=24.9, STD=4.1), were recruited through posters and word-of-mouth, with 24 females and 24 males, and were balanced within each group. All the participants are drivers who have obtained driving licenses for more than one year. Each driver will receive a compensation of 80 Chinese Yuan after completing the experiment. All participants were provided with information explaining the nature and purpose of the procedures involved in this study and signed the consent form before the start of the experiment. This research was approved by the Human and Artefacts Research Ethics Committee [HKUST(GZ)-HSP-2024-0076] of the Hong Kong University of Science and Technology (Guangzhou).

3.2 Data Processing

Apparatus. For data acquisition, we utilize the PhysioLAB platform[2], which is installed on our experimental laptop. This platform integrates a three-electrode ECG sensor, a respiratory belt transducer, a Logitech C925e RGB camera, and an NIR camera for synchronous data collection. The Ergoneers platform allows for the simultaneous recording and pausing of multiple devices, providing millisecond-level timestamps based on the laptop's local time. The acquisition frequency for physiological signals was set at 1000 Hz, while the RGB and IR video acquisition frequencies were set at 30 fps. RGB and IR video are recorded in 'MP4' format, and ECG, RESP signals are in 'CSV' file. For BVP and SpO2 data, we use the Contech CMS50E fingertip blood oxygen meter, which connects to the same laptop. The sampling frequencies are 60 Hz for BVP and 1 Hz for SpO2. The Contech CMS50E provides a second-level timestamp to synchronize the start of recording with the laptop. The mmWave radar is configured as an effective bandwidth of 2.6 GHz with a virtual array of 12 antennas, with a range resolution of around 6 cm and an angular resolution of 14°. It transmits 20 frames in one second. Each frame has 64 chirps with a velocity resolution of 7.6 cm/s. The placement and installation positions of each device are shown in Figure 2(e).

Data Synchronization. To ensure robust temporal synchronization across the data acquisition platforms, all systems are hosted on a single laptop, and their respective timestamps are manually verified for consistency before starting each experimental session. An ordered procedure for initiating and terminating recordings is implemented to further facilitate precise timestamp alignment. The PhysioLAB platform, chosen for its superior timestamp accuracy, initiates recording first and ends last, thereby encompassing the data streams from other sensors. The Contech CMS50E, which provides timestamps accurate to the second, is started after PhysioLAB and stopped before it. Lastly, the mmWave radar system begins after the Contech device and is the first to stop recording. The time interval between the start and end of recording across the different platforms typically does not exceed five seconds. Therefore, during the alignment of the collected data, we can ensure that the drift does not exceed one second. It can be seen from Figure 2(d) that, considering the naturally existing time shift between ECG and BVP, the signals we provide are well aligned.

Next, we take the three perceptual modalities as the benchmark and align them respectively with the physiological data. We align the RGB and NIR videos with ECG, RESP, BVP, and SpO2, and unify the aligned physiological signals to 30 Hz after up/down sampling. For mmWave data, we downsample the ECG and RESP signals to 20 Hz.

Data Preprocessing. To verify the validity of the data and to prepare for the subsequent construction of the benchmark, we conducted preprocessing on the dataset. This dataset is intended for academic use only and is not allowed to be used for commercial purposes. Due to the privacy sensitivity of the original data and the storage limitations of the data hosting platform, we published the processed mmWave data of all participants, and the raw RGB and NIR data for one individual who agreed to publish without any data release agreement, along with the corresponding aligned physiological labels, on https://www.kaggle.com/datasets/xiaoyang274/physdrive. The way of accessing all of the raw data can be found at https://github.com/WJULYW/PhysDrive-Dataset.

For ECG, RESP, and BVP signals, the preprocessing steps mainly include bandpass filtering, signal credibility monitoring, and trend removal. Specifically, PhysDrive can be directly read and processed using rPPG-Toolbox [41], enabling both intra-dataset and cross-dataset training. Inspired by iBVP [42], we introduce a waveform similarity-based quality assessment method to enhance the preprocessing pipeline, allowing for more rigorous validation of signal integrity and completeness before model training. Further, HR and RR were derived using the NeuroKit2[3] package from the ECG and RESP signals. The distribution of three indicators and one visualization example of processed physiological signals is shown in Figure 2. It is worth noting that due to safety considerations in our experiment, induction for low SpO2 was not carried out, so the distribution of SpO2 was relatively concentrated. Therefore, our subsequent benchmarks will not be targeted at SpO2.

For RGB and NIR video, the resolution of the raw RGB and NIR frames is 1024×576 and 1920x1080, respectively. In order to meet the input requirements of most baselines and reduce the reading time during the model training process, we follow [19, 43], first locate the face, and then crop all frames to a size of 128x128 and enlarge the face area. We take it as the input of the baseline for directly

[2]https://www.infoinstruments.cn/product/physiolab/

[3]https://neuropsychology.github.io/NeuroKit/

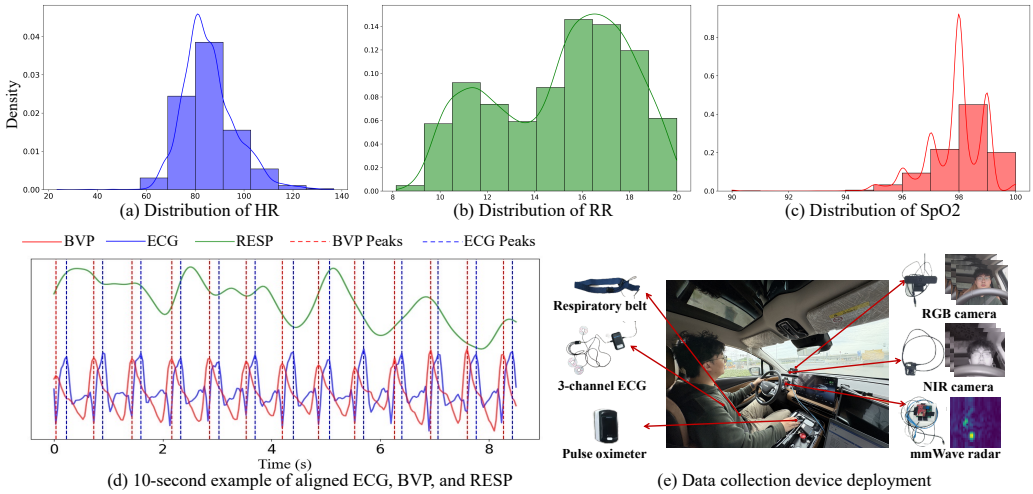

Figure 2: Visualization of processed physiological signals and deployment of data collection devices.

inputting the video. Further, we generate the frames of the cropped images as STMap as the input of some baselines according to Appendix A.

