Smart Health 28 (2023) 100396

Contents lists available at ScienceDirect

Smart Health

journal homepage: www.elsevier.com/locate/smhl

SimPPG: Self-supervised photoplethysmography-based heart-rate estimation via similarity-enhanced instance discrimination

Soumyadeep Bhattachrjee [a,*], Huining Li [b], Jun Xia [b,c], Wenyao Xu [b]

[a] Williamsville East High School, NY, United States of America
[b] Department of Computer Sc. & Engg., University at Buffalo, United States of America
[c] Department of Biomedical Engg., University at Buffalo, United States of America

ARTICLE INFO

Keywords:
Remote photoplethysmography (rPPG) approximation
Heart-rate estimation
Evaluation
Self-supervised learning
Instance discrimination

ABSTRACT

Photoplethysmography (PPG) is an optical measurement technique to detect blood volume changes in the microvascular bed of target tissues and has been a widely used technique adopted by wearable devices to evaluate an individual's health condition. However, as motion and noise artifacts continue to pose manifold challenges toward the task of remote estimation of this signal (popularly known as rPPG), an authentic approximation of rPPG signals is of immense interest and convenience. In this work, we present a self-supervised learning-based regression framework that can reliably estimate the heart rate by accurately approximating a PPG signal from a participant's videos with sufficient coverage of their skin regions, in the presence of different types of local noises. The idea is that the PPG signals generated from the different parts of an individual will be nearly identical, while these signals from two different individuals may be comparably different. Though the severity of difference may vary based on the variance in their health conditions. Motivated by this intuition, we augment an rPPG signal extracted from a given facial landmark of an individual, using various other rPPG signals extracted from the neighboring facial landmarks of the same individual, to align their unique personalized patterns. Specifically, we develop a robust *Sim*ilarity-aware *PPG* (*SimPPG*) based heart-rate estimation scheme that adopts the instance discrimination method within a self-supervised learning setting to perform two-fold objectives: (1) discriminating every positive pair (rPPGs from different local skin regions of a given individual) from all negative pairs (rPPGs from different local skin regions of different individuals in a batch); (2) enforcing the alignments in the individual's respective heart-rate predictions with that of the corresponding ground truth PPG signals in parallel. Experiments using the large-scale UBFC-Phys dataset and our in-house data collection not only show a remarkable performance of *SimPPG* (MAE_{hr}=1.89bpm, which is 77% improvement compared to the existing baseline and MSE=2.91 in approximating the PPG signal), but also show its effectiveness in terms of handling various skin features across demographics. In a limited data environment, *SimPPG* is reported to have used only a random 75% of the available training collection to attain a competitive performance compared to several state-of-the-art models.

[*] Corresponding author.
E-mail addresses: sbhattac@buffalo.edu (S. Bhattachrjee), wenyao.xu@buffalo.edu (W. Xu).

https://doi.org/10.1016/j.smhl.2023.100396
Received 14 March 2023; Accepted 15 March 2023
Available online 25 March 2023

1. Introduction

While the Heart Rate Variability (HRV) of an individual often captures some critical insights into their health conditions, Photoplethysmography (PPG) is a popular technique adopted in various applications of medicine, health, and sports, to track heart activity and proactively detect an anomalous condition (if any) (Gil et al., 2010). As such, Remote estimation of PPG signals (rPPG) is a vascular optical measurement method, which estimates the reflected light and its variation to monitor the blood volume changes in the microvascular bed of the skin tissues (Allen, 2007). In contrast to the traditional PPG technique that utilizes a contact or near-field pulse oximeter (Hassan et al., 2017), rPPG utilizes a low-cost RGB camera under visible light and does not require any physical contact. Typically, a video containing an individual's skin regions (often facial videos) is captured by a webcam. The existing facial landmark localization algorithms are then leveraged to mark the regions of interest (ROIs) (Chen et al., 2018; Choi et al., 2022; Patil, Wang, Gao, Xu, & Jin, 2018a, 2018b; Ryu et al., 2021). The sequences of average pixel information obtained from several ROIs across three color channels are defined as the initial raw rPPG signals, which are later post-processed to eliminate the effects of motion illumination variances. Finally, Heart Rates (HR) are estimated from these filter signals by identifying a bunch of reliable peaks (Lin, Zheng, Li, Zhou, & Chen, 2021; Zaunseder, Vehkaoja, Fleischhauer, & Antink, 2022). Not only in evaluating an individual's health condition, but this technology has also been used in several other critical use case applications, including driver status assessment, affective state evaluation, and in-vivo detection (Chen et al., 2018).

However, the majority of these studies demonstrate their effectiveness under certain assumptions being valid: (1) requiring visibility to some predefined skin regions (e.g. finger, toe); (2) availability of large-scale training data demonstrating diverse data characteristics; (3) satisfying some restricted experimental conditions (e.g. in the absence of noises like motion artifacts); (4) lack of physical explainabilities (e.g. higher pulsatile strength may not always translate into the region's potential for more accurate rPPG extraction). For example, a considerable amount of works have attempted to estimate heart rates from the rPPG, for which signals are first approximated from the face videos (Perepelkina, Artemyev, Churikova, & Grinenko, 2020; Yang, Yang, Jin, & Wu, 2019) and later aggregated to appraise a gross signal. Significant attention is also invested in establishing the correlation between the region's position and rPPG quality. Kwon, Kim, Lee, and Park (2015) find the forehead and both cheeks as potential regions for accurate pulse extraction, while the mouth and chin regions provide a comparably less accurate estimation. Zhao, Mei, Xu, Li, and Feng (2019) use the ROIs below eye lines for the estimation task. Another set of works select a predefined number of face regions, which are later evaluated in terms of their comparative pulsatile strength. However, due to the frequent non-rigid motion artifacts, a higher pulsatile strength may not always translate into the region's potential for more accurate rPPG extraction. Furthermore, as we observe, most of these algorithms rely on handcrafted features and multiple complex post-processing steps, which are difficult to reproduce. Additionally, a wide variety of skin colors, the number of melanocytes, and textural changes due to the age of the individual (Shao, Tsow, Liu, Yang, & Tao, 2016), may also pose extra challenges to the optical skin feature extractor module of the system. To improve the representation scheme, a few recent works (Lampier et al., 2022; Ni, Azarang, & Kehtarnavaz, 2021; Perepelkina et al., 2020) leverage the power of deep learning-based techniques for the task of remote heart-rate estimation. However, most of these existing methods still depend on the availability of massive amounts of data, which may not always be a feasible assumption in this specific scenario. A wide variety of data characteristics influenced by an individual's age group, racial/ethnic, and other background details, may not have sufficient representatives in the training collection.

