# Contrast-Phys: Unsupervised Video-based Remote Physiological Measurement via Spatiotemporal Contrast

Zhaodong Sun and Xiaobai Li

Center for Machine Vision and Signal Analysis, University of Oulu, Oulu, Finland
{zhaodong.sun, xiaobai.li}@oulu.fi

**Abstract.** Video-based remote physiological measurement utilizes face videos to measure the blood volume change signal, which is also called remote photoplethysmography (rPPG). Supervised methods for rPPG measurements achieve state-of-the-art performance. However, supervised rPPG methods require face videos and ground truth physiological signals for model training. In this paper, we propose an unsupervised rPPG measurement method that does not require ground truth signals for training. We use a 3DCNN model to generate multiple rPPG signals from each video in different spatiotemporal locations and train the model with a contrastive loss where rPPG signals from the same video are pulled together while those from different videos are pushed away. We test on five public datasets, including RGB videos and NIR videos. The results show that our method outperforms the previous unsupervised baseline and achieves accuracies very close to the current best supervised rPPG methods on all five datasets. Furthermore, we also demonstrate that our approach can run at a much faster speed and is more robust to noises than the previous unsupervised baseline. Our code is available at https://github.com/zhaodongsun/contrast-phys.

**Keywords:** Remote Photoplethysmography, Face Video, Unsupervised Learning, Contrastive Learning

## 1 Introduction

Traditional physiological measurement requires skin-contact sensors to measure physiological signals such as contact photoplethysmography (PPG) and electrocardiography (ECG). Some physiological parameters like heart rate (HR), respiration frequency (RF), and heart rate variability (HRV) can be derived from PPG signals for healthcare [42,55] and emotion analysis [57,29,40]. However, the skin-contact physiological measurement requires specific biomedical equipment like pulse oximeters, and contact sensors may cause discomfort and skin irritation. Remote physiological measurement uses a camera to record face videos for measuring remote photoplethysmography (rPPG). The weak color change in faces can be captured by cameras to obtain the rPPG signal from which several

physiological parameters such as HR, RF, and HRV [38] can be measured. Video-based physiological measurement only requires off-the-shelf cameras rather than professional biomedical sensors, and is not constrained by physical distance, which has a great potential for remote healthcare [42,55] and emotion analysis applications [57,29,40].

In earlier rPPG studies [49,38,8,52], researchers proposed handcrafted features to extract rPPG signals. Later, some deep learning (DL)-based methods [7,45,56,58,18,23,32,33,25,35] were proposed, which employed supervised approaches with various network architectures for measuring rPPG signals. On one side, under certain circumstances, e.g., when head motions are involved or the videos are heterogeneous, DL-based methods could be more robust than the traditional handcrafted approaches. On the other side, DL-based rPPG methods require a large-scale dataset including face videos and ground truth physiological signals. Although face videos are comparatively easy to obtain in large amount, it is expensive to get the ground truth physiological signals which are measured by contact sensors, and synchronized with the face videos.

Can we only use face videos without ground truth physiological signals to train models for rPPG measurement? Gideon and Stent [10] proposed a self-supervised method to train rPPG measurement models without labels. They first downsample a video to get the downsampled rPPG (negative sample) and upsample the downsampled rPPG to get the reconstructed rPPG (positive sample). They also used the original video to get the anchor rPPG. Then a triplet loss is used to pull together the positive and anchor samples, and push away negative and anchor samples. However, their method has three problems. 1) They have to forward one video into their backbone model twice, which causes an extra computation burden. 2) There is still a significant gap between the performance of their unsupervised method and the state-of-the-art supervised rPPG methods [45,56,34]. 3) They showed in their paper [10] that their method is easily impacted by external periodic noise.

We propose a new unsupervised method (Contrast-Phys) to tackle the problems above. Our method was built on four observations about rPPG. 1) **rPPG spatial similarity**: rPPG signals measured from different facial areas have similar power spectrum densities (PSDs) 2) **rPPG temporal similarity**: In a short time, two rPPG signals (e.g., two consecutive 5s clips) usually present similar PSDs as the HR tends to take smooth transits in most cases. 3) **Cross-video rPPG dissimilarity**: The PSDs of rPPG signals from different videos are different. 4) **HR range constraint**: The HR should fall between 40 and 250 beats per minute (bpm), so we only care about the PSD in the frequency interval between 0.66 Hz and 4.16 Hz.

We propose to use a 3D convolutional neural network (3DCNN) to process an input video to get a spatiotemporal rPPG (ST-rPPG) block. The ST-rPPG block contains multiple rPPG signals along three dimensions of height, width, and time. According to the rPPG spatiotemporal similarity, we can randomly sample rPPG signals from the same video in different spatiotemporal locations and pull them together. According to the cross-video rPPG dissimilarity, the sampled

rPPG signals from different videos are pushed away. The whole procedures are shown in Fig. 4 and Fig. 5.

The contributions of this work are 1) We propose a novel rPPG representation called spatiotemporal rPPG (ST-rPPG) block to obtain rPPG signals in spatiotemporal dimensions. 2) Based on four observations about rPPG, including rPPG spatiotemporal similarity and cross-video rPPG dissimilarity, we propose an unsupervised method based on contrastive learning. 3) We conduct experiments on five rPPG datasets (PURE [46], UBFC-rPPG [4], OBF [19], MR-NIRP [26], and MMSE-HR [59]) including RGB and NIR videos under various scenarios. Our method outperforms the previous unsupervised baseline [10] and achieves very close performance to supervised rPPG methods. Contrast-Phys also shows significant advantages with fast running speed and noise robustness compared to the previous unsupervised baseline.

## 2 Related Work

**Video-Based Remote Physiological Measurement.** Verkruysse et al. [49] first proposed that rPPG can be measured from face videos from the green channel. Several traditional handcraft methods [38,8,9,52,17,48,53] were proposed to further improve rPPG signal quality. Most rPPG methods proposed in earlier years used handcrafted procedures and did not need datasets for training, which are referred to as traditional methods. Deep learning (DL) methods for rPPG measurement are rapidly emerging. Several studies [7,45,23,35] used a 2D convolutional neural network (2DCNN) with two consecutive video frames as the input for rPPG measurement. Another type of DL-based methods [32,33,25] used a spatial-temporal signal map extracted from different facial areas as the input to feed into a 2DCNN model. Recently, 3DCNN-based methods [58,56,10] were proposed to achieved good performance on compressed videos [58]. The DL-based methods require both face videos and ground truth physiological signals, so we refer to them as supervised methods. Recently, Gideon and Stent [10] proposed an unsupervised method to train a DL model without ground truth physiological signals. However, their method falls behind some supervised methods and is not robust to external noise. In addition, the running speed is not satisfactory.

