PULSE: Physiological Understanding with Liquid Signal Extraction

Shahzad Ahmad[1] shahzaa@hiof.no
Sania Bano[2] sania.22eez0012@iitrpr.ac.in
Sachin Verma[3] sachin.verma@ntnu.no
Yogesh Singh Rawat[4] yogesh@crcv.ucf.edu
Sukalpa Chanda[1] sukalpa@ieee.org
Santosh Kumar Vipparthi[2] skvipparthi@iitrpr.ac.in
Subrahmanyam Murala[3] muralas@tcd.ie
[1]Østfold University College, Norway [2]Indian Institute of Technology Ropar, India [5]Trinity College Dublin, Ireland
[3]Norwegian University of Science and Technology, Norway [4]University of Central Florida, USA

Abstract

The non-contact estimation of vital signs, particularly heart rate, from video data is a promising method for remote health monitoring. 3D convolutional layers are widely used for this task due to their ability to capture both spatial and temporal features. However, traditional 3D convolutions, while effective in many cases, lack the capacity to adjust dynamically to the temporal variability inherent in physiological signals such as remote photoplethysmography (rPPG), which are characterized by subtle frequency changes over time. To address this, we propose ***PULSE (Physiological Understanding with Liquid Signal Extraction)****, a framework that employs Liquid Time-Constant (LTC) models with 3D convolutional layers to enhance temporal sensitivity and improve the extraction of these fine-grained rPPG signals. In PULSE, traditional 3D-conv layers are deployed for initial feature extraction, while LTC-based 3D-conv layers dynamically adapt and guide the temporal processing, allowing the model to better track and interpret the subtle variations in heart rate signals under different conditions, such as motion artifacts and lighting changes. We evaluated the effectiveness of PULSE in an unsupervised training setting, demonstrating that our solution performs well even in the absence of labeled datasets a common challenge in rPPG signal extraction. Experimental evaluations on three public datasets confirm that PULSE achieves comparable or superior results to existing methods, proving its robustness and efficacy for real-world, non-contact health monitoring applications.*

1. Introduction

Monitoring vital signs such as heart rate (HR), respiratory frequency (RF), and heart rate variability (HRV) is crucial in healthcare and wellness domains [42, 56]. Traditionally, these signals are measured using skin-contact sensors like photoplethysmography (PPG) and electrocardiography (ECG), which track blood volume changes and electrical activity in the body [2, 22]. Although accurate, these methods often require specialized equipment such as pulse oximeters or ECG monitors, which can be cumbersome and uncomfortable for continuous use [30]. Issues such as skin irritation from ECG electrodes and the unsuitability of pulse oximeters for long-term monitoring, especially in active patients, have driven interest in non-invasive alternatives [33, 58]. These alternatives are particularly valuable in telemedicine and remote monitoring, where real-time data is critical but constant sensor contact is impractical [6, 15].

Camera-based remote photoplethysmography (rPPG) has emerged as a promising non-contact method for monitoring vital signs by analyzing subtle color changes in facial videos caused by blood flow [42, 56]. This approach can be easily integrated into consumer electronics like smartphones and webcams, expanding its potential applications in healthcare and well-being. However, the development of accurate rPPG extraction models is challenging due to noise from lighting variations [10], head movements, and skin tone differences, as well as the scarcity of large, diverse datasets with synchronized video and physiological recordings.

To address these challenges, researchers have explored various solutions [17, 44, 48, 51] , with 3D Convolutional Neural Networks (3D CNN) becoming widely adopted in video-based rPPG signal extraction [17, 35, 43, 59] . 3D CNN are powerful tools for capturing both spatial and temporal features simultaneously, making them a natural choice for rPPG, where temporal changes in pixel intensity are directly related to physiological signals like heart rate. These layers allow the model to process video sequences holistically, extracting spatio-temporal patterns critical for accurate rPPG signal estimation.

Despite their success, traditional 3D CNN face limitations when applied to rPPG signal extraction. The fixed temporal receptive fields of 3D-conv layers makes it difficult for them to adapt to the subtle and variable nature of physiological signals . rPPG signals vary in frequency and

amplitude over time, exhibiting both quick fluctuations over short periods and slower trends over longer periods, adding another layer of complexity to modeling them. Traditional 3D-conv layers may struggle to capture these fine-grained, dynamic temporal patterns, leading to reduced accuracy in scenarios where lighting, head motion, or other external factors introduce noise [9, 27, 45, 54].

We propose the PULSE (Physiological Understanding with Liquid Signal Extraction) framework, which integrates Liquid Time-Constant (LTC) technique [18] alongside 3D CNNs to improve rPPG signal extraction. In this hybrid approach, traditional 3D-conv layers handle initial spatio-temporal feature extraction, while LTC-based 3D-conv layers, designed specifically for processing time-varying data, guide the temporal adaptation. This combination allows the model to effectively capture diverse temporal dependencies, making it more robust to variations in rPPG signals, such as heart rate fluctuations, even under challenging conditions like head movements or lighting changes. By incorporating LTC-based 3D-conv layers in the final stages, PULSE enhances the system's ability to track subtle physiological signals while preserving the spatio-temporal structure extracted by the initial 3D-conv layers.

We evaluated the effectiveness of PULSE in an unsupervised training setting using the SiNC [44] framework, demonstrating that our solution performs well even in the absence of large labeled datasets, a common challenge in rPPG signal extraction. Experimental evaluations on three public datasets confirm that PULSE achieves comparable or superior results to existing methods, proving its robustness and efficacy for real-world, non-contact health monitoring applications.

The motivation behind PULSE is to address the inherent temporal variability in rPPG signals by dynamically adapting to changing temporal patterns while maintaining robust spatial feature extraction. Through extensive experiments, we show that PULSE improves the accuracy and stability of rPPG signal extraction, paving the way for future innovations in remote health monitoring and telemedicine.

Our contributions can be summarized as follows:

- We introduce the PULSE framework, which combines 3D-conv layer based blocks with Liquid Time-Constant (LTC) 3D-convolutional layer based block . This hybrid approach leverages the strengths of 3D-convolutional layers for initial spatio-temporal feature extraction and LTC-based 3D-convolutional layers for dynamic temporal adaptation, enhancing the model's ability to accurately estimate rPPG signals under varying conditions such as head movements and lighting changes.
- Extensive experiments on three public datasets demonstrate that our approach achieves comparable or superior results to state-of-the-art methods, significantly improving camera-based heart-rate estimation.

