Contrast-Phys+: Unsupervised and Weakly-supervised Video-based Remote Physiological Measurement via Spatiotemporal Contrast

Zhaodong Sun and Xiaobai Li

Abstract—Video-based remote physiological measurement utilizes facial videos to measure the blood volume change signal, which is also called remote photoplethysmography (rPPG). Supervised methods for rPPG measurements have been shown to achieve good performance. However, the drawback of these methods is that they require facial videos with ground truth (GT) physiological signals, which are often costly and difficult to obtain. In this paper, we propose Contrast-Phys+, a method that can be trained in both unsupervised and weakly-supervised settings. We employ a 3DCNN model to generate multiple spatiotemporal rPPG signals and incorporate prior knowledge of rPPG into a contrastive loss function. We further incorporate the GT signals into contrastive learning to adapt to partial or misaligned labels. The contrastive loss encourages rPPG/GT signals from the same video to be grouped together, while pushing those from different videos apart. We evaluate our methods on five publicly available datasets that include both RGB and Near-infrared videos. Contrast-Phys+ outperforms the state-of-the-art supervised methods, even when using partially available or misaligned GT signals, or no labels at all. Additionally, we highlight the advantages of our methods in terms of computational efficiency, noise robustness, and generalization. Our code is available at https://github.com/zhaodongsun/contrast-phys.

Index Terms—Remote Photoplethysmography, Face Video, Unsupervised Learning, Weakly-supervised Learning, Semi-supervised Learning, Contrastive Learning

1 Introduction

In the realm of traditional physiological measurement, skin-contact sensors are commonly employed to capture physiological signals. Examples of such sensors include contact photoplethysmography (PPG) and electrocardiography (ECG). These sensors enable the derivation of crucial physiological parameters such as heart rate (HR), respiration frequency (RF), and heart rate variability (HRV). However, the reliance on skin-contact sensors necessitates specialized biomedical equipment like pulse oximeters, which can lead to discomfort and skin irritation. An alternative approach is remote physiological measurement, which employs cameras to record facial videos for the measurement of remote photoplethysmography (rPPG). This technique harnesses the ability of cameras to capture subtle color changes in the human face, from which multiple physiological parameters including HR, RF, and HRV can be extracted [1]. Unlike traditional methods, video-based physiological measurement relies on readily available cameras rather than specialized biomedical sensors. This approach offers the advantage of not being constrained by physical proximity, rendering it particularly promising for applications in remote healthcare [2], [3], [4], emotion analysis [5], [6], [7], [8], and face security [9], [10], [11].

- *Z. Sun is with the Center for Machine Vision and Signal Analysis, University of Oulu, Oulu, Finland. E-mail: zhaodong.sun@oulu.fi*
- *X. Li is with The State Key Laboratory of Blockchain and Data Security, Zhejiang University, Hangzhou China, and the Center for Machine Vision and Signal Analysis, University of Oulu, Oulu, Finland. E-mail: xiaobai.li@oulu.fi (Corresponding author: Xiaobai Li)*

In earlier studies related to rPPG [1], [12], [13], [14], researchers devised handcrafted features to extract rPPG signals. Subsequently, several deep learning (DL)-based methods [15], [16], [17], [18], [19], [20], [21], [22], [23], [24] were introduced. These DL-based approaches utilize supervised techniques and diverse network architectures to measure rPPG signals. Under certain conditions, such as when head movements are present or the videos exhibit heterogeneity, DL-based methods tend to exhibit greater robustness compared to traditional handcrafted approaches. However, it's important to note that DL-based rPPG methods heavily rely on extensive datasets comprising face videos and ground truth (GT) physiological signals. Acquiring GT physiological signals, typically measured by contact sensors and synchronized with facial videos, can be a costly endeavor. Issues like missing GT signals or misalignment with facial videos during data collection are common challenges encountered in this context.

Considering the cost and challenges associated with obtaining GT physiological signals, we propose an unsupervised and weakly-supervised method for rPPG measurement, particularly when dealing with data that lacks complete or high-quality labels. The unsupervised method can effectively process facial videos that lack GT signals, while the weakly-supervised method can be employed when dealing with data containing incomplete or low-quality labels, where GT signals may be missing or misaligned.

In our approach, we leverage four key rPPG observations as foundational knowledge. 1) **rPPG spatial similarity**: rPPG signals obtained from different facial regions tend to

exhibit similar power spectrum densities (PSDs). 2) **rPPG temporal similarity**: Segments of rPPG data taken within short time intervals (e.g., two consecutive 5-second clips) typically display similar PSDs, as HR tends to transition smoothly in most cases. 3) **Cross-video rPPG dissimilarity**: PSDs of rPPG signals from different videos often exhibit variations. 4) **HR range constraint**: The HR typically falls within the range of 40 to 250 beats per minute (bpm).

In our prior ECCV 2022 publication [25], we introduced Contrast-Phys, an unsupervised learning framework. The present work, referred to as Contrast-Phys+, represents an extension of our earlier research. This work contributes significantly in the following ways:

- We propose Contrast-Phys+, a versatile model capable of adapting to diverse data conditions, including scenarios with no labels, partial labels, or misaligned labels. Importantly, Contrast-Phys+ operates effectively in both unsupervised and weakly-supervised settings. To the best of our knowledge, Contrast-Phys+ is the first work to train an rPPG model in both weakly-supervised and unsupervised settings.
- We showcase the efficacy of Contrast-Phys+ in weakly-supervised scenarios, where some ground truth signals may be missing or lack synchronization. Remarkably, Contrast-Phys+ with missing labels exhibits performance that can surpass that of fully supervised methods employing complete label sets. Moreover, Contrast-Phys+ demonstrates significantly enhanced robustness when faced with ground truth signal desynchronization, outperforming other fully supervised methods.
- We conduct extensive experiments and analyses pertaining to Contrast-Phys+. A comprehensive performance comparison is also offered, contrasting the capabilities of Contrast-Phys+ against recent state-of-the-art baselines. Additional experiments also demonstrate that Contrast-Phys+ can use unlabeled data to expand and diversify the training dataset for improved generalization. We also offer a thorough analysis of the reasons why Contrast-Phys+ can be effective in unsupervised and weakly-supervised scenarios. Besides, we present statistical analysis to validate the proposed rPPG observations and include detailed ablation studies to substantiate the effectiveness of Contrast-Phys+.

2 Related Work

2.1 Video-Based Remote Physiological Measurement

The concept of measuring remote photoplethysmography (rPPG) from facial videos via the green channel was initially introduced by Verkruysse et al. [12]. Subsequently, various handcrafted methods [1], [13], [14], [26], [27], [28], [29], [30] were proposed to enhance the quality of rPPG signals. These methods, predominantly developed in the earlier years, relied on manual procedures and did not necessitate training datasets, earning them the label of "traditional methods." In recent years, deep learning (DL) techniques have surged in rPPG measurement. Some studies [15], [16], [20], [24], [31] employed a 2D convolutional neural network (2DCNN) with two consecutive video frames as input for rPPG estimation. Another category of DL-based methods [21], [22], [23], [32], [33] utilized spatial-temporal signal maps extracted from various facial regions as input for 2DCNN models. Additionally, 3DCNN-based methods [17], [18], [34] were introduced to achieve high performance, particularly on compressed videos [18]. These DL-based approaches, categorized as supervised methods, demand both facial videos and ground truth (GT) physiological signals for training. More recently, Wang et al. [35] proposed a self-supervised rPPG method to acquire rPPG representations, although it still necessitates heart rate (HR) labels for fine-tuning the rPPG model. Gideon et al. [34] introduced the first unsupervised rPPG method, which does not rely on GT physiological signals for training. However, this method, while pioneering, exhibits lower accuracy compared to state-of-the-art supervised methods and can be sensitive to external noise. Subsequent to these developments, multiple unsupervised rPPG techniques have emerged [25], [36], [37], [38]. These unsupervised rPPG methods have gained attention because they solely require facial videos for training, eliminating the need for GT signals, yet they achieve performance levels similar to those of supervised methods. This is particularly advantageous given the expense associated with collecting GT signals alongside facial videos. However, none of the methods above considered utilizing partial or low-quality labels to further refine rPPG signal quality.

2.2 Contrastive Learning

Contrastive learning, a widely adopted self-supervised learning technique in computer vision tasks, empowers deep learning models to map high-dimensional images or videos into lower-dimensional feature embeddings without the need for labeled data [39], [40], [41], [42], [43], [44], [45], [46], [47]. Its primary objective is to ensure that features derived from different perspectives of the same sample (referred to as positive pairs) are brought closer together, while features from different samples (referred to as negative pairs) are pushed apart. This approach finds extensive utility in pre-training models, thereby facilitating subsequent task-specific training in domains such as image classification [42], video analysis [47], [48], face recognition [40], and face detection [49]. This is particularly advantageous in situations characterized by limited access to labeled data. In our research, we leverage prior knowledge related to remote photoplethysmography (rPPG) to generate suitable positive and negative pairs of rPPG signal instances for contrastive learning. Diverging from prior methodologies that focus on feature embedding, our proposed method, Contrast-Phys+, possesses the capability to directly generate rPPG signals without the need for labeled data, thereby enabling unsupervised learning. Additionally, we harness ground truth (GT) signals to construct positive/negative pairs for contrastive learning, thus facilitating end-to-end weakly-supervised training even in scenarios where labels are missing or of suboptimal quality.

3 Observations about rPPG

This section describes four observations about rPPG, which are the preconditions to designing Contrast-Phys+ and en-

abling unsupervised and weakly-supervised learning.

3.1 rPPG Spatial Similarity

rPPG signals originating from various facial regions exhibit analogous waveforms, accompanied by the similarity in their Power Spectrum Densities (PSDs). This spatial coherence in rPPG signals has been leveraged in the design of multiple methodologies, as demonstrated in prior works [27], [28], [29], [50], [51], [52], [53]. While subtle phase and amplitude disparities may exist in the temporal domain when comparing rPPG signals from distinct skin areas [54], [55], these distinctions become inconsequential when rPPG waveforms are analyzed in the frequency domain, where PSDs are normalized. As illustrated in Fig. 1, the rPPG waveforms derived from four distinct spatial regions share a striking resemblance, characterized by identical peaks in their respective PSDs.

3.2 rPPG Temporal Similarity

The heart rate (HR) undergoes gradual changes within short time frames, as noted by Gideon et al. [34]. A similar finding was reported by Stricker et al. [56], who observed slight HR variations in their dataset over short time intervals. Given that HR is prominently represented by a dominant peak in the PSD, it follows that the PSD experiences minimal fluctuations as well. Therefore, when randomly selecting small temporal windows from a brief rPPG segment (e.g., 10 seconds), one can anticipate that the PSDs of these windows will exhibit similarity. As depicted in Fig. 2, we illustrate this by sampling two 5-second windows from a 10-second rPPG signal and comparing the PSDs of these windows. Indeed, the two PSDs demonstrate similarity, with dominant peaks occurring at identical frequencies. It is important to note that this observation holds true when dealing with short-term rPPG signals. We will delve into the impact of signal duration on our model's performance in the forthcoming ablation study.

We can summarize spatiotemporal rPPG similarity using the following relation.

$$\mathrm{PSD}\{G(v(t_1 \rightarrow t_1 + \Delta t, \mathcal{H}_1, \mathcal{W}_1))\} \approx \mathrm{PSD}\{G(v(t_2 \rightarrow t_2 + \Delta t, \mathcal{H}_2, \mathcal{W}_2))\} \tag{1}$$

In this relation, $v \in \mathbb{R}^{T\times H\times W\times 3}$ represents a facial video, and G signifies an rPPG measurement algorithm. We can select a facial region defined by height $\mathcal{H}_1$ and width $\mathcal{W}_1$ and a time interval $t_1 \rightarrow t_1 + \Delta t$ from video v to derive one rPPG signal. A similar rPPG signal can be obtained from the same video, utilizing parameters $\mathcal{H}_2$, $\mathcal{W}_2$, and $t_2 \rightarrow t_2 + \Delta t$. To meet the criteria for short-term rPPG signals, the temporal separation $|t_1 - t_2|$ should remain small.

3.3 Cross-video rPPG Dissimilarity

rPPG signals obtained from different facial videos exhibit distinct PSDs. This divergence arises from the fact that each video features distinct individuals with varying physiological conditions, such as physical activity and emotional states, which are known to influence HRs [57]. Even in cases where HRs between two videos may appear similar, disparities in the PSDs can persist. This is due to the presence of additional physiological factors within the PSDs, such as respiration rate [58] and HRV [59], which are unlikely to align entirely across different videos. To substantiate this observation, we conducted an analysis involving the calculation of mean squared errors for cross-video PSD pairs within the OBF dataset [60]. The results, illustrated in Fig. 3, underscore the primary dissimilarity in cross-video PSDs as being centered around the heart rate peak.

