Self-rPPG: Learning the Optical & Physiological Mechanics of *R*emote *P*hoto*p*lethysmo*g*raphy with *Self*-Supervision

Zahid Hasan*
Abu Zaher MD Faridee
Masud Ahmed
Nirmalya Roy
zhasan3@umbc.edu
faridee1@umbc.edu
mahmed10@umbc.edu
nroy@umbc.edu
University of Maryland, Baltimore County
Baltimore, Maryland, USA

ABSTRACT

Remote Photoplethysmography (rPPG) systems provide a contactless, low-cost, ubiquitous mechanism for regular heart rate (HR) monitoring by leveraging the diffused reflection from blood volumetric variations of human skin tissues (i.e. PPG). However, they have achieved limited adoption due to the lack of a generalized methodology to estimate HR from skin videos under various practical scenarios. Traditional supervised approaches require a large amount of synchronized ground truth annotations between video and rPPG signals, which have severely limited end-to-end generalized rPPG model development. In this paper, we propose *Self-rPPG*, which directly learns the optical and physiological mechanics of rPPG from the unlabeled videos without any synchronized rPPG signal stream annotation. We design a self-supervised contrastive learning-based pretraining strategy to learn the representation of the underlying diffusion signals' frequency, phase, and the video frames' temporal coherence from unlabeled video frame sequences collected over multiple public datasets. We run extensive experiments on the optimal contrastive learning schemes (loss functions, sampling strategy), and the saliency of the features learned by *Self-rPPG*, and show that our self-supervised presentations can successfully encode the diffusion signals' frequency and phase while demonstrating robustness against temporal corruption. The performance of *Self-rPPG* is validated on three public datasets where *Self-rPPG* outperforms the supervised state-of-the-art methods in PPG reconstruction and HR estimation using only 10% of the labeled data.

KEYWORDS

rPPG, Self-supervised learning, Contrastive learning

*Corresponding author

CHASE' 22, November 17–19, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9476-5/22/11…$15.00
https://doi.org/10.1145/3551455.3559609

ACM Reference Format:
Zahid Hasan, Abu Zaher MD Faridee, Masud Ahmed, and Nirmalya Roy. 2022. *Self-rPPG*: Learning the Optical & Physiological Mechanics of *R*emote *P*hoto*p*lethysmography with *Self*-Supervision. In *ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE' 22), November 17–19, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3551455.3559609

1 INTRODUCTION

Heart rate (HR) and heart rate variability (HRV) are vital biomarkers of human physiological health. Monitoring regular HR and heart rate variability (HRV) can provide early signs of coronary artery diseases, risk of stroke, cardiovascular diseases [37], asthma, anemia, or respiratory diseases [26]. Photoplethysmography (PPG), an optical technique to detect volumetric changes in the peripheral blood circulation, provides a non-invasive, low-cost HR, HRV measurement [1] methods compared to the more traditional electrocardiography approach and has been widely adopted into wearable sensor-based HR and HRV monitoring systems [3]. However, wearable-based PPG measurement systems still require special optical sensors and close skin contact [39] to encode skin blood volumetric variation. Alternatively, remote Photoplethysmography (rPPG) can utilize off-the-shelf video-camera sensors to estimate PPG from the captured variations in the reflected light off human skin and assess cardiovascular activity, HR, and HRV without any form of skin contact [30]. Due to its sensors' accessible nature and widespread deployment prospect, rPPG has the potential to supplant other invasive HR monitoring systems.

A typical rPPG system operates in two stages: (1) utilize optical sensors to capture videos of the subjects that encode PPG signal and (2) a *model* to extract PPG from videos. Facial videos, captured by stable-positioned HD cameras focused on the subjects' faces, have been shown to encode sufficient information to derive PPG [30]. However, the PPG derived from the captured videos possess a low signal-to-noise ratio (SNR) and is often inconsistent due to practical factors such as variability in sensor-subject distances, relative motion, background light, camera properties, and inherent video compression artifacts [21]. Hence, the fundamental research task in rPPG literature is to drive the development of reliable models to extract volumetric blood variation (PPG) from videos captured under varying realistic conditions.

Traditional rPPG models detect and localize the unexposed skin, and extract the underlying PPG using a number of signal processing techniques. But these models often rely on a number of heuristics for skin/face detection and PPG extraction which often leads to inconsistent results [24, 30]. Deep learning (DL) based approaches have recently garnered attention in the research community for their potential in developing data-driven PPG extraction from video sequences [22, 36]. However, we note that these modern deep learning-based models have their own set of drawbacks when it comes to developing generalized end-to-end rPPG systems. The performance of these data-driven supervised approaches is heavily reliant on the *quality* and *amount* of the ground truth labels. A typical rPPG dataset comprises a number of video streams (capture of exposed facial zone/skin) and synchronously captured streams of PPG readings captured through wrist/finger-mounted sensors. Unfortunately, most publicly available datasets do not contain enough varieties in the lighting (e.g., indoor, outdoor), occlusion (e.g., sunglasses), skin tone, and complexion variation in the subjects (due to age, race, and sex), variations in the placement of the physical PPG sensor (e.g. different wrist/finger position) and camera/PPG sensor models and makes. Hence, the supervised models trained on these datasets cannot provide accurate predictions when any previously unseen variations are encountered. Moreover, synchronized large-scale capture and pinpoint alignment of video and physical PPG signal can often be a time and resource-consuming task.

In this paper, we propose a novel and generalized framework called *Self-rPPG* to overcome both the existing *dataset* and *model* level limitations of building end-to-end deep-learning models for video-based rPPG prediction. We present a self-supervised contrastive learning (CL) based sampling and training strategy to learn the optical & physiological mechanics of rPPG from a diverse set of unlabelled rPPG video snippets. The CL pretrained network enables us to extract robust rPPG signals from only a few labeled video clips. Based on our experimentation with self-supervised rPPG learning methods, we pose the following contributions.

- **Learning rPPG Representation from Unlabeled Video Streams with Self-supervision Requiring No Synchronized PPG Ground Truth Annotations:** We introduce a self-supervised CL framework, *Self-rPPG* that can learn salient representations of rPPG signals from unlabeled facial video data. We develop the rPPG optical and physiological mechanics-based optimal positive and negative samples selection strategy for the CL self-supervision task that automatically incorporates the inductive biases corresponding to the PPG frequency, phase, and temporal coherence. By learning the similarity and dissimilarity among the carefully sampled instances, *Self-rPPG* can incorporate the relevant rPPG *priors* to learn a robust rPPG representation from unlabeled video streams. Our self-supervised task does not require any synchronized ground truth PPG annotations. Hence *Self-rPPG* can incorporate video streams from multiple datasets to further improve the learned representations, which is in stark contrast to traditional supervised state-of-the-arts.
- **Few-shot rPPG Extraction Only 2 minutes (i.e. 10%) of Labeled Samples:** Our simple fine-tuning mechanism takes the learned self-supervised model (from the unlabeled data) and trains an end-to-end PPG extraction model with only 2 minutes (i.e., 10%) of labeled samples to attain similar or less rPPG prediction RMSE scores compared to state-of-the-art supervised alternatives. We perform an extensive analysis with three state-of-the-art supervised and self-supervised rPPG extraction models and enumerate the other peripheral advantages *Self-rPPG* provides over them.
- **Experimental Evaluation of the Efficacy and Robustness of *Self-rPPG*** We conduct extensive experiments and perform qualitative and quantitative evaluations of our rPPG representation learning and extraction methods using three diverse public rPPG datasets. We demonstrate that *Self-rPPG* overcomes rPPG dataset-specific alignment (between video and PPG ground truth) issues and is compatible with heterogeneous cameras (e.g., DSLR, action-camera), video modalities (RGB, RGB mosaic, near-infrared), and PPG sensors (e.g., finger, wrist, sampling rate). We experiment with three possible loss functions to optimize the CL objective (e.g., InfoNCE, Max-Margin, and Triplet loss) and provide insight into the performances of each of the losses in different training scenarios. We further perform an ablation study on the positive (face cropping, channel selection, shifting) and negative sampling (frame repeating, random shuffle, phase, and frequency alteration) strategies and the choices of the network architecture design (e.g., projection head, negative sample size, loss functions). We provide the readers with an in-depth understanding of the proposed self-supervised rPPG extraction system.