2. Related Work

Remote Photoplethysmography (rPPG):
Conventional rPPG techniques [11, 25, 49, 50, 53] estimate pulse signals from facial videos by detecting and analyzing slight skin color changes caused by the heartbeat. Remote pulse estimation methods have evolved significantly, progressing from blind source separation [39, 40] to linear color transformations [12, 40, 52, 53], and more recently, to supervised deep learning models [9, 23, 27, 29, 34, 35, 43, 59, 60, 62]. To evaluate their effectiveness, rPPG datasets have been developed to include various interferences such as head motion [47], facial expressions [24, 49], video compression [20], and skin tone variations [55]. Additionally, rPPG estimation can be performed using pre-processed representations like normalized differences [9, 26] and spatial-temporal maps [29]. To further improve robustness, self-adaptive [9] and background-guided [37] attention mechanisms have been introduced to emphasize important facial regions in the physiological representation. In deep learning-based remote photoplethysmography (rPPG) measurement, diverse architectures have been utilized, including 2D convolutional neural networks (2DCNN) that use consecutive video frames as input [9, 27, 37, 46], spatial-temporal signal maps [29, 35, 36], and more recently, 3DCNN-based methods [17, 59] designed for optimal performance on compressed videos.

Unsupervised Learning for rPPG:
Recent unsupervised rPPG approaches leverage a contrastive learning framework [4, 8, 32], training models with video pairs to minimize prediction distances for similar videos and maximize them for dissimilar ones. Gideon et al. [17] introduced a contrastive method incorporating frequency resampling for negative samples, calculating mean square errors between power spectral densities, but their reliance on known resampled frequencies undermines accuracy. Yuzhe et al. [57] modified the InfoNCE loss [38] by adding a resampling factor to adjust pair similarity based on sampling rates, although their framework still needs post-self-supervised fine-tuning with PPG labels. Unlike Gideon's approach, Contrast-Phys [48] and SLF-RPM [51] consider all non-anchor samples as negatives, facing similar challenges due to potential pulse rate similarities among individuals. Yue et al. [61] proposed an advanced self-supervised contrastive framework featuring a learnable frequency augmentation module, local rPPG expert aggregation, and frequency-inspired losses. In contrast, Speth et al. [44] developed the SiNC framework, a non-contrastive method that employs periodic signal priors and frequency domain filtering as a loss function, presenting a novel alternative to traditional contrastive techniques.

Liquid Neural Network:
Liquid Neural Networks (LNNs), introduced by Hasani et

al. [1], represent a significant advancement in neural network architecture, drawing inspiration from the nervous system of the C. elegans nematode. These networks are designed to process time-series data more effectively than traditional neural networks, making them particularly valuable for applications involving continuous sequential information [18, 41]. This approach builds upon recent advancements in neural network architectures designed to handle temporal data. For instance, recurrent neural networks (RNNs) and long short-term memory (LSTM) networks have shown promise in capturing temporal dependencies in various time series analysis tasks [21].LNNs offer several technical advantages over conventional neural networks. They operate in continuous time, potentially capturing subtle temporal variations that discrete-time models might miss to potentially capturing subtle temporal variations in blood flow that are crucial for rPPG. Their sparse connectivity, utilizing fewer but more expressive neurons, can lead to more efficient processing. Additionally, LNNs exhibit improved interpretability, allowing for easier understanding of the network's decision-making process [5, 19, 41]. LNNs are capable to learn "on the job" adapting beyond the initial training phase to handle changing conditions and real-time data more effectively [41]. This adaptability makes LNNs particularly well-suited for tasks such as processing and forecasting time series data, image and video processing, and natural language understanding [18]. The key innovation of LNNs lies in their Liquid Time Constant (LTC) model, which allows each artificial neuron's time constant to adapt based on input data, enabling real-time adjustment of network dynamics [41]. The LTC module enables the network to operate in continuous time, which is critical for capturing the fine-grained temporal dynamics often missed by discrete-time models. This continuous-time operation allows the LTC module to maintain a persistent memory of past inputs, which is essential for tasks that require long-term dependencies and real-time adaptability, such as physiological signal processing, including remote photoplethysmography (rPPG).

3. Proposed Method

Our goal is to extract remote photoplethysmography (rPPG) signals from facial videos using a fully unsupervised framework, without relying on labeled data. We present **PULSE (Physiological Understanding with Liquid Signal Extraction)**, an architecture designed to capture both spatial and temporal features from video frames for robust rPPG signal estimation. PULSE is composed of two main types of blocks: the first utilizes traditional 3D-convolutional layers for initial spatio-temporal feature extraction, while the second employs LTC-based 3D-convolutional layers to dynamically model temporal dependencies. For signal prediction, we use a frequency-domain loss function from the SiNC [44] framework . The input to the model is an RGB facial video $V \in \mathbb{R}^{C\times W\times H\times T}$, where C is the number of color channels, $W \times H$ are the frame dimensions, and T is the number of frames. The output of the model $\mathcal{M}$ is a predicted rPPG waveform $Y = \mathcal{M}(V)$, where $Y \in \mathbb{R}^T$, representing the estimated signal across the T frames.

3.1. Overview of PULSE Architecture

The PULSE architecture leverages a combination of blocks with 3D-convolutional layers and block with Liquid Time-Constant (LTC) 3D-convolutional layers to ensure robust spatio-temporal feature extraction and temporal consistency, as shown in Figure 1.

3D-Convolutional Layer Based Blocks: The input facial video is processed through a series of spatial downsampling blocks (B_1, B_2, B_3, B_4) using a 3D-CNN architecture inspired by [43]. These blocks reduce the spatial dimensions while preserving temporal resolution, enabling efficient extraction of abstract spatial features from the input frames.

LTC 3D-Convolutional Layer Based Block: After the spatial downsampling blocks, the model transitions to LTC-based convolutional block B_{LTC}, where LTC-based 3D convolution layers replace standard 3D convolutions. This block capture temporal patterns across various time scales, enabling the model to track subtle changes crucial for rPPG signal estimation. The concept of LTC-based 3D convolution layers is explained further in section 3.2.

Motivation for Design: PULSE is designed to capture both spatial and temporal information for effective rPPG signal estimation. The initial 3D-conv blocks (B_1, B_2, B_3, B_4) extract stable spatial features, reducing variability early on. The LTC-based block B_{LTC} then refine the temporal aspects, enabling the model to adapt to fluctuations in the physiological signals while maintaining the spatio-temporal structure from the earlier layers.

3.2. Liquid Time Constant (LTC) based 3D Convolutional Layer

To capture the varying temporal patterns inherent in physiological signals, we integrate a Liquid Time Constant (LTC) model with 3D convolutional layers. Before delving into the details, we provide a brief introduction to the underlying principles of the LTC model.

Concept Behind the LTC Model:

The Liquid Time-Constant (LTC) model describes the temporal evolution of the hidden state $h(t)$ in response to inputs and internal state dynamics using an ordinary differential equation (ODE) [16]

$$\tau \frac{dh}{dt} = -h + I_{\text{input}} \tag{1}$$

where:

- τ is the time constant, representing how quickly the potential decays.
- h is hidden state of model.