This cross-video rPPG dissimilarity is described by the following relation:

$$\mathrm{PSD}\{G(v(t_1 \rightarrow t_1 + \Delta t, \mathcal{H}_1, \mathcal{W}_1))\} \neq \mathrm{PSD}\{G(v'(t_2 \rightarrow t_2 + \Delta t, \mathcal{H}_2, \mathcal{W}_2))\} \tag{2}$$

where v and v' represent two distinct videos. By selecting specific facial areas and time intervals from these videos, one can expect the PSDs of the two resulting rPPG signals to exhibit noticeable differences.

3.4 HR Range Constraint

The typical HR range for the majority of individuals falls within the interval of 40 to 250 beats per minute (bpm) [61]. In line with established practices [1], [62], this HR range serves as the basis for rPPG signal filtering, with the highest peak identified within this range to estimate HR. Consequently, our method will primarily concentrate on PSD within the frequency band of 0.66 Hz to 4.16 Hz.

4 METHOD

In this section, we propose Contrast-Phys+ for weakly-supervised and unsupervised rPPG learning as shown in Fig. 4. We describe the face preprocessing in Sec. 4.1, the ST-rPPG block representation in Sec. 4.2, the rPPG spatiotemporal sampling in Sec. 4.3, and the contrastive loss function in Sec. 4.5.

4.1 Preprocessing

The initial step involves preprocessing the original video, and the primary task is facial cropping. Utilizing OpenFace [63], we generate facial landmarks. To determine the central facial point for each frame, we compute the minimum and maximum horizontal and vertical coordinates of these landmarks. Subsequently, a bounding box is established, sized at 1.2 times the vertical coordinate range of the landmarks observed in the initial frame, and this size remains constant for all subsequent frames. With the central facial point and bounding box size determined for each frame, we proceed to crop the face in every frame. These cropped facial regions are then resized to dimensions of 128×128, rendering them ready for input into our model.

4.2 Spatiotemporal rPPG (ST-rPPG) Block Representation

We have adapted the 3DCNN-based PhysNet [17] to compute the ST-rPPG block representation. Our modified model takes as input an RGB video with dimensions $T\times 128\times 128\times 3$, where T represents the number of frames. In the final

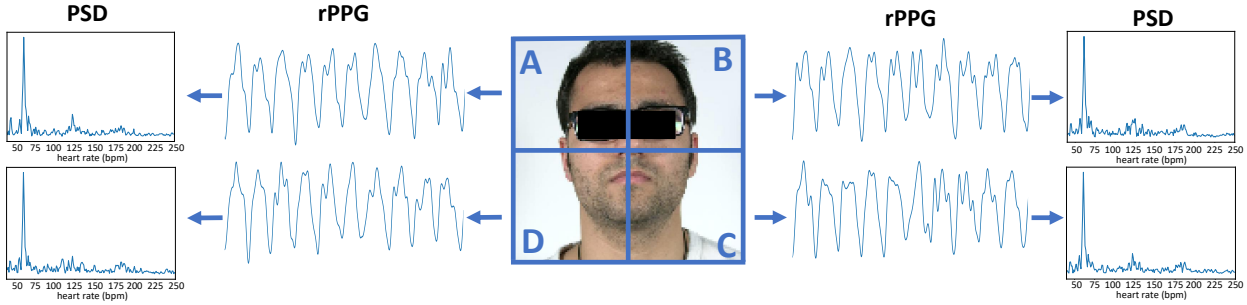


Fig. 1. Illustration of rPPG spatial similarity. The rPPG signals from four facial areas (A, B, C, D) have similar waveforms and power spectrum densities (PSDs).

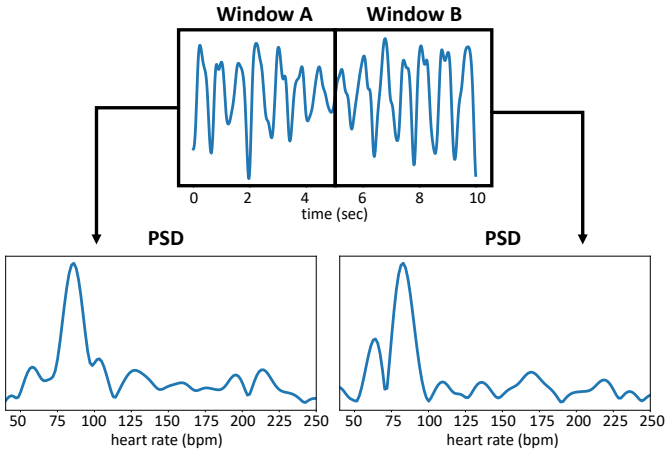


Fig. 2. Illustration of rPPG temporal similarity. The rPPG signals from two temporal windows (A, B) have similar PSDs.

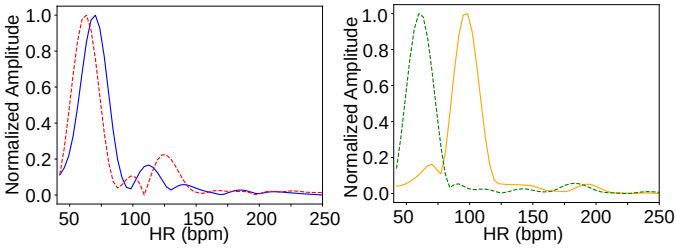


Fig. 3. The most similar (left) and most different (right) cross-video PSD pairs in the OBF dataset.

stage of our model, we employ adaptive average pooling to perform downsampling along spatial dimensions, enabling control over the output spatial size. This alteration facilitates the generation of a spatiotemporal rPPG block with dimensions $T \times S \times S$, where S denotes the length of the spatial dimension, as depicted in Fig. 5. Further elaboration on the 3DCNN model is available in the supplementary material.

The ST-rPPG block is essentially a collection of rPPG signals embedded within spatiotemporal dimensions. To denote this ST-rPPG block, we use $P \in \mathbb{R}^{T \times S \times S}$. When selecting a specific spatial location (h, w) within the ST-rPPG block, the corresponding rPPG signal $P(\cdot, h, w)$ is extracted from the receptive field associated with that spatial position in the input video. It is worth noting that when the spatial dimension length S is small, each spatial position within the ST-rPPG block encompasses a larger receptive field, albeit with fewer rPPG signals contained within the block. Importantly, the receptive field of each spatial position in the ST-rPPG block encompasses a portion of the facial region, ensuring that all spatial positions in the ST-rPPG block encompass valuable rPPG information.

4.3 rPPG Spatiotemporal Sampling

Algorithm 1 rPPG Spatiotemporal Sampling

Input: ST-rPPG block: P with shape $T \times S \times S$, Number of rPPG samples per spatial location: K, The default rPPG sample length $\Delta t = T/2$

1: Initialze an empty list H for storing all rPPG samples
2: **for** $h, w \in \{1, ..., S\}, \{1, ..., S\}$ **do** ▷ Loop over all spatial locations
3: **for** $k \in \{1, ..., K\}$ **do** ▷ K rPPG samples per spatial location
4: Randomly choose a starting time t between 0 and $T - \Delta t$
5: Append the rPPG sample $P(t \rightarrow t + \Delta t, h, w)$ into the list H
6: **end for**
7: **end for**

Output: The list $H = [p_1, ..., p_N]$ containing rPPG samples

In the process of generating rPPG samples from the ST-rPPG block, as depicted in Fig. 5, which is the spatial and temporal sampler illustrated in Fig. 4, we employ both spatial and temporal sampling techniques. For spatial sampling, we extract the rPPG signal denoted as $P(\cdot, h, w)$ from a specific spatial position. In the case of temporal sampling, we select a short time interval from $P(\cdot, h, w)$, resulting in the final spatiotemporal sample, denoted as $P(t \rightarrow t + \Delta t, h, w)$, where h and w represent the spatial position, t signifies the starting time, and Δt signifies the duration of the time interval.

Given an ST-rPPG block, we iterate through all spatial positions and extract K rPPG clips, each with a randomly selected starting time t, for each spatial position as shown in Algorithm 1. Consequently, we obtain a total of $N = S \cdot S \cdot K$ rPPG clips from the ST-rPPG block. It is important to note that these sampling procedures are employed to generate multiple rPPG samples for use in contrastive learning during the model training phase. During inference, the ST-rPPG block is spatially averaged to yield the final rPPG signal.

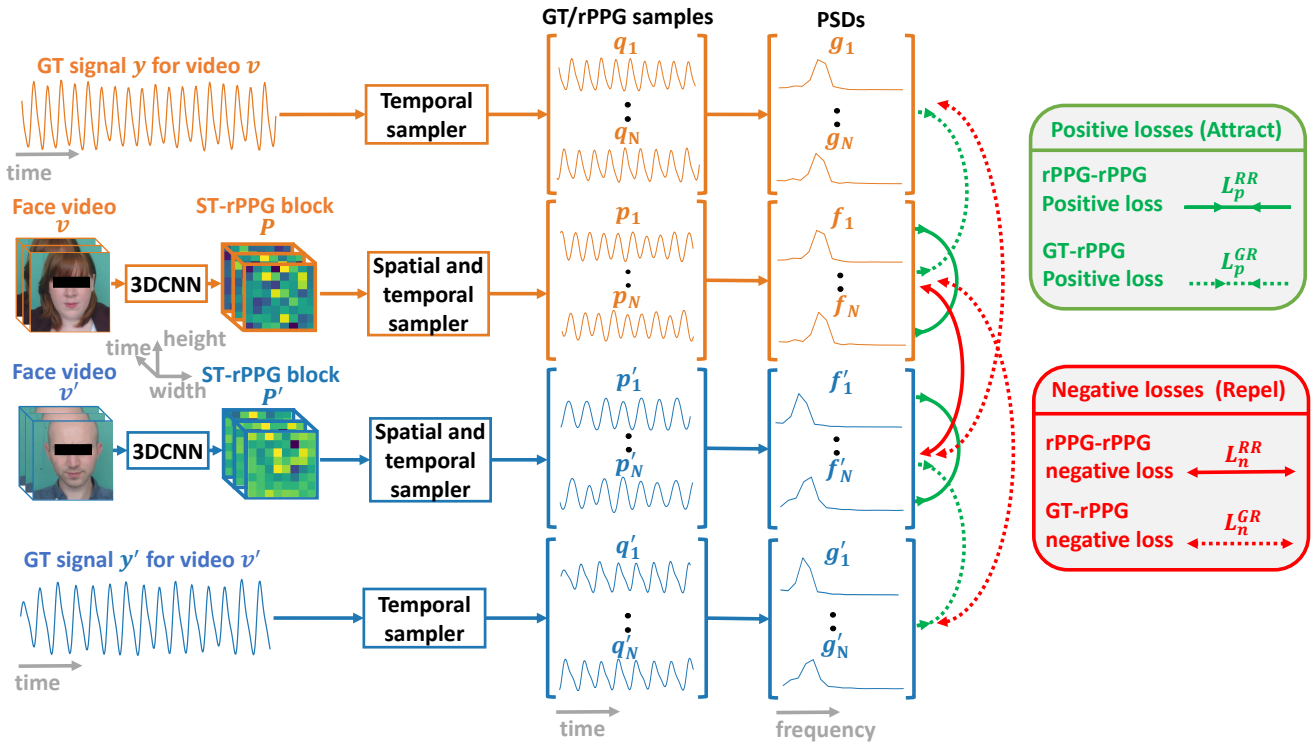


Fig. 4. The diagram of Contrast-Phys+ for weakly-supervised or unsupervised learning.

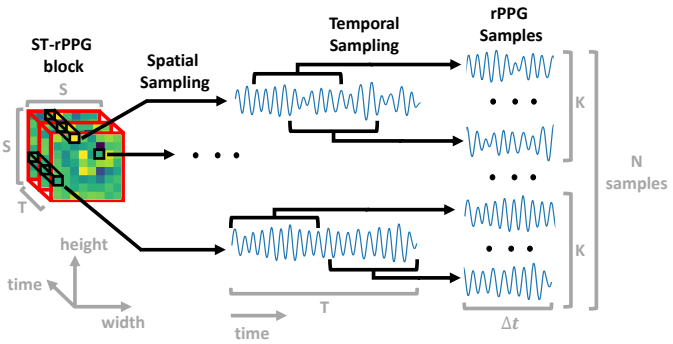


Fig. 5. Spatial and temporal Sampler for an ST-rPPG Block.