The raw mmWave data is first organized as a tensor $X \in \mathbb{R}^{N_f \times 2 \times N_r \times N_a \times N_d}$ with range-angle-doppler FFT operations. Here, N_f represents the number of time frames, the dimension of size '2' accounts for the real and imaginary components of the complex radar signals, and N_r, N_a, N_d denote the number of bins partitioning the signal by range (distance to reflectors), angle (spatial direction of reflections), and Doppler (velocity of reflectors), respectively. To isolate subtle dynamic physiological signals, we first apply static reflection removal techniques to filter out interference from stationary objects within the environment and then localize the human subject [44]. Subsequently, to enhance computational efficiency and concentrate on the subject's immediate vicinity, we crop the data around the localized region. This results in a focused tensor with $N_r = 8$ range bins (a physical range of 48 cm to cover typical human torso), $N_a = 16$ angle bins (45° angle coverage, to cover frontal human presence), and $N_d = 8$ Doppler bins (a velocity range of 60.8 cm/s on cardiorespiratory motions without gross body movements). Finally, we segment the data into sequences of $N_f = 200$ frames, equivalent to 10 seconds of observation, ensuring temporal alignment with other sensor modalities.

4 Benchmarks

4.1 Benchmark Setup

Baselines. We choose five non-learning RGB-based traditional methods: CHROM [45], POS [46], GREEN [11], ICA [23], and ARM-RR [47]. Besides, we select two unsupervised DL methods: Contrast-Phys+ [26] and SiNC [24], which are trained on unlabelled PhysDrive RGB video and tested on the test set. Other seven single-task DL methods (including DeepPhys [27], PhysNet [28], PhysFormer [48], EfficientPhys [49], Rhythmformer [50], BVPNet [51], RhythmNet [19]) and four MTL baselines (including MTTS-CAN [52], BigSmall [53], BaseNet [25], PhysMLE [25]). Among them, Contrast-Phys+ [26], DeepPhys [27], and PhysNet [28] also serve as the baselines for NIR sensing. For mmWave sensing, we evaluate three baselines (including IQ-MVED [54], VitaNet [55], and mmFormer [44]). Note that, since we didn't find the proper MTL mmWave method, we make minor changes to the above method, including: (1) aggregating multiple Doppler bins with a linear layer; (2) adding new estimation heads after the backbone network. Besides, we remove the AU head from BigSmall for fair comparison. RhythmNet, BaseNet, and PhysMLE take STMap as input.

Implementation. All models used in the evaluation are implemented in PyTorch. The implementation of the baselines is primarily sourced from the rPPG-toolbox [41], otherwise from their open-source repositories. Our experiments are conducted on two servers, one equipped with 8 Nvidia RTX 3090 cards and the other with 8 Nvidia RTX A6000 cards. For the parameter settings of each baseline, we adhere to the descriptions provided in their respective source papers or the default configurations found in the rPPG-toolbox.

Table 2: **Intra-dataset.** HR estimation performance on RGB, NIR, and mmWave modalities.

Modality	Method	MAE↓	RMSE↓	P↑
RGB	CHROM [45]	12.23	15.97	0.11
	POS [46]	12.42	16.15	0.10
	SiNC [24]	13.49	16.57	0.03
	PhysNet [28]	6.29	8.93	0.61
	RhythmFormer [50]	7.21	9.84	0.45
	RhythmNet* [19]	6.84	9.01	0.58
NIR	PhysNet [28]	10.69	13.21	0.12
	Contrast-Phys+ [26]	13.65	16.08	0.05
mmWave	VitaNet [55]	4.94	7.15	0.94
	mmFormer [44]	3.65	5.09	0.97

Table 3: **Intra-dataset.** Multi-task estimation performance of MTL methods.

Method	HR			RR		
	MAE↓	RMSE↓	P↑	MAE↓	RMSE↓	P↑
CHROM [45]	12.23	15.97	0.11	N/A	N/A	N/A
ARM-RR [47]	N/A	N/A	N/A	4.63	5.88	0.08
MTTS-CAN [52]	8.75	11.02	0.26	3.01	4.14	0.12
BigSmall [53]	9.21	11.57	0.24	3.18	4.29	0.10
BaseNet* [25]	6.97	9.32	0.53	2.70	3.29	0.15
PhysMLE* [25]	7.02	9.90	0.55	2.21	3.59	0.16
IQ-MVED [54]	14.17	18.21	0.45	3.71	4.69	0.04
-(Wave Recovery)	29.18	37.31	0.10	2.20	3.15	0.32
VitaNet [55]	4.94	7.15	0.94	2.56	3.64	0.80
-(Wave Recovery)	35.21	43.1	0.07	2.88	3.71	0.12
mmFormer [44]	3.65	5.09	0.97	1.49	2.41	0.83
-(Wave Recovery)	33.87	41.96	0.03	2.1	3.12	0.37

Metrics. The goal of this dataset is to study RPM. Depending on our different sensing methods, we have outlined the following specific tasks: For the RGB and NIR sensing methods, we aim to fit the BVP and RESP signals, while evaluating HR and RR extracted from BVP and RESP, respectively. The mmWave methods in our task focus on HR and RR values regression, as well as ECG and RESP signals recovery, simultaneously. We have adopted evaluation metrics from previous studies [56, 57], including mean absolute error (MAE), root mean square error (RMSE), and Pearson's correlation coefficient (P) to assess the estimated indicators.

