Preprints.org

Article

iBVP Dataset: RGB-Thermal rPPG Dataset With High Resolution Signal Quality Labels

Jitesh Narendra Joshi and Youngjun Cho [*]

Article

iBVP Dataset: RGB-Thermal rPPG Dataset with High Resolution Signal Quality Labels

Jitesh Joshi [1,†,‡] (ID) **and Youngjun Cho** [2,*] (ID)

Department of Computer Science, University College London, NW1 2AE, UK. jitesh.joshi.20@ucl.ac.uk (J.J.)

*   Correspondence: youngjun.cho@ucl.ac.uk (Y.C.)

†   Current address: UCLIC and Computer Science, 169 Euston Road, Room: 403, London NW1 2AE, UK

Abstract: Remote Photoplethysmography (rPPG) has emerged as a non-intrusive and promising physiological sensing capability in HCI research, gradually extending its applications in health-monitoring and clinical care contexts. With advanced machine learning models, recent datasets collected in real-world conditions have gradually enhanced the performance of rPPG methods in recovering heart-rate and heart-rate variability metrics. However, the signal quality of reference ground-truth PPG data in existing datasets is by and large neglected, while poor quality references negatively influence models. Here, this work introduces a new imaging blood volume pulse (iBVP) dataset of synchronized RGB and thermal infrared videos with PPG ground-truth signals from the ear and its high resolution signal quality labels, for the first time. Participants perform rhythmic breathing, head-movement, and stress-inducing tasks, which help reflect real-world variations in psycho-physiological states. This work conducts dense (per sample) signal quality assessment to discard noisy segments of ground-truth and corresponding video frames. We further present a novel end-to-end machine learning framework, iBVPNet that features an efficient and effective spatio-temporal feature aggregation for reliable estimation of BVP signals. Finally, this work examines the feasibility of extracting BVP signals from thermal video frames, which is underexplored. The iBVP dataset and source codes are publicly available for research use.

Keywords: remote PPG; RGB-thermal dataset; signal quality labels

---

1. Introduction

The foundation of optical sensing for blood volume pulse signal laid by [1] proved significantly useful in clinical care settings, and past decade also witnessed proliferation of health-tracking devices and smart-watches that monitor heart rate and heart-rate variability metrics. Since Verkruysse [2]'s pioneering investigation on the feasibility of extracting photo-plethysmography signals from RGB cameras in a contact less manner, increasing attention has been given to a wide range of imaging-based physiological sensing methods and their promising applications and contexts where non-invasive and contactless measurement techniques are preferred, such as stress and mental workload recognition [3,4] and biometric authentication [5].

Several rPPG datasets have been made available for academic use, and this recent review [6] overviews these datasets. Some of the widely used datasets include MANHOB-HCI [7], PURE [8], MMSE-HR [9], VIPL-HR [10], UBFC-rPPG [11], UBFC-Phys [12], V4V [13], and SCAMPS [14]. These datasets consists of RGB videos with resolution ranging from 320x240 to 1920x1080, and frame-rates from 20 frames per second (FPS) to 120 FPS. The ground-truth data often consists of photoplethysmography (PPG) signal and/or electro-cardiography (ECG) signal, along with their computed pulse rate (PR) or heart rate (HR) metrics. The majority of these datasets are acquired in laboratory settings with controlled lighting conditions [7,9,11,12], and varying head movement, pose and and emotional changes. PURE [8], ECG-Fitness [15] and VIPL-HR [10] datasets introduce illumination changes for recording videos. VIPL-HR [10] and MANHOB-HCI [7] deploy multiple cameras to capture videos with different resolution as well as face poses. Unlike other datasets, SCAMPS [14] dataset consists of synthetically generated videos with randomly sampled appearance

attributes such as skin texture, hair, clothing, lighting, and environment, making it more suited to train supervised methods, rather than to evaluate rPPG methods. Most of the datasets focus on the RGB imaging modality.

As the ground-truth signals for rPPG datasets is collected using contact-based PPG or ECG sensors, it is essential to screen noise artifacts [16,17] present in such signals. Data collection scenarios including altered physiological states, varying ambient conditions and head-movement offer rich real-world representations enabling robust training of supervised methods, as well as realistic validation of both supervised and unsupervised rPPG extraction methods. However, these scenarios actively involving participants' movement result in noises in the reference PPG signals as well due to varying light issues on the contact points. Representative noise artifacts present in ground-truth PPG signals of the existing datasets can be observed in Figure 1.
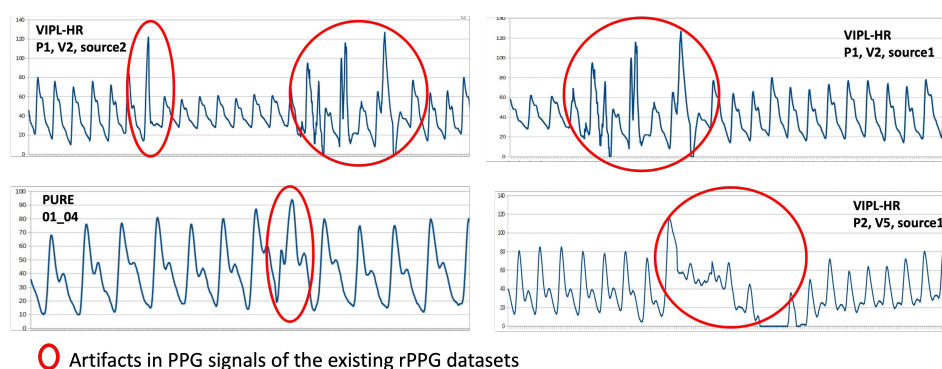

Artifacts in PPG signals of the existing rPPG datasets

Figure 1. Noise artifacts present in the illustrative samples of PPG signals from existing rPPG datasets.

While researchers have proposed de-noising algorithms to remove artifacts from ECG as well as PPG signals [18,19], such methods are often limited to the signals that are not severely corrupted and are therefore, reparable, while insufficient to denoise signals with substantial artifacts. Despite the availability of several rPPG datasets, the signal quality of the ground-truth signals is often neglected. Besides misleading the training of supervised methods, poor quality ground-truth PPG signals can lead to inappropriate evaluation of rPPG methods.

In this work, our collected RGB-Thermal rPPG dataset (which we call iBVP dataset) is first assessed for signal quality of the ground-truth PPG signals. Signal quality assessment method deployed for assessing PPG signals is adapted from a recent work [20], in which the CNN based decoder module is replaced with more efficient matrix decomposition based module [21] and the inference is made per sample, making it a dense 1D segmentation task. The noisy segments are then removed from the ground-truth PPG signals as well as the corresponding video frames. The compiled dataset, therefore can serve as a training dataset as well as a reliable bench-marking dataset for evaluating rPPG methods. We also make the original dataset available with high-resolution signal quality labels used for our processing.