4.4 GT Signal Temporal Sampling

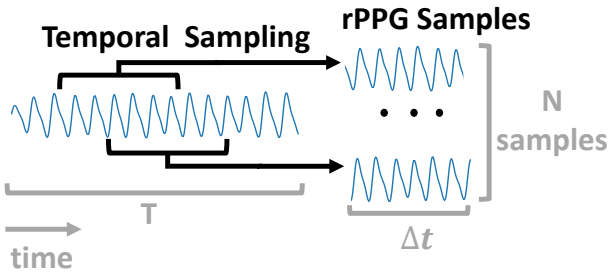


Fig. 6. Temporal Sampler for a GT Signal.

Unlike ST-rPPG blocks, which encompass spatiotemporal signals, GT signals, which are one-dimensional temporal signals, necessitate different sampling approaches. Given the dimensional disparity between GT signals and ST-rPPG blocks, we employ temporal sampling for GT signals and spatiotemporal sampling for ST-rPPG blocks. The GT signal temporal sampling process, as depicted in Fig. 6, entails selecting a short time interval from the GT signal y, resulting in a temporal sample denoted as $y(t \rightarrow t + \Delta t)$, where t represents the starting time, and Δt signifies the duration of the time interval. For a single GT signal, we sample N GT clips, each with a randomly determined starting time t.

As illustrated in both the top and bottom branches of Fig. 4, the GT signals y and y' corresponding to the facial videos v and v' undergo temporal sampling, generating two sets of GT samples, namely $[q_1, ..., q_N]$ and $[q'_1, ..., q'_N]$, respectively. Subsequently, these two sets of GT samples are transformed into two sets of PSDs, denoted as $[g_1, ..., g_N]$ and $[g'_1, ...g'_N]$, respectively.

4.5 Contrastive Loss for Contrast-Phys+

As illustrated in Fig. 4, our process begins with the selection of two distinct videos randomly chosen from a dataset as input. For each video, we derive an ST-rPPG block denoted as P, a set of rPPG samples $[p_1, \ldots, p_N]$, and their corresponding rPPG PSDs $[f_1, \ldots, f_N]$. If the GT signal is available for this video, we additionally obtain a set of GT signal samples $[q_1, \ldots, q_N]$ and their corresponding GT PSDs $[g_1, \ldots, g_N]$. This procedure is repeated for the second video. Note that we normalize each PSD by dividing it by its summation, ensuring all PSDs are on the same scale.

As shown in Fig. 4 (right), the underlying principle of our contrastive loss lies in the alignment of GT-rPPG or rPPG-rPPG PSD pairs stemming from the same video, while simultaneously pushing apart GT-rPPG or rPPG-rPPG PSD pairs originating from different videos. The HR range constraint is imposed after the rPPG signals are converted to PSDs. We only keep PSD values within the HR range of 0.66

Hz to 4.16 Hz (corresponding to 40 to 250 beats per minute) while removing the PSD values outside the HR range, so only the PSD values within the HR range are used in our method. The HR range constraint has no impact on the input videos.

rPPG-rPPG Positive Loss. In accordance with the rPPG spatiotemporal similarity, it is expected that the rPPG PSDs resulting from spatiotemporal sampling of the same ST-rPPG block should exhibit similarity. The following relations outline this property for the two input videos:

For one video:

$$\begin{aligned}&\text{PSD}\big\{P(t_1 \to t_1+\Delta t, h_1, w_1)\big\} \approx \text{PSD}\big\{P(t_2 \to t_2+\Delta t, h_2, w_2)\big\}\\ &\implies f_i \approx f_j, i \neq j\end{aligned} \tag{3}$$

For the other video:

$$\begin{aligned}&\text{PSD}\big\{P'(t_1 \to t_1+\Delta t, h_1, w_1)\big\} \approx \text{PSD}\big\{P'(t_2 \to t_2+\Delta t, h_2, w_2)\big\}\\ &\implies f'_i \approx f'_j, i \neq j\end{aligned} \tag{4}$$

To bring together the rPPG PSDs from the same video, we employ the mean squared error as the loss function for rPPG-rPPG positive pairs, denoted as (f_i, f_j). The rPPG-rPPG positive loss term, L_p^{RR}, is presented below, and it is normalized with respect to the total number of rPPG-rPPG positive pairs.

$$L_p^{RR} = \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} \frac{\parallel f_i - f_j \parallel^2 + \parallel f'_i - f'_j \parallel^2}{2N(N-1)} \tag{5}$$

rPPG-rPPG Negative Loss. In accordance with the cross-video rPPG dissimilarity, it is expected that the rPPG PSDs resulting from spatiotemporal sampling of two different ST-rPPG blocks should differ. We can employ the following relation to describe this property for the two input videos:

$$\begin{aligned}&\text{PSD}\big\{P(t_1 \to t_1+\Delta t, h_1, w_1)\big\} \neq \text{PSD}\big\{P'(t_2 \to t_2+\Delta t, h_2, w_2)\big\}\\ &\implies f_i \neq f'_j\end{aligned} \tag{6}$$

To separate the rPPG PSDs originating from two different videos, we utilize the negative mean squared error as the loss function for rPPG-rPPG negative pairs, represented as (f_i, f'_j). The rPPG-rPPG negative loss term, denoted as L_n^{RR}, is presented below, and it is normalized with respect to the total number of rPPG-rPPG negative pairs.

$$L_n^{RR} = -\sum_{i=1}^{N}\sum_{j=1}^{N} \parallel f_i - f'_j \parallel^2 / N^2 \tag{7}$$

GT-rPPG Positive Loss. Inspired by the rPPG temporal similarity, it is expected that the rPPG PSDs from temporal sampling of the ST-rPPG block and the GT PSDs from temporal sampling of the corresponding GT signal should be similar since GT signals are the reference of rPPG signals. The following relations outline this property.

For one input video and the corresponding GT signal:

$$\begin{aligned}&\text{PSD}\big\{P(t_1 \to t_1+\Delta t, h_1, w_1)\big\} \approx \text{PSD}\big\{y(t_2 \to t_2+\Delta t)\big\}\\ &\implies f_i \approx g_j\end{aligned} \tag{8}$$

For the other video and the corresponding GT signal:

$$\begin{aligned}&\text{PSD}\big\{P'(t_1 \to t_1+\Delta t, h_1, w_1)\big\} \approx \text{PSD}\big\{y'(t_2 \to t_2+\Delta t)\big\}\\ &\implies f'_i \approx g'_j\end{aligned} \tag{9}$$

The GT-rPPG positive loss L_p^{GR} is to pull together rPPG PSDs from one ST-rPPG block and GT PSDs from the corresponding GT signal (GT-rPPG positive pairs, e.g., (f_i, g_j) where f_i is from ST-rPPG block P and g_j is from the corresponding GT signal y) so that the model is encouraged to output rPPG signals similar to the corresponding GT signals. Note that this GT-rPPG positive loss does not require exactly synchronized GT signals since rPPG PSDs and GT PSDs are from rPPG samples and GT samples which are randomly temporally sampled from the ST-rPPG block and the GT signal. This indicates that GT-rPPG positive loss does not need the alignment information between the GT signal and the video. Since it is assumed that some videos may not have GT signals in weakly-supervised learning, the function ϕ is defined below to return whether a video has a GT signal.

$$\phi(v) = \begin{cases}1, & \text{video } v \text{ has a GT signal}\\ 0, & \text{otherwise}\end{cases} \tag{10}$$

The GT-rPPG positive loss term L_p^{GR} is defined below, which is normalized by the number of GT-rPPG positive pairs.

$$L_p^{GR} = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\phi(v) \parallel f_i - g_j \parallel^2 + \phi(v') \parallel f'_i - g'_j \parallel^2}{\big(\phi(v)+\phi(v')\big)N^2} \tag{11}$$

GT-rPPG Negative Loss. Like the cross-video rPPG dissimilarity, it is expected that the rPPG PSDs sampled from the ST-rPPG block and the GT PSDs from temporal sampling of the non-corresponding GT signal should be different. The following relations illustrate this property.

For one input video and the non-corresponding GT signal:

$$\begin{aligned}&\text{PSD}\big\{P(t_1 \to t_1+\Delta t, h_1, w_1)\big\} \neq \text{PSD}\big\{y'(t_2 \to t_2+\Delta t)\big\}\\ &\implies f_i \neq g'_j\end{aligned} \tag{12}$$

For the other input video and the non-corresponding GT signal:

$$\begin{aligned}&\text{PSD}\big\{P'(t_1 \to t_1+\Delta t, h_1, w_1)\big\} \neq \text{PSD}\big\{y(t_2 \to t_2+\Delta t)\big\}\\ &\implies f'_i \neq g_j\end{aligned} \tag{13}$$

The GT-rPPG negative loss term L_n^{GR} pushes away PSDs from one ST-rPPG block and a non-corresponding GT signal (GT-rPPG negative pairs, e.g., (f_i, g'_j) where f_i is from ST-rPPG block P and g'_j is from the non-corresponding GT signal y') so that more negative pairs can be involved during the contrastive learning. [42] has demonstrated that more negative samples in contrastive learning can improve performance and facilitate convergence. The GT-rPPG negative loss L_n^{GR} is defined below, which is normalized by the number of GT-rPPG negative pairs.

$$L_n^{GR} = -\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\phi(v) \parallel f'_i - g_j \parallel^2 + \phi(v') \parallel f_i - g'_j \parallel^2}{\big(\phi(v)+\phi(v')\big)N^2} \tag{14}$$

Overalle Loss. The overall loss function for Contrast-Phys+ is the combination of the four losses, which can adapt to both unsupervised and weakly-supervised settings.

$$L_{CP+} = L_p^{RR} + L_n^{RR} + L_p^{GR} + L_n^{GR} \tag{15}$$

4.6 Why Contrast-Phys+ Works with Missing or Unsynchronized Labels

The four rPPG observations are used as constraints to make the model learn the target rPPG signal and exclude noises since noises do not satisfy all observations. Noises that appear in a small local region, such as periodical eye blinking, are excluded since the noises violate rPPG spatial similarity. Noises such as head motions/facial expressions that do not have a temporal constant frequency are excluded since they violate rPPG temporal similarity. The rPPG spatiotemporal similarity is satisfied by minimizing rPPG-rPPG positive loss L_p^{RR}. Cross-video rPPG dissimilarity can make two videos' PSDs discriminative and show distinguishable heart rate peaks between two videos' PSDs since heart rate peaks are the main features to distinguish two videos' PSDs as shown in Fig. 3. Cross-video rPPG dissimilarity is fulfilled by minimizing rPPG-rPPG negative loss L_n^{RR}. In addition, PSD values during the heart rate range are used so that noises such as light flickering exceeding the heart rate range are excluded due to the heart rate range constraint.

The loss function L_{CP+} can always be used even though some GT signals are missing. rPPG-rPPG positive loss L_p^{RR} and rPPG-rPPG negative loss L_n^{RR} using rPPG observations do not require GT signals. GT-rPPG positive loss L_p^{GR} and GT-rPPG negative loss L_n^{GR} using GT signals can be adapted to different situations (e.g., Both videos have GT signals, only one video has a GT signal, or neither video has a GT signal).

Contrast-Phys+ is also robust to unsynchronized GT signals. GT-rPPG negative loss L_n^{GR} is only intended to increase negative pairs using GT samples and rPPG samples for improved contrastive learning [42], so the loss does not require synchronization between facial videos and GT signals. GT-rPPG positive loss L_p^{GR} encourages the rPPG PSD to be similar to the GT PSD. When the GT signal is not precisely synchronized with the facial video, temporally sampled GT/rPPG for the same video can still share similar PSDs since PSDs do not change rapidly in a short time interval as shown in Fig. 7. The temporal sampling of GT/rPPG also removes alignment between the GT signal and the video to some extent, making GT-rPPG positive loss L_p^{GR} independent of the exact synchronization. Therefore, temporally sampled GT/rPPG for the same video can be pulled together in the unsynchronized case. We can also use the following relations to demonstrate that GT-rPPG positive loss L_p^{GR} is robust to GT signal misalignment. Suppose that the GT signal $y(t)$ has a small misalignment u, resulting $y(t+u)$. The PSDs of temporal samples of $y(t)$ and $y(t+u)$ are $\text{PSD}\{y(t_2 \rightarrow t_2+\Delta t)\}$ and $\text{PSD}\{y(t_2+u \rightarrow t_2+u+\Delta t)\}$, respectively. According to the temporal similarity in Sec. 3.2,

$$\text{PSD}\{y(t_2 \rightarrow t_2+\Delta t)\} \approx \text{PSD}\{y(t_2+u \rightarrow t_2+u+\Delta t)\} \tag{16}$$

holds if $|t_2 + u - t_2| = |u|$ is small where u is the small misalignment. Combine the above relation with relation 8, we get

$$\begin{aligned}\text{PSD}\{P(t_1 \rightarrow t_1+\Delta t, h_1, w_1)\} &\approx \text{PSD}\{y(t_2 \rightarrow t_2+\Delta t)\} \\ &\approx \text{PSD}\{y(t_2+u \rightarrow t_2+u+\Delta t)\}\end{aligned} \tag{17}$$

which indicates that rPPG samples from the ST-rPPG block P are similar to the GT samples from the misaligned GT signal $y(t+u)$. Therefore, our method is robust to GT signal misalignment.