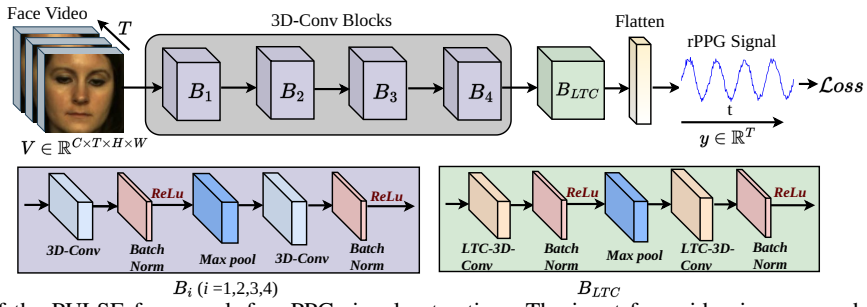


Figure 1. Overview of the PULSE framework for rPPG signal extraction. The input face video is processed through a series of 3D-convolutional blocks (B_i, $i = 1, 2, 3, 4$), where each block performs spatial downsampling while maintaining the temporal dimension. Following this, the LTC-based 3D-convolutional block B_{LTC} are applied to capture and stabilize temporal dynamics, ensuring robust waveform reconstruction in the temporal domain. Finally, the extracted features are flattened and used to predict the rPPG signal over time, which is then optimized using a loss function to enhance accuracy. The details of the LTC-based convolutional layers are explained in Figure 2.

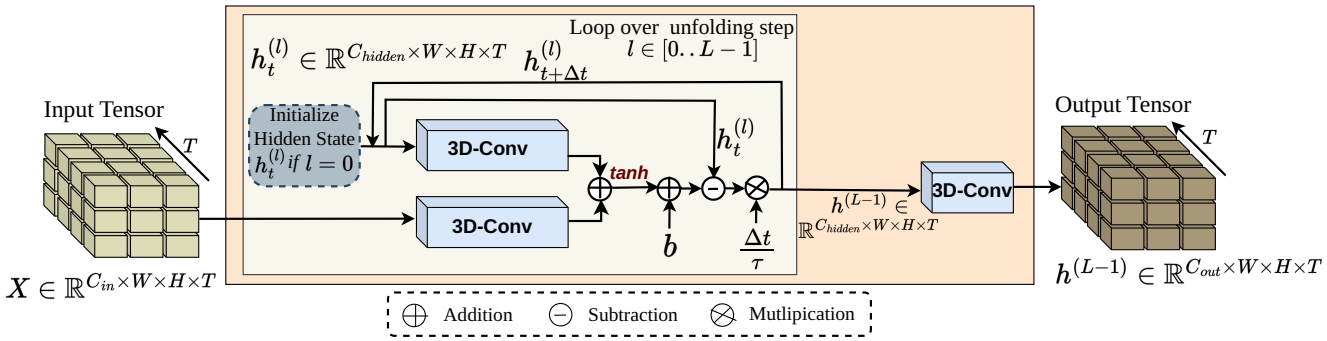


Figure 2. Overview of the LTC-based 3D Convolutional Layer. The input tensor $\mathbf{X} \in \mathbb{R}^{C_{in}\times W\times H\times T}$ is processed through multiple unfolding steps l (ranging from 0 to $L-1$). At each step, the hidden state $\mathbf{h}_t^{(l)}$ is updated using 3D convolution operations combined with the Liquid Time-Constant (LTC) model, as described in Algorithm 1. The process begins by initializing the hidden state $\mathbf{h}_t^{(0)}$. Each subsequent hidden state $\mathbf{h}_{t+\Delta t}^{(l)}$ is computed by applying 3D convolutions to the input tensor and the previous hidden state, followed by a `tanh` nonlinearity. After all steps, the final hidden state $\mathbf{h}^{(L-1)} \in \mathbb{R}^{C_{hidden}\times W\times H\times T}$ undergoes an additional 3D convolution to produce the output tensor $\mathbf{h}^{(L-1)} \in \mathbb{R}^{C_{out}\times W\times H\times T}$.

- I_{input} is the input current driving the neuron.

Incorporating External Inputs and Internal States:
In a more general neural network model, the input current I_{input} can be decomposed into contributions from external inputs x and recurrent connections from other neurons (or from the same neuron):

$$I_{\text{input}} = W_{ih}x + W_{hh}h + b \tag{2}$$

where: $W_{ih}x$ represents the weighted input, $W_{hh}h$ represents the weighted recurrent input from the neuron's own state or other neurons' states and b is a bias term. Substituting this into the equation 1, we get:

$$\tau\frac{dh}{dt} = -h + W_{ih}x + W_{hh}h + b \tag{3}$$

This equation can be generalized as a function $f(x, h)$:

$$f(x, h) = W_{ih}x + W_{hh}h \tag{4}$$

Thus, the equation becomes:

$$\tau\frac{dh}{dt} = -h + f(x, h) + b \tag{5}$$

Using Euler's Method numerical technique for solving ordinary differential equations (ODEs)

$$h_{n+\Delta t} = h_n + \frac{\Delta t}{\tau}\left(-h_n + f(x, h_n) + b\right). \tag{6}$$

See detail derivation in Supplementary(Section: 1.1)

Integration of LTC-based Model with 3D Convolutional Layer:
The LTC-based 3D convolutional layer in our model is essential for effectively capturing both spatial and temporal dynamics. This layer applies 3D convolutional operations across the spatial dimensions (height and width) and the temporal dimension. These operations serve as a replacement for traditional linear transformations $W_{ih}\mathbf{X}$ and $W_{hh}\mathbf{h}$, utilizing 3D convolutions instead, as depicted in Figure 2.

The nonlinear transformation function $f(\mathbf{X}, \mathbf{h})$ within this layer is defined by:

$$f(\mathbf{X}, \mathbf{h}) = \tanh(\text{conv3d}(\mathbf{X}, \mathbf{W}_{ih}) + \text{conv3d}(\mathbf{h}, \mathbf{W}_{hh})) \tag{7}$$

where the 3D convolution operations are applied to the input tensor $\mathbf{X} = \{x_0, x_1, \ldots, x_{T-1}\} \in \mathbb{R}^{C_{in}\times W\times H\times T}$ with T frames. The hidden states of the LTC model at each time step $t \in [0, T-1]$ are represented by $\mathbf{h} =$

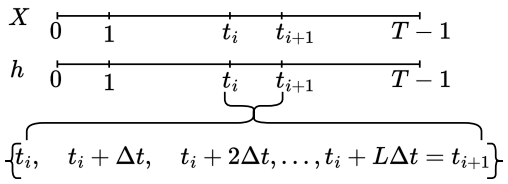
$\{h_0, h_1, \ldots, h_{T-1}\} \in \mathbb{R}^{C_{hidden} \times W \times H \times T}$, with the corresponding weights $\mathbf{W}_{ih}$ and $\mathbf{W}_{hh}$.

The evolution of the hidden state in the LTC convolutional layer is governed by the following ordinary differential equation (ODE):

$$\tau \frac{d\mathbf{h}}{dt} = -\mathbf{h} + \tanh(\text{conv3d}(\mathbf{X}, \mathbf{W}_{ih}) + \text{conv3d}(\mathbf{h}, \mathbf{W}_{hh})) + \mathbf{b} \tag{8}$$

where τ is the time constant, $\mathbf{b}$ is the bias term, and $\mathbf{h}$ represents the hidden state.