We present primary evaluation of this dataset with 3D CNN based end-to-end learning approaches [22,23] for estimating PPG signals. We further propose a 3D-CNN architecture, iBVPNet, to effectively capture blood volume pulse (BVP) related spatial and temporal features from video frames. To evaluate iBVPNet and existing state-of-the-art (SOTA) models, we leverage maximum amplitude of cross-correlation (MACC) [24] as a metrics that is well-suited for comparing the estimated BVP signals with ground-truth signals. In spite of MACC being highly relevant metric to evaluate rPPG estimation, it has not been leveraged sufficiently in the literature. In summary, we make the following contributions:

- introducing the iBVP dataset comprising RGB and thermal facial video data with signal quality assessed ground-truth PPG signals.

- presenting and validating a new rPPG framework iBVPNet for estimating BVP signal from RGB as well as thermal video frames.
- discovering MACC [24] as an effective evaluation metric to assess rPPG methods.

2. iBVP Dataset

2.1. Data Collection Protocol

The data acquisition was conducted with an objective of inducing variations in physiological states, as well as head movement. Each participant experienced four conditions, including (a) a rhythmic slow breathing and rest, (b) an easy math task, (c) a difficult math task, and (d) a guided head movement task, as depicted in Figure 2A. While a higher agreement between the rPPG and the ground-truth PPG can be achieved in the absence of these variances, inclusion of the same in the data acquisition protocol enable simulating real-world physiological variations.

Cognitively challenging math tasks with varying degrees of difficulty levels were chosen, as these have been reported to alter the physiological responses [25–27]. The achieved distribution of heart rate computed from ground-truth PPG signals can be observed in Figure 2B. Furthermore, as wearable sensors are less reliable under significant motion conditions [28], we added an experimental condition that involved guided head movement. Each condition lasted for 3 minutes, with 1 minute of rest after each condition. To randomize the sequence of conditions, we inter-changed "A" with "D" and "B" with "C". The study protocol was approved by the University College London Interaction Centre ethics committee (ID Number: UCLIC/1920/006/Staff/Cho).
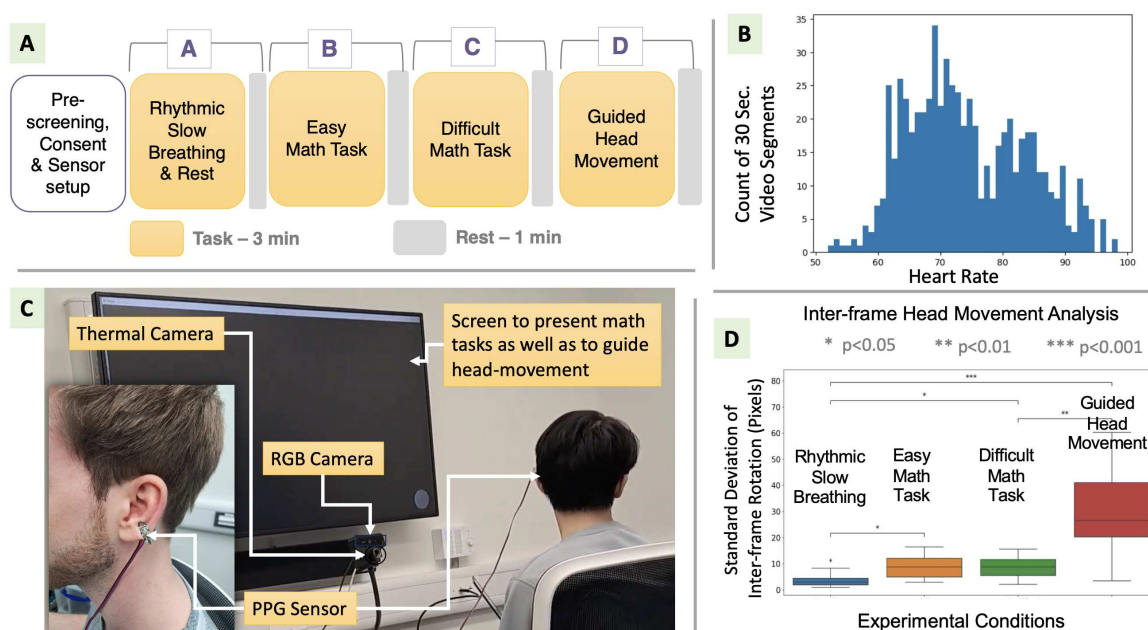

Figure 2. (**A**): Setup for acquiring iBVP dataset; (**B**): Analysis showing the magnitude of inter-frame head movement under different conditions involved in the data acquisition; (**C**): Data acquisition protocol; (**D**): Histogram showing the variations in Heart Rate.

2.2. Participants

PPG signals were collected from 33 participants (adults, 23 females) recruited through an online recruitment platform for research. All participants reported having no known health conditions, provided informed consent ahead of the study, and were compensated for their time following the study. After being welcomed and briefed, participants were asked to remove any bulky clothing (e.g., winter coats, jackets) and seated comfortably in front of a 65 by 37 inch screen, where they were fitted

with Physiokit sensors. PPG sensor was attached to participants' left ear with a metal clip. Of the 33 participants, 3 were excluded from the dataset due to the fitment issue of the PPG sensor, resulting in sensor drop during the acquisition or significant noise artifacts in the ground-truth PPG signals.

2.3. Data Acquisition

As depicted in Figure 2C, RGB and thermal cameras were positioned in front of the participant at around 1 meter distance. Logitech's Brio 4K webcam [29] was used to capture RGB video frames, while thermal infrared frames were captured using FLIR A65SC camera [30]. Key considerations in acquiring ground-truth PPG signal include the close resemblance of morphology with rPPG signals and minimum time-delay or phase-difference. With these considerations, we carefully chose ear for the sensor placement and attached the sensor clip to the upper or lower lobe of the ear based on the best fitment and comfort of the participants. PhysioKit toolkit [20] was adapted to acquire the ground-truth PPG signals in synchronization with RGB and thermal frames. Implementation was adapted such that RGB frames, thermal frames and PPG signal were acquired in separate and dedicated threads, while sharing common onset trigger and a timer to stop the acquisition in synchronized manner. RGB and thermal frames were acquired with a frame rate of 30 frames per second (FPS), while PPG signals were acquired with a sampling rate of 250.

2.4. Morphology and Time-Delay of PPG Signals

Majority of existing rPPG datasets have ground-truth PPG signals acquired using finger -probe or wrist-watch, making it challenging to match the morphology as well as phase [31] of ground-truth and extracted rPPG signals from facial video frames. The morphology of PPG waveforms is site-dependent [32] and therefore it is crucial to acquire ground-truth signals for rPPG from a site that is closest to the face. In addition, a recent study highlights the significant delay, equivalent to half pulse duration between PPG signal acquired from finger and the rPPG signals [31]. With these considerations, for the introduced iBVP dataset, we carefully chose ear as the sensor-site for acquiring ground-truth PPG signals, resulting in close resemblance of the morphology as well as minimum time-delay. This makes iBVP dataset highly suitable for training as well as evaluating the deep-learning-based models that estimate BVP signals. It can be argued that the models that can reliably estimate the BVP signals offer significant advantages over the models trained to directly estimate heart-rate or other BVP derived metrics.