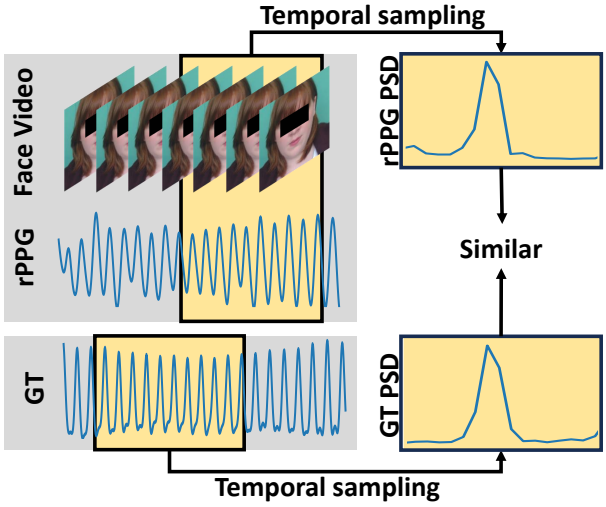


Fig. 7. Illustration showing that temporally sampled GT/rPPG are similar and independent of the exact synchronization.

5 Experiments

5.1 Experimental Setup and Metrics

5.1.1 Datasets

We conducted experiments using five common rPPG datasets, encompassing RGB and NIR videos recorded under diverse scenarios. Specifically, we employed the PURE dataset [56], UBFC-rPPG dataset [64], OBF dataset [60], and MR-NIRP dataset [30], [65] for intra-dataset evaluations. Additionally, we employed the MMSE-HR dataset [66] for both intra-dataset and cross-dataset evaluations. We also use the EquiPleth dataset [67] to evaluate the fairness performance on different skin tones. In addition, we also compare our method with other baselines when training on synthetic rPPG datasets including UCLA-Synth [68] and SCAMPS [69].

PURE [56] comprises facial videos from ten subjects recorded in six distinct setups, encompassing both static and dynamic tasks. To ensure consistency, we followed the same experimental protocol used in previous studies [16], [23] for partitioning the training and test sets.

UBFC-rPPG [64] comprises facial videos from 42 subjects who participated in a mathematical game designed to elevate their heart rates. For evaluation, we adhered to the protocol outlined in [23] for train-test split.

OBF [60] encompasses 200 videos from 100 healthy subjects recorded both before and after exercise sessions. To

facilitate a fair comparison with prior work, we conducted subject-independent ten-fold cross-validation, as previously described in [17], [18], [22].

MR-NIRP [30], [65] contains NIR videos of eight subjects, capturing instances of subjects remaining stationary as well as engaging in motion tasks. Due to its limited scale and the inherent challenge of weak rPPG signals in NIR [70], [71], we employed a leave-one-subject-out cross-validation protocol for our experiments. Both stationary and motion videos are used.

MMSE-HR [66] includes 102 videos from 40 subjects recorded during emotion elicitation experiments. Given the presence of spontaneous facial expressions and head movements, we conducted subject-independent 5-fold cross-validation for intra-dataset testing on the MMSE-HR dataset. Further details regarding these datasets are available in the supplementary material.

EquiPleth [67] includes 91 participants (28 light. 49 medium, and 14 dark skin tone participants). Each subject has 6 recordings, and each recording has a 30s RGB video and radar data. The GT PPG signals are recorded and synchronized with the RGB videos and radar data. We use the predefined train, validation, and test splits and report the results on the test set. We use the RGB data and GT PPG signals to train and test our method on different skin tones.

There are two synthetic rPPG datasets including **UCLA-Synth** [68] and **SCAMPS** [69]. The avatar facial videos are rendered by graphic pipelines. The GT PPG signals are temporally modulated with skin albedo maps to produce the skin color variations induced by rPPG. In addition, head motions and facial expressions are added during rendering. UCLA-Synth has 476 avatar subjects including 120 African, 118 Asian, 120 Caucasian, and 118 Indian avatar subjects. Each avatar subject has a 70-second facial video and GT signal. SCAMPS has 2800 avatar subjects with diverse skin tones and facial appearances. Each avatar subject has a 20-second facial video and GT signal.

5.1.2 Experimental Setup

During each training iteration, the model receives two 10-second clips from two different videos as inputs. If available, ground truth (GT) signals are incorporated; for instance, 20% of the videos contain GT signals, or in some cases, all videos possess unsynchronized GT signals. To train Contrast-Phys+ effectively, we employ the AdamW optimizer [72] with a learning rate of 10^{-5}, training the model for 30 epochs on a single NVIDIA Tesla V100 GPU. Following the approach in [34], we select the model with the lowest irrelevant power ratio (IPR) on the training set to achieve model selection (for further insights into IPR, refer to the supplementary materials).

During the testing phase, we segment each test video into non-overlapping 30-second clips and extract the rPPG signal from each clip. To compute the heart rate (HR), we identify the HR peak in the PSD of the rPPG signal. Additionally, we employ Neurokit2 [73] to locate systolic peaks within the rPPG signals, allowing us to derive heart rate variability (HRV) metrics.

According to our ablation study for the ST-rPPG blocks in Sec. 5.7.1, we set the spatial resolution of the ST-rPPG block to be 2×2, with the time duration of 10 seconds. For the rPPG spatiotemporal sampling process, we use $K = 4$, indicating that, for each spatial position within the ST-rPPG block, four rPPG samples are randomly selected. The time interval Δt between each rPPG sample is set to 5 seconds, which is half of the time duration of the ST-rPPG block. Consequently, we obtain 16 rPPG samples ($N = 16$) from each ST-rPPG block. Regarding the temporal sampling of GT signals, we maintain the same Δt of 5 seconds, resulting in the selection of 16 GT samples ($N = 16$) from a GT signal.

5.1.3 Evaluation Metrics

In line with prior research [18], [22], [62], we use three metrics to assess the accuracy of heart rate (HR) measurement: the mean absolute error (MAE), root mean squared error (RMSE), and Pearson correlation coefficient (R). Additionally, we utilize the signal-to-noise ratio (SNR) [13] to evaluate the quality of the rPPG signal. For the evaluation of HRV features, which encompass respiration frequency (RF), low-frequency power (LF) in normalized units (n.u.), high-frequency power (HF) in normalized units (n.u.), and the LF/HF power ratio, we follow the approach outlined in [23] and employ the standard deviation (STD), RMSE, and R as evaluation metrics. In the context of MAE, RMSE, and STD, smaller values indicate lower errors, whereas for R, higher values approaching one denote reduced errors. For SNR, larger values indicate higher-quality rPPG signals. For a more comprehensive understanding of these evaluation metrics, please refer to the supplementary material.

5.2 Intra-dataset Testing

5.2.1 HR Estimation

We conducted intra-dataset testing for HR estimation on four datasets: PURE, UBFC-rPPG, OBF, and MR-NIRP. Contrast-Phys+ was trained under various conditions, including scenarios where 0%, 20%, or 60% of the videos contain GT signals. These settings represent the unsupervised and semi-supervised paradigms, with the semi-supervised setup encompassing partially available labels. Additionally, Contrast-Phys+ was trained with 100% of the labels, representing the supervised setting.

The results of HR estimation for Contrast-Phys+ are presented in Table 1 and compared against multiple baseline methods. These baselines include traditional methods, supervised methods, semi-supervised methods, and recent unsupervised methods. Notably, Contrast-Phys+ (0%) outperforms several unsupervised baselines [34], [36], [38] and comes remarkably close to the performance of supervised methods [22], [23], [24]. In the semi-supervised setting, when partial GT signals are available (Contrast-Phys+ 20% and 60%), the performance improves further, often surpassing recent supervised methods [22], [23], [24]. In the supervised setting (Contrast-Phys+ (100%)), Contrast-Phys+ achieves the best performance among supervised methods across most evaluation metrics. This underscores the advantage of Contrast+Phys+ as it learns from both labels and rPPG observations, whereas previous supervised methods only rely on labels. The consistently superior performance of Contrast-Phys+ holds across all four datasets, including the MR-NIRP dataset containing NIR videos.

TABLE 1
Intra-dataset HR results. The best results are in bold, and the second-best results are underlined.

Method Types	Methods	UBFC-rPPG			PURE			OBF			MR-NIRP (NIR)		
		MAE (bpm)	RMSE (bpm)	R	MAE (bpm)	RMSE (bpm)	R	MAE (bpm)	RMSE (bpm)	R	MAE (bpm)	RMSE (bpm)	R
Traditional	GREEN [12]	7.50	14.41	0.62	-	-	-	-	2.162	0.99	-	-	-
	ICA [1]	5.17	11.76	0.65	-	-	-	-	-	-	-	-	-
	CHROM [13]	2.37	4.91	0.89	2.07	9.92	0.99	-	2.733	0.98	-	-	-
	2SR [74]	-	-	-	2.44	3.06	0.98	-	-	-	-	-	-
	POS [14]	4.05	8.75	0.78	-	-	-	-	1.906	0.991	-	-	-
Supervised	CAN [15]	-	-	-	-	-	-	-	-	-	7.78	16.8	-0.03
	HR-CNN [16]	-	-	-	1.84	2.37	0.98	-	-	-	-	-	-
	SynRhythm [75]	5.59	6.82	0.72	-	-	-	-	-	-	-	-	-
	PhysNet [17]	-	-	-	2.1	2.6	0.99	-	1.812	0.992	3.07	7.55	0.655
	rPPGNet [18]	-	-	-	-	-	-	-	1.8	0.992	-	-	-
	CVD [22]	-	-	-	-	-	-	-	1.26	0.996	-	-	-
	PulseGAN [76]	1.19	2.10	0.98	-	-	-	-	-	-	-	-	-
	Dual-GAN [23]	0.44	**0.67**	**0.99**	0.82	1.31	0.99	-	-	-	-	-	-
	Nowara2021 [24]	-	-	-	-	-	-	-	-	-	2.34	4.46	0.85
	Contrast-Phys+ (100%)	**0.21**	0.80	**0.99**	**0.48**	**0.98**	0.99	**0.34**	**0.75**	**0.998**	**1.96**	**3.02**	**0.93**
Semi-supervised	**Contrast-Phys+ (60%)**	0.22	0.81	**0.99**	0.52	1.02	0.99	0.35	0.79	0.997	2.58	3.65	0.89
	Contrast-Phys+ (20%)	0.24	0.87	**0.99**	0.61	1.18	0.99	0.37	0.84	0.997	2.57	4.02	0.88
Unsupervised	**Contrast-Phys+ (0%)**	0.64	1.00	**0.99**	1.00	1.40	0.99	0.51	1.39	0.994	2.68	4.77	0.85
	Gideon2021 [34]	1.85	4.28	0.93	2.3	2.9	0.99	2.83	7.88	0.825	4.75	9.14	0.61
	SiNC [36]	0.59	1.83	**0.99**	0.61	1.84	**1.00**	-	-	-	-	-	-
	Yue *et al.* [38]	0.58	0.94	**0.99**	1.23	2.01	0.99	-	-	-	-	-	-

TABLE 2
HRV results on UBFC-rPPG. The best results are in bold, and the second-best results are underlined.