4.2 Cross-subject Evaluation

Evaluation Protocol. To evaluate the dataset's usability, we first conduct an intra-dataset cross-subject evaluation within PhysDrive. We utilize data from 48 drivers for training, validation, and testing in a distribution of 80%, 10%, and 10%, respectively. By controlling random seeds, we perform five independent evaluations, and we report the average values of all indicators across these five evaluations.

Results of HR Estimation. Since there are page limitations, we present main findings in Table 2. The complete version can be found in Table 6 in Appendix B. As shown in Table 2, from the perspective of data validity, our results align closely with those from previous similar datasets. The baseline methods using RGB, NIR, and mmWave modalities achieve reasonable results on our dataset, particularly the DL method. This indicates that our dataset effectively validates the input modalities, such as physiological signals (*i.e.*, ECG and BVP), and the correlations between inputs and outputs.

Regarding the baseline results across multiple modalities, it's noteworthy that RGB techniques significantly outperform the NIR methods. This finding is consistent with results reported in [37]. However, the vision-based solutions do not match the accuracy levels achieved in the indoor rPPG dataset under simpler protocols, such as intra-dataset evaluations. For example, results from [26] indicate that PhysNet achieves a MAE of 2.1 and a P of 0.99 in intra-dataset evaluations on PURE. Furthermore, the mmWave methods substantially outperform all vision-based approaches, particularly in terms of P value. For instance, VitaNet exceeded the best RGB method (PhysNet) by about 54.1%.

Results of Multi-task Estimation. In Table 3, we evaluate the performance of the MTL methods of three modalities. The results demonstrate not only the validity of HR-related BVP and ECG signals but also confirm the correlation between these inputs and RESP signals in the context of RR monitoring. In terms of baseline performance, we observe that among vision-based methods, MTL techniques built on STMap (i.e., BaseNet and PhysMLE) outperform those that directly use face videos. Notably, for the RR estimation task, the average P value improved by approximately 25%. While mmWave methods show strengths, their performance notably drops for HR/RR extraction via recovered ECG/RESP waveforms compared to direct parameter regression. This underperformance is plausibly attributed to mmWave's acute sensitivity to temporal misalignment during the critical waveform regression step. Consequently, a key research direction is to develop training schemes that confer temporal robustness to mmWave models, thereby alleviating stringent data synchronization demands and potentially broadening practical applicability.

Results of Cross-scenario Evaluation. To address the two challenging factors that affect the RPM methods (*i.e.,* driver motions and road conditions), we perform cross-scenario evaluations, assessing data from different segments of the test set individually. The results are displayed in Table 4.

Table 4: **Intra-dataset Scenario Evaluation.** HR and RR estimation performances under driver motion conditions and road scenarios. PhysMLE: RGB, PhysNet: NIR, VitaNet and mmFormer: mmWave.

Method	Stationary				Talking				Flat & Unobstructed				Flat & Congested				Bumpy & Congested			
	HR		RR		HR		RR		HR		RR		HR		RR		HR		RR	
	MAE↓	P↑	MAE↓	P↑	MAE↓	P↑	MAE↓	P↑	MAE↓	P↑	MAE↓	P↑	MAE↓	P↑	MAE↓	P↑	MAE↓	P↑	MAE↓	P↑
PhysMLE [25]	6.55	0.58	2.12	0.18	6.98	0.54	2.30	0.15	6.95	0.56	2.16	0.17	7.01	0.55	2.22	0.15	7.16	0.53	2.32	0.14
PhysNet (NIR) [28]	10.35	0.15	N/A	N/A	11.13	0.11	N/A	N/A	10.59	0.14	N/A	N/A	10.67	0.12	N/A	N/A	10.73	0.11	N/A	N/A
VitaNet [55]	4.92	0.93	2.47	0.80	4.99	0.92	2.80	0.77	4.87	0.94	2.44	0.81	5.17	0.92	2.56	0.80	5.97	0.90	2.86	0.78
mmFormer [44]	3.32	0.97	1.24	0.86	3.75	0.94	1.51	0.84	3.32	0.97	1.43	0.85	3.39	0.97	1.51	0.86	3.77	0.94	1.55	0.84

Table 5: **Cross-dataset Scenario Evaluation.** HR estimation performance of RGB-based methods under varying lighting and motion conditions when trained on PURE and UBFC-rPPG. Here, 'E.M.&D.' indicates early morning&dusk; and 'R.&C.' is rainy&cloudy.

Method	Train Set	E.M.&D.		Noon		Night		R.&C.		Stationary		Talking		All	
		MAE↓	P↑	MAE↓	P↑	MAE↓	P↑	MAE↓	P↑	MAE↓	P↑	MAE↓	P↑	MAE↓	P↑
SiNC [24]	PURE	11.16	0.06	10.06	0.17	13.24	0.01	11.39	0.11	11.90	0.11	11.02	0.13	11.46	0.11
	UBFC-rPPG	15.99	0.03	11.62	0.26	15.42	-0.01	13.99	0.08	14.07	0.14	14.39	0.08	14.26	0.12
RhythmFormer [50]	PURE	12.51	0.04	10.40	0.19	13.65	-0.05	11.73	0.04	11.45	0.13	12.09	0.07	11.72	0.11
	UBFC-rPPG	12.98	0.01	11.04	0.15	12.94	0.05	12.18	-0.01	12.16	0.11	12.57	0.05	12.53	0.08
RhythmNet* [19]	PURE	7.91	0.04	8.87	0.23	11.96	-0.06	8.40	0.14	9.61	0.15	8.99	0.11	9.30	0.14
	UBFC-rPPG	6.83	0.03	10.56	0.15	12.25	-0.03	9.95	0.01	10.18	0.11	9.56	0.17	9.86	0.14

Firstly, as shown in Table 4, aligned with Table 3, mmWave methods outperform RGB and NIR methods, while the impact of motions is larger on mmWave methods. Specifically, regarding driver actions, as expected, the performance of all methods in both HR and RR monitoring tasks declines when drivers are in the talking state compared to the stationary state.