2.5. Pre-Processing and Signal Quality Assessment

A band-pass filter (0.5–2.5 Hz) of the second order was applied to PPG signals which were then re-sampled to a sampling rate of 30, to match it with the FPS of RGB and thermal video frames. The band-pass filter was further applied after re-sampling the signal to reduce the sampling artifacts. The cut-off for higher frequency was chosen as 2.5 Hz to preserve only the pulsating waveforms with systolic peaks, while discarding the features related to dicrotic notch and diastolic peak as rPPG signals may not contain these characteristic features.

While PPG signals acquired from the ear tend to have good signal quality [20,33], it is still prone to noise artifacts due to head-movement. So, it is important to assess the the quality of ground-truth PPG signals. Figure 2D shows comparison of head movement across different experimental conditions, computed as inter-frame rotation of facial frames. The conventional signal quality assessment methods for PPG signals rely on extracting i) frequency components to compute signal to noise ratio [34], ii) different measures of signal quality indices (SQI) [35] including relative power SQI [27], or iii) analyzing morphological features and compare with template signal [36]. In several real-world settings, frequency based SQI measures of signal quality can be misleading due to overlapping frequency components of noise artifacts [37]. Morphological features based signal quality assessment is challenging owing to several factors [36,38] that include following: i) it is required to accurately segment the pulses to match the template, ii) as morphological features vary significantly between different individuals,

generalized pulse template can not be used as reference to match, and iii) some noise artifacts resemble the pulse morphology making it further challenging to discriminate between good quality PPG signal and noise artifacts.

Machine-learning and deep-learning based methods for PPG signal quality assessment have recently attracted wider attention among the researchers [38–42]. These developments have been captured in a recent survey [43] that reviews signal quality assessment methods for contact-based as well as imaging based PPG. While the majority of works focus on developing a classifier models [39,44–46] with binary inference for quality of a length of PPG signal, a few recent works have proposed models that offer high temporal resolution signal quality assessment [20,40,42,47]. In this work, we first extend SQA-Phys, the signal quality assessment method deployed in [20] such that the inference for signal quality is made per each sample, resulting in 1D dense segmentation task and secondly we replace the CNN based decoder of the encoder-decoder architecture with a decoder based on matrix decomposition [21].

A recent work on 2D semantic segmentation [21] shows the efficacy of matrix decomposition based decoder in capturing the global context. Inspired from the approach of low-rank discovery through matrix decomposition [21], this work adapts the hamburger module by implementing the Non-negative Matrix Factorization (NMF) for 1D features. Combining the 1D-CNN encoder module of SQA-Phys [20] and a matrix decomposition based decoder [21], we refer to this new architecture as *SQA-PhysMD*, as depicted in Figure 3. Dataset used for the training of *SQA-PhysMD* model include PPG DaLiA [48] training set, WESAD dataset [49] and TROIKA [17] dataset. The signal quality labels for this training dataset are provided by the authors of [40] and are available as a part of their repository [50]. Training parameters and the model validation was conducted in line with the recent state of the art works [40,42], through which the performance of the trained *SQA-PhysMD* model was found to be at par with the SOTA (results not shown in this work).
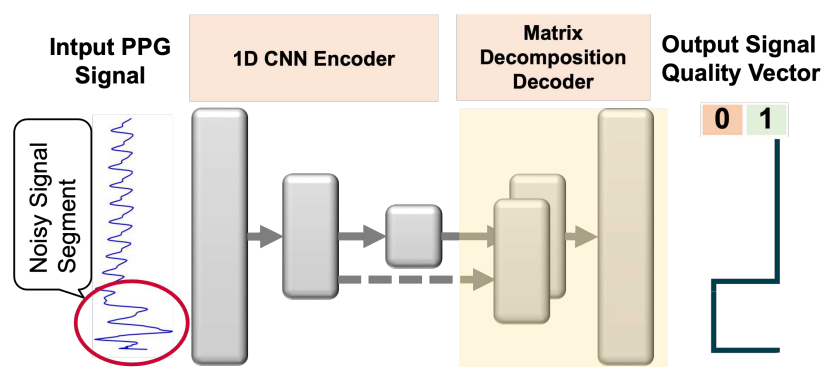

Figure 3. SQA-PhysMD: Signal quality assessment module for PPG signals. Noisy PPG signal segments along with corresponding video frames are eliminated from the iBVP Dataset.

The ground-truth PPG signals of the our iBVP dataset were first assessed with the *SQA-PhysMD* model and then the clean and non-overlapping segments of 30 seconds were each were prepared. Noisy segments along with their corresponding video frames were discarded. If segments of duration less than minimum duration of 20 seconds remained towards the end of signal, these were also discarded. This resulted in a total of 689 video segments and the noise-free ground-truth PPG signals. SQA-PhysMD can be further be used with any existing rPPG dataset to clean the ground-truth signal and thereby eliminate the corresponding video frames.

The resolution of the acquired video frames is 640 x 480 pixels for RGB camera, and 640 x 512 pixels for the thermal camera. As most of the rPPG methods deploy face detection as initial step of the processing pipeline, we prepare the dataset with cropped facial regions. We use Python Facial Expression Analysis Toolbox (Py-Feat) [51] along with RetinaFace [52] to detect facial frame in the RGB images. We then pick a cropping pixel dimension as 256 x 256 as it could contain the largest detected facial frame dimension with margins. We apply this cropping to reduce the overall size of the dataset

without any loss of information in temporal or spatial dimensions. Inspired for a recent work that explored different resolution of images for rPPG estimation using 3D CNN networks (RTrPPG) [22], two additional versions of the dataset are prepared, each with 128 x128 and 64 x 64 resolution. As thermal video frames were well acquired in alignment with the RGB video frames, the same cropping was used to pre-process the thermal video frames. The compiled dataset, therefore can serve as a training dataset as well as a reliable bench-marking dataset for evaluating rPPG methods.

2.6. Comparison with Existing Datasets

Table 1 presents a comparison of different rPPG datasets (non exhaustive), highlighting various aspects of each dataset. The advancements in rPPG research have been made owing to the availability of these datasets. Higher number of participants, and varying scenarios including illumination conditions and tasks performed by the participants offer several advantages including reliable validation of rPPG methods, as well as robust training of supervised rPPG algorithms. The key highlight of the iBVP dataset is its labels that are assessed for the signal quality, making it highly reliable bench-marking dataset as well as a good candidate to train supervised models. Additionally, most of the existing datasets have captured ground-truth PPG signals from finger or wrist, introducing not just a phase difference but also morphological differences [32] with the rPPG signal that is extracted from facial regions. While the phase difference can be easily adjusted, when combined with morphological differences, it can not be optimally synchronized with the facial rPPG signals. iBVP dataset therefore is more suitable for evaluating as well as training the models that estimate PPG signals in contrast to the models trained to estimate heart rate or related metrics with end-to-end approach. For exhaustive discourse and description of different rPPG datasets, it it recommended to refer to a recent review article [6].