Method Types	Methods	LF (n.u.)			HF (n.u.)			LF/HF			RF(Hz)		
		STD	RMSE	R	STD	RMSE	R	STD	RMSE	R	STD	RMSE	R
Traditional	GREEN [12]	0.186	0.186	0.280	0.186	0.186	0.280	0.361	0.365	0.492	0.087	0.086	0.111
	ICA [1]	0.243	0.240	0.159	0.243	0.240	0.159	0.655	0.645	0.226	0.086	0.089	0.102
	POS [14]	0.171	0.169	0.479	0.171	0.169	0.479	0.405	0.399	0.518	0.109	0.107	0.087
Supervised	CVD [22]	0.053	0.065	0.740	0.053	0.065	0.740	0.169	0.168	0.812	0.017	0.018	0.252
	Dual-GAN [23]	0.034	0.035	0.891	0.034	0.035	0.891	0.131	0.136	0.881	**0.010**	**0.010**	0.395
	Contrast-Phys+ (100%)	**0.025**	**0.025**	**0.947**	**0.025**	**0.025**	**0.947**	**0.064**	**0.066**	**0.963**	0.029	0.029	**0.803**
Semi-supervised	**Contrast-Phys+ (60%)**	0.035	0.035	0.908	0.035	0.035	0.908	0.100	0.105	0.906	0.036	0.043	0.746
	Contrast-Phys+ (20%)	0.037	0.037	0.882	0.037	0.037	0.882	0.119	0.120	0.866	0.037	0.040	0.662
Unsupervised	**Contrast-Phys+ (0%)**	0.096	0.098	0.798	0.096	0.098	0.798	0.391	0.395	0.782	0.085	0.083	0.347
	Gideon2021 [34]	0.142	0.139	0.694	0.142	0.139	0.694	0.687	0.691	0.684	0.098	0.098	0.103

5.2.2 HRV Estimation

Intra-dataset testing for heart rate variability (HRV) evaluation was conducted on the UBFC-rPPG dataset, and the results are presented in Table 2. HRV analysis demands precisely measured, high-quality rPPG signals for accurate systolic peak detection. Notably, Contrast-Phys+ significantly outperforms traditional methods and the previous unsupervised baseline [34] in terms of HRV results. When partial GT signals are incorporated, the performance of Contrast-Phys+ closely approaches that of supervised methods. In the case of Contrast-Phys+ utilizing all labels (100%), it achieves the best results across most HRV metrics. These findings underscore the capability of Contrast-Phys+ to yield high-quality rPPG signals with accurate systolic peaks, enabling the derivation of HRV features. This feature makes it a promising candidate for applications in emotion understanding [5], [6], [7] and healthcare [2], [3]. Additionally, Contrast-Phys+ has the potential to further refine its understanding of rPPG signals by leveraging GT information, as illustrated in Section 5.10.2.

5.3 Cross-dataset Testing

We perform cross-dataset testing on MMSE-HR to test the generalization of the proposed methods. We train recent supervised methods [17], [20], [77], [78], the unsupervised baseline [34], and Contrast-Phys+ on UBFC and test the models on MMSE-HR. In addition, we also provide intra-dataset results by training and testing the models on MMSE-HR as a reference to be compared with the cross-dataset results. Tab. 3 shows the cross-dataset and intra-dataset results on MMSE-HR, which can be summarized in four aspects as below. **1)** First, Contrast-Phys+ achieves good cross-dataset results compared with other supervised and unsupervised baselines, which means the proposed method can generalize well to a new dataset. The results are very promising, as in practical applications, we might potentially use enormous facial videos from different sources with no/partial GT signals to train Contrast-Phys/Contrast-Phys+ and then

TABLE 3
Cross-dataset and intra-dataset HR results for MMSE-HR. The best results are in bold, and the second-best results are underlined.

Method Types	Methods	Cross-dataset (UBFC → MMSE-HR)				Intra-dataset (MMSE-HR → MMSE-HR)			
		MAE (bpm)	RMSE (bpm)	R	SNR (dB)	MAE (bpm)	RMSE (bpm)	R	SNR (dB)
Traditional	Li2014 [1]	-	19.95	0.38	-	-	19.95	0.38	-
	CHROM [13]	-	13.97	0.55	-	-	13.97	0.55	-
	SAMC [28]	-	11.37	0.71	-	-	11.37	0.71	-
Supervised	PhysNet [17]	2.04	6.85	0.86	1.17	1.22	4.49	0.94	2.8
	TS-CAN [20]	3.41	9.29	0.76	-1.18	2.89	7.18	0.86	-2.01
	PhysFormer [77]	2.68	7.01	0.86	1.2	1.48	4.22	0.95	2.55
	Contrast-Phys+ (100%)	**1.76**	**5.34**	**0.92**	**1.37**	**1.11**	**3.83**	**0.96**	**3.72**
Semi-supervised	**Contrast-Phys+ (60%)**	2.30	6.32	0.89	1.25	1.20	3.89	**0.96**	3.51
	Contrast-Phys+ (20%)	2.28	6.51	0.88	1.15	1.51	4.15	0.95	2.93
Unsupervised	**Contrast-Phys+ (0%)**	2.43	7.34	0.86	1.09	1.82	6.69	0.87	2.64
	Gideon2021 [34]	4.10	11.55	0.70	0.26	3.98	9.65	0.85	0.67

TABLE 4
Cross-dataset results of Contrast-Phys+ when additional unlabeled videos (PURE and OBF) are used for training. The best results are in bold.

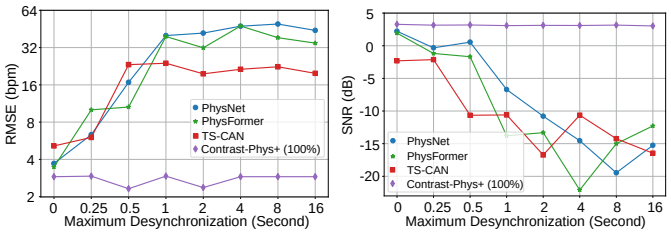
Training Sets			Cross-dataset Results (test on MMSE-HR)			
Labeled UBFC	Unlabeled PURE	Unlabeled OBF	MAE (BPM)	RMSE (bpm)	R	SNR (dB)
✓			1.76	5.34	0.92	1.37
✓	✓		1.13	**3.71**	**0.96**	2.37
✓		✓	1.47	4.55	0.94	**3.17**

apply them to the target data. **2)** Second, more labels from Contrast-Phys+ (0%, unsupervised) to Contrast-Phys+ (100%, fully supervised) can provide better performance for both cross- and intra-dataset results, which means additional GT signals can help fit rPPG signals and improve generalization. **3)** Third, for both cross- and intra-dataset results, Contrast-Phys+ (100%) using both label information and rPPG observations achieves better performance than other supervised methods that only utilize label information. Therefore, rPPG observations as the prior knowledge play an important role in improving rPPG measurement performance in the fully supervised setting. **4)** Last, comparing cross- and intra-dataset results, performance for intra-dataset is generally better than for cross-dataset for each deep learning-based method, so training and testing on the same dataset are preferred to keep good performance. Compared with previous supervised methods, Contrast-Phys+ lowers the requirement of intra-dataset training since it only needs facial videos with no or partial labels.

Contrast-Phys+ exhibits the capability to adapt to both labeled and unlabeled videos during training, allowing for the expansion and diversification of the training dataset by incorporating unlabeled videos from other sources. This augmentation strategy aims to enhance the model's generalization. To this end, we employed all labeled UBFC videos alongside additional unlabeled videos from PURE or OBF to train Contrast-Phys+ and evaluated the model's performance on MMSE-HR.

The results in Table 4 demonstrate that the inclusion of additional unlabeled videos for training results in improved performance compared to training solely with labeled UBFC data. When additional unlabeled training data is introduced, the cross-dataset testing performance even approaches the levels achieved by the best intra-dataset testing performance, as demonstrated in Table 3. This suggests that Contrast-Phys+ can seamlessly expand its training dataset by incorporating unlabeled videos from different domains, thereby enhancing generalization and achieving performance levels close to intra-dataset results. Such a capability was not feasible with previous supervised methods, highlighting the strengths of Contrast-Phys+.

5.4 Training with Unsynchronized GT Signals



(a) RMSE (b) SNR

Fig. 8. rPPG measurement performance ((a) RMSE and (b) SNR) with respect to maximum desynchronization of GT signals.

In our experiments, we explored scenarios where ground truth (GT) signals are desynchronized from the facial videos, which is a noisy label case in weakly-supervised learning. We introduced a parameter, the maximum desynchronization $D_{\max}$, and temporally shifted each GT signal by a random offset within the range of $-D_{\max}$ to $D_{\max}$, ensuring that the GT signals were no longer synchronized with the corresponding facial videos. This desynchronization was applied to GT signals in the training set, after which we trained the model and evaluated its performance on the test set. This experiment was tested on a single fold of the MMSE-HR dataset and trained on the other 4 folds.

As depicted in Fig. 8, we analyzed the RMSE and SNR across various levels of maximum desynchronization. Notably, as the maximum desynchronization increased, the performance of previous supervised methods exhibited significant deterioration. Even a small maximum desynchronization of 0.25 seconds, which is realistic and likely to occur during data collection, considerably impacted their performance.

In contrast, Contrast-Phys+ (100%) demonstrated robust and stable performance in terms of both RMSE and SNR across different maximum desynchronization values. These results underscore the robustness of Contrast-Phys+ to GT signal desynchronization, while previous supervised methods proved to be highly susceptible to even minor misalignments. This robustness can be attributed to Contrast-Phys+'s use of PSD instead of pulse curves in the temporal domain, which is comparatively stable over short time intervals. Consequently, learning an rPPG signal with misaligned GT signals in the frequency domain, aided by the rPPG observation constraint, is a viable approach as demonstrated in Sec. 4.6. These results indicate that Contrast-Phys+ offers greater tolerance when facial videos and GT signals are not perfectly aligned, streamlining the rPPG data collection process.

5.5 Evaluation of Skin Tone Fairness

TABLE 5
HR performance and fairness on EquiPleth dataset. RMSE measures the HR performance on the entire dataset. Fairness is the RMSE difference between dark and light skin groups (Lower values mean better fairness.). The best results are in bold, and the second-best results are underlined.

Modalities	Methods	RMSE ↓ (bpm)	Fairness↓ (bpm)
RGB	PhysNet [17]	5.26	4.05
	Contrast-Phys+ (0%)	5.16	3.39
	Contrast-Phys+ (50%)	4.97	3.05
	Contrast-Phys+ (100%)	<u>4.40</u>	1.96
Radar	Vilesov et al. [67]	6.12	**0.85**
RGB+Radar	Vilesov et al. [67]	**3.42**	<u>1.44</u>

Previous studies [67], [79], [80] have pointed out that rPPG algorithms based on RGB videos have performance bias in different skin tones. RGB video-based measurement is less accurate on subjects with darker skin, since darker skin absorbs more light resulting in weaker skin color changes. We test our method on the EquiPleth dataset [67] to study the performance gap between dark and light skin tones. We also add baselines from Vilesov et al. [67] that utilized radar and RGB+radar to improve skin tone fairness and overall performance. Table 5 shows the overall performance and fairness on the EquiPleth dataset. For the overall RMSE performance, RGB+Radar fusion achieves the best performance, and Contrast-phys+ (100%) achieves the second best. The results indicate that RGB+Radar multimodal fusion can improve the overall performance. For fairness (the RMSE difference between dark and light skin groups), radar and radar+RGB achieve the best and second-best fairness since radar detects chest movement for heart rate measurement and does not depend on skin color. The results agree with the previous finding [67] that, in general, RGB video-based methods show larger skin tone bias than the radar approach. On the other hand, among RGB methods, Contrast-Phys+ achieves better fairness than PhysNet, which means we can develop new video-based rPPG algorithms to improve skin tone fairness.

5.6 Training with Synthetic rPPG Data

To solve the GT lacking issue, one solution is to develop unsupervised or semi-supervised methods like Contrast-Phys+, and another is to train supervised models on synthetic rPPG data [68], [69]. We take the real UBFC-rPPG dataset as the test set, and compare the two solutions in three settings: 1) previous supervised methods and supervised Contrast-Phys+ trained on labeled synthetic data, 2) semi-supervised Contrast-Phys+ trained on labeled synthetic data and unlabeled real data (unlabeled MMSE-HR), 3) unsupervised Contrast-Phys+ trained on unlabeled real data (unlabeled MMSE-HR).

TABLE 6
Performance on UBFC-rPPG when training with synthetic rPPG datasets (SCAMPS [69] and UCLA-Synth [68]). Note that since Contrast-Phys+ can work in a semi-supervised setting, it can be trained on a labeled synthetic dataset and an unlabeled real dataset (unlabeled MMSE-HR [66]). The best results are in bold.

(1) Training with SCAMPS [69]

Methods	Training sets	Test on UBFC-rPPG		
		MAE (bpm)	RMSE (bpm)	R
TS-CAN [20]	labeled SCAMPS	3.62	6.92	0.93
PhysNet [17]	labeled SCAMPS	5.40	10.89	0.82
Contrast-Phys+	labeled SCAMPS	0.89	3.25	0.98
	labeled SCAMPS and unlabeled MMSE-HR	**0.51**	**2.16**	**0.99**
	unlabeled MMSE-HR	1.03	2.70	**0.99**

(b) Training with UCLA-Synth [68]

Methods	Training sets	Test on UBFC-rPPG		
		MAE (bpm)	RMSE (bpm)	R
PRN [81]	labeled UCLA-Synth	1.09	1.99	0.83
PhysNet [17]	labeled UCLA-Synth	0.84	**1.76**	0.83
Contrast-Phys+	labeled UCLA-Synth	0.74	2.29	**0.99**
	labeled UCLA-Synth and unlabeled MMSE-HR	**0.60**	2.07	**0.99**
	unlabeled MMSE-HR	1.03	2.70	**0.99**

Table 6 shows the HR results on UBFC-rPPG when models are trained with synthetic datasets. We can conclude the following three points corresponding to the three settings. 1) When trained on a labeled synthetic dataset, supervised Contrast-Phys+ outperforms previous supervised methods. 2) Semi-supervised Contrast-Phys+ trained on a labeled synthetic dataset and an unlabeled real dataset (unlabeled MMSE-HR) achieves the best performance while previous supervised methods can only use labeled synthetic data and cannot utilize unlabeled real data for training. 3) Unsupervised Contrast-Phys+ trained on an unlabeled real dataset (unlabeled MMSE-HR) achieves better or comparable performance than previous supervised methods trained on synthetic data. Overall, the results demonstrate that Contrast-Phys+ allows integrating the two solutions via merging both labeled synthetic data and unlabeled real data which can achieve better performance than using either one solution alone.