**Contrastive Learning.** Contrastive learning is a self-supervised learning method widely used in video and image feature embedding, which facilitates downstream task training and small dataset fine-tuning [12,41,36,6,47,13,11,30,39]. A DL model working as a feature extractor maps a high-dimensional image/video into a low dimensional feature vector. To train this DL feature extractor, features from different views of the same sample (positive pairs) are pulled together, while features from views of different samples (negative pairs) are pushed away. Data augmentations (such as cropping, blurring [6], and temporal sampling [39]) are used to obtain different views of the same sample so that the learned features are invariant to some augmentations. Previous works mentioned above use

contrastive learning to let a DL model produce abstract features for downstream tasks such as image classification [6], video classification [39], face recognition [41]. On the other hand, our work uses contrastive learning to directly let a DL model produce rPPG signals, enabling unsupervised learning without ground truth physiological signals.

## 3 Observations about rPPG

This section describes four observations about rPPG, which are the precondition to design our method and enable unsupervised learning.

**rPPG Spatial Similarity.** rPPG signals from different facial areas have similar waveforms, and their PSDs are also similar. Several works [17,48,16,53,51,22,21] also exploited rPPG spatial similarity to design their methods. There might be small phase and amplitude differences between two rPPG signals from two different body skin areas [14,15]. However, when rPPG waveforms are transformed to PSDs, the phase information is erased, and the amplitude can be normalized to cancel the amplitude difference. In Fig. 1, the rPPG waveforms from four spatial areas are similar, and they have the same peaks in PSDs.

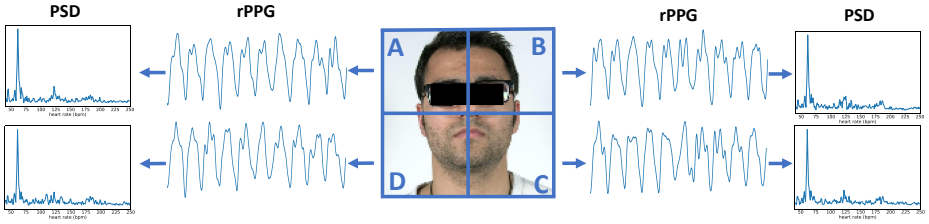


Fig. 1: Illustration of rPPG spatial similarity. The rPPG signals from four facial areas (A, B, C, D) have similar waveforms and power spectrum densities (PSDs)

**rPPG Temporal Similarity.** The HR does not change rapidly in a short term [10]. Stricker et al. [46] also found that the HR varies slightly in a short time interval in their dataset. Since the HR has a dominant peak in PSD, the PSD does not change rapidly, either. If we randomly sample several small windows from a short rPPG clip (e.g., 10s), the PSDs of these windows should be similar. In Fig. 2, we sample two 5s windows from a short 10s rPPG signal and get the PSDs of these two windows. The two PSDs are similar and have sharp peaks at the same frequency. Since this observation is only valid under the condition of short-term rPPG signals, in the following sensitivity analysis part, we will discuss the influence of the signal length on our model performance. Overall, we can use the equation $\mathrm{PSD}\{G(v(t_1 \rightarrow t_1 + \Delta t, \mathcal{H}_1, \mathcal{W}_1))\} \approx \mathrm{PSD}\{G(v(t_2 \rightarrow$

$t_2 + \Delta t, \mathcal{H}_2, \mathcal{W}_2)) \}$ to describe spatiotemporal rPPG similarity. $v \in \mathbb{R}^{T \times H \times W \times 3}$ is a facial video, $G$ is an rPPG measurement algorithm. We can choose one facial area with a set of height $\mathcal{H}_1$ and width $\mathcal{W}_1$, and a time interval $t_1 \rightarrow t_1 + \Delta t$ from video $v$ to achieve one rPPG signal. We can achieve another rPPG signal similarly from the same video with $\mathcal{H}_2$, $\mathcal{W}_2$, and $t_2 \rightarrow t_2 + \Delta t$. $|t_1 - t_2|$ should be small to satisfy the condition of short-term rPPG signals.

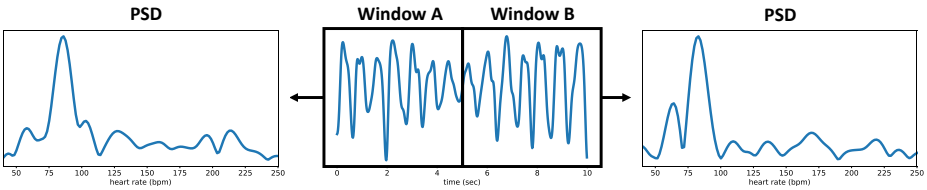


Fig. 2: Illustration of rPPG temporal similarity. The rPPG signals from two temporal windows (A, B) have similar PSDs

**Cross-video rPPG Dissimilarity.** We assume rPPG signals from different face videos have different PSDs. Each video is recorded with different people and different physiological states (such as exercises and emotion status), so the HRs across different videos are likely different [1]. Even though the HRs might be similar between two videos, the PSDs might still be different since PSD also contains other physiological factors such as respiration rate [5] and HRV [37] which are unlikely to be all the same between two videos. To further validate the observation, we calculate the mean squared error for all cross-video PSD pairs in the OBF dataset [19] and show the most similar and most different cross-video PSD pairs in Fig. 3. It can be observed that the main cross-video PSD difference is the heart rate peak. The following equation describes cross-video rPPG dissimilarity. $\text{PSD}\{G(v(t_1 \rightarrow t_1 + \Delta t, \mathcal{H}_1, \mathcal{W}_1))\} \neq \text{PSD}\{G(v'(t_2 \rightarrow t_2 + \Delta t, \mathcal{H}_2, \mathcal{W}_2))\}$ where $v$ and $v'$ are two different videos. We can choose facial areas and time intervals from these two videos. The PSDs of the two rPPG signals should be different.