2 RELATED WORK

In the last two decades, several research works have laid down the foundations of rPPG extraction from video cameras of multiple modalities (e.g., RGB video frame [24, 30, 34], Near-infrared (NIR) [19, 31, 38]). Numerous approaches (e.g., chrominance signal analysis method (CHROM) [7], source separation [18, 20], frequency filtering [11], Singular Spectrum Analysis (SSA) for compressed video data [40], Independent Component Analysis (ICA) [25], nonlinear mode decomposition [8], etc.) utilize signal processing techniques to extract the underlying PPG (diffusion signal intensity) from video by recognizing and analyzing the bare skin region from successive video frames.

In recent years, data-driven deep learning methods have demonstrated their success in learning the implicit function to obtain PPG from the face or bare skin videos. Convolutional Attention Networks have successfully learned face characteristics and extracted PPG from RGB and IR videos [5]. The research of *VitaMon* [15], *CamSense* [13, 14] proposed CNN based network to reconstruct PPG signals from facial videos using ECG peak and PPG signal as the target in an end-to-end setting. Further, the authors of [17] proposed meta-learning based rPPG framework.

However, the nature of PPG sensors and asynchronous alignment between sensors, PPG, and the corresponding video causes ground truth misalignment and consistency problems between different rPPG data samples. It hinders training an end-end network in a traditional setting by combining data samples. The authors of *CamSense* [14] proposed multi-task learning (MTL) and transfer learning (TX) based approach to avoid the misalignment issues in small scale to develop an end-to-end supervised model. However, these are not scalable for large-scale data integration and

generalized model development. Instead, we investigate the self-supervised representation learning domain to learn meaningful rPPG representation from raw video input without considering the ground truth label. The work of [9] proposed self-supervised PPG learning by incorporating a frequency-contrastive approach. However, their method lacks phase and temporal sequence learning in self-supervision, which is fundamental in system robustness and reconstructing PPG from video segments.

3 BACKGROUND

3.1 Optical & Physiological Mechanics of Remote Photoplethysmography (rPPG)

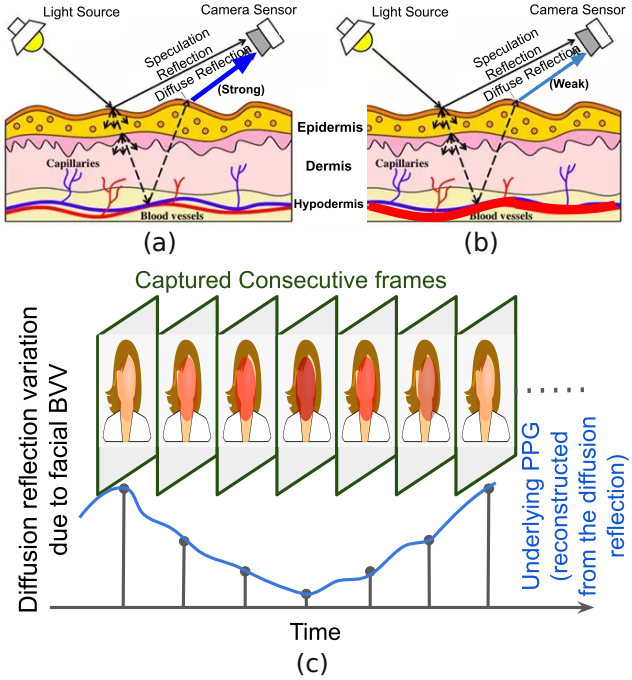

Figure 1: (a) Strong Diffuse reflection from micro bed of tissues due to low BV [32], (b) Weak Diffuse reflection from micro bed of tissues due to high BV, (c) Pictorial visualization of the video frames and corresponding diffusion signal. The overlapped red color intensity variation indicates blood volume in the peripheral tissues.

Under ambient light condition, video cameras capture two types of light reflections from the exposed skin: Specular and Diffused reflection (figure 1 (a)). *Specular reflection* refers to the light reflected from the skin surface and is responsible for visual information. The *diffused reflection* carries the remaining light after the absorption and scattering from the skin-surface blood vessels and tissues [32]. The amplitude of diffuse reflection varies temporally and depends on the blood volume in the microvascular bed of tissues under the skin (figure 1 (a) and (b)).

The microvascular tissues in the facial skin experience blood volume variations (BVV/PPG) synchronous to periodic heart activities. The face-focused video cameras pick the cyclic PPG information in terms of periodically varying diffusion signals and encode them in the consecutive temporal frames' RGB channels (figure 1(c)). However, the temporal channel variation contains low PPG SNR and variability due to randomnesses such as micro facial movement, camera movement, light source variation, randomly scattered reflection, etc.

The video contains near-constant visual specular reflection, and varying diffuse reflection throughout its consecutive frames, capturing the cyclic blood circulation to facial veins. Leveraging the prior knowledge about rPPG encoding and properties of PPG signal, we can create different diffusion patterns using the unlabeled video frames while keeping the visual signal constant. For example, repeating frames contain no diffusion variation, and shuffling consecutive frames shuffles BVV accordingly. Both cases destroy the underlying BVV signal despite having similar visual information. Shifting the starting position of the video snippet changes the phase of the concurrent PPG signals (figure 2 (a)). Further, utilizing the frame per rate (fps) and PPG frequency information, we can generate video snippets with opposite/similar phases and frequency (figure 5). Moreover, the cropped versions of the same video snippet contain similar PPG information while varying the visual information.

The rPPG systems aim to extract the BVV-stimulated information by comprehending the diffusion reflection and filtering out specular information from the sequence of consecutive video-frame channels. Hence, a robust system should be able to separate consecutive video frames (consistent PPG) from the temporally corrupted shuffled/repeated version (deconstructed PPG) (figure 2 (c), (d)). The model also should return a matching representation for video snippets where the underlying PPG information stays similar, e.g., for different face-cropped of the same video snippet, for video snippets with similar PPG frequency and phases disregarding any visuals variation (figure 2 (b)). Further, the model should have a different understanding of the video snippets with different underlying PPG frequency and phases, although they have similar visual information.

We hypothesize that the self-supervised approach can incorporate the rPPG priors to the model by encouraging it to distinguish the BVV variations from various carefully curated unlabeled video snippets. Hence, the model would become sensitive to BVV while indifferent to other variations without requiring any corresponding PPG label. The rPPG prior would also reduce the models' requirement of the massive labeled dataset for rPPG reconstruction.

3.2 Self-supervised Contrastive Learning

Self-supervised approaches utilize designed pretext tasks to learn representative features from unlabeled data. Contrastive learning (CL) approaches utilized notion of similarity to define similar and dissimilar items in the unlabeled data and enforces close representation for similar samples in the embedding space. Based on the notion of similarity CL enables network to learn discriminative data features in prior and eventually reduce the label-cost for the relevant downstream task.