5.7 Ablation Study

5.7.1 ST-rPPG Block Parameters

In our ablation study, we investigated the impact of two key parameters of the ST-rPPG block: spatial resolution (S) and temporal length (T).

Table 7(a) presents the heart rate (HR) results for Contrast-Phys+ (0%) on UBFC-rPPG when varying the spatial resolution (S) of the ST-rPPG block across four levels: 1x1, 2x2, 4x4, and 8x8. It's important to note that 1x1 implies that rPPG spatial similarity is not considered. As evident from the results, the performance with a spatial resolution of 1x1 is inferior to the other resolutions, indicating that rPPG spatial similarity enhances performance. Furthermore, a spatial resolution of 2x2 yields satisfactory results, and larger resolutions do not substantially improve HR estimation. This is because larger resolutions, such as 8x8 or 4x4, provide more rPPG samples, but each block has a smaller receptive field, leading to noisier rPPG samples.

Table 7(b) demonstrates the HR results for Contrast-Phys+ (0%) on UBFC-rPPG while varying the temporal length (T) of the ST-rPPG block across three levels: 5 seconds, 10 seconds, and 30 seconds. The rPPG sample length (Δt) is the default value (T/2). The results highlight that a temporal length of 10 seconds yields the best performance. A shorter time length (5 seconds) results in coarse PSD estimation, while a longer time length (30 seconds) might violate the conditions for rPPG temporal similarity. As a result, we opted for $S = 2$ and $T = 10$ seconds in our experiments, as these settings strike a balance and offer optimal performance.

Table 7(c) shows the HR results for Contrast-Phys+ (0%) on UBFC-rPPG while varying the rPPG sample lengths (Δt) when T = 10s. Three levels of Δt are selected: T/4 (2.5s), T/2 (5s), and 3T/4 (7.5s). The findings emphasize that both T/2 and 3T/4 exhibit comparable performance, whereas the shorter T/4 demonstrates lower performance. A shorter Δt like T/4 leads to inaccurate PSDs and heart rate peaks, while a longer Δt such as T/2 and 3T/4 can offer more precise PSDs, aiding in sample comparisons within contrastive learning. However, long Δt like 3T/4 also increases computational costs. Therefore, we adopt T/2 as the default value.

TABLE 7
Ablation study for ST-rPPG block parameters: (a) HR results of Contrast-Phys+ on UBFC-rPPG with different ST-rPPG block spatial resolutions (S). (b) HR results of Contrast-Phys+ on UBFC-rPPG with different ST-rPPG block time lengths (T) when $\Delta t = T/2$. (c) HR results of Contrast-Phys+ on UBFC-rPPG with different rPPG sample lengths (Δt) when T = 10s (The best results are in bold.)

(a)

Spatial Resolution (S)	MAE (bpm)	RMSE (bpm)	R
1×1	3.14	4.06	0.963
2×2	**0.64**	**1.00**	**0.995**
4×4	0.55	1.06	0.994
8×8	0.60	1.09	0.993

(b)

Time Length (T)	MAE (bpm)	RMSE (bpm)	R
5s	0.68	1.36	0.990
10s	**0.64**	**1.00**	**0.995**
30s	1.97	3.58	0.942

(c)

rPPG Sample Length (Δt)	MAE (bpm)	RMSE (bpm)	R
T/4	0.98	1.74	0.986
T/2	**0.64**	1.00	**0.995**
3T/4	0.65	**0.98**	**0.995**

5.7.2 rPPG Observations

In our ablation study, we examined the individual impact of each of the four rPPG observations on the performance of Contrast-Phys+. These observations include rPPG spatial and temporal similarity (represented by rPPG spatial and temporal sampling), rPPG cross-video dissimilarity (represented by the rPPG-rPPG negative loss L_n^{RR}), and the HR range constraint (utilizing PSDs in the HR frequency range).

Table 8 showcases the results for Contrast-Phys+ (0%) when one of the rPPG observations is removed, as well as the results when all observations are utilized. The findings indicate that Contrast-Phys+ achieves its best performance when all rPPG observations are enabled. When rPPG spatial or temporal similarity is disabled, the performance experiences a slight decrease. However, when rPPG cross-video dissimilarity or the HR range constraint is disabled, the performance deteriorates significantly. The HR range constraint plays a crucial role in preventing the model from learning irrelevant periodic noises, such as light flickering, which can interfere with accurate HR estimation. Additionally, rPPG cross-video dissimilarity, represented by the rPPG-rPPG negative loss L_n^{RR}, is essential in contrastive learning as it prevents the model from collapsing into trivial solutions, as discussed in [39].

These results underscore the importance of all four rPPG observations in enhancing the performance of Contrast-Phys+ and emphasize their individual contributions to accurate and robust rPPG signal extraction.

5.7.3 The Influence of GT Signals

GT-related Losses. We conducted an ablation study to assess the influence of GT-related losses on our model's performance. Table 9 presents the results of the ablation study performed on the MMSE-HR dataset using Contrast-Phys+ (100%). When we exclude all GT-related terms, the model effectively undergoes unsupervised training, resulting in the lowest performance. However, when we include only the GT-rPPG negative term, the model's performance improves, as it generates more negative pairs from both GT signals and ST-rPPG blocks. Subsequently, utilizing solely the GT-rPPG positive term further enhances performance, as it enforces consistency between ST-rPPG blocks and their corresponding GT signals, effectively incorporating GT information into the model's training. The combined use of both terms yields the highest performance, which is the top-performing configuration.

GT Signal Ratios. Since Contrast-Phys+ is capable of adapting to different availability of data labels, we conducted an ablation study to examine the impact of different GT signal ratios. Specifically, we trained Contrast-Phys+ using 0%, 20%, 40%, 60%, 80%, and 100% labels from the MMSE-HR dataset. The performance variation of Contrast-Phys+ under different label ratios is illustrated in Fig. 9.

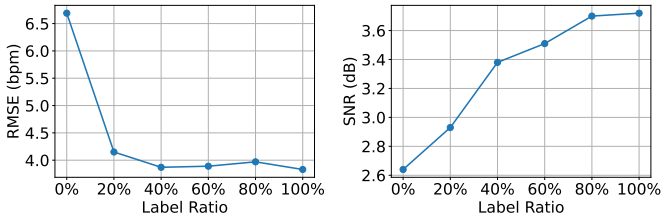
TABLE 8
Ablation Study for rPPG Observations on UBFC-rPPG dataset. The best results are in bold.

rPPG Spatial Similarity	rPPG Temporal Similarity	rPPG Cross-video Dissimilarity	HR Range Constraint	MAE (bpm)	RMSE (bpm)	R
✓	✓	✓		39.66	44.49	-0.401
✓	✓		✓	22.11	33.84	0.281
✓		✓	✓	1.26	3.64	0.948
	✓	✓	✓	3.14	4.06	0.963
✓	✓	✓	✓	**0.64**	**1.00**	**0.995**

TABLE 9
Ablation Study for GT-related Positive Loss Term L_p^{GR} and Negative Loss Term L_n^{GR} on MMSE-HR dataset. The best results are in bold.

L_p^{GR}	L_n^{GR}	MAE (bpm)	RMSE (bpm)	R
		1.82	6.69	0.87
	✓	1.77	5.30	0.88
✓		1.39	4.46	0.91
✓	✓	**1.11**	**3.83**	**0.96**

Regarding RMSE, the performance reaches a plateau at 40% label ratio, and the HR error does not significantly decrease when using more than 40% labels. On the other hand, SNR, which serves as a metric for rPPG signal quality, exhibits continuous improvement with an increasing number of labels. These findings suggest that while employing more labels (beyond 40%) may not lead to a substantial reduction in HR measurement error, they do contribute to refining the quality of the output rPPG signals. We will further demonstrate this through waveform visualization in Sec. 5.10.2.



(a) RMSE (b) SNR

Fig. 9. rPPG measurement performance ((a) RMSE and (b) SNR) with respect to label ratios.

5.8 Statistical Validation for rPPG Observations

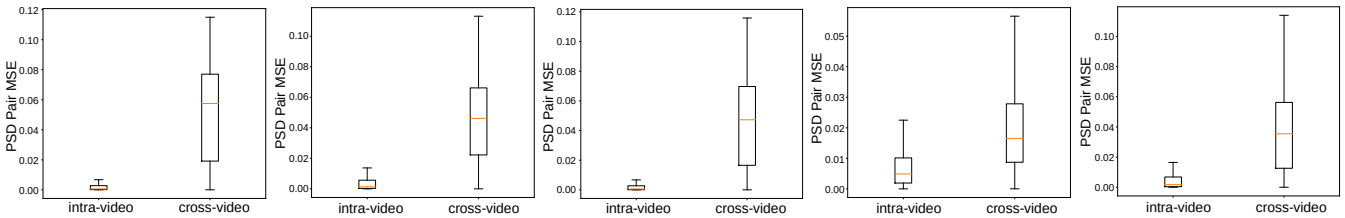
We conducted a statistical analysis to validate both the spatiotemporal similarity of rPPG signals within the same video (referred to as "intra-video") and the dissimilarity of rPPG signals between different videos (referred to as "cross-video"). The spatiotemporal similarity of rPPG signals refers to the similarity in the PSDs of rPPG signals measured at different spatiotemporal locations within the same video. Conversely, the cross-video rPPG dissimilarity refers to the differences in the PSDs of rPPG signals measured at different spatiotemporal locations between two different videos.

To quantify these observations, we calculated the mean squared errors (MSE) of PSD pairs for both the intra-video and cross-video cases. Figure 10 illustrates that the PSD pair MSE for the intra-video case is significantly smaller compared to the PSD pair MSE for the cross-video case. To assess the significance of these differences, we employed the two-sample Kolmogorov-Smirnov test [6], [82]. The results indicate that the PSD pair MSE for the cross-video case is significantly higher than for the intra-video case ($p < 0.001$) across all five rPPG datasets. These statistical test results provide solid evidence supporting the validity of both the rPPG spatiotemporal similarity and the cross-video rPPG dissimilarity observations.