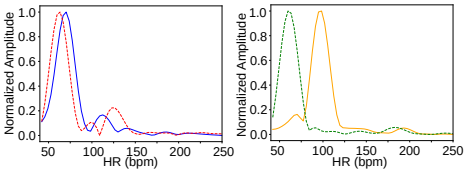


Fig. 3: The most similar (left) and most different (right) cross-video PSD pairs in the OBF dataset.

**HR Range Constraint.** The HR range for most people is between 40 and 250 bpm [2]. Most works [38,20] use this HR range for rPPG signal filtering and find

the highest peak to estimate the HR. Therefore, our method will focus on PSD between 0.66 Hz and 4.16 Hz.

## 4 Method

The overview of Contrast-Phys is shown in Fig. 4 and Fig. 5. We describe the procedures of Contrast-Phys in this section.

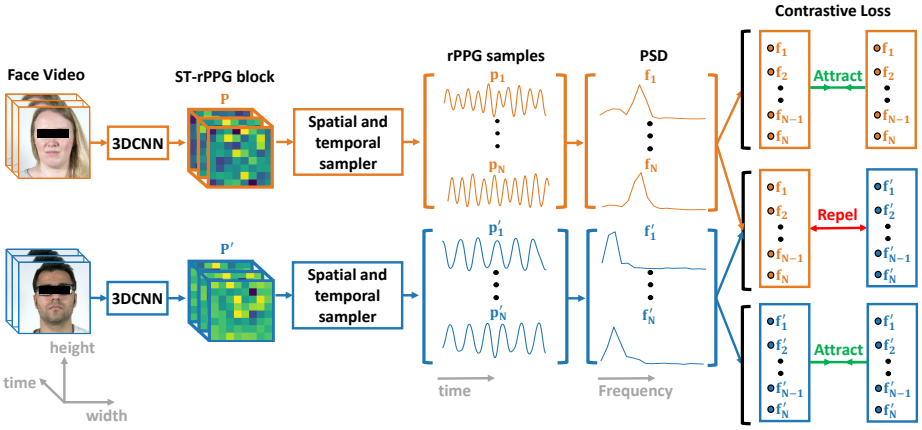


Fig. 4: Contrast-Phys Diagram. A pair of videos are fed into the same 3DCNN to generate a pair of ST-rPPG blocks. Multiple rPPG samples are sampled from the ST-rPPG blocks (The spatiotemporal sampler is illustrated in Fig. 5) and converted to PSDs. The PSDs from the same video are attracted while the PSDs from different videos are repelled.

### 4.1 Preprocessing

The original videos are firstly preprocessed to crop the face to get the face video shown in Fig. 4 left. Facial landmarks are generated using OpenFace [3]. We first get the minimum and maximum horizontal and vertical coordinates of the landmarks to locate the central facial point for each frame. The bounding box size is 1.2 times the vertical coordinate range of landmarks from the first frame and is fixed for the following frames. After getting the central facial point of each frame and the size of the bounding box, we crop the face from each frame. The cropped faces are resized to $128 \times 128$, which are ready to be fed into our model.

### 4.2 Spatiotemporal rPPG (ST-rPPG) Block Representation

We modify 3DCNN-based PhysNet [56] to get the ST-rPPG block representation. The modified model has an input RGB video with the shape of $T \times 128 \times$

$128 \times 3$ where $T$ is the number of frames. In the last stage of our model, we use adaptive average pooling to downsample along spatial dimensions, which can control the output spatial dimension length. This modification allows our model to output a spatiotemporal rPPG block with the shape of $T \times S \times S$ where $S$ is spatial dimension length as shown in Fig. 5. More details about the 3DCNN model are described in the supplementary material.

The ST-rPPG block is a collection of rPPG signals in spatiotemporal dimensions. We use $P \in \mathbb{R}^{T \times S \times S}$ to denote the ST-rPPG block. Suppose we choose a spatial location $(h, w)$ in the ST-rPPG block. In that case, the corresponding rPPG signal in this position is $P(\cdot, h, w)$ which is extracted from the receptive field of this spatial position in the original video. We can deduce that when the spatial dimension length $S$ is small, each spatial position in the ST-rPPG block has a larger receptive field. The receptive field of each spatial position in the ST-rPPG block can cover part of the facial region, which means all spatial positions in the ST-rPPG block can include rPPG information.

### 4.3 rPPG Spatiotemporal Sampling

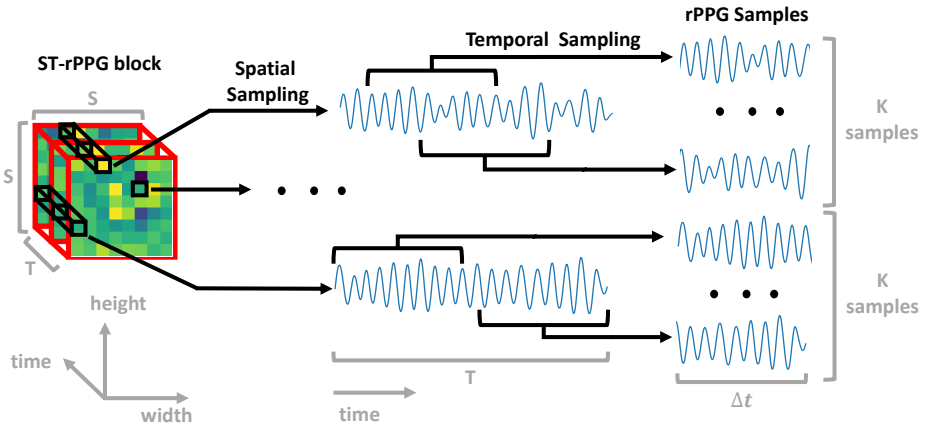


Fig. 5: Spatiotemporal Sampler

Several rPPG signals are sampled from the ST-rPPG block as illustrated in Fig. 5. For spatial sampling, we can get the rPPG signal $P(\cdot, h, w)$ at one spatial position. For temporal sampling, we can sample a short time interval from $P(\cdot, h, w)$, and the final spatiotemporal sample is $P(t \rightarrow t + \Delta t, h, w)$ where $h$ and $w$ are the spatial position, $t$ is the starting time, and $\Delta t$ is the time interval length. For one ST-rPPG block, we will loop over all spatial positions and sample $K$ rPPG clips with a randomly chosen starting time $t$ for each spatial position. Therefore, we can get $S \cdot S \cdot K$ rPPG clips from the ST-rPPG block. The more detailed sampling procedures are in our supplementary material. The sampling

procedures above are used during model training. After our model is trained and used for testing, we can directly average ST-rPPG over spatial dimensions to get the rPPG signal.