Given the available unlabeled dataset, $D^u = \{x_i; i = 1....M\}, x_i \in \mathcal{X}$, where N is the number of data samples and the samples are from data distribution $\mathcal{X}$. The CL framework has three main components.

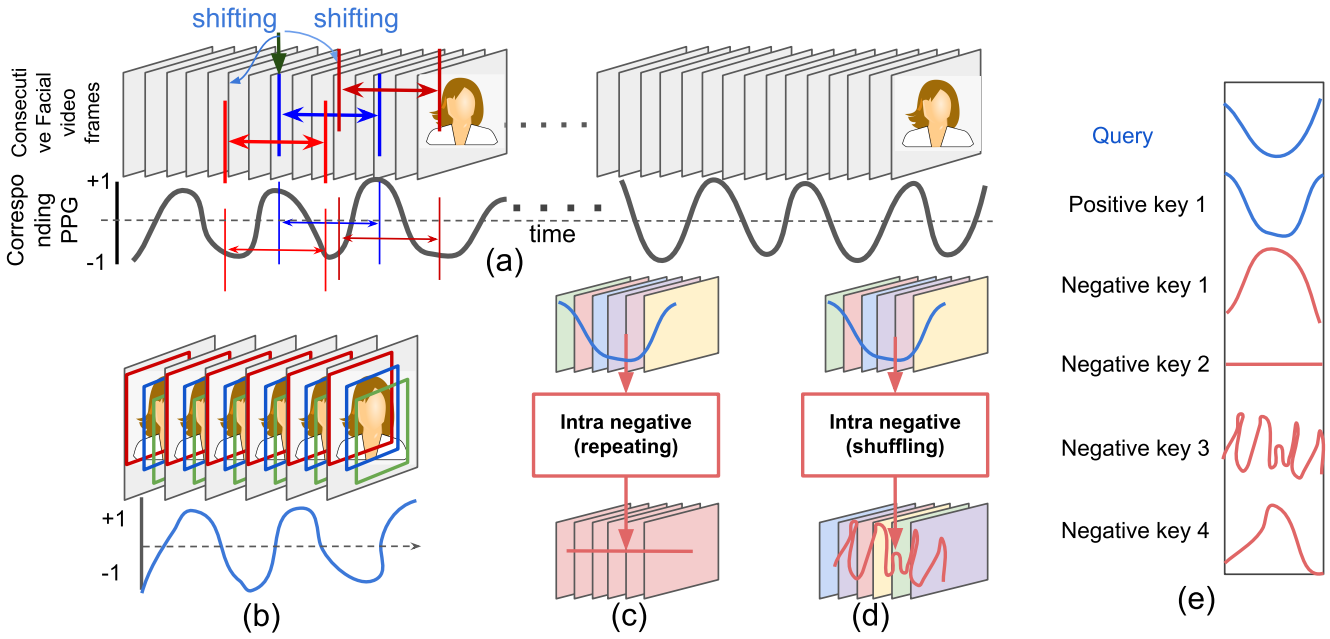

Figure 2: (a) Video and corresponding PPG, (b) The face cropped copy of the same temporal video contains similar PPG information, (c) Formation of negative by repeating one frame across the time, (d) Formation of negative by shuffling frames, (e) Visualization of Sample Negative and positive sampling output (all are scaled between -1 to 1).

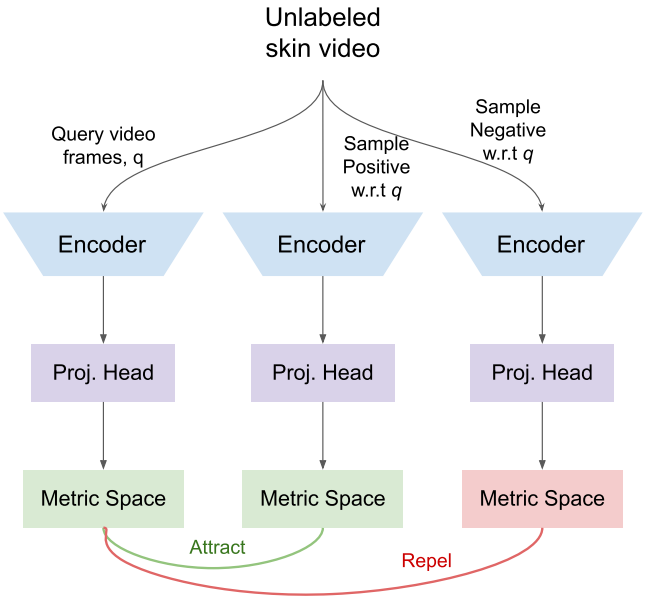

Figure 3: Overview framework for contrastive learning pre-training.

Firstly, CL requires sampling query, q_i, positive, $p_i \in \mathcal{P}$, negative $n_i \in \mathcal{N}$ samples based on designed similarity from the dataset D^u, where, $\{q_i, p_i, n_i\}$ are all data instances or augmented data instance. The designed similarity selection mechanism determines the types of inductive bias model will focus during the similarity learning and requires careful design to align with the intended final tasks for best results. Secondly, CL requires a capable DL based parametric model to learn the relevant features for CL task, suitable for downstream applications, and compatible with the data. The literature uses data compatible base encoder [14, 15], e_θ, and fully connected projection head, g_ψ, to learn the CL tasks and project the data instances in low dimensional embedding $r \in \mathbb{R}^d$, where $r_{x_i} = g_\psi(e_\theta(x_i))$. Finally, the Contrastive loss applies metric based loss to enforce the similarity/dissimilarity among the samples by pulling the positive embedding space and query embedding space r_{q_i}, r_{p_i} together and pushing away the query embedding space and negative embedding space r_{q_i}, r_{n_i} apart (figure 3).

4 METHODOLOGIES

We develop the rPPG extraction network in two steps. Firstly, the network learns rPPG embedding in a self-supervised fashion to represent the rPPG information in low dimensional metric space. Secondly, we fine-tune the self-supervised pretrained network to reconstruct the PPG by training on a small amount of PPG labeled video with a supervised objective.

4.1 Representation Learning through Contrastive Pretraining

In this stage, CL trains the network to learn the mechanism of rPPG representation, BVV induced frequency (f) and phase (ϕ), from a large unlabeled facial videos. To incorporate rPPG prior, we design positive samples by keeping the underlying PPG f and

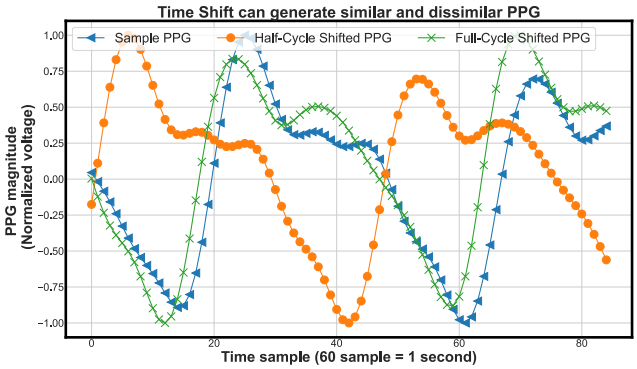

Figure 4: The controlled time-shift of 12 video frames and 24 video frames inhibit dissimilar and similar PPG behavior with respect to the original position.

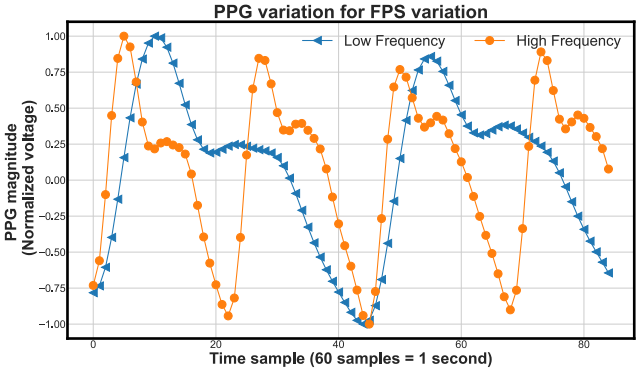

Figure 5: Varying sampling rate generates high frequency PPG from the same video stream.