5.9 Running Speed

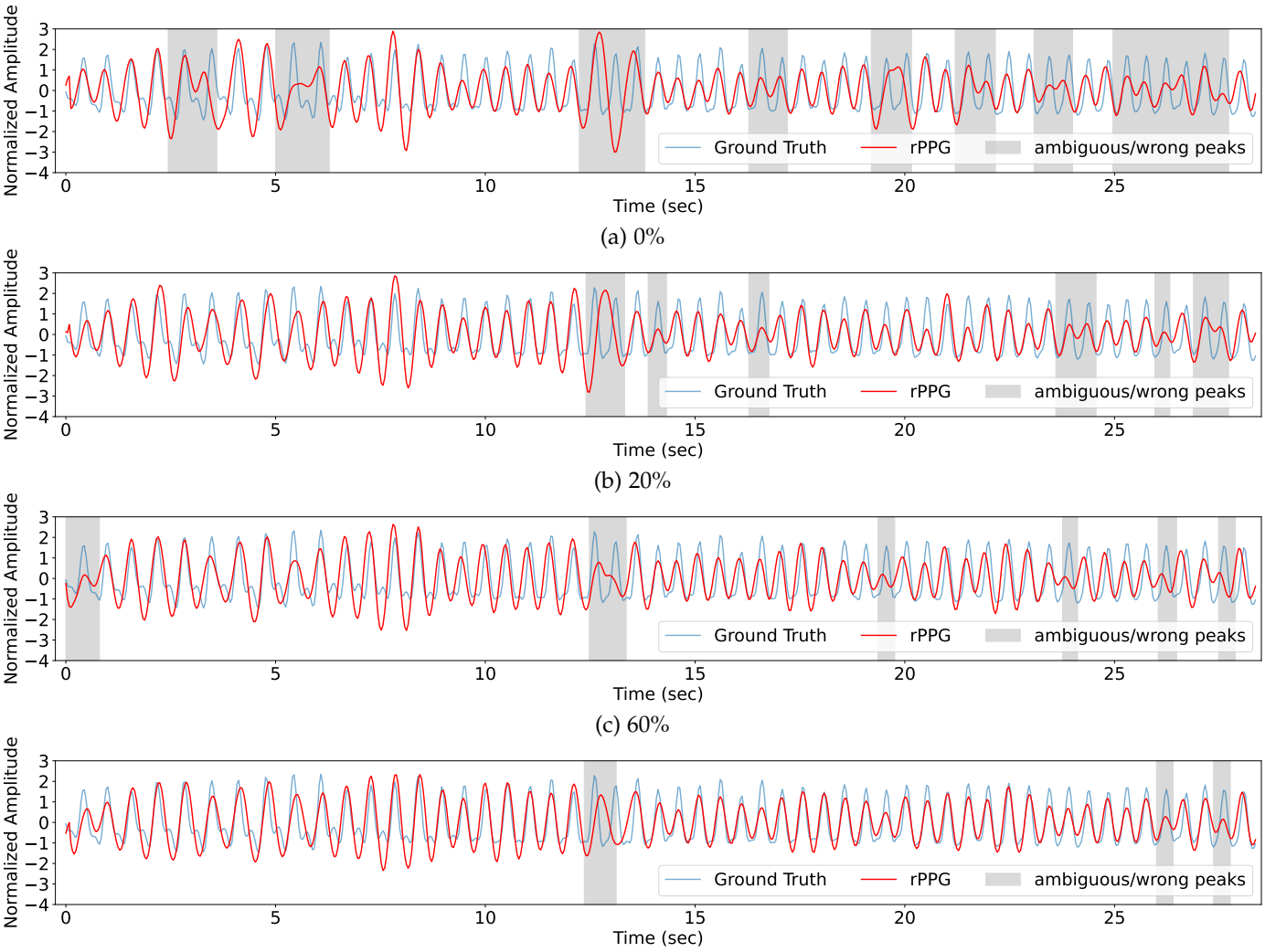
We conducted experiments to compare the running speed of Contrast-Phys+ and Gideon2021 [34]. During training, the running speed of Contrast-Phys+ (0%) was measured at **802.45** frames per second (fps), while Gideon2021 achieved a speed of **387.87** fps, which is approximately half of Contrast-Phys+'s speed. This significant difference in speed can be attributed to the different method designs employed by the two models. In Gideon2021, the input video is fed into the model twice, first as the original video and then as a temporally resampled video, resulting in double computation. On the other hand, Contrast-Phys+ only requires the input video to be fed into the model once, leading to a substantial decrease in computational cost. Additionally, the running speed of Contrast-Phys+ for label ratios of 60% and 100% was measured at **792.70** fps and **776.19** fps, respectively. When compared to Contrast-Phys+ (0%) (802.45 fps), incorporating GT signals in Contrast-Phys+ (60%, 100%) only resulted in a slight decrease in speed.

Furthermore, we compared the convergence speed using the metric of Irrelevant power ratio (IPR). IPR is used in [34] to evaluate signal quality during training with lower values indicating higher signal quality. More details about IPR can be found in the supplementary materials. Figure 12 illustrates the IPR values over time during training on the OBF dataset. The results demonstrate that Contrast-Phys+ achieves faster convergence to a lower IPR compared to Gideon2021. While Contrast-Phys+ (60%, 100%) takes slightly longer to reach the lowest IPR compared to Contrast-Phys+ (0%), it ultimately achieves a lower IPR due to its ability to utilize GT signals to further enhance the rPPG signal quality.



(a) PURE (b) UBFC (c) OBF (d) MR-NIRP (e) MMSE-HR

Fig. 10. Boxplots of PSD pair MSE for intra-video and cross-video for rPPG datasets: (a) PURE, (b) UBFC, (c) OBF, (d) MR-NIRP, and (e) MMSE-HR.



(a) 0%

(b) 20%

(c) 60%

(d) 100%

Fig. 11. rPPG waveforms from Contrast-Phys+ trained with different label ratios: (a) 0%, (b) 20%, (c) 60%, and (d) 100%. The ambiguous/wrong peaks from rPPG are highlighted in gray areas.

5.10 Result Visualization

5.10.1 Saliency Maps

To demonstrate the interpretability of Contrast-Phys+, we present saliency maps. These saliency maps are generated using a gradient-based method proposed in [83]. We keep the weights of the trained model fixed and calculate the gradient of the Pearson correlation with respect to the input video. More detailed information can be found in the supplementary materials. Saliency maps are useful for highlighting the spatial regions that contribute to the estimation of rPPG signals by the model. A saliency map of a good rPPG model should exhibit a strong response in skin regions, as demonstrated in previous works such as [15], [17], [18], [24], [34].

Fig. 13 presents saliency maps in two scenarios to showcase the robustness of our method against interferences: 1) when periodic noise is manually injected, and 2) when head motion is involved. In the presence of a periodic noise patch injected into the upper-left corner of the videos, Contrast-Phys+ remains unaffected by the noise and continues to

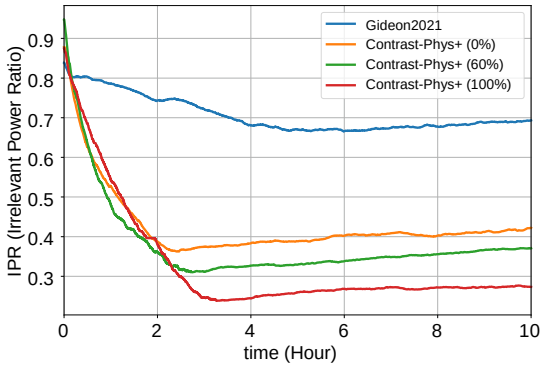


Fig. 12. Irrelevant power ratio (IPR) with respect to training time.

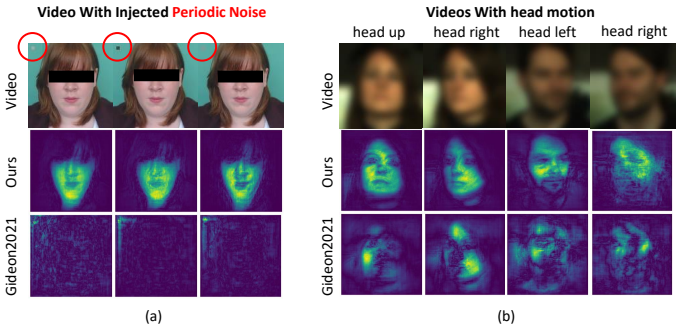


Fig. 13. Saliency maps for Contrast-Phys+ (0%) and Gideon2021 [34] (a) In this scenario, we introduce a random flashing block with a heart rate (HR) range between 40-250 bpm in the top-left corner of all UBFC-rPPG videos. Both Contrast-Phys+ (0%) and Gideon2021 are trained on these videos with the added noise. The saliency maps of Contrast-Phys+ (0%) exhibit a strong response in facial regions, indicating its focus on these areas. On the other hand, Gideon2021 primarily focuses on the injected periodic noise. The quantitative results in Table 10 further support the robustness of Contrast-Phys+ (0%) to the noise. (b) For this scenario, we select moments of head motion from the PURE dataset (Video frames shown here are blurred due to privacy concerns.) and generate the corresponding saliency maps for Contrast-Phys+ (0%) and Gideon2021.

TABLE 10
HR results trained on UBFC-rPPG with/without injected periodic noise shown in Fig. 13(a).

Methods	Injected Periodic Noise	MAE (bpm)	RMSE (bpm)	R
Gideon2021 [34]	w/o	1.85	4.28	0.939
	w/	22.47	25.41	0.244
Contrast-Phys+ (0%)	w/o	0.64	1.00	0.995
	w/	0.74	1.34	0.991

focus on skin areas. In contrast, Gideon2021 is completely distracted by the noise block. We also evaluate the performance of both methods on UBFC-rPPG videos with the injected noise, and the results are summarized in Table 10. These results align with the saliency map analysis, confirming that Contrast-Phys+ is not impacted by the periodic noise, while Gideon2021 fails to handle it effectively. The robustness of Contrast-Phys+ to noise can be attributed to the rPPG spatial similarity constraint, which helps to filter out noise. Fig. 13(b) displays the saliency maps when head motion is involved. The saliency maps of Contrast-Phys+ primarily focus on and activate skin areas, indicating its ability to handle head motions effectively. In contrast, the saliency maps of Gideon2021 exhibit noise and are scattered, covering only partial facial areas during head motions.

5.10.2 rPPG Waveforms

Fig. 11 displays the rPPG waveforms obtained from Contrast-Phys+ trained with different label ratios on the MMSE-HR dataset. As more labels are available during training, ranging from 0% to 100%, the rPPG waveforms become more similar to the ground truth (GT) signal and exhibit fewer ambiguous or incorrect peaks. The waveform corresponding to the 0% label ratio contains noisy components highlighted by gray areas indicating ambiguous or incorrect peaks. In contrast, the waveform at the 100% label ratio is well aligned with the GT signal, with almost all peaks clearly distinguishable. The presence of distinguishable peaks in the rPPG waveform also facilitates the accurate calculation of HRV. The visualization of rPPG waveforms demonstrates that incorporating more labels during the training of Contrast-Phys+ improves the quality of the rPPG signal. This finding is consistent with the signal-to-noise ratio (SNR) results discussed in Sec. 5.7.3.

6 Conclusion

We propose Contrast-Phys+, which can be trained in unsupervised and weakly-supervised settings and achieve accurate rPPG measurement. Contrast-Phys+ is based on four rPPG observations and utilizes spatiotemporal contrast to enable unsupervised and weakly-supervised learning including missing and unsynchronized GT signals, or even no labels. By combining rPPG prior knowledge and additional GT information, Contrast-Phys+ outperforms both unsupervised and supervised state-of-the-art methods and achieves good generalization to unseen data. Besides, the proposed method is robust against noise interference and computationally efficient. For future studies, the proposed method can be extended to learn other periodic signals such as respiration signals.

Acknowledgments

This work was supported by the Research Council of Finland (former Academy of Finland) Academy Professor project EmotionAI (grants 336116, 345122), ICT 2023 project TrustFace (grant 345948), the University of Oulu & Research Council of Finland Profi 7 (grant 352788), and by Infotech Oulu. The work was also supported by the Spearhead project 'Gaze on Lips' funded by the Eudaimonia Institute of the University of Oulu, Finland. The authors also acknowledge CSC-IT Center for Science, Finland, for providing computational resources.

References

[1] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE transactions on biomedical engineering*, vol. 58, no. 1, pp. 7–11, 2010.

[2] J. Shi, I. Alikhani, X. Li, Z. Yu, T. Seppänen, and G. Zhao, "Atrial fibrillation detection from face videos by fusing subtle variations," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 8, pp. 2781–2795, 2019.

[3] B. P. Yan, W. H. Lai, C. K. Chan, S. C.-H. Chan, L.-H. Chan, K.-M. Lam, H.-W. Lau, C.-M. Ng, L.-Y. Tai, K.-W. Yip *et al.*, "Contact-free screening of atrial fibrillation by a smartphone using facial pulsatile photoplethysmographic signals," *Journal of the American Heart Association*, vol. 7, no. 8, p. e008585, 2018.

[4] Z. Sun, J. Junttila, M. Tulppo, T. Seppänen, and X. Li, "Non-contact atrial fibrillation detection from face videos by learning systolic peaks," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 9, pp. 4587–4598, 2022.

[5] Z. Yu, X. Li, and G. Zhao, "Facial-video-based physiological signal measurement: Recent advances and affective applications," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 50–58, 2021.

[6] D. McDuff, S. Gontarek, and R. Picard, "Remote measurement of cognitive stress via heart rate variability," in *2014 36th annual international conference of the IEEE engineering in medicine and biology society*. IEEE, 2014, pp. 2957–2960.

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

[8] Z. Sun, A. Vedernikov, V.-L. Kykyri, M. Pohjola, M. Nokia, and X. Li, "Estimating stress in online meetings by remote physiological signal and behavioral features," in *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers*, 2022, pp. 216–220.

[9] F. Juefei-Xu, R. Wang, Y. Huang, Q. Guo, L. Ma, and Y. Liu, "Countering malicious deepfakes: Survey, battleground, and horizon," *International Journal of Computer Vision*, vol. 130, no. 7, pp. 1678–1734, 2022.

[10] U. A. Ciftci, I. Demir, and L. Yin, "Fakecatcher: Detection of synthetic portrait videos using biological signals," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[11] Z. Sun and X. Li, "Privacy-phys: Facial video-based physiological modification for privacy protection," *IEEE Signal Processing Letters*, vol. 29, pp. 1507–1511, 2022.