Similarly, seeing Table 4, as road conditions more challenging due to jolting and congestion, the accuracy of the mmWave solution decreases. Specifically, compared to flat and unobstructed roads, the performance on more congested flat roads experience a minor decline, usually not exceeding 5%. However, when the road surface becomes bumpy, the degradation is more pronounced, with the MAE increasing by about 13% compared to flat and unobstructed roads.

4.3 Cross-dataset Evaluation

Evaluation Protocol. To assess the capability of our dataset as both a training and a test set for evaluating the generalization ability of the DL model, we conduct extensive evaluations based on a cross-dataset protocol. We select the PURE [17] and UBFC-rPPG [18] datasets, which are commonly used in previous studies [56, 43], alongside PhysDrive for our evaluation. It is important to note that neither PURE nor UBFC-rPPG provides RESP, and neither PURE nor PhysDrive is specifically designed for measuring blood oxygen levels. Thus, our evaluations focus solely on HR estimation tasks. The main results are shown in Table 5, while the remaining results can be referenced in Table 9, 10 in Appendix B.3, B.4.

Results of Cross-scenario Evaluation. We evaluate methods trained on other datasets across various challenging scenarios in PhysDrive. As shown in Table 5, we can conclude that, first, all methods achieve the best performance around noon and the lowest performance at night. Second, conditions on cloudy and rainy days, which have varying brightness but maintain stability, outperform those in the early morning and at dusk, when lighting fluctuates with the driving direction. As expected, MAE may vary when drivers engage in additional conversational behaviors.

5 Discussion and Limitations

In this paper, we present a novel multimodal in-vehicle driver RPM dataset called PhysDrive. Extensive evaluations have confirmed the validity of the data, revealing some interesting findings.

Modality Comparison. Our evaluation of RGB, NIR, and mmWave modalities reveals distinct strengths that depend on the modality and scenario. In almost all driving conditions, mmWave radar

methods demonstrate superior accuracy in directly estimating indicators (i.e., HR, RR) and recovering RESP. However, they consistently struggle to reconstruct ECG waveforms. This discrepancy likely arises because ECG captures detailed cardiac features [58], that are inherently more difficult to recover and require stricter time synchronization than BVP signals. Vision-based methods excel at recovering BVP waveforms, particularly for RGB methods with the highest SNR in well-lit conditions. Although NIR methods are theoretically more stable than RGB in low-light conditions, their performances do not exceed that of RGB due to a lower SNR as aligned with [37].

Our analyses show that the smoothness of the road surface has the greatest effect on mmWave's performance, followed by driver movements and traffic congestion. In contrast, vision-based methods are particularly sensitive to brightness and quick changes in lighting. This suggests the possibility of dynamic modality selection or adaptive fusion to leverage each sensor's strengths effectively [59, 35].

Insights for RPM Models. Compared with traditional signal-processing baselines, fully supervised deep-learning models, and unsupervised (or self-supervised) approaches under the same evaluation protocols, distinct training paradigms emerge. Supervised DL models, while effective in-domain, often exhibit poor generalization across different vehicle, lighting, or road scenarios, likely due to label noise and environmental overfitting. Conversely, unsupervised pretraining on large volumes of unlabeled driving data leads to representations that generalize more robustly across datasets. Therefore, we recommend a two-stage training strategy: initially, learn robust feature extractors using unsupervised/self-supervised methods on diverse, unlabeled real-world driving videos, followed by supervised fine-tuning on annotated data tailored to the specific application context.

Architecturally, we have observed that STMap methods for RGB video outperform direct video-input networks in multi-task estimation and generalization. By converting subtle changes in facial color into structured maps, STMap reduces the effects of noise from head motion and variations in illumination. However, this preprocessing pipeline introduces additional latency, which may hinder real-time applications. Future research should explore learnable or more efficient preprocessing techniques that maintain the robustness of STMap while minimizing computational overhead.

Future Works Enabled by PhysDrive. The PhysDrive dataset supports future research in remote physiological monitoring and intelligent in-vehicle systems. Its rich multimodal data encourages the development of sensor fusion techniques and advanced representation learning for robust physiological measurement across diverse driving conditions. For mmWave sensing, PhysDrive's provision of raw data will be invaluable for creating training schemes that enhance temporal robustness, potentially reducing the strictness of synchronization requirements in practical data collection.

Limitations. Despite its comprehensiveness, the PhysDrive dataset still has several limitations. First, our participant cohort predominantly consists of individuals of East Asian descent, which limits the evaluation of camera-based methods across a diverse range of skin tones [15]. Future data collection should prioritize the inclusion of individuals with different skin tones to assess and mitigate rPPG biases. Secondly, we currently focus on drivers as the first step. We envision the principles within PhysDrive can inspire initial explorations into passenger monitoring, adapting models to new challenges such as varied seating positions and occlusions [14]. Third, although we have SpO2 recordings, the lack of a dedicated acquisition protocol results in limited variability, diminishing its utility. Targeted experiments, such as those under simulated hypoxia or high-altitude conditions, are necessary to develop reliable in-vehicle SpO2 estimation methods. Finally, the current one-second synchronization tolerance between modalities may obscure transient cardiac events; future datasets should implement hardware-level timestamping or higher-precision synchronization to support analyses of rapid physiological changes.