ϕ constant and negative samples by varying f and ϕ using rPPG physics domain knowledge and train a capable neural network to represent f and ϕ.

4.1.1 Architecture. We select a base encoder from [14] network class, e_θ, that has shown success in extracting PPG from video streams and apply multi-layer projection head, g_ψ, on top of encoder output. The selected e_θ network takes 40 consecutive video frames to learn the rPPG representation from the unlabeled video data and the g_ψ projects the encoder output in the low-dimensional metric space for contrastive loss optimization.

4.1.2 Data Sampling. Different tasks require different inductive biases and carefully designed sampling in CL training can provide the missing inductive biases [35]. In the rPPG reconstruction context, the query and positive samples have similar f and ϕ, and negative samples have differing f or ϕ or both. We can leverage rPPG domain specific knowledge and video properties to find the q_i, p_i and n_i by designing the rPPG specific data augmentation.

Positive samples (Similar f and ϕ). We sample i–th query, q_i, by taking 40 consecutive green channel frame from a randomly starting index, idx_q, of the videos and sample positive, p_i, based on idx_q and q_i by one of the three methods. i) *Face Cropping:* We crop same face portion alone the frames of q_i. ii) *Channel Selection:* We select either red or blue channel of the q_i. iii) *Shifting:* We leverage the periodic properties of target PPG and information about video fps to temporally sample the video frames with similar f and ϕ by shifting the idx_q one cycle. The exact shifting parameter depends on the video fps and underlying HR. We utilize the approximate regular resting HR 72 BPM for a healthy adult and exact fps information. For example, one cycle would require $\frac{60}{72} = 0.833$ seconds and at 30 fps positive requires to shift approximately $0.833 \times 30 = 25$ frames (figure 4).

Negative samples (Dissimilar f, ϕ). We sample negative, n_i, by altering the underlying query PPG of q_i by two approaches to contrast: *Intra negative* (destroys both f and ϕ) and *inter negative* (changes either ϕ or f) [29]. In intra negative we destroy the PPG information of the q_i frames by corrupting their temporal order either by shuffling frames or frame repeating (figure 2 (c) and (d)). On the other hand, we sample inter negative samples by altering ϕ or f.

ϕ altering. We selectively shift the query index idx_q by half-cycle in the corresponding videos to alter the q_i PPG's ϕ by 180°. Based on the previous discussion 3.1, we can shift half-cycle by shifting approximately $12 \sim 13$ frames from idx_q, to get n_i with different ϕ but same f (figure 4). Further, we can crop either the negative samples or the non-face regions to create more negative ϕ samples.

f altering. We sample n_i with differing f than the q_i by carefully downsampling the consecutive video frames. By limiting the maximum PPG frequency at 240 BPM or $4Hz$, the Nyquist rate would be $8Hz$. Since most rPPG videos are captured at 30 fps, the Nyquist rate allows to drop the sampling rate by half to 15 fps without aliasing problems in the PPG frequency domain. However, with dropping sampling rate underlying PPG moves faster through the temporal frames. For example, with 15 fps the 40 seconds video contains two full cycle compared to one cycle in 30 fps (figure 5).

4.1.3 Loss objective. The CL loss function cluster put together the embedding of r_{q_i} and r_{p_i} and push apart the r_{q_i} and r_{n_i} by applying metric based objective in the embedding space (figure 6). The CL loss uses similarity function, $s(r_q, r_x)$, to measure the closeness between embedding and optimize CL loss, $\mathcal{L}_{cl}(r_q, r_p, r_n)$, to enforce similar embedding for positive pairs. The CL loss uses the negative pairs to avoid the trivial solution of collapse representation.

4.2 Fine-tuning

We utilize the CL pretrained base encoder, e_θ, with a randomly initialized task specific head, g_β, to perform the final downstream task of rPPG reconstruction. In this stage, we fine-tune the network with supervised regression loss objective, l_sup, using few labeled rPPG data, D^l containing video instances, $v_i \in \mathcal{X}$, and corresponding PPG label, $l_i \in \mathcal{R}^{d1}$, of (v_i, l_i), and $D^l = \{(v_i, l_i)\}_{i=1}^N$.

We provide our pseudo-algorithm for our self-supervised rPPG network development in algorithm 1.

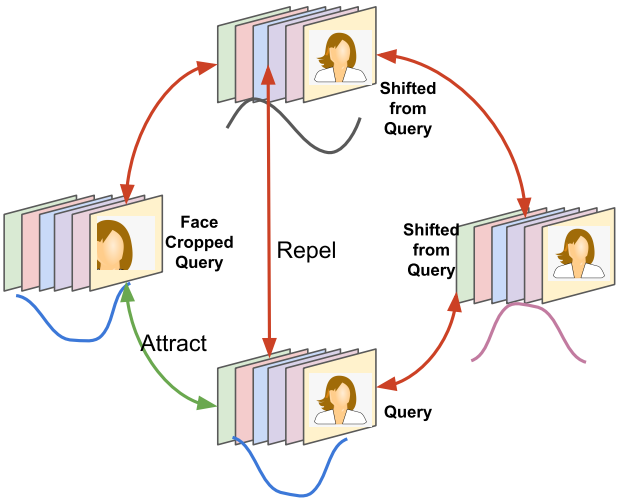

Figure 6: Video segments requires to be segmented in the metric space. The green arrow enforces close representation and red arrows pushes the representation away.

5 EXPERIMENTS

We have experimented with three rPPG datasets (varying camera properties), and three CL settings to demonstrate the adaptability of *Self-rPPG* as generalized PPG extraction setting.

5.1 Datasets

5.1.1 MPSC-rPPG dataset [14]. The dataset contains facial videos and corresponding simultaneous *wrist* PPG record at $64Hz$. The dataset contains videos from DSLR, action camera, and web-camera under different artificial lights and multiple realistic environments like natural lights, angular exposure, and sensor variabilities. The dataset contains volunteers with different skin color, gender, and facial characteristics.

5.1.2 MERL [23]. MERL-Rice NIR Pulse (MR-NIRP) dataset consists of 8 subjects of near-infrared (NIR), RGB raw, and RGB demosaic modalities at 30 fps and simultaneous *finger* PPG at $60Hz$.

5.1.3 UBFC-rPPG [2]. The UBFC-rPPG dataset contains about 8 simple and 42 realistic RGB videos with varying fps from 28 to 30 with simultaneous PPG varies from $30Hz$ to $62.5Hz$.

5.2 Network Architecture

5.2.1 Encoder, e_θ. We use the *CamSense* architecture CNN parts as base CNN encoder. The *CamSense* uses four CNN layers with an l2 regularization. Each convolutional layer is followed by Batch Normalization (BN), Rectified Linear Unit (ReLU) activation function, and max-pooling layers. The Inception layers with multi-scale parallel convolutional filters have shown to be better feature selection in the case of facial images [27] and the global average pooling (GAP) helps to localize the object [10]. Subsequently, we use two naive inception layers [28] each followed by a GAP layer. Finally, we flatten the GAP layer output for projection head input.