[12] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light." *Optics express*, vol. 16, no. 26, pp. 21 434–21 445, 2008.

[13] G. De Haan and V. Jeanne, "Robust pulse rate from chrominance-based rppg," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 10, pp. 2878–2886, 2013.

[14] W. Wang, A. C. den Brinker, S. Stuijk, and G. De Haan, "Algorithmic principles of remote ppg," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1479–1491, 2016.

[15] W. Chen and D. McDuff, "Deepphys: Video-based physiological measurement using convolutional attention networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 349–365.

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

[17] Z. Yu, X. Li, and G. Zhao, "Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks," in *30th British Machine Vision Conference 2019*. BMVA Press, 2019, p. 277.

[18] Z. Yu, W. Peng, X. Li, X. Hong, and G. Zhao, "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*, 2019, pp. 151–160.

[19] E. Lee, E. Chen, and C.-Y. Lee, "Meta-rppg: Remote heart rate estimation using a transductive meta-learner," in *European Conference on Computer Vision*. Springer, 2020, pp. 392–409.

[20] X. Liu, J. Fromm, S. Patel, and D. McDuff, "Multi-task temporal shift attention networks for on-device contactless vitals measurement," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33, 2020, pp. 19 400–19 411.

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

[22] X. Niu, Z. Yu, H. Han, X. Li, S. Shan, and G. Zhao, "Video-based remote physiological measurement via cross-verified feature disentangling," in *European Conference on Computer Vision*. Springer, 2020, pp. 295–310.

[23] H. Lu, H. Han, and S. K. Zhou, "Dual-gan: Joint bvp and noise modeling for remote physiological measurement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 404–12 413.

[24] E. M. Nowara, D. McDuff, and A. Veeraraghavan, "The benefit of distraction: Denoising camera-based physiological measurements using inverse attention," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4955–4964.

[25] Z. Sun and X. Li, "Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast," in *European Conference on Computer Vision*. Springer, 2022, pp. 492–510.

[26] G. De Haan and A. Van Leest, "Improved motion robustness of remote-ppg by using the blood volume pulse signature," *Physiological measurement*, vol. 35, no. 9, p. 1913, 2014.

[27] A. Lam and Y. Kuno, "Robust heart rate measurement from video using select random patches," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3640–3648.

[28] S. Tulyakov, X. Alameda-Pineda, E. Ricci, L. Yin, J. F. Cohn, and N. Sebe, "Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2396–2404.

[29] W. Wang, S. Stuijk, and G. De Haan, "Exploiting spatial redundancy of image sensor for motion robust rppg," *IEEE transactions on Biomedical Engineering*, vol. 62, no. 2, pp. 415–425, 2014.

[30] E. M. Nowara, T. K. Marks, H. Mansour, and A. Veeraraghavan, "Near-infrared imaging photoplethysmography during driving," *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[31] J. Li, Z. Yu, and J. Shi, "Learning motion-robust remote photoplethysmography through arbitrary resolution videos," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 1334–1342.

[32] H. Lu, Z. Yu, X. Niu, and Y.-C. Chen, "Neuron structure modeling for generalizable remote physiological measurement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 589–18 599.

[33] J. Du, S.-Q. Liu, B. Zhang, and P. C. Yuen, "Dual-bridging with adversarial noise generation for domain adaptive rppg estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 355–10 364.

[34] J. Gideon and S. Stent, "The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3995–4004.

[35] H. Wang, E. Ahn, and J. Kim, "Self-supervised representation learning framework for remote physiological measurement using spatiotemporal augmentation loss," *AAAI*, 2022.

[36] J. Speth, N. Vance, P. Flynn, and A. Czajka, "Non-contrastive unsupervised learning of physiological signals from video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 464–14 474.

[37] Y. Yang, X. Liu, J. Wu, S. Borac, D. Katabi, M.-Z. Poh, and D. McDuff, "Simper: Simple self-supervised learning of periodic targets," in *The Eleventh International Conference on Learning Representations*, 2022.

[38] Z. Yue, M. Shi, and S. Ding, "Facial video-based remote physiological measurement via self-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[39] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[40] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[41] A. Van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv e-prints*, pp. arXiv–1807, 2018.

[42] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in

International conference on machine learning. PMLR, 2020, pp. 1597–1607.

[43] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *European conference on computer vision*. Springer, 2020, pp. 776–794.

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

[45] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 271–21 284, 2020.

[46] I. Misra and L. v. d. Maaten, "Self-supervised learning of pretext-invariant representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6707–6717.

[47] R. Qian, T. Meng, B. Gong, M.-H. Yang, H. Wang, S. Belongie, and Y. Cui, "Spatiotemporal contrastive video representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6964–6974.

[48] Z. Qi, S. Wang, C. Su, L. Su, Q. Huang, and Q. Tian, "Self-regulated learning for egocentric video activity anticipation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[49] W. Wang, X. Wang, W. Yang, and J. Liu, "Unsupervised face detection in the dark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 1250–1266, 2022.

[50] M. Kumar, A. Veeraraghavan, and A. Sabharwal, "Distanceppg: Robust non-contact vital signs monitoring using a camera," *Biomedical optics express*, vol. 6, no. 5, pp. 1565–1588, 2015.

[51] W. Wang, A. C. den Brinker, and G. De Haan, "Discriminative signatures for remote-ppg," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 5, pp. 1462–1473, 2019.

[52] S. Liu, P. C. Yuen, S. Zhang, and G. Zhao, "3d mask face anti-spoofing with remote photoplethysmography," in *European Conference on Computer Vision*. Springer, 2016, pp. 85–100.

[53] S.-Q. Liu, X. Lan, and P. C. Yuen, "Remote photoplethysmography correspondence feature for 3d mask face presentation attack detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 558–573.

[54] A. A. Kamshilin, S. Miridonov, V. Teplov, R. Saarenheimo, and E. Nippolainen, "Photoplethysmographic imaging of high spatial resolution," *Biomedical optics express*, vol. 2, no. 4, pp. 996–1006, 2011.

[55] A. A. Kamshilin, V. Teplov, E. Nippolainen, S. Miridonov, and R. Giniatullin, "Variability of microcirculation detected by blood pulsation imaging," *PloS one*, vol. 8, no. 2, p. e57117, 2013.

[56] R. Stricker, S. Müller, and H.-M. Gross, "Non-contact video-based pulse rate measurement on a mobile service robot," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2014, pp. 1056–1062.

[57] A. H. Association. (2015) All about heart rate (pulse). [Online]. Available: https://www.heart.org/en/health-topics/high-blood-pressure/the-facts-about-high-blood-pressure/all-about-heart-rate-pulse

[58] M. Chen, Q. Zhu, M. Wu, and Q. Wang, "Modulation model of the photoplethysmography signal for vital sign extraction," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 4, pp. 969–977, 2020.

[59] A. Pai, A. Veeraraghavan, and A. Sabharwal, "Hrvcam: robust camera-based measurement of heart rate variability," *Journal of Biomedical Optics*, vol. 26, no. 2, p. 022707, 2021.

[60] X. Li, I. Alikhani, J. Shi, T. Seppanen, J. Junttila, K. Majamaa-Voltti, M. Tulppo, and G. Zhao, "The obf database: A large face video database for remote physiological signal measurement and atrial fibrillation detection," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 242–249.

[61] A. H. Association. (2021) Target heart rates chart. [Online]. Available: https://www.heart.org/en/healthy-living/fitness/fitness-basics/target-heart-rates

[62] X. Li, J. Chen, G. Zhao, and M. Pietikainen, "Remote heart rate measurement from face videos under realistic situations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 4264–4271.

[63] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 59–66.

[64] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois, "Unsupervised skin tissue segmentation for remote photoplethysmography," *Pattern Recognition Letters*, vol. 124, pp. 82–90, 2019.

[65] E. Magdalena Nowara, T. K. Marks, H. Mansour, and A. Veeraraghavan, "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*, 2018, p. 1272–1281.

[66] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang *et al.*, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3438–3446.

[67] A. Vilesov, P. Chari, A. Armouti, A. B. Harish, K. Kulkarni, A. Deoghare, L. Jalilian, and A. Kadambi, "Blending camera and 77 ghz radar sensing for equitable, robust plethysmography," *ACM Trans. Graph.(SIGGRAPH)*, vol. 41, no. 4, pp. 1–14, 2022.

[68] Z. Wang, Y. Ba, P. Chari, O. D. Bozkurt, G. Brown, P. Patwa, N. Vaddi, L. Jalilian, and A. Kadambi, "Synthetic generation of face videos with plethysmograph physiology," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 20 587–20 596.

[69] D. McDuff, M. Wander, X. Liu, B. Hill, J. Hernandez, J. Lester, and T. Baltrusaitis, "Scamps: Synthetics for camera measurement of physiological signals," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3744–3757, 2022.

[70] L. F. C. Martinez, G. Paez, and M. Strojnik, "Optimal wavelength selection for noncontact reflection photoplethysmography," in *22nd Congress of the International Commission for Optics: Light for the Development of the World*, vol. 8011. International Society for Optics and Photonics, 2011, p. 801191.

[71] V. Vizbara, "Comparison of green, blue and infrared light in wrist and forehead photoplethysmography," *BIOMEDICAL ENGINEERING 2016*, vol. 17, no. 1, 2013.

[72] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

[73] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. Chen, "Neurokit2: A python toolbox for neurophysiological signal processing," *Behavior research methods*, vol. 53, no. 4, pp. 1689–1696, 2021.

[74] W. Wang, S. Stuijk, and G. De Haan, "A novel algorithm for remote photoplethysmography: Spatial subspace rotation," *IEEE transactions on biomedical engineering*, vol. 63, no. 9, pp. 1974–1984, 2015.

[75] X. Niu, H. Han, S. Shan, and X. Chen, "Synrhythm: Learning a deep heart rate estimator from general to specific," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 3580–3585.

[76] R. Song, H. Chen, J. Cheng, C. Li, Y. Liu, and X. Chen, "Pulsegan: Learning to generate realistic pulse waveforms in remote photoplethysmography," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1373–1384, 2021.

[77] Z. Yu, Y. Shen, J. Shi, H. Zhao, P. H. Torr, and G. Zhao, "Physformer: facial video-based physiological measurement with temporal difference transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4186–4196.

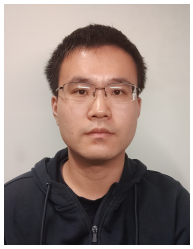
[78] Z. Yu, Y. Shen, J. Shi, H. Zhao, Y. Cui, J. Zhang, P. Torr, and G. Zhao, "Physformer++: Facial video-based physiological measurement with slowfast temporal difference transformer," *International Journal of Computer Vision*, 2023.

[79] A. Kadambi, "Achieving fairness in medical devices," *Science*, vol. 372, no. 6537, pp. 30–31, 2021.

[80] E. M. Nowara, D. McDuff, and A. Veeraraghavan, "A meta-analysis of the impact of skin tone and gender on non-contact photoplethysmography measurements," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 284–285.

[81] Y. Ba, Z. Wang, K. D. Karinca, O. D. Bozkurt, and A. Kadambi, "Style transfer with bio-realistic appearance manipulation for skin-tone inclusive rppg," in *2022 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2022, pp. 1–12.

[82] J. Hodges Jr, "The significance probability of the smirnov two-sample test," *Arkiv för matematik*, vol. 3, no. 5, pp. 469–486, 1958.
[83] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.



Zhaodong Sun is a final-year doctoral student at the Center for Machine Vision and Signal Analysis, University of Oulu. He received M.Sc. and B.E from Swiss Federal Institute of Technology Lausanne and University of Electronic Science and Technology of China in 2020 and 2018. His research interests include computer vision, biomedical signal processing, remote physiological measurement, and affective computing.

Xiaobai Li (Senior Member, IEEE) received the BSc degree in psychology from Peking University, the MSc degree in biophysics from the Chinese Academy of Science, and the PhD degree in computer science from the University of Oulu. She is currently a ZJU100 professor with the School of Cyber Science and Technology, Zhejiang University, and she is also an Adjunct Professor with the Center for Machine Vision and Signal Analysis, University of Oulu. Her research of interests include facial expression recognition, micro-expression analysis, remote physiological signal measurement from facial videos, and related applications in affective computing, healthcare and biometrics. She is an associate editor of IEEE Transactions on Circuits and Systems for Video Technology, Frontiers in Psychology, and Image and Vision Computing. She was a co-chair of several international workshops in CVPR, ICCV, FG, and ACM Multimedia.