6 Conclusion

In this paper, we introduce PhysDrive, a multimodal dataset designed for the RPM of in-vehicle drivers. PhysDrive is the first to focus on and meticulously design driving scenarios, simultaneously providing RGB, NIR, and mmWave sensing data as well as a large-scale dataset of various physiological signals (including ECG, RESP, BVP, HR, RR, and SpO2). It aims to meet the demand for a relatively complete public dataset in the previous research communities of different monitoring technologies. The designs of various scenarios provided in PhysDrive can offer a relatively comprehensive public benchmark for subsequent work. It also provides an opportunity to integrate multiple sensing technologies and communities, accelerating the development of the next-generation intelligent cockpit.

References

[1] Lu, J., Z. Peng, S. Yang, et al. A review of sensory interactions between autonomous vehicles and drivers. *Journal of Systems Architecture*, 141:102932, 2023.

[2] Wang, J., A. Wang, H. Hu, et al. Multi-source domain generalization for ecg-based cognitive load estimation: Adversarial invariant and plausible uncertainty learning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1631–1635. IEEE, 2024.

[3] Wang, J., X. Yang, Z. Wang, et al. Efficient mixture-of-expert for video-based driver state and physiological multi-task estimation in conditional autonomous driving. *arXiv preprint arXiv:2410.21086*, 2024.

[4] Gharamohammadi, A., A. Khajepour, G. Shaker. In-vehicle monitoring by radar: A review. *IEEE Sensors Journal*, 23(21):25650–25672, 2023.

[5] Detjen, H., S. Faltaous, B. Pfleging, et al. How to increase automated vehicles' acceptance through in-vehicle interaction design: A review. *International Journal of Human–Computer Interaction*, 37(4):308–330, 2021.

[6] Yang, L., H. Yang, H. Wei, et al. Video-based driver drowsiness detection with optimised utilization of key facial features. *IEEE Transactions on Intelligent Transportation Systems*, 25(7):6938–6950, 2024.

[7] Liang, K., J. Chen, T. He, et al. Review of the open data sets for contactless sensing. *IEEE Internet of Things Journal*, 11(11):19000–19022, 2024.

[8] Aarts, L. A., V. Jeanne, J. P. Cleary, et al. Non-contact heart rate monitoring utilizing camera photoplethysmography in the neonatal intensive care unit—a pilot study. *Early human development*, 89(12):943–948, 2013.

[9] Yan, B. P., W. H. Lai, C. K. Chan, et al. High-throughput, contact-free detection of atrial fibrillation from video with deep learning. *JAMA cardiology*, 5(1):105–107, 2020.

[10] Chiu, L.-W., Y.-R. Chou, Y.-C. Wu, et al. Deep-learning-based remote photoplethysmography measurement in driving scenarios with color and near-infrared images. *IEEE Transactions on Instrumentation and Measurement*, 72:1–12, 2023.

[11] Verkruysse, W., L. O. Svaasand, J. S. Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008.

[12] Vizbara, V. Comparison of green, blue and infrared light in wrist and forehead photoplethysmography. *BIOMEDICAL ENGINEERING 2016*, 17(1), 2013.

[13] Morabet, F., A. Lazaro, M. Lazaro, et al. Driver activity monitoring based on modulated frequency selective surface and millimeter-wave radar. *IEEE Sensors Journal*, 2025.

[14] Van Marter, J. P., A. G. Dabak, A. V. Mani, et al. A deep learning approach for in-vehicle multi-occupant detection and classification using mmwave radar. *IEEE Sensors Journal*, 2024.

[15] Tang, J., K. Chen, Y. Wang, et al. Mmpd: Multi-domain mobile video physiology dataset. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–5. IEEE, 2023.

[16] Liu, K., J. Tang, Z. Jiang, et al. Summit vitals: Multi-camera and multi-signal biosensing at high altitudes. In *The 21st IEEE International Conference on Ubiquitous Intelligence and Computing (UIC 2024)*. 2024.

[17] Stricker, R., S. Müller, H.-M. 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.

[18] Bobbia, S., R. Macwan, Y. Benezeth, et al. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124:82–90, 2019.

[19] Niu, X., S. Shan, H. Han, et al. Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Transactions on Image Processing*, 29:2409–2423, 2019.

[20] Li, G.-H., H.-C. Chiang, Y.-C. Li, et al. A driver activity dataset with multiple rgb-d cameras and mmwave radars. In *Proceedings of the 15th ACM Multimedia Systems Conference*, pages 360–366. 2024.

[21] Wang, J., J. M. Warnecke, M. Haghi, et al. Unobtrusive health monitoring in private spaces: The smart vehicle. *Sensors*, 20(9):2442, 2020.

[22] Melders, L., R. Smigins, A. Birkavs. Recent advances in vehicle driver health monitoring systems. *Sensors (Basel, Switzerland)*, 25(6):1812, 2025.

[23] Poh, M.-Z., D. J. McDuff, R. W. Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, 58(1):7–11, 2010.

[24] Speth, J., N. Vance, P. Flynn, et al. Non-contrastive unsupervised learning of physiological signals from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14464–14474. 2023.

[25] Wang, J., H. Lu, A. Wang, et al. Physmle: Generalizable and priors-inclusive multi-task remote physiological measurement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[26] Sun, Z., X. Li. Contrast-phys+: Unsupervised and weakly-supervised video-based remote physiological measurement via spatiotemporal contrast. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[27] Chen, W., D. McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the european conference on computer vision (ECCV)*, pages 349–365. 2018.

[28] Yu, Z., X. Li, G. Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In *30th British Machine Visison Conference: BMVC 2019. 9th-12th September 2019, Cardiff, UK*. The British Machine Vision Conference (BMVC), 2019.

[29] Magdalena Nowara, E., T. K. Marks, H. Mansour, et al. Sparseppg: Towards driver monitoring using camera-based vital signs estimation in near-infrared. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1272–1281. 2018.

[30] Heusch, G., A. Anjos, S. Marcel. A reproducible study on remote heart rate measurement. *arXiv preprint arXiv:1709.00962*, 2017.

[31] Soleymani, M., J. Lichtenauer, T. Pun, et al. A multimodal database for affect recognition and implicit tagging. *IEEE transactions on affective computing*, 3(1):42–55, 2011.