Toward this, we present a self-supervised learning-based (Devlin, Chang, Lee, & Toutanova, 2018; Ge et al., 2022) regression framework that can reliably approximate the PPG signal (and thereby estimate the heart rate) from a participant's videos with sufficient coverage of their skin regions. Self-supervised learning (Devlin et al., 2018; Ge et al., 2022) is a form of semi-supervised learning, which has shown tremendous promise in recent years. In fact, even in challenging datasets like Imagenet, self-supervised learning utilizing contrastive losses (Chen, Kornblith, Norouzi & Hinton, 2020; He, Fan, Wu, Xie, & Girshick, 2020) has recently outperformed supervised pre-training, which requires a large collection of labeled training data necessitating an intensive human annotation effort. In this work, we develop a robust *Sim*ilarity-aware *PPG* (*SimPPG*) based heart-rate estimation scheme that leverages the instance discrimination approach, which matches features from multiple representations/views of the same instance while distinguishing these features from those of other instances. Within a self-supervised learning setting, the proposed *SimPPG* learns to discriminate every positive pair (two rPPGs from different local skin regions of a given individual are considered to be a positive pair) from all negative pairs (two rPPGs from different local skin regions of different individuals are considered to be a negative pair) by utilizing contrastive losses (Chen, Kornblith, Norouzi et al., 2020; He et al., 2020), while also enforcing the alignments in their respective heart-rate predictions with that of the corresponding ground truth PPG signals. Actually, our approach to instance discrimination stems from a basic observation that the vital health signals generated from the different parts of an individual will be nearly identical to their ground truth PPG obtained via wearable devices (like pulse oximeter), whereas these signals from two different individuals may still be comparably different and the severity of these difference may vary based on the variance in their health conditions. Thus, if the model may learn the instance-specific discriminative characteristics of vital rPPG signals without requiring any extra annotation effort, it may end up delivering a learning representation scheme that can capture the unique personalized pattern variances for such physiological signals, while also remaining equally effective in depicting its class-wise patterns (e.g. healthy Vs. non-healthy). The overview of the proposed method is shown in Fig. 1. To summarize, the primary contributions of the work include.

1. The proposed *SimPPG* introduces a self-supervised learning-based regression framework that may reliably reconstruct the PPG signal and thereby estimate the heart rate from an individual's video stream with sufficient coverage of their skin regions. The proposed estimation scheme may not only enable a robust and proactive health condition evaluation but also may facilitate a real-time solution for several user case applications (e.g., emotion analysis, and biometric recognition).

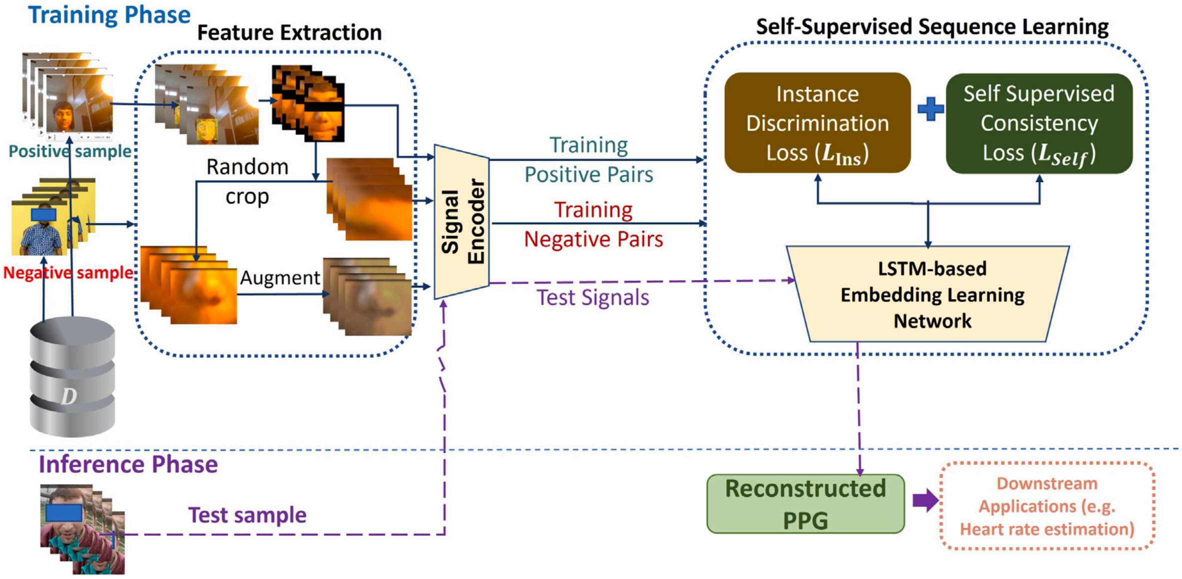
Training Phase
Feature Extraction
Self-Supervised Sequence Learning
Positive sample
Negative sample
D
Random crop
Augment
Signal Encoder
Training Positive Pairs
Training Negative Pairs
Test Signals
Instance Discrimination Loss (L_{Ins})
Self Supervised Consistency Loss (L_{Self})
LSTM-based Embedding Learning Network
Inference Phase
Test sample
Reconstructed PPG
Downstream Applications (e.g. Heart rate estimation)

Fig. 1. Overview of the Proposed *SimPPG* Method, in which the system aims to learn a self-supervised sequence learning model that may approximate a PPG response from a test video sequence, which may then be used for several downstream tasks like heart rate estimation.

2. By formulating an instance discrimination-based loss component, *SimPPG* explicitly emphasizes highlighting the instance-specific similarity observed within two rPPG signals generated from the different skin regions of a given individual covered in a video, while enforcing each of their learned intermediate representations to be closely aligned in their prediction of heart rate, which makes the model more robust to local noises and the training module more effective to the limited data environment.
3. An extensive experimental analysis using a large-scale UBFC dataset and our in-house test collection demonstrate the efficiency of the proposed *SimPPG* in estimating the heart-rate, while also demonstrating its effectiveness in terms of handling various facial features across demographics. As reported in the experiments, the proposed *SimPPG* can effectively reconstruct the PPG signals (with an average approximation error of MSE=2.91) and reports an impressive Mean Absolute Error (MAE) of 1.89 bpm, which is an improvement of 77% compared to that reported by the best existing baseline. In fact, we observe that *SimPPG* only uses a random 75% of the training collection to attain an improved performance compared to several state-of-the-art models.

The rest of the paper is organized as follows: Section 2 briefly describes the related works. The proposed method is described in Section 3. Sections 4 and 5 respectively present the experimental results and conclusion

2. Related works

2.1. Remote photoplethysmography estimation

To identify the critical PPG features, several algorithms are introduced. These include: adaptive methods (Argüello-Prada, 2019), digital filter based models (Jang, Park, Hahn, & Park, 2014), wavelet transformation (Vadrevu & Manikandan, 2018). However, since non-rigid motion artifacts are found to be one of the main sources of noises contaminating the extracted rPPG signals (Siontis, Noseworthy, Attia, & Friedman, 2021), blind source separation methods emphasize the denoising efforts to extract the BVP signal (e.g. Principle Component Analysis (PCA) (Chen et al., 2018), Independent Component Analysis (ICA) (Wei, He, Zhang, & Wu, 2017)). Another set of works rely upon the optical or physiological principles to develop skin reflection-based models (Wang, Den Brinker, Stuijk, & De Haan, 2016). These algorithms perform the signal estimation task by projecting the components related to specular and diffusion reflectance, which are later fine-tuned, and the resulting signal along every color channel is represented as the linear combination of all corresponding projected components. While at most two independent sources may be eliminated, in a practical scenario, the component may still carry more independent sources (Wang, den Brinker, Stuijk, & de Haan, 2017). Though such methods perform comparably better both in stationary and motion situations, strict reliance on hand-crafted features is not sufficiently generalizable.