Solving Eq:8 analytically is challenging due to the nonlinearity introduced by the LTC model [18]. However, the state of the system at any time point t can be determined using a numerical ODE solver, which simulates the system starting from an initial state h_0 to h_{T-1}. The ODE solver discretizes the continuous time interval $[0, T-1]$ into discrete steps $[t_0, t_1, \ldots, t_{n-1}]$, with the task of updating the hidden states from t_i to t_{i+1}. Let $\mathbf{L}$ represent the number of unfolding (discretization) steps that the solver needs to process within each unit time interval $[t_i, t_{i+1}]$, such that $t_{i+1} = t_i + \mathbf{L}\Delta t$, where Δt is the unfolding time step. Using Euler's method to numerically solve this ODE, the hidden states $\mathbf{h}_t$ where $\forall t \in [0, T-1]$ are iteratively updated at l^{th} unfolding time step as follows:



$$\mathbf{h}_{t+\Delta t}^{(l)} = \mathbf{h}_t^{(l)} + \frac{\Delta t}{\tau}\Big(-\mathbf{h}_t^{(l)} + \tanh(\text{conv3d}(\mathbf{X}, \mathbf{W}_{ih}) + \text{conv3d}(\mathbf{h}_t^{(l)}, \mathbf{W}_{hh})) + \mathbf{b}\Big) \tag{9}$$

Algorithm 1 details the iterative update process for the LTC-based 3D Convolutional Layer. The input sequence $\mathbf{X} = \{x_0, x_1, \ldots, x_{T-1}\} \in \mathbb{R}^{C_{in} \times W \times H \times T}$ is processed to produce the final hidden states $\mathbf{h}^{(L-1)} \in \mathbb{R}^{C_{out} \times W \times H \times T}$. The computational complexity of the LTC Convolutional Layer for an input sequence of length T is $O(\mathbf{L} \times T)$, where $\mathbf{L}$ is the number of discretization steps per unit time.

4. Loss Functions

We employ the frequency domain loss functions from SiNC [44], which are designed to guide the model in capturing and preserving essential frequency characteristics during training.

$$\mathcal{L} = \mathcal{L}_b + \mathcal{L}_s + \mathcal{L}_v \tag{10}$$

where, $\mathcal{L}_b$ = Bandwidth Loss , $\mathcal{L}_s$ = Sparsity Loss , $\mathcal{L}_v$ = Variance Loss

$$\mathcal{L}_b = \frac{1}{\sum_{i=-\infty}^{\infty} F_i}\left[\sum_{i=-\infty}^{a} F_i + \sum_{i=b}^{\infty} F_i\right] \tag{11}$$

Algorithm 1: Iterative Update Rule for the LTC-based 3D Convolutional Layer Using Euler's Method

Input: Initial hidden state $\mathbf{h}_0$, time constant τ, LTC bias $\mathbf{b}$, unfolding time step Δt,
$\mathbf{L}$ = Number of unfolding steps, convolutional weights $\mathbf{W}_{ih}$, $\mathbf{W}_{hh}$, $\mathbf{W}_{out}$, input sequence $\mathbf{X} = \{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{T-1}\} \in \mathbb{R}^{C_{in} \times W \times H \times T}$
Output: Final hidden states $\mathbf{h}^{(L-1)} \in \mathbb{R}^{C_{out} \times W \times H \times T}$
Initialization:
Set the initial hidden states:
$\mathbf{h} = \{h_0, h_1, \ldots, h_{T-1}\} \in \mathbb{R}^{C_{hidden} \times W \times H \times T}$
for $l = 0$ ***to*** $\mathbf{L} - 1$ **do**
  Compute the next hidden state $\forall t \in [0, T-1]$**:**
  $\mathbf{h}_{t+\Delta t}^{(l)} = \mathbf{h}_t^{(l)} + \frac{\Delta t}{\tau}\Big(-\mathbf{h}_t^{(l)} + \tanh\Big(\text{conv3d}(\mathbf{X}, \mathbf{W}_{ih}) + \text{conv3d}(\mathbf{h}_t^{(l)}, \mathbf{W}_{hh}) + \mathbf{b}\Big)\Big)$
end
$\mathbf{h}^{(L-1)} = \text{conv3d}(\mathbf{h}^{(L-1)}, \mathbf{W}_{out})$
return $\mathbf{h}^{(L-1)}$

$$\mathcal{L}_s = \frac{1}{\sum_{i=a}^{b} F_i}\left[\sum_{i=a}^{F^*-\Delta_F} F_i + \sum_{i=F^*+\Delta_F}^{b} F_i\right] \tag{12}$$

$$\mathcal{L}_v = \frac{1}{d}\sum_{i=1}^{d}(CDF_i(Q) - (CDF_i(P))^2 \tag{13}$$

$F_i = \text{FFT}(Y)$ be the power in the i-th frequency bin of the predicted signal. Here, $a = 0.66$ Hz and $b = 3$ Hz. F^* represents argmax(F), and Δ_F denotes the frequencies of the spectral peak and the padding around the peak, respectively. CDF refers to the cumulative distribution function.

5. Gradient Flow of the Loss in the LTC Convolution

The gradient of the loss $\mathcal{L}$ with respect to W_{ih},W_{hh}, τ and b can be computed from Eq:9

Gradient of Loss with respect to W_{ih}:

$$\frac{\partial \mathcal{L}}{\partial W_{ih}} = \sum_{l=0}^{L-1} \frac{\partial \mathcal{L}}{\partial h_t^{(l+1)}} \cdot \frac{\partial h_t^{(l+1)}}{\partial W_{ih}} \tag{14}$$

Gradient of Loss with respect to W_{hhV}:

$$\frac{\partial \mathcal{L}}{\partial W_{hh}} = \sum_{l=0}^{L-1} \frac{\partial \mathcal{L}}{\partial h_t^{(l+1)}} \cdot \frac{\partial h_t^{(l+1)}}{\partial W_{hh}} \tag{15}$$

Gradient of Loss with respect to τ:

$$\frac{\partial \mathcal{L}}{\partial \tau} = \sum_{l=0}^{L-1} \frac{\partial \mathcal{L}}{\partial h_t^{(l+1)}} \cdot \frac{\partial h_t^{(l+1)}}{\partial \tau} \tag{16}$$

Gradient of Loss with respect to b:

$$\frac{\partial \mathcal{L}}{\partial b} = \sum_{l=0}^{L-1} \frac{\partial \mathcal{L}}{\partial h_t^{(l+1)}} \cdot \frac{\partial h_t^{(l+1)}}{\partial b} \tag{17}$$

For detail derivation of loss gradient flow see Supplementary(Section: 1.2).