### 4.4 Contrastive Loss Function

As illustrated in Fig. 4, we have two different videos randomly chosen from a dataset as the input. For one video, we can get one ST-rPPG block $P$, a set of rPPG samples $[p_1, \ldots, p_N]$ and the corresponding PSDs $[f_1, \ldots, f_N]$. For another video, we can get one ST-rPPG block $P'$, one set of rPPG samples $[p'_1, \ldots, p'_N]$ and the corresponding PSDs $[f'_1, \ldots, f'_N]$ in the same way. As shown in Fig. 4 right, the principle of our contrastive loss is to pull together PSDs from the same video and push away PSDs from different videos. It is noted that we only use the PSD between 0.66 Hz and 4.16 Hz according to the HR range constraint in Sec. 3.

**Positive Loss Term.** According to rPPG spatiotemporal similarity, we can conclude that the PSDs from the spatiotemporal sampling of the same ST-rPPG block should be similar. We can use the following equations to describe this property for the two input videos. For one video, $\mathrm{PSD}\big\{P(t_1 \rightarrow t_1 + \Delta t, h_1, w_1)\big\} \approx \mathrm{PSD}\big\{P(t_2 \rightarrow t_2 + \Delta t, h_2, w_2)\big\} \implies f_i \approx f_j, i \neq j$. For another video, $\mathrm{PSD}\big\{P'(t_1 \rightarrow t_1 + \Delta t, h_1, w_1)\big\} \approx \mathrm{PSD}\big\{P'(t_2 \rightarrow t_2 + \Delta t, h_2, w_2)\big\} \implies f'_i \approx f'_j, i \neq j$.

We can use the mean squared error as the loss function to pull together PSDs (positive pairs) from the same video. The positive loss term $L_p$ is shown below, which is normalized with the total number of positive pairs.

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

**Negative Loss Term.** According to the cross-video rPPG dissimilarity, we can conclude that the PSDs from the spatiotemporal sampling of the two different ST-rPPG blocks should be different. We can use the following equation to describe this property for the two input videos. $\mathrm{PSD}\big\{P(t_1 \rightarrow t_1 + \Delta t, h_1, w_1)\big\} \neq \mathrm{PSD}\big\{P'(t_2 \rightarrow t_2 + \Delta t, h_2, w_2)\big\} \implies f_i \neq f'_j$

We use the negative mean squared error as the loss function to push away PSDs (negative pairs) from two different videos. The negative loss term $L_n$ is shown below, which is normalized with the total number of negative pairs.

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

The overall loss function is $L = L_p + L_n$, which is the sum of the positive and negative loss terms.

**Why Our Method Works.** Our four rPPG observations are constraints to make the model learn to keep rPPG and exclude noises since noises do not satisfy the 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. Noises such as light flickering that exceed the heart rate range are also excluded due to the heart rate range constraint. Cross-video dissimilarity in the loss can make two videos' PSDs discriminative and show clear heart rate peaks since heart rate peaks are one of the discriminative clues between two videos' PSDs.

## 5 Experiments

### 5.1 Experimental Setup and Metrics

**Datasets.** We test five commonly used rPPG datasets covering RGB and NIR videos recorded under various scenarios. PURE [46], UBFC-rPPG [4], OBF [19] and MR-NIRP [26] are used for the intra-dataset testing. MMSE-HR [59] is used for cross-dataset testing. **PURE** has ten subjects' face videos recorded in six different setups, including steady and motion tasks. We use the same experimental protocol as in [45,25] to divide the training and test set. **UBFC-rPPG** includes facial videos from 42 subjects who were playing a mathematical game to increase their HRs. We use the same protocol as in [25] for the train-test split and evaluation. **OBF** has 100 healthy subjects' videos recorded before and after exercises. We use subject-independent ten-fold cross-validation as used in [56,58,33] to make fair comparison with previous results. **MR-NIRP** has NIR videos from eight subjects sitting still or doing motion tasks. The dataset is challenging due to its small scale, and weak rPPG signals in NIR [28,50]. We will use a leave-one-subject-out cross-validation protocol for our experiments. **MMSE-HR** has 102 videos from 40 subjects recorded in emotion elicitation experiments. This dataset is also challenging since spontaneous facial expressions, and head motions are involved. More details about these datasets can be found in the supplementary material.

**Experimental Setup.** This part shows our experimental setup during training and testing. For the spatiotemporal sampler, we evaluate different spatial resolutions and time lengths of ST-rPPG blocks in the sensitivity analysis part. According to the results, we fix the parameters in the other experiments as follows. We set $K = 4$, which means, for each spatial position in the ST-rPPG block, four rPPG samples are randomly chosen. We set the spatial resolution of the ST-rPPG block as $2 \times 2$, and the time length of the ST-rPPG block as 10s. The time interval $\Delta t$ of each rPPG sample is half of the time length of the ST-rPPG block. We use AdamW optimizer [24] to train our model with a learning rate of $10^{-5}$ for 30 epochs on one NVIDIA Tesla V100 GPU. For each

training iteration, the inputs are two 10s clips from two different videos, respectively. During testing, we broke each test video into non-overlapping 30s clips and computed rPPG for each clip. We locate the highest peak in the PSD of an rPPG signal to calculate the HR. We use Neurokit2 [27] to calculate HRV metrics for the reported HRV results.