A set of recent research works (Almarshad, Islam, Al-Ahmadi, & BaHammam, 2022; Cheng, Wong, Chin, Chan, & So, 2021) have adopted the deep learning algorithms for analyzing vital health signals like electrocardiogram (ECG) and PPG to perform several machine learning tasks like classification, waveform segmentation toward facilitating early disease diagnosis or monitoring other health conditions. Chen and McDuff (2018) designed a Convolutional Neural Network (CNN) based on an attention-enhanced skin reflection model for video-based HR estimation. Špetlšík, Franc, and Matas (2018) perform a 2-step CNN model to improve the HR estimation performance. To enhance the reconstruction of rPPG signal from an individual's video, some authors attempt to combine spatiotemporal information via designing a deep Recurrent Neural Network (RNN) and 3D CNN (Tsou, Lee, Hsu, & Chang, 2020; Yu, Li & Zhao, 2019; Yu, Peng, Li, Hong & Zhao, 2019). Yu, Peng et al. (2019) propose a spatio-temporal video enhancement network (STVEN) to enhance the video and an rPPG network (rPPGNet) for accurate recovery of the rPPG signal from the enhanced video. Tsou et al. (2020) proposes a Siamese-rPPG network that combines a Siamese architecture with 3D CNN to contrastively improve the rPPG representation. Hu, Qian, Guo, Wang, He and Ren (2021), Hu, Qian, Wang, He, Guo and Ren (2021) introduce a spatio-temporal attention network to highlight the unique information about the rPPG signal from a video segment that may enhance its long-range temporal details. Song, Chen, Cheng, Li, Liu, and Chen (2021) employ the generative neural network to compensate for the dearth of the data by generating some realistic rPPG signals. To boost the generalization capacity, recent authors also propose attention-enhanced, sequential quality assessment networks (Gao, Wu, Geng, & Lv, 2022; Gao, Wu, Shi, Gao, & Geng, 2021), which may explicitly contribute to accurate HR estimation. In recent work, Gao et al. (2022) predict the HR by focusing more attention to the less noisy signal segments, whereas cues from the other segments, which are comparably more corrupt, are not included in the estimation process. In contrast to these existing algorithms, which depend on a fundamental assumption of the availability of a large-scale data collection, *SimPPG* enables a self-supervised learning mechanism that may leverage only a limited data collection to capture the unique personalized signal patterns within the learned feature descriptor, while also retaining the critical class-specific discriminative patterns.

2.2. *Self-supervised instance discrimination*

Self-supervised learning (SSL) as a form of semi-supervised learning has shown tremendous promise in recent years. In fact, in recent works (Oord, Li, & Vinyals, 2018; Zhang et al., 2021), SSL methods utilizing contrastive losses (Chen, Kornblith, Norouzi et al., 2020; He et al., 2020) have reported a remarkable performance superseding that of supervised pre-training. Various contrastive SSL methods solve an instance discrimination task, wherein the target is to discriminate each positive pair from all negative pairs within a batch of samples. Although some alternatives to contrastive objectives (self-distillation (Caron et al., 2021; Grill et al., 2020), input reconstruction (Bao, Dong, & Wei, 2021; He et al., 2022)) have also been proposed, owing to its impressive performances, self-supervised instance discrimination has been introduced as a prominent pre-training strategy in various problem settings. Although the initial attempts were driven by the challenges specific to visual data (Feng, Xu, & Tao, 2019; Liu, Wu, Hu, & Lin, 2019), recently, several variants of instance discrimination algorithms have also been developed to handle the pre-training tasks for natural language data (El-Nouby et al., 2021; Ericsson, Gouk, Loy, & Hospedales, 2022; Giorgi, Nitski, Bader, & Wang, 2020; Meng et al., 2021; Rethmeier & Augenstein, 2021). In this work, we formulate self-supervised instance discrimination to support twofold objectives: (1) contrastively highlight instance-specific similarity patterns observed within two rPPG signals generated from different skin regions of a given individual covered in a video, compared to the pair obtained from two different individual's videos; (2) the learned representations of rPPGs representing different skin regions of the same individual to be closely aligned with their ground-truth PPG signals.

3. The *SimPPG* framework

In this work, we propose a self-supervised learning-based regression framework that can estimate a Photoplethysmography (PPG) signal from the videos with sufficient coverage of human skin regions. While this may facilitate a real-time, unobtrusive, early screening process of the individual's health condition, videos are often susceptible to noise, such as color channel variances, non-uniform resolution, lightness conditions, and offset of face positions in the visuals. To mitigate the issues related to such noise or variations, we employ a pre-processing stage to crop, reshape, and normalize the image, described in Section 3.1. Unlike existing methods obtaining a 1-D aggregated PPG signal to represent a human face video, the proposed *SimPPG* extracts multiple PPG signals from different local skin regions to deal with the variances in the PPG waveform over time, which may be due to the differences in morphology and dynamics between different peripheral body sites (Huthart, Elgendi, Zheng, Stansby, & Allen, 2020). To leverage the unique personalized pattern variances for such physiological signals that may facilitate an improved system generalization, *SimPPG* adopts the instance discrimination method (Wu, Xiong, Yu, & Lin, 2018), where each video instance is treated as its own distinct class, and discriminative PPG representations are learned to distinguish between such individual instance classes. In a self-supervised contrastive learning setting with two different views of the same video instance, obtained by spatial data augmentation, we train the network with the objective that the skin regions from the same video should be mapped nearby in learned embedding space while all other inputs should be contrasted.

In fact, given an annotated data collection $\mathcal{D} = \{(\mathbf{v}_j, \mathbf{y}_j)\}_{j=1}^{|\mathcal{D}|}$, where $\mathbf{v}_j$ represents a video sequence and the corresponding label $\mathbf{y}_j \in \mathbb{R}^d$ is the ground truth PPG signal collected by the wearable pulse Oximeter in parallel while recording the video, the task is to learn an embedding model that can estimate an accurate PPG signal response from a video sequence $\mathbf{v}$.

3.1. Data preprocessing

While the proposed model is generic enough to capture PPG information from any video displaying sufficient skin regions of any body part of the patient, given the video context of the datasets used in our work, the face was the primary region dominantly representing the skin information of the patients. Therefore, as a part of the preprocessing task, face regions are detected using a RetinaNet network (Lin, Goyal, Girshick, He, & Dollár, 2017) with a backbone MobileNet network (Howard et al., 2017). By means of ROI average pooling, the facial area in each keyframe of the video sequence is resized to a fixed size of $W \times H \times C$, where $W \times H$ represents the spatial dimension of each keyframe and C is the color channel. Thus, the entire video sequence $\mathbf{v}_j$ is represented in terms of a sequence of N key frames $\{\mathbf{v}_j^f\}_{f=1}^{N}$ (i.e. $\mathbf{v}_j \in \mathbb{R}^{W \times H \times C \times N}$). As this preprocessing step is mandatory to our proposed model, we adopt a slight notation abuse and introduce a notation simplification to refer to the entire processed video collection as $\{\mathbf{v}_j\}_{j=1}^{|\mathcal{D}|}$.