6. Training Details

Datasets:
We evaluated our proposed method using three well-established benchmarks for rPPG signal estimation **PURE** [47], **UBFC-rPPG** [3], and **COHFACE** [20]. The PURE dataset comprises 60 face videos from 10 subjects, with each recording lasting around one minute over six sessions. These sessions encompass various head movements, such as steady motion, talking, slow and fast head translations, as well as small and medium head rotations. The videos are recorded at 640×480 resolution and 30 FPS. The UBFC-rPPG dataset contains 42 one-minute face videos of participants engaged in a time-constrained math game. The videos also recorded at 640×480 resolution and 30 FPS, are accompanied by simultaneous PPG signals and heart rate recordings. Lastly, the COHFACE dataset features 160 one-minute videos from 40 subjects, captured under both studio and natural lighting conditions. The videos are compressed using the MPEG-4 Visual codec, which, as highlighted by [31], may degrade the quality of rPPG signals. These videos are recorded at a resolution of 640×480 pixels and a frame rate of 20 FPS.

Data Preprocessing:
For preparing the video clips, we employed RetinaFace [13] to detect and crop faces from each frame, producing 64×64 cropped facial images. Rather than cropping faces based on each predicted bounding box individually, we determined the maximum window across all possible bounding boxes within each frame. This method helps to prevent the artificial jerks that can occur when faces are cropped separately for each frame and stacked together. By adopting this approach, we achieve smoother transitions between frames, thereby enhancing the overall quality of the input images fed into our model.

Augmentations:
We applied the same augmentation techniques as SiNC [44], including Gaussian noise addition and brightness adjustment for image intensity augmentation. Spatial augmentation involved random flips and cropping, followed by interpolation. Temporal augmentation used random sequence flipping and frame rate adjustments, along with resampling to handle temporal variations. These strategies enhance model performance and robustness against diverse, noisy inputs.

Implementation details:
For implementation, the model was trained using a Quadro RTX 8000 GPU for 200 epochs with a batch size of 20. PyTorch served as the development framework, and the AdamW optimizer [28] was used with a learning rate of 10^{-4}. To ensure consistent input processing, a clip length of $T = 120$ frames (4 seconds) was utilized, adjusting the input signal to achieve a frequency resolution of 0.33 beats per minute (bpm). We use a 5-fold cross-validation approach for all three datasets, following the same fold configuration as in [17]. This approach involved using three folds for training, one for validation, and one for testing, instead of separate training and testing sets. To improve model robustness, we trained three models with different initializations, resulting in a total of 15 models across the three datasets. The results provide both the mean and the standard deviation of the errors.

Evaluation Metrics:
We assess heart rate accuracy using mean absolute error (MAE), root mean squared error (RMSE), and Pearson correlation coefficient (r). MAE and RMSE are measured in beats per minute (bpm), with smaller values indicating lower errors. Conversely, a higher Pearson correlation coefficient (r), close to one, reflects lower errors. Detailed information on evaluation metrics can be found in our supplementary material (section: 1.4).

7. Results & Analysis

7.1. Intra-Dataset Evaluation

Table 1 presents the intra-dataset results for various approaches, including traditional, supervised, and unsupervised methods. Our model, PULSE, surpasses the state-of-the-art unsupervised non-contrastive method SiNC [44], achieving the best results among all unsupervised methods and performing on par with the leading supervised approaches. The proposed PULSE recorded the lowest Mean Absolute Error (MAE) among unsupervised methods, with a Pearson correlation coefficient (r) close to 1 for all three datasets PURE , UBFC-rPPG and COHFACE, underscoring its exceptional accuracy and reliability.
Figure 3 presents the rPPG waveform generated by PULSE alongside the ground truth rPPG signal and corresponding feature heat map. The comparison highlights the model's capability to extract stable features and produce smooth waveforms, demonstrating a close similarity between the generated and ground truth signals.

7.2. Cross-Dataset Evaluation

We further evaluated our method PULSE in a cross-dataset setting using the PURE ,UBFC-rPPG and COHFACE datasets to assess the model's robustness and generalization capabilities. These datasets were chosen due to their differences in factors such as head motion, lighting variations, and camera sensors, which are critical for rPPG estimation. Table 2 presents the cross-dataset test results. In this experiment, we trained the model on the PURE dataset and evaluated its performance on UBFC-rPPG, then reversed the process by training on UBFC-rPPG and testing on PURE. In both cases, PULSE exhibited strong performance, outperforming both supervised and unsupervised

Table 1. The table presents intra-dataset Heart Rate (HR) results. An upward arrow (↑) indicates that higher values are better, while a downward arrow (↓) denotes that lower values are preferable. The best results are highlighted in bold, and the second-best are underlined. Abbreviations: MAE stands for Mean Absolute Error, RMSE for Root Mean Square Error, and r for Pearson correlation coefficient.

Method	UBFC-rPPG			PURE			COHFACE		
	MAE ↓	RMSE↓	r ↑	MAE ↓	RMSE↓	r ↑	MAE ↓	RMSE↓	r ↑
Traditional Method									
Verkruysse et al. (GREEN) [50]	7.50	14.41	0.62	7.23	17.05	0.69	-	-	-
Poh et al. (ICA) [39]	5.17	11.76	0.65	3.76	12.60	0.85	-	-	-
Haan et al. (CHROM) [11]	2.36	9.23	0.87	0.75	2.23	1.00	7.8	12.45	0.26
Wang et al. (POS) [53]	2.11	9.11	0.87	0.80	4.11	0.98	13.43	17.05	0.24
Supervised Method									
Špetlík et al. (HR-CNN) (BMVC 18) [46]	-	-	-	1.84	2.37	0.98	10.8	8.1	0.29
Lu et al. (Dual-GAN) (CVPR 21) [29]	0.44	0.67	0.99	0.82	1.31	0.99	-	-	-
Speth et al. (RPNet) (21) [43]	0.53 ± 0.01	1.78 ± 0.02	0.99	1.15 ± 0.27	5.77 ± 1.25	0.96 ± 0.01	-	-	-
Yu et al. (PhysNet) (19) [59]	0.55 ± 0.03	2.03 ± 0.37	0.99	0.99 ± 0.19	5.22 ± 0.93	0.97 ± 0.01	-	-	-
Deshpande et al. (CVPR 23) [14]	-	-	-	-	-	-	2.92	6.128	0.86
Chen et al (CVPR 23) [7]	-	-	-	-	-	-	2.042	3.142	0.959
Gideon et al. (ICCV 21) [17]	-	-	-	2.1	2.6	0.99	2.5	7.8	0.75
Unsupervised Method									
Gideon et al. (ICCV 21) [17]	1.85	4.28	0.93	2.3	2.9	0.99	1.5	4.6	**0.99**
Sun et al. (Contrast-Phys) (ECCV 22) [48]	0.64	1.00	**0.99**	1.00	1.40	0.99	-	-	-
Yue et al. (TPAMI 23) [61]	0.58	**0.94**	**0.99**	1.23	2.01	0.99	-	-	-
Speth et al. (SiNC) (CVPR 23) [44]	0.59	1.83 ± 0.04	**0.99**	0.61 ± 0.66	1.84 ± 0.40	**1.00**	2.44 ± 0.64	6.02 ± 1.07	0.86 ± 0.05
PULSE (Ours)	**0.50 ± 0.04**	1.81 ± 0.19	**0.99**	**0.25 ± 0.01**	**0.38 ± 0.02**	**1.00**	**1.30 ± 0.14**	**3.71 ± 0.56**	0.94 ± 0.01