**Evaluation Metrics.** Following previous work [20,33,58], we use mean absolute error (MAE), root mean squared error (RMSE), and Pearson correlation coefficient (R) to evaluate the accuracy of HR measurement. Following [25], we also use standard deviation (STD), RMSE, and R to evaluate the accuracy of HRV features, including respiration frequency (RF), low-frequency power (LF) in normalized units (n.u.), high-frequency power (HF) in normalized units (n.u.), and the ratio of LF and HF power (LF/HF). For MAE, RMSE, and STD, smaller values mean lower errors, while for R, larger values close to one mean lower errors. Please check our supplementary material for more details about evaluation metrics.

### 5.2 Intra-dataset Testing

**HR Estimation.** We perform intra-dataset testing for HR estimation on PURE [46], UBFC-rPPG [4], OBF [19], and MR-NIRP [26]. Table 1 shows HR estimation results, including traditional methods, supervised methods, and unsupervised methods. The proposed Contrast-Phys largely surpasses the previous unsupervised baseline [10] and almost approaches the best supervised methods. The superior performance of Contrast-Phys is consistent on all four datasets, including the MR-NIRP [26] dataset, which contains NIR videos. The results indicate that the unsupervised Contrast-Phys can achieve reliable HR estimation on both RGB and NIR videos without requiring any ground truth physiological signals for training. Fig. 6 also shows that the rPPG waveform from our unsupervised method is very similar to the ground truth PPG signal.

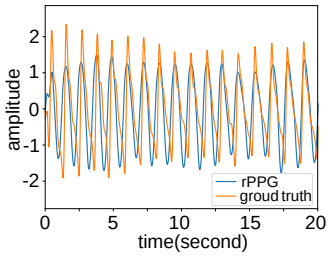


Fig. 6: rPPG waveform and ground truth PPG signal.

**HRV Estimation.** We also perform intra-dataset testing for HRV on UBFC-rPPG [4] as shown in Table 2. HRV results need to locate each systolic peak, which requires high-quality rPPG signals. Our method significantly outperforms traditional methods and the previous unsupervised baseline [10] on HRV results. There is a marginal difference in HRV results between ours and the supervised methods [33,25]. The results indicate that Contrast-Phys can achieve high-quality rPPG signals with accurate systolic peaks to calculate HRV features, which makes it feasible to be used for emotion understanding [57,29,40] and healthcare applications [42,55].

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 [49] | 7.50 | 14.41 | 0.62 | - | - | - | - | 2.162 | 0.99 | - | - | - |
| | ICA [38] | 5.17 | 11.76 | 0.65 | - | - | - | - | - | - | - | - | - |
| | CHROM [8] | 2.37 | 4.91 | 0.89 | 2.07 | 9.92 | **0.99** | - | 2.733 | 0.98 | - | - | - |
| | 2SR [54] | - | - | - | 2.44 | 3.06 | 0.98 | - | - | - | - | - | - |
| | POS [52] | 4.05 | 8.75 | 0.78 | - | - | - | - | 1.906 | 0.991 | - | - | - |
| Supervised | CAN [7] | - | - | - | - | - | - | - | - | - | 7.78 | 16.8 | -0.03 |
| | HR-CNN [45] | - | - | - | 1.84 | 2.37 | 0.98 | - | - | - | - | - | - |
| | SynRhythm [31] | 5.59 | 6.82 | 0.72 | - | - | - | - | - | - | - | - | - |
| | PhysNet [56] | - | - | - | 2.1 | 2.6 | **0.99** | - | 1.812 | 0.992 | 3.07 | 7.55 | 0.655 |
| | rPPGNet [58] | - | - | - | - | - | - | - | 1.8 | 0.992 | - | - | - |
| | CVD [33] | - | - | - | - | - | - | - | **1.26** | **0.996** | - | - | - |
| | PulseGAN [44] | 1.19 | 2.10 | 0.98 | - | - | - | - | - | - | - | - | - |
| | Dual-GAN [25] | **0.44** | **0.67** | **0.99** | **0.82** | **1.31** | **0.99** | - | - | - | - | - | - |
| | Nowara2021 [35] | - | - | - | - | - | - | - | - | - | **2.34** | **4.46** | **0.85** |
| Unsupervised | Gideon2021 [10] | 1.85 | 4.28 | 0.93 | 2.3 | 2.9 | **0.99** | 2.83 | 7.88 | 0.825 | 4.75 | 9.14 | 0.61 |
| | **Ours** | 0.64 | 1.00 | **0.99** | 1.00 | 1.40 | **0.99** | **0.51** | 1.39 | 0.994 | 2.68 | 4.77 | **0.85** |

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 [49] | 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 [38] | 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 [52] | 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 [33] | 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 [25] | **0.034** | **0.035** | **0.891** | **0.034** | **0.035** | **0.891** | **0.131** | **0.136** | **0.881** | **0.010** | **0.010** | **0.395** |
| Unsupervised | Gideon2021 [10] | 0.091 | 0.139 | 0.694 | 0.091 | 0.139 | 0.694 | 0.525 | 0.691 | 0.684 | 0.061 | 0.098 | 0.103 |
| | **Ours** | 0.050 | 0.098 | 0.798 | 0.050 | 0.098 | 0.798 | 0.205 | 0.395 | 0.782 | 0.055 | 0.083 | 0.347 |

### 5.3 Cross-dataset Testing

We conduct cross-dataset testing on MMSE-HR [59] to test the generalization ability of our method. We train Contrast-Phys and Gideon2021 on UBFC-rPPG and test on MMSE-HR. We also show MMSE-HR cross-dataset results reported in the papers of some supervised methods [32,56,33,23,35] with different training sets. Table 3 shows the cross-dataset test results. Our method still outperforms Gideon2021 [10] and is close to supervised methods. The cross-dataset testing results demonstrate that Contrast-Phys can be trained on one dataset without ground truth physiological signals, and then generalize well to a new dataset.