3.2. Feature extraction

For each detected face, affine face alignment based on facial landmarks detection [21] is performed to track the landmarks over time. PPG signals are approximated using the rPPG method (we use the POS algorithm for our experiments) (Wang et al., 2016) at multiple identified facial landmarks (Dong et al., 2018). A bandpass filter for $[45bpm, 180bpm]$ frequencies is applied for each (pixel, channel) pair independently in order to filter out signals not related to pulse cycles. Then, each $\mathbf{v}_j$ is represented in terms of a collection of rPPG signals $\{x_j^i\}_{i=1}^{n_j}$, where n_j denotes the number of identified facial landmarks that can be tracked in the entire video sequence and x_j^i is an initial approximation of the PPG signal (Wang et al., 2016) obtained from a region of interest around the ith landmark.

3.3. Instance discrimination

For each detected face, affine fa

We adopt the instance discrimination approach that formulates an objective function based on the Softmax criterion, to highlight the instance-specific similarity observed within the PPG signals generated from different skin regions of a given individual covered in a video. Each $\mathbf{v}_j$ is assumed to represent a distinct class in itself, and *SimPPG* aims to learn a stacked LSTM-based sequence representation learning framework (described as an embedding function F) that can reconstruct the ground truths $\{\mathbf{y}_j\}_j^{|\mathcal{D}|}$. In fact, if we consider $F(x) = \mathbf{g}$, the probability of the learned representation $\mathbf{g}$ (representing a given rPPG signal x) being assigned into the jth class is:

$$P(j|x) = \frac{exp(\mathbf{y}_j^T \mathbf{g}/\tau)}{\sum_{j=1}^{|\mathcal{D}|} exp(\mathbf{y}_j^T \mathbf{g}/\tau)}, \tag{1}$$

where $\mathbf{y}_j^T \mathbf{g}$ evaluates the fitment of g to the jth instance, τ is a temperature parameter that controls the concentration of the distribution (Hinton, Vinyals, Dean, et al., 2015), and $\mathbf{g}$ is normalized ($\|\mathbf{g}\| = 1$). Our objective is to maximize the joint probability $\prod_{j=1}^{|\mathcal{D}|} P(i|F(x_j^i))$ as:

$$\mathcal{L}_{Ins} = -\sum_{j=1}^{|\mathcal{D}|}\sum_{i=1}^{n_j} log(P(i|F(x_j^i))) = -\sum_{j=1}^{|\mathcal{D}|}\sum_{i=1}^{n_j} log\left(\frac{exp(\mathbf{y}_j^T F(x_j^i)/\tau)}{\sum_{j=1}^{|\mathcal{D}|} exp(\mathbf{y}_j^T F(x_j^i)/\tau)}\right). \tag{2}$$

3.4. Self-supervised learning with positive and negative samples

Given x_j^i, a batch $\mathcal{B}$ of training samples is formed by two types of instances: a set of random b rPPG signals obtained from an augmented sample $\mathbf{v}_j^{aug}$ generated from the same video sequence as $\mathbf{v}_j$; a set of b random negative samples generated from the different video sequences $\mathbf{v}_l \in \mathcal{D}$ (such that $l \neq j$), so that $diff(\mathbf{y}_j, \mathbf{y}_l)$ is greater than θ. In our experiments, we use $\theta = 1bpm$. Given the limited annotated collection $\mathcal{D}$, we adopt a recent self-supervised learning approach SimCLRV2 (Chen, Kornblith, Swersky, Norouzi, & Hinton, 2020) that may contrastively learn a robust descriptor by maximizing the instance-level representation consistency between a sample (x_j^i, representing an instance of $\mathbf{v}_j^{aug}$) and its augmented visual ($x_j^{i,aug}$) compared to an overall representation consistency computed on the samples in the $\mathcal{B}$.

$$\mathcal{L}_{Self} = -\sum_{j=1}^{b} log\left(\frac{exp(F(x_j^{i,aug})^T F(x_j^i)/\tau)}{\sum_{l=1}^{b} \mathbb{I}(l \neq j)\big(exp(F(x_l^i)^T, F(x_j^i)/\tau)\big)}\right), \tag{3}$$

where $\mathbb{I}$ is the indicator function such that $\mathbb{I}(0) = 1$ and $\mathbb{I}(a) = 0, \forall a \in \mathbb{R} - \{0\}$, τ is a temperature parameter.

Sequence Representation Learning: the Long–Short-Term Memory (LSTM) network model (Pascanu, Gulcehre, Cho, & Bengio, 2013), a variant of the Recurrent Network Model (RNN), is used for the first phase binary pathology condition detection module. RNNs form a chain-like neural network architecture that takes into consideration the current input in the context of the information from the past, to propagate the relevant historical information. While RNNs face a vanishing gradient problem and are unable to learn long-term dependencies, in this work, the Long–Short-Term Memory (LSTM) network model (Pascanu et al., 2013), a variant of the Recurrent Network Model (RNN), is used for learning the embedding function F. RNNs form a chain-like neural network architecture

Fig. 2. Some examples of video shots from the UB-PPG dataset with different skin tones and lighting conditions, where all participants are in the age range of 20–80 years and of Asian origin.

that takes into consideration the current input in the context of the information from the past, to propagate the relevant historical information. While RNNs face a vanishing gradient problem and are unable to learn long-term dependencies, LSTM integrates the gating functions into its state dynamics to provide an efficient alternative. We use a stacked LSTM with L_0 layers (Ullah, Ullah, Khan, & Cheikh, 2019) with the total loss as $\mathcal{L}_{total} = \mathcal{L}_{Ins} + \mathcal{L}_{Self}$. While several scaling configurations can be employed to weigh each of these components, we have not used any scaling in our experiments. Each layer is followed by a drop-out layer. The number of hidden units in each of the LSTM layers is set to be 128, and the drop-out ratio for each of their corresponding dropout layers is set as 0.2. The resulting output of the stacked LSTM sequence learning module is fed into a stack of FC layers, with the last layer having d units. The proposed stacked LSTM model (the proposed model has 2 FC layers) uses an rPPG signal x as input and produces a learned embedding as the activation of its last FC layer that is expected to produce the reconstructed $\mathbf{y}_{j,rec}$.

4. Evaluation

4.1. Data preparation

UBFC-Phys Dataset: Reliable estimation of heart rate from an individual's video, despite their various emotion and affective states, is a challenging task and of immense importance. Therefore, to evaluate the performance of the proposed *SimPPG*, we use a large, publicly available UBFC-Phys dataset, which captures the videos and PPG signals of individuals at different stress and emotion states. The dataset contains the videos and corresponding ground truth PPG signals (blood volume pulse and electrodermal activity signals obtained from the E4 wristband recording the PPG signals with a sampling rate of 64 Hz) from the 56 healthy subjects, all aged between 19 and 38 (mean age is 21.8, and the standard deviation is 3.11). The frame rate of the videos is about 35 fps. Among the participants, 46 are female, and 10 are male. During the data collection phase, the participants were given three experience tasks: a resting task (T1), a speech task (T2), and an arithmetic task (T3). To ensure uniformity, each video representing an individual executing a specific task is 3 min long. For the speech task, the sample video in the dataset captures the middle 3 min, while for the speech and the arithmetic tasks, the corresponding videos represent the 3 min at the beginning. As ground truth data, we transformed the blood volume pulse recorded in this dataset into HR by using Python framework for Virtual Heart Rate (pyVHR) (Boccignone et al., 2020).