Table 1. Comparison of different rPPG datasets

Dataset	Modality	Subjects	Tasks	No. of Videos	Duration (min)	Varying Illumination	SQ Labels	Resolution	Compression	FPS	Free Access
PURE [8]	RGB	10	S, M, T	60	60	Y	N	640 x 480	None	30	Yes
OBF* [53]	RGB, NIR	106	M	200	1000	N	N	640 x 480	None	30	No
MANHOB-HCI [7]	RGB	27	E	527	350	N	N	1040 × 1392	None	24	Yes
MMSE-HR [9]	RGB, 3D Thermal	40	E	102	935	N	N	RGB: 1040 × 1392; Thermal: 640 x 480	None	25	No
VIPL-HR [10]	RGB, NIR	107	S, M, T	3130	1235	Y	N	Face-cropped	MJPG	25	Yes
UBFC-rPPG [11]	RGB	43	S, C	43	86	Y	N	640 x 480	None	30	Yes
UBFC-Phys [12]	RGB	56	S, C, T	168	504	N	N	1024 x 1024	JPEG	35	Yes
iBVP (Ours)	RGB, Thermal	30	B, C, M	689	341 (noise-removed)	N	Y	RGB: 640 x 480; Thermal: 640 x 512	None	30	Yes

B: Rhythmic Breathing; CT: Cognitive tasks; E: Facial expression; M: Head movement; S: Stable; T: Talking; C: Controlled; SQ: Signal Quality. *OBF dataset is temporarily unavailable at the time of submission.

3. Validation of iBVP Dataset

To evaluate the iBVP dataset, we chose the models that can be trained to infer BVP signals in an end-to-end manner. Among such models, 3D-CNN architecture based models including PhysNet3D [23] and RTrPPG [22] were found the most suitable to learn spatio-temporal features from facial video frames. Evaluation of the rPPG models trained with iBVP dataset is performed with an objective to validate the iBVP dataset supporting the use of the dataset as benchmarcking as well as training dataset.

We further introduce a novel 3D-CNN framework, iBVPNet as illustrated in Figure 4. iBVPNet is a fully convolutional architecture designed with an objective of effectively learning spatio-temporal features for BVP estimation. It consists of three blocks, with each block distinctly learning spatial and temporal features. The first block aggregates spatial features, while encoding temporal features. The second block deploys large temporal kernels to encode the long range temporal information. The final block further aggregates the spatial dimension while decoding the temporal features. Below, we describe the experiments and present the preliminary results highlighting the efficacy of the proposed iBVPNet model.
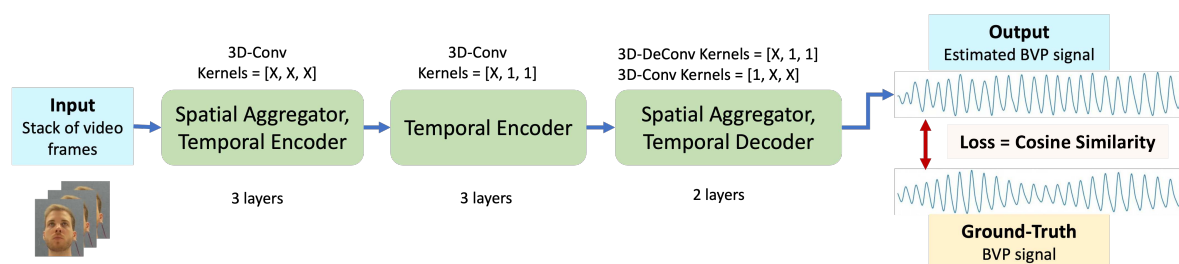

Figure 4. Overview of iBVPNet: Facial region is first cropped for every video frame, and then the frame is resized to 64x64 pixels resolution. The architecture is fully convolutional and comprise three blocks, with first and the last block aggregating the spatial features. Temporal encoding is achieved with first 2 blocks, with second block deploying higher temporal kernels. The final block decodes temporal signal while entirely reducing the spatial dimension.

3.1. Experiments

PhysNet3D [23], RTrPPG [22] and iBVPNet are trained using iBVP dataset with subject-wise 10 fold cross validation approach. Models are separately trained and evaluated for RGB and thermal video frames. In each fold, data of 3 out of 30 participants is left out for validation, and the models are trained with the remaining data of 27 participants. 20 seconds video segment and the corresponding ground-truth PPG signal are used for the training. 600 face-cropped video frames are stacked and provided as input to the models, while the ground-truth PPG signals are resampled to 30 samples per second to match the count of video frames.

Batch size of 8 is used across all the experiments, and the learning rate is initialized to 1e-4, with a step-size of 2 iterations and gamma of 0.95. Models are trained for 100 iterations in each fold, with cosine similarity (CS) loss function. Empirically, CS loss was found to achieve stable convergence in comparison with the negative Pearson correlation , which has been used by earlier works [22,23]. For augmenting the RGB video frames, video AugMix [54,55] to apply transforms that include changes related to contrast, equalization, rotation, shear, translation and brightness. Thermal video frames are augmented using only rotation, shear, and translation transforms.

3.2. Evaluation Metrics

rPPG methods are commonly evaluated for HR measurement [6], whereas the methods aimed at estimating the BVP signals use the metrics that measures similarity between two time series signals. To evaluate the accuracy of HR measurement, widely used metrics include Mean Absolute Error (MAE), Root Mean Square Error and Pearson correlation coefficient [6]. Among the rPPG methods focused on BVP estimation, predominantly used metrics include Template Match Correlation (TMC) [56] and Signal-to Noise-ratio (SNR) [22,23]. The performance of TMC can be affected by the accuracy of segmenting individual pulse waveform from the PPG signals [35,56]. In this work, we propose using the metrics that aligns the two time-series signals without requiring to segment the waveform based on morphological features. Specifically, we compute cross-correlation between the ground-truth PPG signal and the estimated BVP signal at multiple time-lags [24], with an assumption that maximum amplitude of cross correlation (MACC) is achieved at the optimal alignment between the two signals.

We present the evaluation results using MACC and SNR metrics to assess the quality of estimated BVP signals. In addition, we also compute metrics based on HR measurement including RMSE and Pearson correlation between the HR values computed from ground-truth and estimated BVP signals.

3.3. Results

Evaluation metrics are first averaged for each fold out of 10-fold cross-validation and then further averaged across all the folds. Table 2 compares the averaged metrics for different end-to-end rPPG models trained with RGB video frames. For MACC, SNR and RMSE (HR), the proposed iBVPNet shows superior performance, while the PhysNet3D shows the highest correlation for HR. Detailed fold-wise results for models trained with RGB video frames are presented in the Table A1.