Algorithm 1: Pseudo-code for CL Pretraining and Supervised Fine-tuning

Result: e_θ, g_ψ
Input: D^u, D^l ;
Hyper-parameters:
`PostNum`: Number of Positive samples
`NegNum`: Number of negative samples
`ContrLoss`: Contrastive loss function
`PostSamp`: Set of Positive Sampling Method
`NegSamp`: Set of Negative Sampling Method
`pt`: Transformation from Positive Sampling
`nt`: Transformation from Negative Sampling
initialization: e_θ, g_ψ
while *Till Converge* **do**
  Sample q from idx_q from D^u;
  initialization: $p_i, n_i = \{\}, \{\}$
  for $(i = 0, i <$ `PostNum`$, i++)$ **do**
    `pt` $\sim$ `PostSamp`
    $p_i =$ `pt`(q)
  end
  for $(j = 0, j <$ `NegNum`$, j++)$ **do**
    `nt` $\sim$ `NegSamp`
    $n_j =$ `nt`(q)
  end
  $r_q = g_\psi(e_\theta(q))$;
  $r_{p_i} = g_\psi(e_\theta(p_i))$ for $i = 1, ...,$ `PostNum`;
  $r_{n_i} = g_\psi(e_\theta(n_i))$ for $i = 1, ...,$ `NegNum`;
  $l_c =$ `ContrLoss`$(r_q, \{r_{p_i}\}_{i=1}^{\text{PostNum}}, \{r_{n_i}\}_{i=1}^{\text{NegNum}})$;
  update e_θ, g_ψ by descending stochastic gradient $\nabla_{e_\theta, g_\psi} l_c$
end
Return: Pretrained encoder e_θ
Hyper-parameters:
m: Batch Size
l_{sup}: Supervised loss;
initialization: g_β
while *Till Converge* **do**
  Sample minibatch $\{(v_i, l_i)\}_{i=1}^m$ from D^l;
  $y_{p,i} = g_\beta(e_\theta(v_i))$ for $i = 1, ..., m$;
  $l_s = \sum_{i=1}^m \frac{1}{m} l_{sup}(y_{p,i}, l_i)$;
  update f_θ, g_β by descending stochastic gradient $\nabla_{f_\theta, g_\beta} l_s$
end
Return: rPPG Network e_θ, g_β;

5.2.2 Projection Heads, g_ψ. We adopt multi layers projection head following the convention of [4]. Particularly, we implement 3 layers fully connected feed forward (FCFF) network with regularized linear unit activation as projection head. The final layer projects the output to a metric space. In the metric space, we contrast between negative example and force similarity between positive samples.

5.3 Loss Functions

We experiment with three different CL settings.

5.3.1 Energy Based Margin Loss. [12], [6] uses Euclidean similarity function Equation 1 to measure embedding distance and applies pair wise CL loss optimization Equation 2. The loss function tries to enforce same embedding for positive pairs and put negative pairs margin m away.

$$s_e(q,k) = ||q-k||_2 \tag{1}$$

$$\mathcal{L}_{pair}(r_{q,i},k) = \begin{cases} D_E(q,k)^2, & \text{if } k \sim \mathcal{P} \\ \max(0, m - D_E(q,k)^2), & \text{if } k \sim \mathcal{N} \end{cases} \tag{2}$$

5.3.2 Max-Margin Triplet Loss. Utilizes euclidean based similarity measurement and unlike energy based loss the max-margin objective cares about relative distance between positive pairs and negative pair Equation 3 [33].

$$\mathcal{L}(r_q, r_p, r_n) = \max(0, D_E(r_q, r_p) - D_E(r_q, r_n) + m) \tag{3}$$

5.3.3 Probabilistic NCE-based Loss [16]. uses cosine based metrics Equation 4 and facilitates multiple negatives in the loss calculation Equation 5. It maximizes the temperature, τ, scaled cosine distance between positives by putting them near in unit sphere and minimize for the negative by distributing them in the unit sphere.

$$s_c(q,k) = \frac{q^T.k}{||q||.||k||} \tag{4}$$

$$\mathcal{L} = -\log \frac{\exp(s_c(r_q, r_p)/\tau)}{\exp(s_c(r_q, r_p)/\tau) + \sum_{j=1}^{n} \exp(s_c(r_q, r_{n_j})/\tau)} \tag{5}$$

5.4 Fine-tuning

We place randomly initialized multi-layer regression head, $g\beta$, on top of pretrained e_θ to approximate the ground truth sensor PPG, $y_{p,i} = g_\beta(e_\theta(v_i))$ and $y_{p,i} \in \mathbb{R}^{d1}$ as discussed in 4.2. We apply hyperbolic tan (*tanh*) activation function for the final layer projection to bound the output values. We utilize small labeled set D^l and apply supervised, $\mathcal{L}_{sup}$, regression loss of weighted sum, w_r, w_s, of root-mean-square error, $\mathcal{L}_{RMSE}$, and sign-agreement, $\mathcal{L}_{sign}$, [14] between the target PPG and network estimated PPG signal as per following equation. However, in practice, we fine-tune the training samples of the evaluated dataset to avoid comparing heterogeneous sensors (e.g., wrist PPG of MPSC-rPPG, finger PPG of MERL, UBFC-rPPG) for PPG reconstruction. Further, we utilized single video-PPG to avoid misalignment training during fine-tuning.

We will open-source our codes to enable future research and development.

$$\mathcal{L}_{RMSE} = \sqrt{\frac{||l_i - y_{p,i}||_2^2)^2}{d1}} \tag{6}$$

$$\mathcal{L}_{sign} = ||\delta \circ (l_i \odot y_{p,i})||_1 \tag{7}$$

Where $\odot, \circ$ represents element-wise multiplication and element-wise function operation.

$$\delta(j) = \begin{cases} 1 & j > 0 \\ 0 & else \end{cases} \tag{8}$$

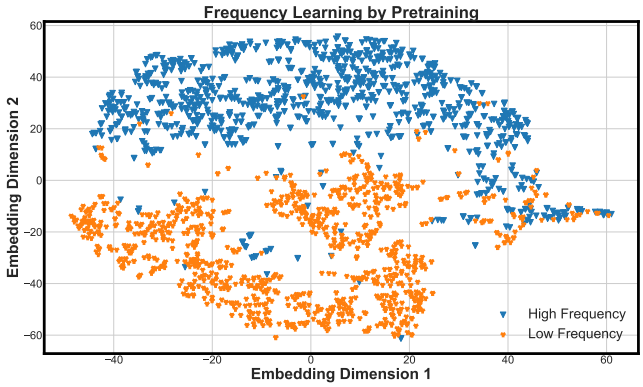

Figure 7: 2-Dimension TNSE plot for different frequency embedding results clustered representation. We can observe that the model model's embedding variation corresponding to the intra-person HR variation.

$$\mathcal{L}_{sup} = w_r \mathcal{L}_{sign} + w_s \mathcal{L}_{RMSE} \tag{9}$$

5.5 Evaluation Criteria

We analyze the representation learning and measure the downstream task performance over the left-out test dataset (unseen subject). We utilize standard t-distributed stochastic neighbor embedding (TSNE) visualization to analyze network-learned the high dimensional representations from unlabeled videos to quantify the self-supervised pretraining. We evaluate the downstream task of PPG estimation by measuring the RMSE between the estimated and target PPG. We calculate HR from the predicted and target PPG by spectrum analysis in the 0.5 to 4Hz frequency range and report their RMSE.

6 RESULT

6.1 Self-supervised Pretraining

We expect our self-supervised approach enables the network to inherently cluster the embedding of the video clips with similar PPG frequency and phase together. We investigate the training set embedding properties across frequency, phase representation, and temporal frame sequence variation to quantify the network performance.