UB-PPG Data Collection: To evaluate the generalization capacity of the proposed system, we also collected an in-house UB-PPG dataset comprising 10 volunteers of both genders of Asian origin, aged from 20 to 70 years old, who participated in the study. Some example video shots are shown in Fig. 2. Following the data collection protocol of the UB-Phys Dataset, each participant was recorded performing two different tasks: rest and speech. The ground truth PPG signal was collected using an FDA-cleared Oximeter device. In a well-lit room environment with uniform background (e.g., the subject standing or sitting with a wall at the back), a 12MP Ultra Wide Camera was placed on a tripod 50 cm away from the face and at the height of the face of the subject. During the video-capturing session, the oximeter was strapped onto the subject's finger. The video recording and oximeter reading device were started simultaneously and timed for 50 s. To ensure the quality of data collection, any sample demonstrating a heart rate less than $60 bpm$ was eliminated, and we repeated the data collection from that subject again.

4.2. SimPPG implementation

As a part of preprocessing, we downsample each video to 30 FPS and segmented them into 40 s clippings with a 10 s sliding window. So, 30 to 40 (depending on the video length) video clips were obtained from each participant. During the *SimPPG* training, we perform 3-fold cross-validation. All the experiments were conducted in a computer server with a 2.90-GHz CPU (Intel Core i7-10700F), 16-G RAM, and an NVIDIA Titan Xp. The proposed *SimPPG* was implemented using the PyTorch2 framework on an NVIDIA GeForce GTX 1650 GPU The Adam optimizer with a learning rate of 0.001 was used in all the implementations of the stacked LSTM model. On average, the testing process took around 5–8 s. While UB-Phys training collection ($\mathcal{D}_1$) was primarily used for training, in different experiment settings, due to the diverse participants' background of UB-PPG, we have also used a part of UB-PPG dataset ($\mathcal{D}_2$) to augment the training collection. We will discuss more details on this in Section 4.4 and in Tables 2 and 3.

Table 1
The evaluation of the *SimPPG* compared to the baseline methods based on MAE metric. The column annotated as $p\%\mathcal{D}_1$ describes the performance of *SimPPG* trained using $p\%$ of the UBFC-Phys dataset ($\mathcal{D}_1$).

Method	MAE_{HR} (in bpm)			
	$25\%\mathcal{D}_1$	$50\%\mathcal{D}_1$	$75\%\mathcal{D}_1$	$\mathcal{D}_1$
Stress-PPG (T1) (Sabour, Benezeth, De Oliveira, Chappe, & Yang, 2021)	–	–	–	3.55
Stress-PPG (T2) (Sabour et al., 2021)	–	–	–	9.26
Stress-PPG (T3) (Sabour et al., 2021)	–	–	–	5.99
Stress-PPG (Avg) (Sabour et al., 2021)	–	–	–	6.27
Green (Verkruysse, Svaasand, & Nelson, 2008)	–	–	–	8.27
ICA (Poh, McDuff, & Picard, 2010)	–	–	–	6.71
SQA-rPPG (Gao et al., 2022)	–	–	–	6.01
LSTM-rPPG Wim et al. (Verkruysse et al., 2008)	–	–	–	6.48
CHROM Haan et al. (De Haan and Jeanne, 2013)	–	–	–	4.39
POS (Wang et al., 2016)	–	–	–	5.98
1D-CNN Radim et al. (Špetlšík et al., 2018)	–	–	–	5.41
SimPPG (T1)	6.95	4.03	1.77	**0.84**
SimPPG (T2)	12.47	10.29	5.25	3.53
SimPPG (T3)	8.62	5.71	2.96	1.29
SimPPG (Avg)	9.35	6.68	3.32	**1.89**

Table 2
The evaluation of the *SimPPG* under *Cross Subjects* and *Cross Dataset* experimental settings. *SimPPG* is trained using different data collections from UBFC and UB-PPG and tested on the remaining videos from UB-PPG dataset, which were not included in the training sub-collection. The experimental setting TS_i that reports the performance of *SimPPG* trained on the ith ($i = 1, 2, 3$) task-specific sub-collection of UBFC-Phys dataset $\mathcal{D}_1$ (and $\mathcal{D}_1 \cup p\%\mathcal{D}_2$), which includes the task-specific sub-collection of $\mathcal{D}_1$ (combined with $p\%$ of $\mathcal{D}_2$). The experimental setting TS_{Comb} that uses the training collection $\mathcal{D}_1$ (and $\mathcal{D}_1 \cup p\%\mathcal{D}_2$), reports the performance of *SimPPG* trained on the entire UBFC-Phys dataset (combined with $p\%$ of $\mathcal{D}_2$). The performance is reported using MAE metric.

Experimental Setting Training Collection	$\mathcal{D}_1$	$\mathcal{D}_1 \cup 20\%\mathcal{D}_2$	$\mathcal{D}_1 \cup 40\%\mathcal{D}_2$	$\mathcal{D}_1 \cup 60\%\mathcal{D}_2$
TS_1	7.81	4.35	3.28	2.44
TS_2	8.76	8.32	5.29	3.12
TS_3	11.22	9.37	8.26	5.06
TS_{Comb}	10.94	10.54	7.11	4.98

Evaluation Metric: To investigate the performance of the proposed approach, we leverage the Mean Absolute Error (MAE) of the heart-rate estimation as the evaluation metric, which allows comparing two variables that have the same scale by evaluating their difference. More specifically, given $v \in \mathcal{D}_{test}$, where $\mathcal{D}_{test}$ represents the test collection, MAE in beats per minute (bpm) is obtained by calculating the mean difference between the remote heart rate estimated by *SimPPG* (y_v^{pred}) and that obtained from the ground truth signal (y_v) obtained by the Oximeter and computed as:

$$MAE_{HR} = \frac{\sum_{v \in \mathcal{D}_{test}} |y_v^{pred} - y_v|}{|\mathcal{D}_{test}|} \tag{4}$$

4.3. Experimental setup

To evaluate the generalization capacity of *SimPPG* to different environments and different video acquisition sources, we perform two types of tests: *Cross Subject Experiments* and *Cross Dataset Experiments*. In ***Cross Subject Experiments***, we use a set of random video recordings from the same dataset to perform the testing. Thus, following a 3-fold cross-validation technique, the entire data is divided into 3 subsets, each subset serves iteratively as a test set, and the rest constitutes the training set, and the MAE error is computed for each fold. We perform this set of experiments using two types of data collections: (1) UBFC-Phys and (2) a combined collection that is comprised of the UBFC-Phys and a random subset of subjects from the UB-PPG. In ***Cross Dataset Experiments***, we use the publicly available UBFC-Phys dataset for training and prepare a separate UB-PPG dataset for testing.