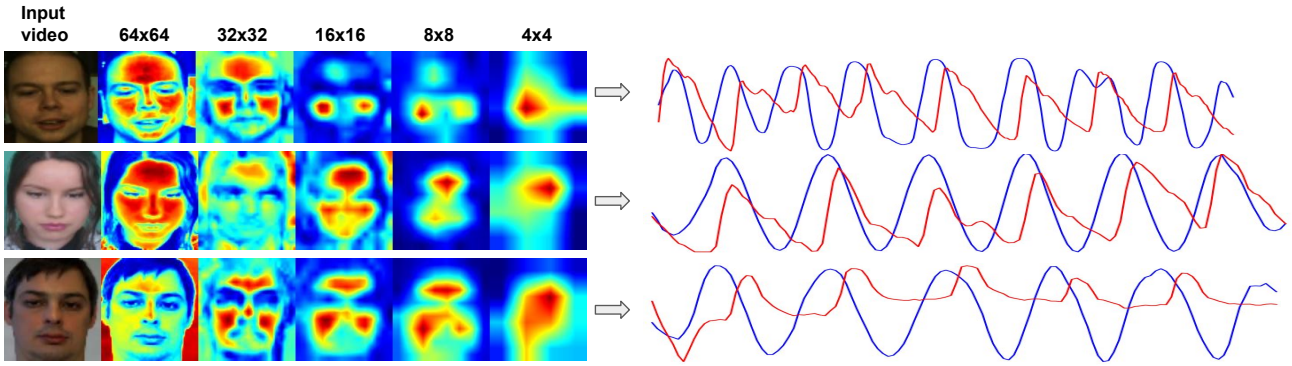


Figure 3. Feature heat maps and waveform extractions from the PURE, UBFC-rPPG, and COHFACE datasets illustrate feature extraction at different spatial resolutions: block (B_1) with 64×64, block (B_2) with 32×32, block (B_3) with 16×16, block (B_4) with 8×8, and block B_{LTC} with 4×4. The heat maps show stable feature extraction across varying resolutions, with low-level resolutions maintaining consistency. The waveforms on the right compare the extracted rPPG signals (blue) to the ground truth (red), demonstrating smooth signal extraction. The first row shows the sample from the PURE dataset, the second from UBFC-rPPG, and the third from COHFACE.

Table 2. The table presents cross-dataset Heart Rate (HR) results. An upward arrow (↑) indicates that higher values are favorable, while a downward arrow (↓) suggests that lower values are better.

Training Dataset	Testing Dataset	Method	MAE (bpm)↓	r ↑
PURE	UBFC	PhysNet [59]	7.02 ± 3.35	0.60 ± 0.13
	UBFC	Contrast-Pys [48]	10.22 ± 0.38	0.45 ± 0.04
	UBFC	SiNC [44]	6.64 ± 1.76	0.59 ± 0.10
	UBFC	**PULSE**	**3.06 ± 0.07**	**0.81 ± 0.01**
	COHFACE	**PULSE**	**14.73 ± 0.23**	**0.21 ± 0.02**
UBFC	PURE	PhysNet [59]	3.81 ± 0.34	0.87 ± 0.02
	PURE	Contrast-Pys [48]	19.61 ± 2.01	0.33 ± 0.06
	PURE	SiNC [44]	4.02 ± 0.06	0.86 ± 0.00
	PURE	**PULSE**	**2.27 ± 0.46**	**0.86 ± 0.01**
	COHFACE	**PULSE**	**9.50 ± 0.06**	**0.20 ± 0.02**
COHFACE	PURE	**PULSE**	**12.16 ± 0.27**	**0.44 ± 0.02**
	UBFC	**PULSE**	**2.61 ± 0.52**	**0.80 ± 0.05**

approaches. However, its performance decreased when tested on COHFACE, likely due to differences in frame rates (COHFACE operates at 20 FPS). Interestingly, while training on COHFACE resulted in less favorable outcomes when tested on PURE, it performed well on UBFC. These results highlight the model's adaptability across different datasets, while also emphasizing the influence of FPS discrepancies on cross-dataset generalization.

7.3. Robustness to Test-Time Perturbations

This experiment evaluates the robustness of PULSE, particularly its LTC-based convolutional layers, compared to SiNC under test-time perturbations. A percentage of video frames (5%, 10%, 15%, 20%) were corrupted with Gaussian noise, exposure changes, or dark frames.

Table 5 shows that PULSE consistently outperforms SiNC across all perturbations. The dynamic time constants of the LTC layers help PULSE handle temporal disruptions effectively, maintaining stability and accuracy even with

Table 3. Comparison of PULSE and SiNC on the PURE dataset under conditions like Steady, Talking/Laughing, Motion, and Low Light.

Method	Steady			Talking , Laughing			Small body movement			Dark light			Head movement		
	MAE ↓	RMSE ↓	r ↑	MAE ↓	RMSE ↓	r ↑	MAE ↓	RMSE ↓	r ↑	MAE ↓	RMSE ↓	r ↑	MAE ↓	RMSE ↓	r ↑
SiNC [44]	0.50	1.50	0.99	0.65	2.00	0.98	0.55	1.80	0.98	0.70	1.90	0.99	0.65	2.00	1.0
PULSE(Ours)	0.18	0.27	0.99	0.32	0.46	0.99	0.21	0.37	0.99	0.24	0.35	0.99	0.24	0.34	0.99

Table 4. Performance comparison between PULSE and SiNC on the COHFACE dataset under various conditions including light, gender, skin tone, and motion.

Method	Normal Light			Low Light			Male (28 subjects)			Female (12 subjets)			Light Skin			Dark Skin(2 subjects)			With motion		
	MAE ↓	RMSE ↓	r ↑	MAE ↓	RMSE ↓	r ↑	MAE ↓	RMSE ↓	r ↑	MAE ↓	RMSE ↓	r ↑	MAE ↓	RMSE ↓	r ↑	MAE ↓	RMSE ↓	r ↑	MAE ↓	RMSE ↓	r ↑
SiNC [44]	2.22	6.12	0.86	2.71	6.85	0.85	2.23	5.48	0.87	2.98	8.24	0.77	2.12	5.34	0.88	6.79	13.97	0.43	2.34	6.11	0.85
PULSE	1.11	3.81	0.96	1.62	4.78	0.91	1.38	3.81	0.93	1.06	3.31	0.96	1.18	3.32	0.95	3.96	9.38	0.72	1.20	3.41	0.95

Table 5. This experiment evaluates the robustness of our PULSE method, particularly the advantage of LTC-based convolutional block, compared to the SiNC [44] method under test-time perturbations. The perturbations included Gaussian noise, exposure changes, and dark frames, with the 'Perturbation level' column indicating the percentage of frames that are randomly corrupted in the video. The results suggest that PULSE demonstrates improved resilience in handling temporal frame corruption.