6.1.1 Frequency Representation. Our self-supervised approach successfully learns the distinguished embedding for video frames with low frequency and high-frequency rPPG. We sample test video-clips with underlying low and high frequency PPG (figure 5) contents and visualize their low-dimensional scatter plot in Figure 7. The network representation of low-frequency and high-frequency samples are linearly separable even in the low dimensional embedding space.

6.1.2 Phase Representation. The underlying PPG phase varies periodically. The PPG dissimilarity increasingly varies with the phase to a maximum at half-cycle (180°) and overlaps with the (0°) at full-cycle (360°) (figure 4). We qualitative analysis of the phase

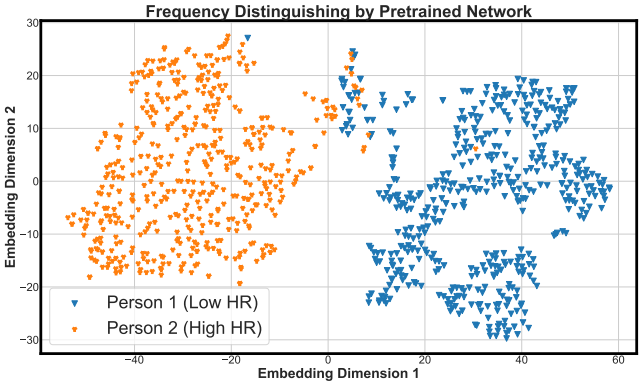

Figure 8: 2-dimension TNSE plot for different person's PPG embedding clusters together. We can observe that the model model's embedding variation corresponding to the inter-person HR variation.

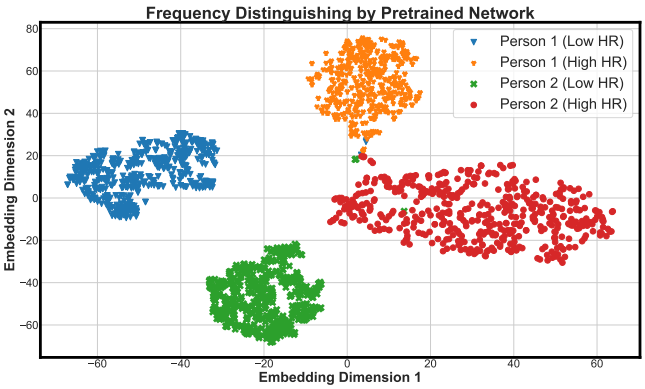

Figure 9: 2-dimension TNSE plot for different person's varying HR PPG embedding clusters together. We can observe that the model separated both the intra-person HR variation and the inter-person HR variation.

represented by the pretrained network. Our self-supervised pretrained model demonstrates the ability to understand the concept PPG phase without explicit PPG reconstruction supervision. We analyze and compare the embedding for multiple video clips of various shifted starting positions and observe the correlation between shifting and embedding. The network representation supports our hypothesis as the embedding dissimilarity continuously increases with the time-shifting reaching a maximum around half-cycle and decreasing afterward to full-cycle shift where the shifted clips' embedding matches with the initial sample clip (figure 10). We can observe the similarity decreases with frame shift, reaches the minimum at half-cycle position ($\sim 24 frames$), and increases from there to the full-cycle ($\sim 48 frames$). The embedding similarity pattern repeats in a cyclic pattern, supporting the phase learning by the pretrained network.

6.1.3 Temporal sequence representation. Our pretraining approach successfully enables the network to understand the temporal frame

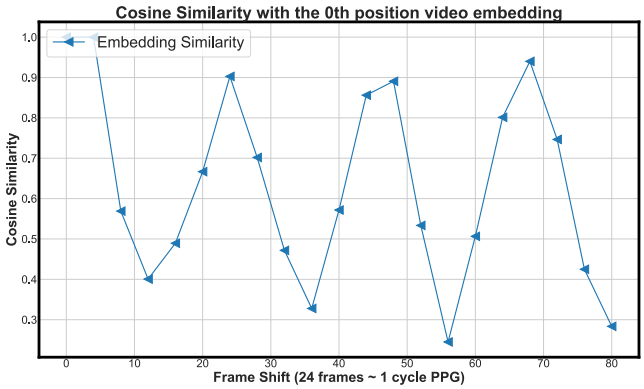

Figure 10: Representation similarity between embedding of varying shifted videos. The cyclic pattern demonstrate that with varying phase the embedding similarity varies in a periodic manner.

sequence and is robust to the example of frame sequence manipulation (repeating or random shuffling). The low-dimensional TSNE embedding cluster the regular frames sequence and sequence of repeated one of the random frames in different groups (Figure 11). We observe the regular frames' sequence embedding cluster together in the scatter plot. However, we find a curious and counter-intuitive multi-cluster behavior of repeated frames' embedding. However, the multi-cluster resembles a better representation, as we created the repeated frames' sequence by randomly selecting one frame for repetition. Since each frame has a different diffusion signal, the network groups them into different clusters based on their underlying diffusion signal.

Our network further identifies the temporal disturbance in the frame sequences. The network generates different embedding for the shuffled frames compared to the regular frame sequences, although both shuffled and the temporal sequential frames have identical spatial information. The scatter plot of figure 12 demonstrates the network's ability to separate the temporal order corrupted frames.

6.2 Downstream Task Performance

6.2.1 Fine-tuning results. We experiment with fine-tuning 4.2 the pretrained model using different amounts of the labeled dataset. Throughout the experiment, only 2 minutes of labeled data (10 % of fully supervised setup) are sufficient to fine-tune the model for PPG estimation without getting overfit. We utilize the trained model to reconstruct the PPG of the left-out test data samples and compare point-wise reconstruction RMSE between the predicted PPG and target PPG to evaluate time-domain reconstruction. Further, we evaluate the model's performance in the frequency domain by calculating the RMSE between the estimated HR of both target and predicted PPG. Table 1 summarizes our overall findings and compares the result with previous supervised and self-supervised approaches. It demonstrates the consistently improved performance over the supervised counterparts for the three datasets (RGB, NIR, RGB mosaic modalities) in both the time and frequency domains.

Table 1: Performance (RMSE) Comparison with Baseline for three Different Dataset.

Approach		SSL Pretraining					Multi-data Tripletaining	Fine-tune (video min.)	RMSE							
		Loss	# of negatives	f	ϕ	t			MPSC-RPPG		MERL RGB, RGB Mosaic)		MERL NIR		UBFC-rPPG	
									PPG	HR	PPG	HR	PPG	HR	PPG	HR
Supervised	CamSense [14]	N/A	N/A	N/A	N/A	N/A	N/A	20	0.1	4.5	0.13	4.9	0.16	4.9	0.12	5.3
	CamSense (Tx) [14]	N/A	N/A	N/A	N/A	N/A	Yes	20	0.08	3.8	0.1	4.2	0.13	4.2	0.1	4.5
	CamSense (MTL) [14]	N/A	N/A	N/A	N/A	N/A	Yes	20	0.13	4.1	0.21	3.6	0.15	3.6	0.12	3.9
	VitaMon [15]	N/A	N/A	N/A	N/A	N/A	N/A	20	N/A	7.6	N/A	5.8	N/A	5.8	N/A	7.2
SSL	[9]	Triplet	Vary	Yes	N/A	N/A	Yes	5	0.2	3.1	0.21	3.5	0.3	3.2	0.2	4.3
	Self-rPPG	Max-Margin	1	Yes	Yes	Yes	Yes	2	0.12	3.8	0.15	3.7	0.18	3.8	0.11	4.3
	Self-rPPG	Triplet	1	Yes	Yes	Yes	Yes	2	0.13	3.5	0.13	3.2	0.12	3.2	0.11	4
	Self-rPPG	InfoNCE	10	Yes	Yes	Yes	Yes	2	0.08	2.5	0.09	2.3	0.1	2.5	0.1	3.1
	Self-rPPG	InfoNCE	15	Yes	Yes	Yes	Yes	2	0.08	2.2	0.08	2.1	0.09	2.2	0.1	2.9
	Self-rPPG	InfoNCE	20	Yes	Yes	Yes	Yes	2	0.09	2.1	0.08	1.9	0.09	2.2	0.1	3.2