4.4. Experimental results and analysis

Comparative Study: The proposed *SimPPG* is compared against the baselines (Green (Verkruysse et al., 2008), ICA (Poh et al., 2010), Stress-PPG (Sabour et al., 2021)) for non-contact HR estimation using the UBFC-Phys ($\mathcal{D}_1$) dataset. The task-specific performances of *SimPPG* compared to that of Stress-PPG (Sabour et al., 2021) in an identical experimental setting is consistently high. As observed in the 5th column of the Table 1, in these *Cross Subject Experiments*, the proposed *SimPPG* has attained an average

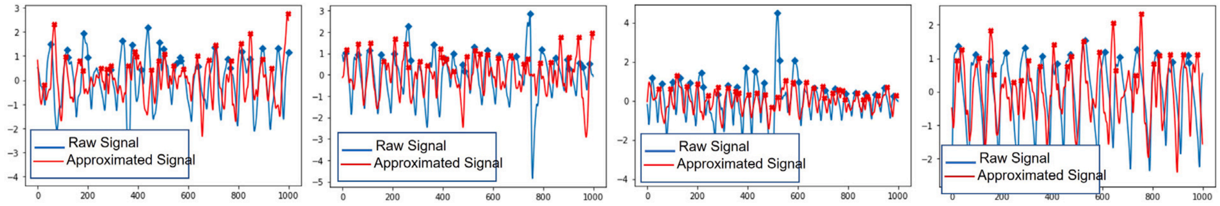

Fig. 3. Illustration of the PPG Signal Estimation Performance by the proposed *SimPPG*. In each figure, a raw signal (i.e. the ground truth PPG) $\mathbf{y}_j$ (shown in Blue) is paired with its corresponding approximated signal (i.e. *SimPPG*) $\mathbf{y}_{j,rec}$ (shown in Red).

of MAE_{HR} of 0.84 bpm for the task T1, 3.53 bpm for the task T2, and 1.29 bpm for the task T3, which in turn has allowed the model to reach an average MAE_{HR} of 1.89 bpm, an improvement of 77% over the best baseline performance (an MAE of 3.86) reported by Poh et al. (2010). In fact, in Task T1 (rest task) *SimPPG* appears to be most effective. In fact, limited movements during task T1, compared to the other two tasks (i.e., T2 and T3) contributed to the most accurate estimation objective.

Performance in a limited Data Environment: To evaluate the performance of the proposed *SimPPG* in the limited data environment, we several subsets of the dataset $\mathcal{D}_1$ containing a given fraction (p%) of the entire data collection to train the model. For example, the column annotated as $p\mathcal{D}_1$ reports the performance of *SimPPG* trained using p% of the dataset $\mathcal{D}_1$. As observed from the 2nd column of the Table 1, the average MAE of *SimPPG* is usually high when we utilize only 25% of the sample collection available in $\mathcal{D}_1$. By comparing the 2nd–4th columns of the 8th–10th rows in the Table 1, depicting the performances of *SimPPG* in several task-specific environments, we also note that as the size of the task-specific training set is increased (i.e. we leverage a larger sub-collection of $\mathcal{D}_1$), the MAE steadily improves, a fact that is also evident from the average performances of *SimPPG* reported in the last row of the table. In the *Cross Subject Experiments*, while *SimPPG* continues to demonstrate a more reliable performance in task T1, it is also interesting to find that in a self-supervised learning environment, *SimPPG* only uses a random 75% of the training collection in $\mathcal{D}_1$ to attain an improved performance compared to several state-of-the-art models (Poh et al., 2010; Sabour et al., 2021; Verkruysse et al., 2008). In fact, the proposed *SimPPG* requires only 50% of $\mathcal{D}_1$ to achieve an MAE of 6.68, a nearly equivalent performance reported by Stress-PPG (Sabour et al., 2021). Finally, using 75% of $\mathcal{D}_1$, *SimPPG* reports an MAE of 3.32, which is competitive to the best performing baseline (Poh et al., 2010) that uses the entire training collection of $\mathcal{D}_1$ to achieve an MAE of 3.36. To validate the performance, in each testing configuration described in 2nd–4th columns of the 8th–10th rows in the Table 1, the experiments are repeated 5 times using 5 random choices of $p\mathcal{D}_1$, and the average result is reported in Table 1.

Performances in the *cross-subject* and *cross-dataset* Environments: While the performance of the PPG-based heart-rate prediction techniques has been evaluated in large-scale data collections, their generalization abilities across demographic specification of the users (where diversity occurs due to the participant's appearances, skin tones, age, and cultural artifacts) have not been analyzed sufficiently yet (Dasari, Prakash, Jeni, & Tucker, 2021). In fact, the data acquisition environment for the existing public datasets is often tailored to eliminate the potential sources of noise as much as possible. Toward this, we investigate the performance of *SimPPG* in several combined collections of UBFC-Phys ($\mathcal{D}_1$) and UB-PPG dataset ($\mathcal{D}_2$) and the results are detailed in Table 2. For example, the experimental setting TS_i that reports the performance of *SimPPG* trained on the ith ($i = 1, 2, 3$) task-specific subcollection of UBFC-Phys dataset, when tested using the samples from the UB-PPG dataset. This represents the generalization capacity of *SimPPG* in the *Cross Dataset Experiment* setting without any retraining. Similarly, the experimental setting TS_i that uses the training collection $\mathcal{D}_1 \cup p\mathcal{D}_2$ reports the performance of *SimPPG* trained on a combined dataset that uses the ith ($i = 1, 2, 3$) task-specific subcollection of UBFC-Phys dataset and p% of the UB-PPG dataset ($\mathcal{D}_2$), when tested using the samples from the UB-PPG dataset. While the second column reports the performance of *SimPPG* in a *Cross Dataset Experiment* setting, third, fourth, and fifth columns describe its performance in several *Cross Subject Experiment* settings, wherein the subjects in training collection and the test collection represent the population from a diverse socio-econo and demographic backgrounds. As the existing literature mostly does not address this challenge, in Table 2, we investigate the performance of *SimPPG* using different *cross dataset* configurations. The experimental setting TS_{Comb} that uses the training collection $\mathcal{D}_1$ (and $\mathcal{D}_1 \cup p\mathcal{D}_2$) reports the performance of *SimPPG* trained on the entire UBFC-Phys dataset (combined with p% of $\mathcal{D}_2$). As observed in the table, *SimPPG* reports a comparably more stable performance in the *Cross Dataset Experiment* setting with no retraining, when learned using the T_1 task specific sub-collection of UBFC-Phys. We note that with finetuning using a smaller sub-collection of the UB-PPG dataset, the performance improves. In fact, for every experimental setting TS_i, with the growing size of the UB-PPG sub-collection, the performance demonstrates a steadily improving pattern.

PPG Signal Estimation Evaluation: While existing methods primarily focus only on estimating the heart rates, in order to show the robustness of the proposed *SimPPG*, we also compute the Mean Squared Error (MSE) of the averaged learned signal $\mathbf{y}_{j,rec} = \frac{\sum_i \mathbf{y}^i_{j,rec}}{n_j}$ compared to the ground truth signal $\mathbf{y}_j$ and a baseline is presented in Table 3, which reports the results in different training settings. The Fig. 3 illustrates the performance of the proposed system in approximating the ground truth $\mathbf{y}_j$ effectively. As reported in the Table 3, in different experiment sessions we use several subsets of the dataset $\mathcal{D}_1$ containing a given fraction (p%) of the entire data collection to train the model. As observed from the 2nd column of the Table 3, the average MSE of *SimPPG* is high when we utilize only 25% of the sample collection available in $\mathcal{D}_1$. Then as the size of the task-specific training set is increased (i.e. we leverage a larger sub-collection of $\mathcal{D}_1$), the MSE continues to improves and this trend is observed across different task-specific experimental

Table 3
The evaluation of the PPG signal estimation task using Mean Squared Error (MSE) metric in the *Cross Subjects* experimental setting. The column annotated as $p\%\mathcal{D}_1$ describes the performance of *SimPPG* trained using $p\%$ of the UBFC-Phys dataset ($\mathcal{D}_1$).