[32] Tulyakov, S., X. Alameda-Pineda, E. Ricci, et al. Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2396–2404. 2016.

[33] Špetlík, R., V. Franc, J. Matas. Visual heart rate estimation with convolutional neural network. In *Proceedings of the british machine vision conference, Newcastle, UK*, pages 3–6. 2018.

[34] Xiang, G., S. Yao, H. Deng, et al. A multi-modal driver emotion dataset and study: Including facial expressions and synchronized physiological signals. *Engineering Applications of Artificial Intelligence*, 130:107772, 2024.

[35] Vilesov, A., P. Chari, A. Armouti, et al. Blending camera and 77 ghz radar sensing for equitable, robust plethysmography. *ACM Trans. Graph.*, 41(4):36–1, 2022.

[36] Sadeghi, E., K. Skurule, A. Chiumento, et al. Comprehensive mm-wave fmcw radar dataset for vital sign monitoring: Embracing extreme physiological scenarios. *arXiv preprint arXiv:2405.12659*, 2024.

[37] Nowara, E. M., T. K. Marks, H. Mansour, et al. Near-infrared imaging photoplethysmography during driving. *IEEE transactions on intelligent transportation systems*, 23(4):3589–3600, 2020.

[38] Wu, B.-F., Y.-C. Wu, Y.-W. Chou. A compensation network with error mapping for robust remote photoplethysmography in noise-heavy conditions. *IEEE Transactions on Instrumentation and Measurement*, 71:1–11, 2022.

[39] Juncen, Z., J. Cao, Y. Yang, et al. mmdrive: Fine-grained fatigue driving detection using mmwave radar. *ACM Transactions on Internet of Things*, 4(4):1–30, 2023.

[40] Wang, F., X. Zeng, C. Wu, et al. Driver vital signs monitoring using millimeter wave radio. *IEEE Internet of Things Journal*, 9(13):11283–11298, 2021.

[41] Liu, X., G. Narayanswamy, A. Paruchuri, et al. rppg-toolbox: Deep remote ppg toolbox. *Advances in Neural Information Processing Systems*, 36:68485–68510, 2023.

[42] Joshi, J., Y. Cho. ibvp dataset: Rgb-thermal rppg dataset with high resolution signal quality labels. *Electronics*, 13(7), 2024.

[43] Yu, Z., Y. Shen, J. Shi, et al. Physformer++: Facial video-based physiological measurement with slowfast temporal difference transformer. *International Journal of Computer Vision*, 131(6):1307–1330, 2023.

[44] Hu, Q., Q. Zhang, H. Lu, et al. Contactless arterial blood pressure waveform monitoring with mmwave radar. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(4):1–29, 2024.

[45] De Haan, G., V. Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE transactions on biomedical engineering*, 60(10):2878–2886, 2013.

[46] Wang, W., A. C. Den Brinker, S. Stuijk, et al. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2016.

[47] Tarassenko, L., M. Villarroel, A. Guazzi, et al. Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. *Physiological measurement*, 35(5):807, 2014.

[48] Yu, Z., Y. Shen, J. Shi, et al. Physformer: Facial video-based physiological measurement with temporal difference transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4186–4196. 2022.

[49] Liu, X., B. Hill, Z. Jiang, et al. 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.

[50] Zou, B., Z. Guo, J. Chen, et al. Rhythmformer: Extracting rppg signals based on hierarchical temporal periodic transformer. *arXiv e-prints*, pages arXiv–2402, 2024.

[51] Das, A., H. Lu, H. Han, et al. Bvpnet: Video-to-bvp signal prediction for remote heart rate estimation. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE, 2021.

[52] Liu, X., J. Fromm, S. Patel, et al. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems*, 33:19400–19411, 2020.

[53] Narayanswamy, G., Y. Liu, Y. Yang, et al. 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.

[54] Zheng, T., Z. Chen, S. Zhang, et al. More-fi: Motion-robust and fine-grained respiration monitoring via deep-learning uwb radar. In *Proceedings of the 19th ACM conference on embedded networked sensor systems*, pages 111–124. 2021.

[55] Khan, U. M., L. Rigazio, M. Shahzad. Contactless monitoring of ppg using radar. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(3):1–30, 2022.

[56] Wang, J., H. Lu, H. Han, et al. Generalizable remote physiological measurement via semantic-sheltered alignment and plausible style randomization. *IEEE Transactions on Instrumentation and Measurement*, 2024.

[57] Wang, J., X. Wei, H. Lu, et al. Condiff-rppg: Robust remote physiological measurement to heterogeneous occlusions. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[58] Lu, G., F. Yang, J. A. Taylor, et al. A comparison of photoplethysmography and ecg recording to analyse heart rate variability in healthy subjects. *Journal of medical engineering & technology*, 33(8):634–641, 2009.

[59] Kurihara, K., D. Sugimura, T. Hamamoto. Non-contact heart rate estimation via adaptive rgb/nir signal fusion. *IEEE Transactions on Image Processing*, 30:6528–6543, 2021.

[60] Kartynnik, Y., A. Ablavatski, I. Grishchenko, et al. Real-time facial surface geometry from monocular video on mobile gpus. *arXiv preprint arXiv:1907.06724*, 2019.

A Producing Spatial-Temporal Map from RGB Video

To reduce the computational resources required during the training process and encourage the model to focus more on facial skin brightness changes rather than environmental variations, we adopted the ROI extraction method from [19] and transformed the video into a Spatial-Temporal Map (STMap) to provide a more lightweight input format.

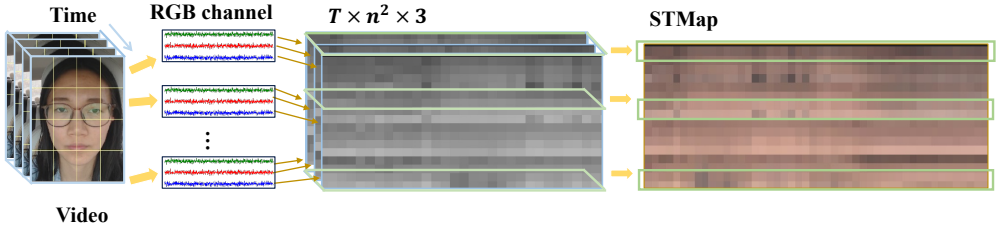

Figure 3: Illustration of producing STMap from RGB video.