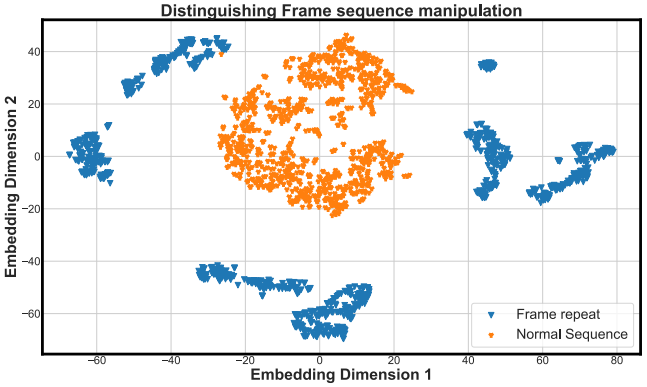

Figure 11: 2-dimension TNSE plot demonstrate the pretrained networks ability to assign different cluster for regular temporal sequence and repeated frames. The multi-cluster separation of repeated frame representation indicates the diffusion signal variation across frames.

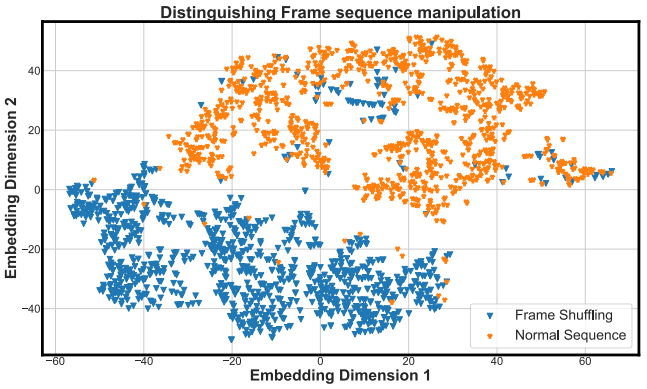

Figure 12: 2-dimension TNSE plot demonstrate the pretrained networks ability to assign different cluster for regular temporal sequence and randomly shuffled frames

7 DISCUSSION

7.1 Benchmark comparison

We compare our *Self-rPPG* with previous end-to-end supervised and self-supervised methods. Our approach provide three major benefit over the supervised end-end setting [14], [15]. Firstly, the self-supervised setting leverage the large unlabeled video data to learn the inherent rPPG mechanism. Secondly, self-supervised methods allow learning rPPG from multiple sample videos without requiring exact video-PPG alignment and PPG sensor homogeneity (similar body position and sampling rate). Finally, the self-supervised pretrained networks perform PPG reconstruction on par with the supervised methods with only 10% of labeled instances compared to the supervised setting. Our approach also improves performances over the previous self-supervised approach [9] in both rPPG representation learning and PPG reconstruction. By incorporating phase and temporal sequence contrastive learning, our approach encodes the complete rPPG physiological mechanisms from the unlabeled facial videos and learns the generalized PPG reconstruction from fewer examples.

7.2 Ablation Study

We ablate over sampling strategies (both positive and negative) and loss-objectives of the self-supervised scheme to better understand their impacts towards generalization and comprehending rPPG physics. We summarize our findings in table 1.

7.2.1 Sampling Strategy. We investigate the impact of each of the three positive (data augmentation) and three proposed negative sample selection techniques (section 4.1.2). We train the network by removing one of the sampling methods and analyzing the learned embedding.

Remarks on Positive Sampling. i) *Face Cropping* augmentation enforces learning PPG information from a partial face view, hence improving generalization and robustness and enabling a partial view-occlusion invariant system. ii) *Channel Selection* provides a data augmentation approach, but in practice, it did not improve the PPG extraction performance. Moreover, mixing the channel selection augmentation, the optimization process slowed down. Since the red and blue channels contain less SNR for PPG and encode PPG differently than the green channel frames and the

network may require different parameters to extract PPG from different channels' frames. iii) *Shifting* contributes in PPG phase understanding 10 and improves the PPG reconstruction in the fine-tune process.

Remark on Negative Sampling. i) *Intra negative* (frame repeating and random frame shuffling) enables the network to understand the temporally correct sequence of the frames. The pretrained network is prone to frame sequence manipulation adversarial attack without intra-negative example training. ii) ϕ *altering* negatives are crucial to avoid collapse under phase variation. They act as hard as the negative for *shifting* positive examples and leads to different representation for differently shifted (hence PPG shift) videos. iii) f *altering* are significant for PPG frequency understanding and contributes to a generalized PPG extraction system compatible with varying HR frequency range. Our proposed approach achieves the best performance by considering all three negative strategies.

7.2.2 Impacts of Loss Function and Architecture. We investigate the impacts of three contrastive loss function on learning rPPG from unlabeled data. We also experiment with modified architecture by enabling multi-layer and single-layer projection heads. We monitor each individual loss function's learning curve, resource usage and down-stream task performances.

Remarks. i) *Loss function:* The InfoNCE loss setting with multiple negative samples outperformed both the energy-based loss functions (max-margin loss and triplet loss) in terms of optimization speed, feature learning, and downstream task performance. We can explain this outcome by the probabilistic nature of InfoNCE loss and its flexibility with sensitivity (temperature τ) design and the ability to incorporate multiple negatives in a single pass. The max-margin loss and triplet loss optimization act comparably as they require much longer iterations to converge by learning to cluster positives and distinguish the negative samples. ii) *Negative sample size:* The optimization performance increases initially with increasing negative sample number and saturated around 15 negative samples with InfoNCE setting. However, the GPU memory usage increases with the negative sample numbers in the InfoNCE. In practice, InfoNCE with 10 negative samples provided a balance between learning good representation efficiently and resource usage. iii) *Projection head:* The multi-layers projection head improves self-supervised performance over the single-layer projection head because of its better learning capacity.

8 LIMITATIONS AND FUTURE WORK

Here, we identify and discuss four potential drawbacks of our self-supervised approaches.

8.1 Learning frame rate instead of frequency

Since we utilize fps variation to provide negative samples for frequency variation, the network may learn to predict the video fps. However, we argue that learning fps variation contributes to better generalization due to the stable nature of the rPPG video. In such steady facial videos, varying fps (a downsampling in our case) does not alter any visual information but significantly changes the underlying PPG incurred diffusion frequency.

8.2 Phase calculation mismatch

We assume a rough HR to calculate the required video shifting to get full-cycle and half-cycle shifted samples. However, due to context (heart activity, persons, exercises), the HR varies in practice, challenging the assumption. A huge HR variation would create a false negative and false positive problem for our self-supervision. However, our datasets have consistent resting HR close to the assumption, enabling phase augmentation without providing false-negative/positive samples.

8.3 Learning to separate person

Besides learning the rPPG representation, the pretrained network unexpectedly learns different embedding for different subjects even if their PPG profiles are close. Since our method contrasts only intra-person video and fails to provide explicit supervision to remove personal information, the model may comprehend the personal information given complete facial video. We can remove personal information using adversarial settings to remove the identifiers to constrain the similarity in embedding distribution for different subjects.