### 5.4 Running Speed

We test the running speed of the proposed Contrast-Phys and compare it with Gideon2021 [10]. During training, the speed of our method is **802.45** frames

Table 3: Cross-dataset HR Estimation on MMSE-HR. The best results are in bold, and the second-best results are underlined.

| Method Types | Methods | MAE (bpm) | RMSE (bpm) | R |
|---|---|---|---|---|
| Traditional | Li2014 [20] | - | 19.95 | 0.38 |
| | CHROM [8] | - | 13.97 | 0.55 |
| | SAMC [48] | - | 11.37 | 0.71 |
| Supervised | RhythmNet [32] | - | 7.33 | 0.78 |
| | PhysNet [56] | - | 13.25 | 0.44 |
| | CVD [33] | - | <u>6.04</u> | 0.84 |
| | TS-CAN [23] | 3.41 | 7.82 | 0.84 |
| | Nowara2021 [35] | **2.27** | **4.90** | **0.94** |
| Unsupervised | Gideon2021 [10] | 4.10 | 11.55 | 0.70 |
| | **Ours** | <u>2.43</u> | 7.34 | <u>0.86</u> |

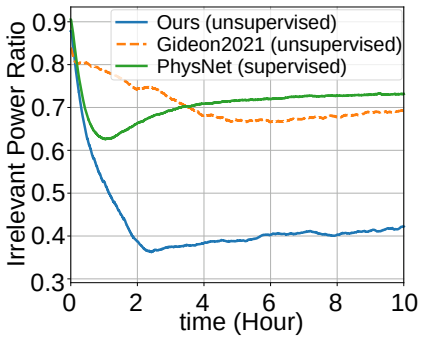


Fig. 7: irrelevant power ratio change during training time for unsupervised and supervised methods

per second (fps), while that of Gideon2021 is **387.87** fps, which is about half of our method's speed. The large difference is due to different method designs. For Gideon2021, one input video has to be fed into the model twice, i.e., firstly as the original video and then as a temporal resampled video for a second time, which causes double computation. For our method, one video is fed into the model once, which can substantially decrease computational cost compared with Gideon2021.

Furthermore, we compare the convergence speed of the two unsupervised methods and one supervised method (PhysNet [56]) using the metric of irrelevant power ratio (IPR). IPR is used in [10] to evaluate the signal quality during training, and lower IPR means higher signal quality. (more details about IPR are in the supplementary materials.) Fig. 7 shows IPR with respect to time during training on the OBF dataset. Our method converges to the lowest IPR point for about 2.5 hours, while Gideon2021 [10] converges to the lowest point for about 5 hours. In addition, the lowest IPR for our method is about 0.36 while that for Gideon2021 [10] is about 0.66 which is higher than ours. The above evidence demonstrates that our method converges faster to a lower IPR than Gideon2021 [10]. In addition, our method also achieves lower IPR than the supervised method (PhysNet [56])

### 5.5 Saliency Maps

We calculate saliency maps to illustrate the interpretability of our method. The saliency maps are obtained using a gradient-based method proposed in [43]. We fix the weights of the trained model and get the gradient of Pearson correlation with respect to the input video (More details are in the supplementary materials.). Saliency maps can highlight spatial regions from which the model estimates the rPPG signals, so the saliency map of a good rPPG model should have a large response on skin regions, as demonstrated in [56,58,10,7,35].

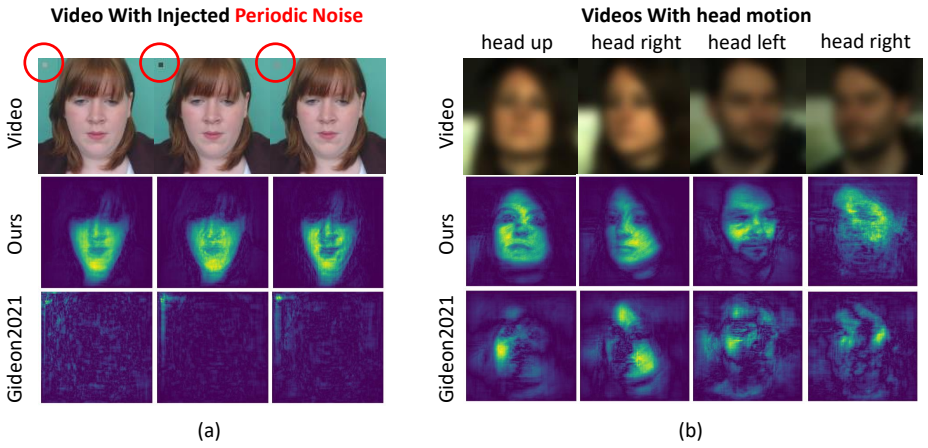


Fig. 8: Saliency maps for our method and Gideon2021 [10] (a) We add a random flashing block with the HR range between 40-250 bpm in the top left corner to all UBFC-rPPG videos. The models for our method and Gideon2021 [10] are trained on these videos with the noise. Our saliency maps have a high response on facial regions, while Gideon2021 focuses on this injected periodic noise. Table 4 also shows that our method is robust to the noise. (b) We choose the head motion moments in PURE dataset (Video frames shown here are blurred due to privacy issues.) and show the saliency maps for our method and Gideon2021 [10].

Fig. 8 shows saliency maps under two cases, 1) periodic noise is manually injected, and 2) head motion is involved. When a periodic noise patch is injected to the left-upper corner of videos, our method is not distracted by the noise and still focuses on skin areas, while Gideon2021 is completely distracted by the noise block. We also calculate the two methods' performance on UBFC-rPPG videos with the injected noise, and the results are listed in Table 4. The results are consistent with the saliency map analysis, that the periodic noise does not impact Contrast-Phys, but fails Gideon2021 completely. Our method is robust to the noise because the noise only exists in one region, which violates rPPG spatial similarity. Fig. 8(b) shows the saliency maps when head motion is involved. The saliency maps for our method focus and activate most skin areas, while saliency maps for Gideon2021 [10] show messy patterns and only partially cover facial areas during head motions.

Table 4: HR results trained on UBFC-rPPG with/without injected periodic noise shown in Fig. 8(a)

| Methods | Injected Periodic Noise | MAE (bpm) | RMSE (bpm) | R |
|---|---|---|---|---|
| Gideon2021 [10] | w/o | 1.85 | 4.28 | 0.939 |
| | w/ | 22.47 | 25.41 | 0.244 |
| Ours | w/o | 0.64 | 1.00 | 0.995 |
| | w/ | 0.74 | 1.34 | 0.991 |

### 5.6 Sensitivity Analysis

We perform sensitivity analysis on two variables: 1) the spatial length $S$ of the ST-rPPG block, and 2) the temporal length $T$ of the ST-rPPG block.