As shown in Figure 3, first, we utilized MediaPipe Face Mesh [60] to perform landmark detection on each frame, obtaining 468 facial key points. Next, we define the width of the facial region as the horizontal distance between the boundary points of the outer cheeks, and the height as 1.2 times the vertical distance between the chin and the eyebrow center, thus cropping the entire facial region. Then, we apply an HSV color range threshold to segment the skin area, effectively removing non-facial regions such as the eyes and background. Specifically, we convert the region of interest (ROI) to the HSV color space, and define the skin color range by setting the hue (H) between 0 and 20, the saturation (S) between 20 and 255, and the value (V) between 70 and 255. Finally, we generate the STMap from the skin regions extracted in each frame of the video. Specifically, for a video with T frames, we divide the skin region in each frame into small blocks of size $(n^2 = 16)$, and compute the average pixel values of the RGB color channels for each block. These average pixel values are then concatenated in the order of the video frames, resulting in an STMap with the shape $(T, 16, 3)$.

Table 6: **Intra-dataset.** HR estimation performance of RPM methods on RGB, NIR, and mmWave modalities.

Modality	**Method**	MAE↓	RMSE↓	P↑
RGB	CHROM [45]	12.23	15.97	0.11
	POS [46]	12.42	16.15	0.10
	GREEN [11]	14.09	17.87	0.02
	ICA [23]	13.31	16.93	0.06
	SiNC [24]	13.49	16.57	0.03
	Contrast-Phys+ [26]	15.93	18.40	0.01
Video	DeepPhys [27]	11.97	13.17	0.20
	PhysNet [28]	6.29	8.93	0.61
	PhysFormer[48]	7.85	10.17	0.41
	EfficientPhys[49]	11.27	13.51	0.20
	RhythmFormer [50]	7.21	9.84	0.45
	BVPnet* [51]	7.95	10.70	0.28
	RhythmNet* [19]	6.84	9.01	0.58
NIR	DeepPhys [27]	15.61	17.89	0.03
	PhysNet [28]	10.69	13.21	0.12
	Contrast-Phys+ [26]	13.65	16.08	0.05
mmWave	IQ-MVED [55]	14.17	18.21	0.45
	VitaNet [55]	4.94	7.15	0.94
	mmFormer [44]	3.65	5.09	0.97

B Other Results of Evaluation

B.1 Intra-dataset Evaluation

We trained baseline methods using RGB, NIR, and millimeter-wave modalities on our dataset and validated the results. As shown in Table 6, our results closely match those from similar datasets, particularly for the DL method, yielding reasonable performance. This effectively validates the physiological signals in our dataset (e.g., ECG and BVP) as input modalities and the correlation between input and output.

B.2 Cross-dataset Evaluation: Trained on Other Datasets

We conducted cross-dataset training, with results shown in Table 7. We observed that when trained on simpler datasets such as PURE and UBFC-rPPG, traditional methods (e.g., CHROM, POS) performed comparably to the DL methods on our dataset. Notably, the DL method using STMap outperformed the direct video input approach.

Table 7: **Cross-dataset.** HR estimation performance of baselines with different training sets.

Method	**Train Set**	MAE↓	RMSE↓	P↑
CHROM [45]	N/A	12.23	15.97	0.11
POS [46]	N/A	12.42	16.15	0.10
ICA [23]	N/A	13.31	16.93	0.06
SiNC [24]	PURE	11.46	15.01	0.11
	UBFC-rPPG	14.26	18.00	0.12
	PhysDrive	6.29	8.93	0.61
PhysNet [28]	PURE	13.52	16.81	0.05
	UBFC-rPPG	14.68	17.98	0.02
	PhysDrive	7.85	10.17	0.41
RhythmFormer [50]	PURE	11.72	15.13	0.11
	UBFC-rPPG	12.53	15.87	0.08
	PhysDrive	7.95	10.70	0.28
BVPnet* [51]	PURE	10.59	14.11	0.11
	UBFC-rPPG	10.46	13.72	0.12
	PhysDrive	6.84	9.01	0.58
RhythmNet* [19]	PURE	9.30	12.21	0.14
	UBFC-rPPG	9.86	13.20	0.14
	PhysDrive	7.21	9.84	0.45
DeepPhys [27]	PURE	18.57	21.67	0.01
	UBFC-rPPG	18.74	21.87	0.02
	PhysDrive	11.97	13.17	0.20
Physformer [48]	PURE	12.93	16.16	0.06
	UBFC-rPPG	14.34	17.49	0.03
	PhysDrive	7.85	10.17	0.41
EfficientPhys [49]	PURE	17.59	20.71	0.03
	UBFC-rPPG	17.42	20.56	0.00
	PhysDrive	11.27	13.51	0.20

B.3 Cross-dataset Evaluation: Tested on Other Datasets

We also trained the DL method on PhysDrive and tested it on the PURE and UBFC-rPPG datasets. The results are displayed in Table 7. We observed that the performance of the same baseline on PURE and UBFC-rPPG was better. This indicates that our dataset can be effectively used for training DL models and further applied in other scenarios. It also highlights that the training difficulty of

PhysDrive is greater than that of PURE and UBFC-rPPG, which is one of the primary reasons for introducing this dataset.

Table 8: **Cross-dataset.** HR estimation performance of baselines when trained in PhysDrive and tested on PURE and UBFC-rPPG.