8.4 Treating all negatives equally

In our notion of negative, we put forth negatives where there is are mismatch either in frequency and phase or both. We provide equal weights for all the negatives and ignore the notion of hard negatives and positives. Practically, this assumption may not hold as the temporal intra-frame augmentation destroy underlying PPG, whereas phase and frequency altering contain different PPG frequency or phase than the query samples. We might look into the ranking-based loss function to weigh the negatives differently and provide better learning information during self-supervision.

9 CONCLUSION

We propose and validate rPPG physics-inspired self-supervised methods to learn relevant rPPG features from unlabeled video data. Our self-supervised pretraining enables the network to perform heterogeneous rPPG reconstruction (finger, wrist) after fine-tuning with limited labeled data. We demonstrate the efficacy of our approach by experimenting with heterogeneous camera sensors (DSRL, HD, NIR) of three public datasets. We further investigate the impact of our self-supervised components to elaborate on their impact on learning. Our insights enable further fine-grain development of generalized, robust rPPG systems using unlabeled video data.

ACKNOWLEDGMENT

This work has been partially supported by NSF CAREER Award #1750936 and U.S.Army Grant #W911NF2120076.

REFERENCES

[1] Ahmed Alqaraawi, Ahmad Alwosheel, and Amr Alasaad. 2016. Heart rate variability estimation in photoplethysmography signals using Bayesian learning approach. *Healthcare technology letters* 3, 2 (2016), 136–142.
[2] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. 2019. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters* 124 (2019), 82–90.
[3] Denisse Castaneda, Aibhlin Esparza, Mohammad Ghamari, Cinna Soltanpur, and Homer Nazeran. 2018. A review on wearable photoplethysmography sensors

and their potential future applications in health care. *International journal of biosensors & bioelectronics* 4, 4 (2018), 195.

[4] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. 2020. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems* 33 (2020), 22243–22255.

[5] Weixuan Chen and Daniel McDuff. 2018. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Computer Vision – ECCV 2018*. Springer International Publishing, Cham, 349–365.

[6] Sumit Chopra, Raia Hadsell, and Yann Lecun. 2005. Learning a similarity metric discriminatively, with application to face verification. *Proc. Computer Vision and Pattern Recognition* 1, 539– 546 vol. 1. https://doi.org/10.1109/CVPR.2005.202

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

[8] Halil Demirezen and Cigdem Eroglu Erdem. 2018. Remote photoplethysmography using nonlinear mode decomposition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1060–1064.

[9] John Gideon and Simon Stent. 2021. 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*. IEEE, 3995–4004.

[10] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, 580–587.

[11] Amogh Gudi, Marian Bittner, Roelof Lochmans, and Jan C. van Gemert. 2019. Efficient Real-Time Camera Based Estimation of Heart Rate and Its Variability. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* (2019), 1570–1579.

[12] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 2. IEEE, 1735–1742.

[13] Zahid Hasan, Emon Dey, Sreenivasan Ramasamy Ramamurthy, Nirmalya Roy, and Archan Misra. 2022. Rhythmedge: Enabling contactless heart rate estimation on the edge. In *2022 IEEE International Conference on Smart Computing (SMARTCOMP)*. IEEE, IEEE, 92–99.

[14] Zahid Hasan, Sreenivasan Ramasamy Ramamurthy, and Nirmalya Roy. 2022. CamSense: A camera-based contact-less heart activity monitoring. *Smart Health* 23 (2022), 100240.

[15] Sinh Huynh, Rajesh Krishna Balan, JeongGil Ko, and Youngki Lee. 2019. VitaMon: measuring heart rate variability using smartphone front camera. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*. 1–14.

[16] Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. 2020. Contrastive representation learning: A framework and review. *IEEE Access* (2020).

[17] Eugene Lee, Evan Chen, and Chen-Yi Lee. 2020. Meta-rppg: Remote heart rate estimation using a transductive meta-learner. In *European Conference on Computer Vision*. Springer, 392–409.

[18] Richard Macwan, Yannick Benezeth, and Alamin Mansouri. 2019. Heart rate estimation using remote photoplethysmography with multi-objective optimization. *Biomedical Signal Processing and Control* 49 (2019), 24–33.

[19] Ewa Magdalena Nowara, Tim K Marks, Hassan Mansour, and Ashok Veeraraghavan. 2018. Sparseppg: Towards driver monitoring using camera-based vital signs estimation in near-infrared. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 1272–1281.

[20] Daniel McDuff, Sarah Gontarek, and Rosalind W Picard. 2014. Remote detection of photoplethysmographic systolic and diastolic peaks using a digital camera. *IEEE Transactions on Biomedical Engineering* 61, 12 (2014), 2948–2954.

[21] Daniel J McDuff, Ethan B Blackford, and Justin R Estepp. 2017. 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)*. IEEE, 63–70.

[22] David M Mirvis and Ary L Goldberger. 2001. Electrocardiography. *Heart disease* 1 (2001), 82–128.

[23] Ewa Magdalena Nowara, Tim K Marks, Hassan Mansour, and Ashok Veeraraghavany. 2018. SparsePPG: towards driver monitoring using camera-based vital signs estimation in near-infrared. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 1353–135309.

[24] Ahmed Osman and et al. 2015. Supervised learning approach to remote heart rate estimation from facial videos. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 1. IEEE, 1–6.

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

[26] Harvard Health Publishing. 2020. How's your heart rate and why it matters? https://www.health.harvard.edu/heart-health/hows-your-heart-rate-and-why-it-matters

[27] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, 815–823.

[28] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, 1–9.

[29] Li Tao, Xueting Wang, and Toshihiko Yamasaki. 2020. Self-supervised video representation learning using inter-intra contrastive framework. In *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, 2193–2201.

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

[31] Wenjin Wang, Albertus C den Brinker, and Gerard De Haan. 2019. Discriminative signatures for remote-PPG. *IEEE Transactions on Biomedical Engineering* 67, 5 (2019), 1462–1473.

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

[33] Kilian Q Weinberger, John Blitzer, and Lawrence K Saul. 2006. Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems*. 1473–1480.

[34] Bing-Fei Wu, Po-Wei Huang, Chun-Hsien Lin, Meng-Liang Chung, Tsong-Yang Tsou, and Yu-Liang Wu. 2018. Motion resistant image-photoplethysmography based on spectral peak tracking algorithm. *IEEE Access* 6 (2018), 21621–21634.

[35] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems* 33 (2020), 6256–6268.

[36] Qi Zhan, Wenjin Wang, and Gerard de Haan. 2020. Analysis of CNN-based remote-PPG to understand limitations and sensitivities. *Biomedical Optics Express* 11, 3 (2020), 1268–1283.

[37] Dongfeng Zhang, Weijing Wang, and Fang Li. 2016. Association between resting heart rate and coronary artery disease, stroke, sudden death and noncardiovascular diseases: a meta-analysis. *Cmaj* 188, 15 (2016), E384–E392.

[38] Qi Zhang, Yimin Zhou, Shuang Song, Guoyuan Liang, and Haiyang Ni. 2018. Heart rate extraction based on near-infrared camera: Towards driver state monitoring. *IEEE Access* 6 (2018), 33076–33087.

[39] Zhilin Zhang. 2015. Photoplethysmography-based heart rate monitoring in physical activities via joint sparse spectrum reconstruction. *IEEE transactions on biomedical engineering* 62, 8 (2015), 1902–1910.

[40] Changchen Zhao, Chun-Liang Lin, Weihai Chen, and Zhengguo Li. 2018. A novel framework for remote photoplethysmography pulse extraction on compressed videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 1299–1308.