Method	MSE			
	$25\%\mathcal{D}_1$	$50\%\mathcal{D}_1$	$75\%\mathcal{D}_1$	$\mathcal{D}_1$
SimPPG (T1)	9.14	6.23	3.48	**1.45**
SimPPG (T2)	11.05	8.97	6.29	4.82
SimPPG (T3)	12.77	7.81	4.35	2.46
SimPPG (Avg)	10.97	7.67	4.71	2.91

settings. As such, the performance of *SimPPG* is more reliable when the individual is resting (i.e. executing the task T1). Also, we note that by using only a random 75% of the training collection in $\mathcal{D}_1$, *SimPPG* attains a reasonably promising 4.71 average MSE. Finally, when in possession of the whole training collection, the system reports an impressive MSE of 2.91.

5. Conclusion

In this work, we present an rPPG-based heart rate monitoring system that leverages a powerful self-supervised learning-based regression framework to accurately estimate the heart rate via approximating the PPG signals using a participant's video with sufficient coverage of their skin regions. To develop a robust prediction framework in a limited data environment, the proposed *SimPPG* adopts an effective instance discrimination approach that matches features from multiple representations/views of the same subject, while distinguishing these features from those of other subjects. An extensive set of experiments performed using the large-scale public as well as our in-house data collection not only show a remarkable performance of *SimPPG* in approximating the heart rates from a large variety of skin features across demographies but also enables an effective training scheme in a limited data environment that can attain competitive performance compared to several state-of-the-art models. In the future, we intend to extend the framework in a multi-modal data environment, where voice signals and visible skin visuals may be combined to ensure further improvement.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request

Acknowledgment

This material is based upon work supported by the National Science Foundation, United States under Grant No. CNS-2050910 and OISE-2106996.

References

Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, *28*(3), R1.

Almarshad, M. A., Islam, M. S., Al-Ahmadi, S., & BaHammam, A. S. (2022). Diagnostic features and potential applications of PPG signal in healthcare: A systematic review. *Vol. 10*, In *Healthcare* (3), (p. 547). MDPI.

Argüello-Prada, E. J. (2019). The mountaineer's method for peak detection in photoplethysmographic signals. *Revista Facultad de Ingeniería Universidad de Antioquia*.

Bao, H., Dong, L., & Wei, F. (2021). Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254.

Boccignone, G., Conte, D., Cuculo, V., d'Amelio, A., Grossi, G., & Lanzarotti, R. (2020). An open framework for remote-PPG methods and their assessment. *IEEE Access*, *8*, 216083–216103.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., et al. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9650–9660).

Chen, X., Cheng, J., Song, R., Liu, Y., Ward, R., & Wang, Z. J. (2018). Video-based heart rate measurement: Recent advances and future prospects. *IEEE Transactions on Instrumentation and Measurement*, *68*(10), 3600–3615.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597–1607). PMLR.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., & Hinton, G. E. (2020). Big self-supervised models are strong semi-supervised learners. *Advances in Neural Information Processing Systems*, *33*, 22243–22255.

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

Cheng, C.-H., Wong, K.-L., Chin, J.-W., Chan, T.-T., & So, R. H. (2021). Deep learning methods for remote heart rate measurement: a review and future research agenda. *Sensors*, *21*(18), 6296.

Choi, S., Gao, Y., Jin, Y., Kim, S. j., Li, J., Xu, W., et al. (2022). Ppgface: Like what you are watching? Earphones can" feel" your facial expressions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *6*(2), 1–32.

Dasari, A., Prakash, S. K. A., Jeni, L. A., & Tucker, C. S. (2021). Evaluation of biases in remote photoplethysmography methods. *NPJ Digital Medicine*, *4*(1), 1–13.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dong, X., Yu, S.-I., Weng, X., Wei, S.-E., Yang, Y., & Sheikh, Y. (2018). Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 360–368).

El-Nouby, A., Izacard, G., Touvron, H., Laptev, I., Jegou, H., & Grave, E. (2021). Are large-scale datasets necessary for self-supervised pre-training? arXiv preprint arXiv:2112.10740.

Ericsson, L., Gouk, H., Loy, C. C., & Hospedales, T. M. (2022). Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, *39*(3), 42–62.

Feng, Z., Xu, C., & Tao, D. (2019). Self-supervised representation learning by rotation feature decoupling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10364–10374).

Gao, H., Wu, X., Geng, J., & Lv, Y. (2022). Remote heart rate estimation by signal quality attention network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2122–2129).

Gao, H., Wu, X., Shi, C., Gao, Q., & Geng, J. (2021). A LSTM-based realtime signal quality assessment for photoplethysmogram and remote photoplethysmogram. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3831–3840).

Ge, Y., Ge, Y., Liu, X., Wang, A. J., Wu, J., Shan, Y., et al. (2022). MILES: Visual BERT pre-training with injected language semantics for video-text retrieval. arXiv preprint arXiv:2204.12408.

Gil, E., Orini, M., Bailon, R., Vergara, J. M., Mainardi, L., & Laguna, P. (2010). Photoplethysmography pulse rate variability as a surrogate measurement of heart rate variability during non-stationary conditions. *Physiological Measurement*, *31*(9), 1271.

Giorgi, J. M., Nitski, O., Bader, G. D., & Wang, B. (2020). Declutr: Deep contrastive learning for unsupervised textual representations. CoRR arXiv:2006.03659, URL https://arxiv.org/abs/2006.03659.

Grill, J., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., et al. (2020). Bootstrap your own latent: A new approach to self-supervised learning. CoRR arXiv:2006.07733, URL https://arxiv.org/abs/2006.07733.

Hassan, M. A., Malik, A. S., Fofi, D., Saad, N., Karasfi, B., Ali, Y. S., et al. (2017). Heart rate estimation using facial video: A review. *Biomedical Signal Processing and Control*, *38*, 346–360.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16000–16009).

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9729–9738).

Hinton, G., Vinyals, O., Dean, J., et al. (2015). Distilling the knowledge in a neural network. *2*, (7), arXiv preprint arXiv:1503.02531.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., et al. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Hu, M., Qian, F., Guo, D., Wang, X., He, L., & Ren, F. (2021). ETA-rppgnet: effective time-domain attention network for remote heart rate measurement. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1–12.

Hu, M., Qian, F., Wang, X., He, L., Guo, D., & Ren, F. (2021). Robust heart rate estimation with spatial-temporal attention network from facial videos. *IEEE Transactions on Cognitive and Developmental Systems*.

Huthart, S., Elgendi, M., Zheng, D., Stansby, G., & Allen, J. (2020). Advancing PPG signal quality and know-how through knowledge translation—from experts to student and researcher. *Frontiers in Digital Health*, *2*, Article 619692.