Dataset	Perturbation level	SiNC			PULSE (Ours)		
		MAE↓	RMSE↓	r↑	MAE↓	RMSE↓	r↑
PURE	5%	2.18 ± 0.26	6.21 ± 0.97	0.86 ± 0.06	0.96 ± 0.17	2.19 ± 0.90	0.93 ± 0.03
	10%	4.09 ± 0.91	12.77 ± 0.87	0.58 ± 0.12	2.58 ± 0.74	8.17 ± 0.97	0.76 ± 0.07
	15%	6.64 ± 1.91	17.30 ± 1.48	0.48 ± 0.14	4.84 ± 0.90	12.59 ± 1.30	0.63 ± 0.10
	20%	7.65 ± 1.79	18.65 ± 1.26	0.44 ± 0.11	6.92 ± 0.86	15.73 ± 1.58	0.58 ± 0.08
UBFC-rPPG	5%	10.69 ± 1.37	18.31 ± 1.28	0.41 ± 0.07	3.57 ± 0.38	10.84 ± 0.90	0.78 ± 0.04
	10%	17.40 ± 1.15	23.98 ± 1.04	0.14 ± 0.04	10.62 ± 1.23	21.07 ± 1.25	0.42 ± 0.07
	15%	19.88 ± 1.18	26.03 ± 1.61	0.09 ± 0.10	16.84 ± 1.51	22.17 ± 1.21	0.23 ± 0.05
	20%	21.64 ± 1.28	27.49 ± 1.78	0.08 ± 0.06	20.53 ± 0.82	25.51 ± 0.03	0.11 ± 0.03

Table 6. Ablation Study on the Effect of Unfolding Steps (L) on the PURE Dataset. The table presents the impact of varying the number of unfolding steps L.

Unfolding Steps(L)	MAE↓	RMSE↓	r↑
40	0.47 ± 0.01	0.61 ± 0.01	0.96±0.003
60	0.41 ± 0.01	0.49 ± 0.04	0.98±0.002
80	0.31 ± 0.02	0.44 ± 0.03	0.99±0.001
100	0.25 ± 0.01	0.38 ± 0.02	1.00

corrupted frames, demonstrating its resilience to real-world video degradation.

7.4. Performance Evaluation Under Different Conditions

We evaluated the PULSE model on the COHFACE and PURE datasets under various challenging conditions to assess its robustness. For COHFACE (Table 4), we tested under conditions like Normal/Low Light, Male/Female, Light/Dark Skin, and Motion. PULSE consistently outperformed SiNC, particularly in difficult scenarios such as Low Light and Dark Skin.

On the PURE dataset (Table 3), we tested under Steady, Talking/Laughing, Small Body Movement, Low Light, and Head Movement conditions. PULSE demonstrated superior performance, especially in motion and low light settings. These results highlight the advantage of LTC-based convolutional block, enabling PULSE to adapt and perform well across different conditions.

7.5. Effect of Unfolding Steps on Performance

We performed an ablation study on the PURE dataset to evaluate the effect of varying the number of unfolding steps L in the LTC-based 3D Convolutional Layer. We tested four settings: 40, 60, 80, and 100 steps. As shown in Table 6, increasing L improved model accuracy, with lower MAE and RMSE values. The best performance was observed at $L = 100$, indicating that more unfolding steps help the model capture temporal dependencies more effectively, enhancing rPPG signal prediction.

8. Conclusion

We introduced PULSE, a framework combining 3D convolutional layers with Liquid Time-Constant (LTC) models to improve rPPG signal extraction. By dynamically capturing temporal variations in physiological signals, PULSE enhances heart rate estimation accuracy, even under challenging conditions like motion and lighting changes. Evaluations on datasets such as PURE and UBFC-rPPG showed that PULSE performs better than existing methods in both accuracy and robustness. Our experiments were conducted within an unsupervised setting using the SiNC [44] framework, demonstrating PULSE's suitability for scenarios with limited labeled data. Future work may extend PULSE to other vital signs and incorporate additional modalities for broader application in non-contact health monitoring.

Acknowledgements

This research was conducted with support from the Eureka Eurostars funding program under Project number 55, with the project acronym BabySensor-Pre.

References

[1] Daniel Ackerman. "liquid" machine-learning system adapts to changing conditions, January 2021. MIT News Office. 3

[2] John Allen. Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, 28(3):R1, 2007. 1

[3] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124:82–90, 2019. 6

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 2

[5] Makram Chahine, Ramin Hasani, Patrick Kao, Aaron Ray, Ryan Shubert, Mathias Lechner, Alexander Amini, and Daniela Rus. Robust flight navigation out of distribution with liquid neural networks. *Science Robotics*, 8(77):eadc8892, 2023. 3

[6] Marie Chan, Daniel Estève, Jean-Yves Fourniols, Christophe Escriba, and Eric Campo. Smart wearable systems: Current status and future challenges. *Artificial Intelligence in Medicine*, 56(3):137–156, 2012. 1

[7] Shutao Chen, Sui Kei Ho, Jing Wei Chin, Kin Ho Luo, Tsz Tai Chan, Richard HY So, and Kwan Long Wong. Deep learning-based image enhancement for robust remote photoplethysmography in various illumination scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6076–6084, 2023. 7

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[9] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the european conference on computer vision (ECCV)*, pages 349–365, 2018. 2

[10] Tamal Chowdhury, Sukalpa Chanda, Saumik Bhattacharya, Soma Biswas, and Umapada Pal. Contact-less heart rate detection in low light videos. In Christian Wallraven, Qingshan Liu, and Hajime Nagahara, editors, *Pattern Recognition - 6th Asian Conference, ACPR 2021, Jeju Island, South Korea, November 9-12, 2021, Revised Selected Papers, Part I*, volume 13188 of *Lecture Notes in Computer Science*, pages 77–91. Springer, 2021. 1

[11] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013. 2, 7

[12] Gerard De Haan and Arno Van Leest. Improved motion robustness of remote-ppg by using the blood volume pulse signature. *Physiological measurement*, 35(9):1913, 2014. 2

[13] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019. 6

[14] Yogesh Deshpande, Surendrabikram Thapa, Abhijit Sarkar, and A Lynn Abbott. Camera-based recovery of cardiovascular signals from unconstrained face videos using an attention network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5974–5983, 2023. 7

[15] Duarte Dias and João Paulo Silva Cunha. Wearable health devices—vital sign monitoring, systems and technologies. *Sensors*, 18(8):2414, 2018. 1

[16] Ken-ichi Funahashi and Yuichi Nakamura. Approximation of dynamical systems by continuous time recurrent neural networks. *Neural networks*, 6(6):801–806, 1993. 3

[17] John Gideon and Simon Stent. The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3995–4004, 2021. 1, 2, 6, 7

[18] R. Hasani, M. Lechner, A. Amini, D. Rus, and R. Grosu. Liquid time-constant networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7657–7666, 2021. 2, 3, 5

[19] Brian Heater. What is a liquid neural network, really?, 8 2023. 3

[20] G Heusch, A Anjos, and S Marcel. A reproducible study on remote heart rate measurement. arxiv 2017. *arXiv preprint arXiv:1709.00962*. 2, 6