Table 2. Performance evaluation for rPPG estimation with RGB frames of iBVP dataset

	MACC (avg)	SNR (avg)	RMSE (HR)	Corr (HR)
PhysNet3D [23]	0.781	0.511	4.283	**0.848**
RTrPPG [22]	0.702	0.283	6.901	0.704
iBVPNet (ours)	**0.784**	**0.677**	**2.717**	0.813

The 33% increase in the SNR for the BVP signals estimated with trained iBVPNet models compared with the existing SOTA method is noteworthy. In Figure 5, we present estimated BVP waveform for rhythmic breathing and head movement conditions to qualitatively compare the outcomes from our proposed models as well as the SOTA methods.
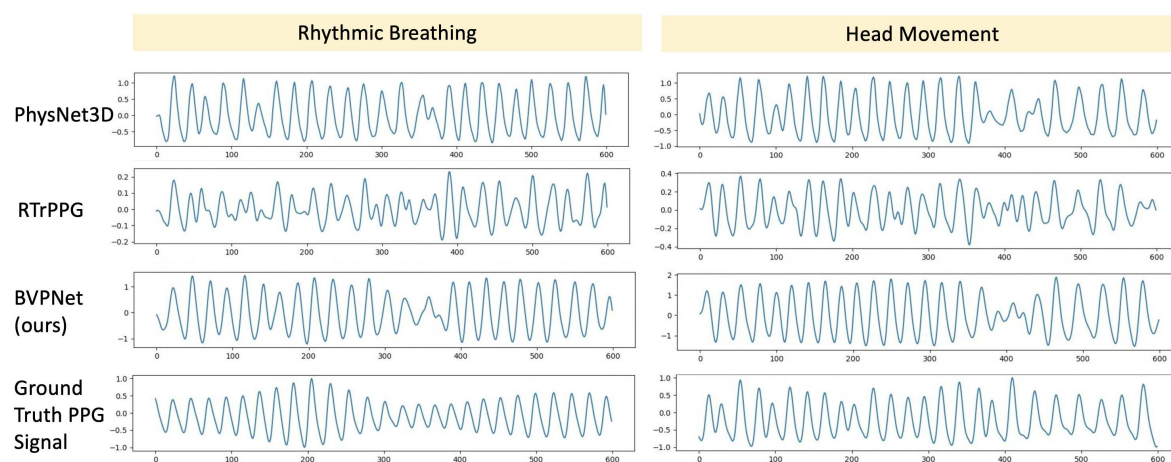

Figure 5. Qualitative comparison of the estimated BVP signals from RGB video frames.

To compare the SOTA models' performance on the iBVP dataset with the same on the existing bench-marking datasets, we have formulated a comparison table based on the data from a recent review article [6]. Table 3 summarizes RMSE and R values of 5 SOTA methods together with our method, which confirms that our newly proposed iBVP dataset is highly compatible with not only the iBVPNet but also existing SOTA methods.

Table 3. Performance comparison of SOTA rPPG methods on existing bench-marking datasets and iBVP datasets (RGB imaging modality only, which is the only common modality in the datasets).
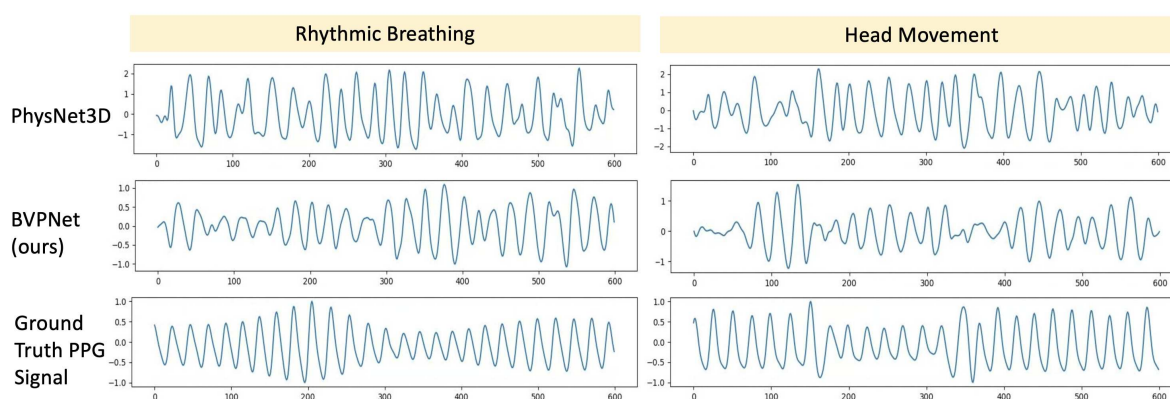
Datasets	rPPG method	RMSE	R
PURE [8]	PhysNet3D [23]	2.60	0.99
	rPPGNet [57]	1.21	1.00
	SAM-rPPGNet [58]	1.21	1.00
MANHOB-HCI [7]	PhysNet3D [23]	8.76	0.69
	rPPGNet [57]	5.93	0.88
VIPL-HR [10]	PhysNet3D [23]	14.80	0.20
	AutoHR [59]	8.68	0.72
iBVP Dataset (ours)	PhysNet3D [23]	4.28	0.85
	RTrPPG [22]	6.90	0.70
	iBVPNet (ours)	2.72	0.81

Note: Only 3D CNN methods that estimate BVP signals are chosen

Lastly, we have performed the same evaluation task on high-temporal resolution infrared thermal image frames, which the iBVP dataset uniquely offers. Table 4 compares the metrics averaged across multiple folds, for different end-to-end rPPG models trained with thermal video frames. Although the iBVPNet showed superior performance across all evaluation metrics as compared with the SOTAs, the overall quality of BVP estimation was not. Figure 6 qualitatively compares the outcomes of different rPPG methods in estimating BVP waveforms from thermal video frames for rhythmic breathing and head movement conditions. This highlights that the BVP information extracted from the thermal frames was not strong. Similar results have been reported in [10] from Near infrared imaging (NIR) based BVP estimation.

Table 4. Performance evaluation for rPPG estimation using thermal frames of iBVP Dataset

	MACC (avg)	SNR (avg)	RMSE (HR)	Corr (HR)
PhysNet3D [23]	0.360	-0.135	6.339	-0.019
iBVPNet (ours)	**0.413**	**0.159**	**5.400**	**0.095**

Figure 6. Qualitative comparison of the estimated BVP signals from Thermal video frames.

4. Discussion and Conclusion

The experiment with the SOTA methods and the proposed iBVPNet model highlights the usefulness of the introduced iBVP dataset in training and validating the rPPG methods. While

most of the rPPG methods estimate HR, it is advantageous to estimate the PPG signal which can then reliably be used to extract various HR and HRV related metrics. The presence of noise artifacts in the ground-truth PPG signals can obscure the training stage for end-to-end supervised models. Here, the SQA-PhysMD implemented in this work for inferring dense signal quality measure for PPG signals (an extended version of *SQA-Phys* [20]) has played a key role in eliminating noisy segments not only from the ground-truth PPG signals but also from the corresponding video frames.

As the SQA-PhysMD can assess signal quality for any types of PPG signals, it can be independently applied to the existing datasets for producing high-resolution signal quality labels as in the iBVP dataset. This can help automatically removing noisy segments in the existing rPPG datasets, reducing tedious manual work and efforts which are otherwise required to be made by researchers [22,60]. Furthermore, the ground-truth PPG signals acquired from the ear lobe closely match the phase and the morphology of the rPPG signals extracted from the facial video frames. Therefore, iBVP dataset can significantly contribute towards improving robustness of rPPG methods.

Some of the existing RGB imaging-based rPPG datasets are available after applying the video compression techniques (e.g., motion JPEG, JPEG). It is noteworthy that the performance of SOTA models can be severely affected owing to the loss of BVP information from the compressed videos [23,58]. To circumvent this, one recent work implemented generative method to reconstruct the original video frames from the compressed video frames as initial step, followed by an architecture to estimate BVP signals [57]. However, this approach adds significant overhead in processing the video frames, and therefore alternative ways are required to address the BVP extraction from the compressed videos. Thus, the iBVP dataset offers raw RGB-Thermal image frames, without the compression methods.