Table 5(a) shows the HR results on UBFC-rPPG, when the ST-rPPG spatial resolution is set in four levels of 1x1, 2x2, 4x4 or 8x8. Note that $1 \times 1$ means that rPPG spatial similarity is not used. The results of $1 \times 1$ are worse than other results with spatial information, which means rPPG spatial similarity improves performance. In addition, $2 \times 2$ is enough to perform well since larger resolutions do not significantly improve the HR estimation. Although 8x8 or 4x4 provides more rPPG samples, it has a smaller receptive field and thus produces noisier rPPG samples than a 2x2 block.

Table 5(b) shows the HR results on UBFC-rPPG, when the ST-rPPG temporal length is set on three levels of 5s, 10s and 30s. The results indicate that 10s is the best choice. A shorter time length (5s) causes coarse PSD estimation, while a long time length (30s) causes a slight violation of short-term signal condition in rPPG temporal similarity. Therefore, in these two cases (5s and 30s), the performance is lower than that of 10s.

Table 5: Sensitivity Analysis: (a) HR results on UBFC-rPPG with different ST-rPPG block spatial resolutions. (b) HR results on UBFC-rPPG with different ST-rPPG block time lengths. (The best results are in bold.)

(a)

| Spatial Resolution | MAE (bpm) | RMSE (bpm) | R |
|---|---|---|---|
| $1 \times 1$ | 3.14 | 4.06 | 0.963 |
| **$2 \times 2$** | **0.64** | **1.00** | **0.995** |
| $4 \times 4$ | 0.55 | 1.06 | 0.994 |
| $8 \times 8$ | 0.60 | 1.09 | 0.993 |

(b)

| Time Length | 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 |

## 6 Conclusion

We propose Contrast-Phys which can be trained without ground truth physiological signals and achieve accurate rPPG measurement. Our method is based on four observations about rPPG and utilizes spatiotemporal contrast to enable unsupervised learning. Contrast-Phys significantly outperforms the previous unsupervised baseline [10] and is on par with the state-of-the-art supervised rPPG methods. In the future work, we would like to combine the supervised and the proposed unsupervised rPPG methods to further improve performance.

**Acknowledgment.** The study was supported by Academy of Finland (Project 323287 and 345948) and the Finnish Work Environment Fund (Project 200414). The authors also acknowledge CSC-IT Center for Science, Finland, for providing computational resources.

## References


1. All about heart rate (pulse), https://www.heart.org/en/health-topics/high-blood-pressure/the-facts-about-high-blood-pressure/all-about-heart-rate-pulse
2. Target heart rates chart, https://www.heart.org/en/healthy-living/fitness/fitness-basics/target-heart-rates
3. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: Facial behavior analysis toolkit. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). pp. 59–66. IEEE (2018)
4. Bobbia, S., Macwan, R., Benezeth, Y., Mansouri, A., Dubois, J.: Unsupervised skin tissue segmentation for remote photoplethysmography. Pattern Recognition Letters **124**, 82–90 (2019)
5. Chen, M., Zhu, Q., Wu, M., Wang, Q.: Modulation model of the photoplethysmography signal for vital sign extraction. IEEE Journal of Biomedical and Health Informatics **25**(4), 969–977 (2020)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
7. Chen, W., McDuff, D.: Deepphys: Video-based physiological measurement using convolutional attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 349–365 (2018)
8. De Haan, G., Jeanne, V.: Robust pulse rate from chrominance-based rppg. IEEE Transactions on Biomedical Engineering **60**(10), 2878–2886 (2013)
9. De Haan, G., Van Leest, A.: Improved motion robustness of remote-ppg by using the blood volume pulse signature. Physiological measurement **35**(9), 1913 (2014)
10. Gideon, J., Stent, S.: 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. pp. 3995–4004 (2021)
11. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in Neural Information Processing Systems **33**, 21271–21284 (2020)
12. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 2, pp. 1735–1742. IEEE (2006)
13. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
14. Kamshilin, A.A., Miridonov, S., Teplov, V., Saarenheimo, R., Nippolainen, E.: Photoplethysmographic imaging of high spatial resolution. Biomedical optics express **2**(4), 996–1006 (2011)
15. Kamshilin, A.A., Teplov, V., Nippolainen, E., Miridonov, S., Giniatullin, R.: Variability of microcirculation detected by blood pulsation imaging. PloS one **8**(2), e57117 (2013)
16. Kumar, M., Veeraraghavan, A., Sabharwal, A.: Distanceppg: Robust non-contact vital signs monitoring using a camera. Biomedical optics express **6**(5), 1565–1588 (2015)
17. Lam, A., Kuno, Y.: Robust heart rate measurement from video using select random patches. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3640–3648 (2015)