[21] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[22] Paul Kligfield, Leonard S Gettes, James J Bailey, Rory Childers, Barbara J Deal, E William Hancock, Gerard Van Herpen, Jan A Kors, Peter Macfarlane, David M Mirvis, et al. Recommendations for the standardization and interpretation of the electrocardiogram. *Journal of the American College of Cardiology*, 49(10):1109–1127, 2007. 1

[23] Eugene Lee, Evan Chen, and Chen-Yi Lee. Meta-rppg: Remote heart rate estimation using a transductive meta-learner. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 392–409. Springer, 2020. 2

[24] Xiaobai Li, Iman Alikhani, Jingang Shi, Tapio Seppanen, Juhani Junttila, Kirsi Majamaa-Voltti, Mikko Tulppo, and Guoying Zhao. The obf database: A large face video database for remote physiological signal measurement and atrial fibrillation detection. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 242–249. IEEE, 2018. 2

[25] Xiaobai Li, Jie Chen, Guoying Zhao, and Matti Pietikainen. Remote heart rate measurement from face videos under realistic situations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4264–4271, 2014. 2

[26] Si-Qi Liu, Xiangyuan Lan, and Pong C Yuen. Multi-channel remote photoplethysmography correspondence feature for 3d mask face presentation attack detection. *IEEE Transactions on Information Forensics and Security*, 16:2683–2696, 2021. 2

[27] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device

contactless vitals measurement. *Advances in Neural Information Processing Systems*, 33:19400–19411, 2020. 2

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[29] Hao Lu, Hu Han, and S Kevin Zhou. Dual-gan: Joint bvp and noise modeling for remote physiological measurement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12404–12413, 2021. 2, 7

[30] Sumit Majumder, Tapas Mondal, and M Jamal Deen. Wearable sensors for remote health monitoring. *Sensors*, 17(1):130, 2017. 1

[31] Daniel J McDuff, Ethan B Blackford, and Justin R Estepp. The impact of video compression on remote cardiac pulse measurement using imaging photoplethysmography. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 63–70. IEEE, 2017. 6

[32] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6707–6717, 2020. 2

[33] Meir Nitzan, Ayal Romem, and Robert Koppel. Pulse oximetry: fundamentals and technology update. *Medical Devices: Evidence and Research*, 7:231, 2014. 1

[34] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V 14*, pages 562–576. Springer, 2019. 2

[35] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Transactions on Image Processing*, 29:2409–2423, 2019. 1, 2

[36] Xuesong Niu, Zitong Yu, Hu Han, Xiaobai Li, Shiguang Shan, and Guoying Zhao. Video-based remote physiological measurement via cross-verified feature disentangling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 295–310. Springer, 2020. 2

[37] Ewa M Nowara, Daniel McDuff, and Ashok Veeraraghavan. The benefit of distraction: Denoising camera-based physiological measurements using inverse attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4955–4964, 2021. 2

[38] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[39] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, 58(1):7–11, 2010. 2, 7

[40] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 18(10):10762–10774, 2010. 2

[41] Rudina Seseri. Ai atlas #24: Liquid neural networks, 2023. 3

[42] Jingang Shi, Iman Alikhani, Xiaobai Li, Zitong Yu, Tapio Seppänen, and Guoying Zhao. Atrial fibrillation detection from face videos by fusing subtle variations. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(8):2781–2795, 2019. 1

[43] Jeremy Speth, Nathan Vance, Patrick Flynn, Kevin Bowyer, and Adam Czajka. Unifying frame rate and temporal dilations for improved remote pulse detection. *Computer Vision and Image Understanding*, 210:103246, 2021. 1, 2, 3, 7

[44] Jeremy Speth, Nathan Vance, Patrick Flynn, and Adam Czajka. 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. 1, 2, 3, 5, 6, 7, 8

[45] Radim Špetlík, Vojtěch Franc, and Jiří Matas. Visual heart rate estimation with convolutional neural network. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 2

[46] Radim Špetlík, Vojtech Franc, and Jirí Matas. Visual heart rate estimation with convolutional neural network. In *Proceedings of the british machine vision conference, Newcastle, UK*, pages 3–6, 2018. 2, 7

[47] Ronny Stricker, Steffen Müller, and Horst-Michael Gross. Non-contact video-based pulse rate measurement on a mobile service robot. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 1056–1062. IEEE, 2014. 2, 6

[48] Zhaodong Sun and Xiaobai Li. Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast. In *European Conference on Computer Vision*, pages 492–510. Springer, 2022. 1, 2, 7

[49] Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, Lijun Yin, Jeffrey F Cohn, and Nicu Sebe. 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. 2

[50] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008. 2, 7

[51] Hao Wang, Euijoon Ahn, and Jinman Kim. Self-supervised representation learning framework for remote physiological measurement using spatiotemporal augmentation loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2431–2439, 2022. 1, 2

[52] Wenjin Wang, Albertus C Den Brinker, and Gerard De Haan. Single-element remote-ppg. *IEEE Transactions on Biomedical Engineering*, 66(7):2032–2043, 2018. 2

[53] Wenjin Wang, Albertus C Den Brinker, Sander Stuijk, and Gerard De Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2016. 2, 7

[54] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard de Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2017. 2

[55] Zhen Wang, Yunhao Ba, Pradyumna Chari, Oyku Deniz Bozkurt, Gianna Brown, Parth Patwa, Niranjan Vaddi, Laleh Jalilian, and Achuta Kadambi. Synthetic generation of face videos with plethysmograph physiology. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20587–20596, 2022. 2

[56] Bryan P Yan, William HS Lai, Christy KY Chan, Stephen Chun-Hin Chan, Lok-Hei Chan, Ka-Ming Lam, Ho-Wang Lau, Chak-Ming Ng, Lok-Yin Tai, Kin-Wai Yip, et al. Contact-free screening of atrial fibrillation by a smartphone using facial pulsatile photoplethysmographic signals. *Journal of the American Heart Association*, 7(8):e008585, 2018. 1

[57] Yuzhe Yang, Xin Liu, Jiang Wu, Silviu Borac, Dina Katabi, Ming-Zher Poh, and Daniel McDuff. Simper: Simple self-supervised learning of periodic targets. *arXiv preprint arXiv:2210.03115*, 2022. 2

[58] Shanshan Yao, Amanda Myers, Abhishek Malhotra, Feiyan Lin, Alper Bozkurt, John F Muth, and Yong Zhu. A wearable hydration sensor with conformal nanowire electrodes. *Advanced Healthcare Materials*, 6(6):1601159, 2017. 1

[59] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. *arXiv preprint arXiv:1905.02419*, 2019. 1, 2, 7

[60] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip HS Torr, and Guoying Zhao. 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. 2

[61] Zijie Yue, Miaojing Shi, and Shuai Ding. Facial video-based remote physiological measurement via self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 7

[62] Yu Zhao, Bochao Zou, Fan Yang, Lin Lu, Abdelkader Nasreddine Belkacem, and Chao Chen. Video-based physiological measurement using 3d central difference convolution attention network. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–6. IEEE, 2021. 2