Lastly, aligned with previous findings on rPPG with infrared (IR) video frames [10], we confirm that the current SOTA rPPG methods as well as ours perform poorly on thermal video frames. It is worth noting that thermal video frames require tailored pre-processing since various factors including ambient temperature and quantization methods [24] can significantly impact the rPPG extraction. Further investigation on assessing the potential of thermal infrared imaging in extracting BVP signals is therefore required, to which our dataset can contribute in future work.

Abbreviations

The following abbreviations are used in this manuscript:

1D-CNN	1-dimensional convolutional neural network
BPM	Beats per minute
BVP	Blood volume pulse
ECG	Electrocardiogram
HR	Heart rate
PPG	Photoplethysmography
RGB	Color images with red, green an blue frames

Appendix A. Detailed Results of Multifold Evaluation

Below, we present the fold-wise comparison between different rPPG methods, separately for the models trained on RGB and thermal video frames.

Appendix A.1. RGB

Table A1. Detailed performance evaluation for rPPG estimation with RGB video frames of iBVP dataset

Folds	MACC (avg)			SNR (avg)			RMSE (HR)			Corr (HR)		
	PhysNet3D	RTrPPG	iBVPNet (Ours)	PhysNet3D	RTrPPG	iBVPNet (Ours)	PhysNet3D	RTrPPG	iBVPNet (Ours)	PhysNet3D	RTrPPG	iBVPNet (Ours)
0	0.767	0.669	**0.790**	0.532	0.250	**0.762**	2.829	6.058	**1.476**	0.846	0.568	**0.860**
1	**0.734**	0.654	0.710	0.373	0.190	**0.423**	8.412	12.480	**5.325**	**0.538**	0.258	0.376
2	0.830	0.773	**0.860**	0.709	0.475	**0.972**	2.937	6.213	**1.412**	0.888	0.587	**0.934**
3	**0.718**	0.637	0.660	**0.305**	0.113	0.291	5.848	7.591	**4.542**	**0.800**	0.674	0.679
4	**0.851**	0.763	0.836	0.637	0.402	**0.740**	2.330	3.993	**1.681**	**0.955**	0.879	0.945
5	**0.867**	0.801	0.853	0.601	0.373	**0.808**	2.092	3.508	**1.113**	0.966	0.905	**0.973**
6	0.780	0.689	**0.824**	0.573	0.297	**0.825**	5.114	7.682	**2.342**	0.898	0.826	**0.945**
7	**0.821**	0.751	**0.821**	0.603	0.342	**0.806**	2.943	5.051	**2.652**	**0.903**	0.781	0.830
8	0.702	0.603	**0.744**	0.329	0.113	**0.604**	4.103	11.395	**2.692**	**0.772**	0.655	0.724
9	0.743	0.680	**0.746**	0.445	0.271	**0.535**	6.222	5.044	**3.932**	0.909	**0.911**	0.870

Appendix A.2. Thermal

Table A2. Detailed performance evaluation for rPPG estimation with thermal video frames of iBVP dataset

Folds	MACC (avg)		SNR (avg)		RMSE (HR)		Corr (HR)	
	PhysNet3D	iBVPNet (Ours)	PhysNet3D	iBVPNet (Ours)	PhysNet3D	iBVPNet (Ours)	PhysNet3D	iBVPNet (Ours)
0	0.377	**0.469**	-0.099	**0.363**	6.496	**3.144**	0.092	**0.136**
1	0.352	**0.403**	-0.110	**0.109**	6.932	**5.557**	**0.286**	0.065
2	0.389	**0.437**	-0.071	**0.266**	5.599	**4.731**	**-0.139**	-0.218
3	0.378	**0.409**	-0.151	**0.171**	5.856	**5.037**	0.093	**0.589**
4	0.367	**0.401**	-0.120	**0.138**	5.475	**5.401**	**0.065**	-0.060
5	0.368	**0.442**	-0.149	**0.232**	5.628	**4.856**	-0.141	**-0.046**
6	0.350	**0.430**	-0.114	**0.213**	6.815	**5.865**	0.014	**0.365**
7	0.338	**0.386**	-0.150	**0.113**	**5.453**	6.015	**-0.238**	-0.247
8	0.358	**0.431**	-0.144	**0.264**	6.409	**4.245**	-0.063	**0.238**
9	**0.326**	0.322	**-0.238**	-0.279	**8.732**	9.152	-0.162	**0.129**

References

1. Hertzman, A.B. The Blood Supply of Various Skin Areas as Estimated by the Photoelectric Plethysmograph. *American Journal of Physiology-Legacy Content* **1938**, *124*, 328–340. doi:10.1152/ajplegacy.1938.124.2.328.
2. Verkruysse, W.; Svaasand, L.O.; Nelson, J.S. Remote Plethysmographic Imaging Using Ambient Light. *Optics Express* **2008**, *16*, 21434–21445. doi:10.1364/OE.16.021434.
3. Cho, Y.; Bianchi-Berthouze, N.; Julier, S.J. DeepBreath: Deep Learning of Breathing Patterns for Automatic Stress Recognition Using Low-Cost Thermal Imaging in Unconstrained Settings. 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), 2017, pp. 456–463. doi:10.1109/ACII.2017.8273639.
4. Cho, Y. Rethinking Eye-blink: Assessing Task Difficulty through Physiological Representation of Spontaneous Blinking. Proceedings of the 2021 CHI Conference on Human Factors in Computing

Systems; Association for Computing Machinery: New York, NY, USA, 2021; CHI '21, pp. 1–12. doi:10.1145/3411764.3445577.

5. Reşit Kavsaoğlu, A.; Polat, K.; Recep Bozkurt, M. A Novel Feature Ranking Algorithm for Biometric Recognition with PPG Signals. *Computers in Biology and Medicine* **2014**, *49*, 1–14. doi:10.1016/j.compbiomed.2014.03.005.

6. Xiao, H.; Liu, T.; Sun, Y.; Li, Y.; Zhao, S.; Avolio, A. Remote Photoplethysmography for Heart Rate Measurement: A Review. *Biomedical Signal Processing and Control* **2024**, *88*, 105608. doi:10.1016/j.bspc.2023.105608.

7. Soleymani, M.; Lichtenauer, J.; Pun, T.; Pantic, M. A Multimodal Database for Affect Recognition and Implicit Tagging. *IEEE Transactions on Affective Computing* **2012**, *3*, 42–55. doi:10.1109/T-AFFC.2011.25.

8. Stricker, R.; Müller, S.; Gross, H.M. Non-Contact Video-Based Pulse Rate Measurement on a Mobile Service Robot. The 23rd IEEE International Symposium on Robot and Human Interactive Communication, 2014, pp. 1056–1062. doi:10.1109/ROMAN.2014.6926392.