Test Set	**PURE**			**UBFC-rPPG**		
Method	MAE↓	RMSE↓	P↑	MAE↓	RMSE↓	P↑
CHROM [45]	9.79	12.76	0.37	7.23	8.92	0.51
POS [46]	9.82	13.44	0.34	7.35	8.04	0.49
SiNC [24]	18.33	21.89	0.14	18.34	21.89	0.13
PhysNet [28]	15.99	19.40	0.08	12.84	15.84	0.13
RhythmFormer [50]	14.19	18.07	0.12	13.64	16.81	0.11
BVPnet* [51]	14.10	17.45	0.13	12.59	15.37	0.13
RhythmNet* [19]	13.76	17.41	0.16	10.01	14.35	0.20

B.4 Cross-dataset Evaluation on Different Scenarios

We compared the performance variations of traditional and DL methods under different lighting and motion scenarios, as shown in Table 10 and Table 9. All methods reached their best performance around noon and had the lowest performance at night. Additionally, although the brightness on cloudy and rainy days fluctuated, it remained stable, and the performance was better than during the morning and evening when lighting fluctuated with the driving direction. As expected, the performance of all methods decreased when the driver engaged in additional conversation.

Table 9: **Cross-dataset Scenario Evaluation with Traditional Methods.** HR estimation performance of RGB-based baselines under varying lighting and motion conditions.

Method	E.M.&D.		Noon		Night		R.&C.		Stationary		Talking		All	
	MAE↓	P↑	MAE↓	P↑	MAE↓	P↑	MAE↓	P↑	MAE↓	P↑	MAE↓	P↑	MAE↓	P↑
CHROM [45]	11.79	0.10	10.27	0.26	14.37	-0.01	12.25	0.07	12.45	0.13	12.35	0.07	12.36	0.12
POS [46]	12.05	0.07	10.44	0.24	14.36	-0.02	12.22	0.09	12.69	0.10	12.20	0.06	12.79	0.07
GREEN [11]	13.56	0.01	13.53	0.03	15.16	-0.02	13.73	0.02	14.13	0.05	13.94	0.01	14.28	0.02
ICA [23]	13.38	0.01	12.57	0.06	14.62	-0.03	13.00	0.03	13.86	0.03	13.41	0.04	13.73	0.03

Table 10: **Cross-dataset Scenario Evaluation When Trained on Other Datasets.** HR estimation performance of RGB vision–based baselines under varying lighting and motion conditions when trained on PURE and UBFC.

Method	Condition	E.M.&D.		Noon		Night		R.&C.		Stationary		Talking		All	
	Train Set	MAE↓	P↑	MAE↓	P↑	MAE↓	P↑	MAE↓	P↑	MAE↓	P↑	MAE↓	P↑	MAE↓	P↑
SiNC [24]	PURE	11.16	0.06	10.06	0.17	13.24	0.01	11.39	0.11	11.90	0.11	11.02	0.13	11.46	0.11
	UBFC	15.99	0.03	11.62	0.26	15.42	-0.01	13.99	0.08	14.07	0.14	14.39	0.08	14.26	0.12
DeepPhys [27]	PURE	22.73	-0.03	17.18	0.00	16.96	-0.02	18.19	0.03	18.71	0.00	18.58	0.00	18.57	0.01
	UBFC	22.05	0.02	17.16	0.02	16.83	-0.01	18.32	0.04	18.47	0.03	18.60	0.03	18.74	0.02
PhysNet [28]	PURE	15.06	0.05	11.99	0.09	13.76	-0.01	13.30	0.03	13.76	0.05	13.34	0.03	13.52	0.05
	UBFC	16.92	0.03	12.79	0.07	13.59	0.04	14.42	0.02	14.50	0.06	14.33	-0.01	14.68	0.02
Physformer [48]	PURE	14.55	0.03	10.98	0.16	12.82	0.02	12.18	0.07	12.90	0.12	12.53	0.07	12.93	0.06
	UBFC	16.39	0.05	12.76	0.06	13.22	0.03	13.57	0.03	14.51	0.05	14.09	0.01	14.34	0.03
EfficientPhys [49]	PURE	21.46	-0.03	15.39	0.03	15.86	0.00	17.06	0.00	17.73	0.03	17.44	-0.02	17.59	0.03
	UBFC	21.25	0.02	15.61	0.03	15.79	-0.01	17.26	-0.02	17.82	0.02	17.44	0.00	17.42	0.00
RhythmFormer [50]	PURE	12.51	0.04	10.40	0.19	13.65	-0.05	11.73	0.04	11.45	0.13	12.09	0.07	11.72	0.11
	UBFC	12.98	0.01	11.04	0.15	12.94	0.05	12.18	-0.01	12.16	0.11	12.57	0.05	12.53	0.08
BVPnet* [51]	PURE	6.72	0.08	10.47	0.14	13.94	-0.04	10.32	0.05	10.89	-0.02	10.29	-0.02	10.59	0.11
	UBFC	13.80	0.01	9.99	0.16	11.14	-0.06	10.74	0.08	11.85	-0.01	11.08	-0.03	10.46	0.12
RhythmNet* [19]	PURE	7.91	0.04	8.87	0.23	11.96	-0.06	8.40	0.14	9.61	0.15	8.99	0.11	9.30	0.14
	UBFC	6.83	0.03	10.56	0.15	12.25	-0.03	9.95	0.01	10.18	0.11	9.56	0.17	9.86	0.14

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please refer to Section 5.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: Did not include theoretical results.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide necessary details of implementation and experiment results in Section 4

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: See Section 3.2, we open preprocessed data and code in https://github.com/WJULYW/PhysDrive-Dataset for result reproduction.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 4.1 and all Evaluation Protocol subctions in 4.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We reported the average values of five independent evaluations. However, since the focus of this paper is not to propose new ones, not conducting statistical significance tests will not affect the findings and contributions of this paper.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Implementation subsection in Section 4.1

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research was approved by the Human and Artefacts Research Ethics Committee of the university.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Section 5.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This article does not involve AI-generated content.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: See Section 3.1.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: See Section 3.1.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]