Jang, D.-G., Park, S., Hahn, M., & Park, S.-H. (2014). A real-time pulse peak detection algorithm for the photoplethysmogram. *International Journal of Electronics and Electrical Engineering*, 45–49.

Kwon, S., Kim, J., Lee, D., & Park, K. (2015). ROI analysis for remote photoplethysmography on facial video. In *2015 37th annual international conference of the IEEE engineering in medicine and biology society* (pp. 4938–4941). http://dx.doi.org/10.1109/EMBC.2015.7319499.

Lampier, L. C., Valadão, C. T., Silva, L. A., Delisle-Rodríguez, D., de Oliveira Caldeira, E. M., & Bastos-Filho, T. F. (2022). A deep learning approach to estimate pulse rate by remote photoplethysmography. *Physiological Measurement*, *43*(7), Article 075012.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980–2988).

Lin, W.-H., Zheng, D., Li, G., Zhou, H., & Chen, F. (2021). Investigation on pulse wave forward peak detection and its applications in cardiovascular health. *IEEE Transactions on Biomedical Engineering*, *69*(2), 700–709.

Liu, B., Wu, Z., Hu, H., & Lin, S. (2019). Deep metric transfer for label propagation with limited annotated data. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*.

Meng, Y., Xiong, C., Bajaj, P., Bennett, P., Han, J., Song, X., et al. (2021). Coco-lm: Correcting and contrasting text sequences for language model pretraining. *Advances in Neural Information Processing Systems*, *34*, 23102–23114.

Ni, A., Azarang, A., & Kehtarnavaz, N. (2021). A review of deep learning-based contactless heart rate measurement methods. *Sensors*, *21*(11), 3719.

Oord, A. v. d., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

Pascanu, R., Gulcehre, C., Cho, K., & Bengio, Y. (2013). How to construct deep recurrent neural networks. arXiv preprint arXiv:1312.6026.

Patil, O. R., Wang, W., Gao, Y., Xu, W., & Jin, Z. (2018a). A low-cost, camera-based continuous PPG monitoring system using Laplacian pyramid. *Smart Health*, *9–10*, 2–11. http://dx.doi.org/10.1016/j.smhl.2018.07.024, CHASE 2018 Special Issue.

Patil, O. R., Wang, W., Gao, Y., Xu, W., & Jin, Z. (2018b). A non-contact PPG biometric system based on deep neural network. In *2018 IEEE 9th international conference on biometrics theory, applications and systems* (pp. 1–7). IEEE.

Perepelkina, O., Artemyev, M., Churikova, M., & Grinenko, M. (2020). HeartTrack: Convolutional neural network for remote video-based heart rate monitoring. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 288–289).

Poh, M.-Z., McDuff, D. J., & Picard, R. W. (2010). Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Transactions on Biomedical Engineering*, *58*(1), 7–11.

Rethmeier, N., & Augenstein, I. (2021). A primer on contrastive pretraining in language processing: Methods, lessons learned and perspectives. arXiv e-prints arXiv–2102.

Ryu, J., Hong, S., Liang, S., Pak, S., Chen, Q., & Yan, S. (2021). A new framework for robust heart rate measurement based on the head motion state estimation. *IEEE Journal of Biomedical and Health Informatics*, *25*(9), 3428–3437.

Sabour, R. M., Benezeth, Y., De Oliveira, P., Chappe, J., & Yang, F. (2021). Ubfc-phys: A multimodal database for psychophysiological studies of social stress. *IEEE Transactions on Affective Computing*.

Shao, D., Tsow, F., Liu, C., Yang, Y., & Tao, N. (2016). Simultaneous monitoring of ballistocardiogram and photoplethysmogram using a camera. *IEEE Transactions on Biomedical Engineering*, *64*(5), 1003–1010.

Siontis, K. C., Noseworthy, P. A., Attia, Z. I., & Friedman, P. A. (2021). Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nature Reviews Cardiology*, *18*(7), 465–478.

Song, R., Chen, H., Cheng, J., Li, C., Liu, Y., & Chen, X. (2021). Pulsegan: Learning to generate realistic pulse waveforms in remote photoplethysmography. *IEEE Journal of Biomedical and Health Informatics*, *25*(5), 1373–1384.

Špetlšík, R., Franc, V., & Matas, J. (2018). Visual heart rate estimation with convolutional neural network. In *Proceedings of the british machine vision conference, Newcastle, UK* (pp. 3–6).

Tsou, Y.-Y., Lee, Y.-A., Hsu, C.-T., & Chang, S.-H. (2020). Siamese-rPPG network: Remote photoplethysmography signal estimation from face videos. In *Proceedings of the 35th annual ACM symposium on applied computing* (pp. 2066–2073).

Ullah, M., Ullah, H., Khan, S. D., & Cheikh, F. A. (2019). Stacked lstm network for human activity recognition using smartphone data. In *2019 8th European workshop on visual information processing* (pp. 175–180). IEEE.

Vadrevu, S., & Manikandan, M. S. (2018). A robust pulse onset and peak detection method for automated PPG signal analysis system. *IEEE Transactions on Instrumentation and Measurement*, *68*(3), 807–817.

Verkruysse, W., Svaasand, L. O., & Nelson, J. S. (2008). Remote plethysmographic imaging using ambient light. *Optics Express*, *16*(26), 21434–21445.

Wang, W., den Brinker, A. C., Stuijk, S., & de Haan, G. (2017). Robust heart rate from fitness videos. *Physiological Measurement*, *38*(6), 1023.

Wang, W., Den Brinker, A. C., Stuijk, S., & De Haan, G. (2016). Algorithmic principles of remote PPG. *IEEE Transactions on Biomedical Engineering*, *64*(7), 1479–1491.

Wei, B., He, X., Zhang, C., & Wu, X. (2017). Non-contact, synchronous dynamic measurement of respiratory rate and heart rate based on dual sensitive regions. *Biomedical Engineering Online*, *16*(1), 1–21.

Wu, Z., Xiong, Y., Yu, S. X., & Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3733–3742).

Yang, Z., Yang, X., Jin, J., & Wu, X. (2019). Motion-resistant heart rate measurement from face videos using patch-based fusion. *Signal, Image and Video Processing*, *13*(3), 423–430.

Yu, Z., Li, X., & Zhao, G. (2019). Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. arXiv preprint arXiv:1905.02419.

Yu, Z., Peng, W., Li, X., Hong, X., & Zhao, G. (2019). Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 151–160).

Zaunseder, S., Vehkaoja, A., Fleischhauer, V., & Antink, C. H. (2022). Signal-to-noise ratio is more important than sampling rate in beat-to-beat interval estimation from optical sensors. *Biomedical Signal Processing and Control*, *74*, Article 103538.

Zhang, D., Li, S.-W., Xiao, W., Zhu, H., Nallapati, R., Arnold, A. O., et al. (2021). Pairwise supervised contrastive learning of sentence representations. arXiv preprint arXiv:2109.05424.

Zhao, C., Mei, P., Xu, S., Li, Y., & Feng, Y. (2019). Performance evaluation of visual object detection and tracking algorithms used in remote photoplethysmography. In *2019 IEEE/CVF international conference on computer vision workshop* (pp. 1646–1655). http://dx.doi.org/10.1109/ICCVW.2019.00204.