9. Zhang, Z.; Girard, J.M.; Wu, Y.; Zhang, X.; Liu, P.; Ciftci, U.; Canavan, S.; Reale, M.; Horowitz, A.; Yang, H.; Cohn, J.F.; Ji, Q.; Yin, L. Multimodal Spontaneous Emotion Corpus for Human Behavior Analysis. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3438–3446.

10. Niu, X.; Han, H.; Shan, S.; Chen, X. VIPL-HR: A Multi-modal Database for Pulse Estimation from Less-Constrained Face Video. Computer Vision – ACCV 2018; Jawahar, C.; Li, H.; Mori, G.; Schindler, K., Eds.; Springer International Publishing: Cham, 2019; Lecture Notes in Computer Science, pp. 562–576. doi:10.1007/978-3-030-20873-8_36.

11. Bobbia, S.; Macwan, R.; Benezeth, Y.; Mansouri, A.; Dubois, J. Unsupervised Skin Tissue Segmentation for Remote Photoplethysmography. *Pattern Recognition Letters* **2019**, *124*, 82–90. doi:10.1016/j.patrec.2017.10.017.

12. Sabour, R.M.; Benezeth, Y.; De Oliveira, P.; Chappé, J.; Yang, F. UBFC-Phys: A Multimodal Database For Psychophysiological Studies of Social Stress. *IEEE Transactions on Affective Computing* **2023**, *14*, 622–636. doi:10.1109/TAFFC.2021.3056960.

13. Revanur, A.; Li, Z.; Ciftci, U.A.; Yin, L.; Jeni, L.A. The First Vision for Vitals (V4V) Challenge for Non-Contact Video-Based Physiological Estimation. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2760–2767.

14. McDuff, D.; Wander, M.; Liu, X.; Hill, B.L.; Hernandez, J.; Lester, J.; Baltrusaitis, T. SCAMPS: Synthetics for Camera Measurement of Physiological Signals, 2022, [arxiv:cs/2206.04197]. doi:10.48550/arXiv.2206.04197.

15. Špetlík, R. Visual Heart Rate Estimation with Convolutional Neural Network. Proceedings of the British Machine Vision Conference; The British Machine Vision Association: Newcastle, UK, 2018.

16. Castaneda, D.; Esparza, A.; Ghamari, M.; Soltanpur, C.; Nazeran, H. A Review on Wearable Photoplethysmography Sensors and Their Potential Future Applications in Health Care. *International journal of biosensors & bioelectronics* **2018**, *4*, 195–202. doi:10.15406/ijbsbe.2018.04.00125.

17. Zhang, Z.; Pi, Z.; Liu, B. TROIKA: A General Framework for Heart Rate Monitoring Using Wrist-Type Photoplethysmographic Signals During Intensive Physical Exercise. *IEEE Transactions on Biomedical Engineering* **2015**, *62*, 522–531. doi:10.1109/TBME.2014.2359372.

18. Chuang, C.H.; Chang, K.Y.; Huang, C.S.; Jung, T.P. IC-U-Net: A U-Net-based Denoising Autoencoder Using Mixtures of Independent Components for Automatic EEG Artifact Removal. *NeuroImage* **2022**, *263*, 119586. doi:10.1016/j.neuroimage.2022.119586.

19. Jain, P.; Ding, C.; Rudin, C.; Hu, X. A Self-Supervised Algorithm for Denoising Photoplethysmography Signals for Heart Rate Estimation from Wearables, 2023, [arxiv:cs, eess/2307.05339]. doi:10.48550/arXiv.2307.05339.

20. Joshi, J.; Wang, K.; Cho, Y. PhysioKit: An Open-Source, Low-Cost Physiological Computing Toolkit for Single- and Multi-User Studies. *Sensors* **2023**, *23*, 8244. doi:10.3390/s23198244.

21. Geng, Z.; Guo, M.H.; Chen, H.; Li, X.; Wei, K.; Lin, Z. Is Attention Better Than Matrix Decomposition?, 2021, [arxiv:cs/2109.04553]. doi:10.48550/arXiv.2109.04553.

22. Botina-Monsalve, D.; Benezeth, Y.; Miteran, J. RTrPPG: An Ultra Light 3DCNN for Real-Time Remote Photoplethysmography. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2146–2154.

23. Yu, Z.; Li, X.; Zhao, G. Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks, 2019, [arxiv:cs/1905.02419]. doi:10.48550/arXiv.1905.02419.

24. Cho, Y.; Julier, S.J.; Marquardt, N.; Bianchi-Berthouze, N. Robust Tracking of Respiratory Rate in High-Dynamic Range Scenes Using Mobile Thermal Imaging. *Biomedical Optics Express* **2017**, *8*, 4480–4503. doi:10.1364/BOE.8.004480.

25. Tonacci, A.; Billeci, L.; Burrai, E.; Sansone, F.; Conte, R. Comparative Evaluation of the Autonomic Response to Cognitive and Sensory Stimulations through Wearable Sensors. *Sensors (Basel, Switzerland)* **2019**, *19*, 4661. doi:10.3390/s19214661.

26. Birkett, M.A. The Trier Social Stress Test Protocol for Inducing Psychological Stress. *Journal of Visualized Experiments : JoVE* **2011**, *56*, 3238. doi:10.3791/3238.

27. Cho, Y.; Julier, S.J.; Bianchi-Berthouze, N. Instant Stress: Detection of Perceived Mental Stress Through Smartphone Photoplethysmography and Thermal Imaging. *JMIR Mental Health* **2019**, *6*, e10140. doi:10.2196/10140.

28. Johnson, K.T.; Narain, J.; Ferguson, C.; Picard, R.; Maes, P. The ECHOS Platform to Enhance Communication for Nonverbal Children with Autism: A Case Study. Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems; Association for Computing Machinery: New York, NY, USA, 2020; CHI EA '20, pp. 1–8. doi:10.1145/3334480.3375206.

29. Logitech BRIO Webcam with 4K Ultra HD Video & HDR. https://www.logitech.com/en-gb/products/webcams/brio-4k-hdr-webcam.html.

30. FLIR A65 IR Temperature Sensor | Teledyne FLIR. https://www.flir.co.uk/products/a65?vertical=rd+science&segment=solutions.

31. Casado, C.Á.; López, M.B. Face2PPG: An Unsupervised Pipeline for Blood Volume Pulse Extraction from Faces. *IEEE Journal of Biomedical and Health Informatics* **2023**, pp. 1–12. doi:10.1109/JBHI.2023.3307942.

32. Allen, J.; Murray, A. Effects of Filtering on Multisite Photoplethysmography Pulse Waveform Characteristics. Computers in Cardiology, 2004, 2004, pp. 485–488. doi:10.1109/CIC.2004.1442980.

33. Patterson, J.A.; McIlwraith, D.C.; Yang, G.Z. A Flexible, Low Noise Reflective PPG Sensor Platform for Ear-Worn Heart Rate Monitoring. 2009 Sixth International Workshop on Wearable and Implantable Body Sensor Networks, 2009, pp. 286–291. doi:10.1109/BSN.2009.16.

34. Huang, F.H.; Yuan, P.J.; Lin, K.P.; Chang, H.H.; Tsai, C.L. Analysis of Reflectance Photoplethysmograph Sensors. *International Journal of Biomedical and Biological Engineering* **2011**, *5*, 622–625.

35. Elgendi, M. Optimal Signal Quality Index for Photoplethysmogram Signals. *Bioengineering* **2016**, *3*, 21. doi:10.3390/bioengineering3040021.

36. Sukor, J.A.; Redmond, S.J.; Lovell, N.H. Signal Quality Measures for Pulse Oximetry through Waveform Morphology Analysis. *Physiological Measurement* **2011**, *32*, 369. doi:10.1088/0967-3334/32/3/008.

37. Song, J.; Li, D.; Ma, X.; Teng, G.; Wei, J. PQR Signal Quality Indexes: A Method for Real-Time Photoplethysmogram Signal Quality Estimation Based on Noise Interferences. *Biomedical Signal Processing and Control* **2019**, *47*, 88–95. doi:10.1016/j.bspc.2018.05.020.

38. Goh, C.H.; Tan, L.K.; Lovell, N.H.; Ng, S.C.; Tan, M.P.; Lim, E. Robust PPG Motion Artifact Detection Using a 1-D Convolution Neural Network. *Computer Methods and Programs in Biomedicine* **2020**, *196*, 105596. doi:10.1016/j.cmpb.2020.105596.

39. Gao, H.; Wu, X.; Shi, C.; Gao, Q.; Geng, J. A LSTM-Based Realtime Signal Quality Assessment for Photoplethysmogram and Remote Photoplethysmogram. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3831–3840.

40. Guo, Z.; Ding, C.; Hu, X.; Rudin, C. A Supervised Machine Learning Semantic Segmentation Approach for Detecting Artifacts in Plethysmography Signals from Wearables. *Physiological Measurement* **2021**, *42*, 125003. doi:10.1088/1361-6579/ac3b3d.

41. Roh, D.; Shin, H. Recurrence Plot and Machine Learning for Signal Quality Assessment of Photoplethysmogram in Mobile Environment. *Sensors* **2021**, *21*, 2188. doi:10.3390/s21062188.

42. Zheng, Y.; Wu, C.; Cai, P.; Zhong, Z.; Huang, H.; Jiang, Y. Tiny-PPG: A Lightweight Deep Neural Network for Real-Time Detection of Motion Artifacts in Photoplethysmogram Signals on Edge Devices. *Internet of Things* **2024**, *25*, 101007. doi:10.1016/j.iot.2023.101007.

43. Desquins, T.; Bousefsaf, F.; Pruski, A.; Maaoui, C. A Survey of Photoplethysmography and Imaging Photoplethysmography Quality Assessment Methods. *Applied Sciences* **2022**, *12*, 9582. doi:10.3390/app12199582.

44. Moscato, S.; Lo Giudice, S.; Massaro, G.; Chiari, L. Wrist Photoplethysmography Signal Quality Assessment for Reliable Heart Rate Estimate and Morphological Analysis. *Sensors* **2022**, *22*, 5831. doi:10.3390/s22155831.

45. Shin, H. Deep Convolutional Neural Network-Based Signal Quality Assessment for Photoplethysmogram. *Computers in Biology and Medicine* **2022**, *145*, 105430. doi:10.1016/j.compbiomed.2022.105430.

46. Feli, M.; Azimi, I.; Anzanpour, A.; Rahmani, A.M.; Liljeberg, P. An Energy-Efficient Semi-Supervised Approach for on-Device Photoplethysmogram Signal Quality Assessment. *Smart Health* **2023**, *28*, 100390. doi:10.1016/j.smhl.2023.100390.

47. Pereira, T.; Ding, C.; Gadhoumi, K.; Tran, N.; Colorado, R.A.; Meisel, K.; Hu, X. Deep Learning Approaches for Plethysmography Signal Quality Assessment in the Presence of Atrial Fibrillation. *Physiological Measurement* **2019**, *40*, 125002. doi:10.1088/1361-6579/ab5b84.

48. Reiss, A.; Indlekofer, I.; Schmidt, P.; Van Laerhoven, K. Deep PPG: Large-Scale Heart Rate Estimation with Convolutional Neural Networks. *Sensors* **2019**, *19*, 3079. doi:10.3390/s19143079.

49. Schmidt, P.; Reiss, A.; Duerichen, R.; Marberger, C.; Van Laerhoven, K. Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. Proceedings of the 20th ACM International Conference on Multimodal Interaction; Association for Computing Machinery: New York, NY, USA, 2018; ICMI '18, pp. 400–408. doi:10.1145/3242969.3242985.

50. Stark, Z. Chengstark/Segade, 2024.

51. Py-Feat: Python Facial Expression Analysis Toolbox — Py-Feat. https://py-feat.org/pages/intro.html#.

52. Deng, J.; Guo, J.; Ververas, E.; Kotsia, I.; Zafeiriou, S. RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5203–5212.

53. Li, X.; Alikhani, I.; Shi, J.; Seppanen, T.; Junttila, J.; Majamaa-Voltti, K.; Tulppo, M.; Zhao, G. The OBF Database: A Large Face Video Database for Remote Physiological Signal Measurement and Atrial Fibrillation Detection. 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), 2018, pp. 242–249. doi:10.1109/FG.2018.00043.

54. Hendrycks, D.; Mu, N.; Cubuk, E.D.; Zoph, B.; Gilmer, J.; Lakshminarayanan, B. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty, 2020, [arxiv:cs, stat/1912.02781]. doi:10.48550/arXiv.1912.02781.

55. Pytorchvideo.Transforms — PyTorchVideo Documentation. https://pytorchvideo.readthedocs.io/en/latest/api/transforms/transforms.html.

56. Orphanidou, C.; Bonnici, T.; Charlton, P.; Clifton, D.; Vallance, D.; Tarassenko, L. Signal-Quality Indices for the Electrocardiogram and Photoplethysmogram: Derivation and Applications to Wireless Monitoring. *IEEE Journal of Biomedical and Health Informatics* **2015**, *19*, 832–838. doi:10.1109/JBHI.2014.2338351.

57. Yu, Z.; Peng, W.; Li, X.; Hong, X.; Zhao, G. Remote Heart Rate Measurement From Highly Compressed Facial Videos: An End-to-End Deep Learning Solution With Video Enhancement. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 151–160.

58. Hu, M.; Qian, F.; Wang, X.; He, L.; Guo, D.; Ren, F. Robust Heart Rate Estimation With Spatial–Temporal Attention Network From Facial Videos. *IEEE Transactions on Cognitive and Developmental Systems* **2022**, *14*, 639–647. doi:10.1109/TCDS.2021.3062370.

59. Yu, Z.; Li, X.; Niu, X.; Shi, J.; Zhao, G. AutoHR: A Strong End-to-End Baseline for Remote Heart Rate Measurement With Neural Searching. *IEEE Signal Processing Letters* **2020**, *27*, 1245–1249. doi:10.1109/LSP.2020.3007086.

60. Deividbotina-Alv/Rtrppg: Python Implementation of the 3DCNN-based Real-Time rPPG Network (RTrPPG). https://github.com/deividbotina-alv/rtrppg.