18. Lee, E., Chen, E., Lee, C.Y.: Meta-rppg: Remote heart rate estimation using a transductive meta-learner. In: European Conference on Computer Vision. pp. 392–409. Springer (2020)
19. Li, X., Alikhani, I., Shi, J., Seppanen, T., Junttila, J., Majamaa-Voltti, K., Tulppo, M., Zhao, G.: The obf database: A large face video database for remote physiological signal measurement and atrial fibrillation detection. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 242–249. IEEE (2018)
20. Li, X., Chen, J., Zhao, G., Pietikainen, M.: Remote heart rate measurement from face videos under realistic situations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4264–4271 (2014)
21. Liu, S.Q., Lan, X., Yuen, P.C.: Remote photoplethysmography correspondence feature for 3d mask face presentation attack detection. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 558–573 (2018)
22. Liu, S., Yuen, P.C., Zhang, S., Zhao, G.: 3d mask face anti-spoofing with remote photoplethysmography. In: European Conference on Computer Vision. pp. 85–100. Springer (2016)
23. Liu, X., Fromm, J., Patel, S., McDuff, D.: Multi-task temporal shift attention networks for on-device contactless vitals measurement. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 19400–19411 (2020)
24. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019), `https://openreview.net/forum?id=Bkg6RiCqY7`
25. Lu, H., Han, H., Zhou, S.K.: Dual-gan: Joint bvp and noise modeling for remote physiological measurement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12404–12413 (2021)
26. Magdalena Nowara, E., Marks, T.K., Mansour, H., Veeraraghavan, A.: 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. p. 1272–1281 (2018)
27. Makowski, D., Pham, T., Lau, Z.J., Brammer, J.C., Lespinasse, F., Pham, H., Schölzel, C., Chen, S.: Neurokit2: A python toolbox for neurophysiological signal processing. Behavior research methods **53**(4), 1689–1696 (2021)
28. Martinez, L.F.C., Paez, G., Strojnik, M.: Optimal wavelength selection for non-contact reflection photoplethysmography. In: 22nd Congress of the International Commission for Optics: Light for the Development of the World. vol. 8011, p. 801191. International Society for Optics and Photonics (2011)
29. McDuff, D., Gontarek, S., Picard, R.: Remote measurement of cognitive stress via heart rate variability. In: 2014 36th annual international conference of the IEEE engineering in medicine and biology society. pp. 2957–2960. IEEE (2014)
30. Misra, I., Maaten, L.v.d.: Self-supervised learning of pretext-invariant representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6707–6717 (2020)
31. Niu, X., Han, H., Shan, S., Chen, X.: Synrhythm: Learning a deep heart rate estimator from general to specific. In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 3580–3585. IEEE (2018)
32. Niu, X., Shan, S., Han, H., Chen, X.: Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. IEEE Transactions on Image Processing **29**, 2409–2423 (2019)

33. Niu, X., Yu, Z., Han, H., Li, X., Shan, S., Zhao, G.: Video-based remote physiological measurement via cross-verified feature disentangling. In: European Conference on Computer Vision. pp. 295–310. Springer (2020)
34. Nowara, E.M., Marks, T.K., Mansour, H., Veeraraghavan, A.: Near-infrared imaging photoplethysmography during driving. IEEE Transactions on Intelligent Transportation Systems (2020)
35. Nowara, E.M., McDuff, D., Veeraraghavan, A.: The benefit of distraction: Denoising camera-based physiological measurements using inverse attention. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4955–4964 (2021)
36. Van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv e-prints pp. arXiv–1807 (2018)
37. Pai, A., Veeraraghavan, A., Sabharwal, A.: Hrvcam: robust camera-based measurement of heart rate variability. Journal of Biomedical Optics **26**(2), 022707 (2021)
38. Poh, M.Z., McDuff, D.J., Picard, R.W.: Advancements in noncontact, multiparameter physiological measurements using a webcam. IEEE transactions on biomedical engineering **58**(1), 7–11 (2010)
39. Qian, R., Meng, T., Gong, B., Yang, M.H., Wang, H., Belongie, S., Cui, Y.: Spatiotemporal contrastive video representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6964–6974 (2021)
40. Sabour, R.M., Benezeth, Y., De Oliveira, P., Chappe, J., Yang, F.: Ubfc-phys: A multimodal database for psychophysiological studies of social stress. IEEE Transactions on Affective Computing (2021)
41. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
42. Shi, J., Alikhani, I., Li, X., Yu, Z., Seppänen, T., Zhao, G.: Atrial fibrillation detection from face videos by fusing subtle variations. IEEE Transactions on Circuits and Systems for Video Technology **30**(8), 2781–2795 (2019)
43. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
44. Song, R., Chen, H., Cheng, J., Li, C., Liu, Y., Chen, X.: Pulsegan: Learning to generate realistic pulse waveforms in remote photoplethysmography. IEEE Journal of Biomedical and Health Informatics **25**(5), 1373–1384 (2021)
45. Špetlík, R., Franc, V., Matas, J.: Visual heart rate estimation with convolutional neural network. In: Proceedings of the british machine vision conference, Newcastle, UK. pp. 3–6 (2018)
46. Stricker, R., Müller, S., Gross, H.M.: Non-contact video-based pulse rate measurement on a mobile service robot. In: The 23rd IEEE International Symposium on Robot and Human Interactive Communication. pp. 1056–1062. IEEE (2014)
47. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In: European conference on computer vision. pp. 776–794. Springer (2020)
48. Tulyakov, S., Alameda-Pineda, X., Ricci, E., Yin, L., Cohn, J.F., Sebe, N.: 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. pp. 2396–2404 (2016)
49. Verkruysse, W., Svaasand, L.O., Nelson, J.S.: Remote plethysmographic imaging using ambient light. Optics express **16**(26), 21434–21445 (2008)

50. Vizbara, V.: Comparison of green, blue and infrared light in wrist and forehead photoplethysmography. BIOMEDICAL ENGINEERING 2016 **17**(1) (2013)
51. Wang, W., den Brinker, A.C., De Haan, G.: Discriminative signatures for remote-ppg. IEEE Transactions on Biomedical Engineering **67**(5), 1462–1473 (2019)
52. Wang, W., den Brinker, A.C., Stuijk, S., De Haan, G.: Algorithmic principles of remote ppg. IEEE Transactions on Biomedical Engineering **64**(7), 1479–1491 (2016)
53. Wang, W., Stuijk, S., De Haan, G.: Exploiting spatial redundancy of image sensor for motion robust rppg. IEEE transactions on Biomedical Engineering **62**(2), 415–425 (2014)
54. Wang, W., Stuijk, S., De Haan, G.: A novel algorithm for remote photoplethysmography: Spatial subspace rotation. IEEE transactions on biomedical engineering **63**(9), 1974–1984 (2015)
55. Yan, B.P., Lai, W.H., Chan, C.K., Chan, S.C.H., Chan, L.H., Lam, K.M., Lau, H.W., Ng, C.M., Tai, L.Y., Yip, K.W., et al.: Contact-free screening of atrial fibrillation by a smartphone using facial pulsatile photoplethysmographic signals. Journal of the American Heart Association **7**(8), e008585 (2018)
56. Yu, Z., Li, X., Zhao, G.: Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In: 30th British Machine Vision Conference 2019. p. 277. BMVA Press (2019)
57. Yu, Z., Li, X., Zhao, G.: Facial-video-based physiological signal measurement: Recent advances and affective applications. IEEE Signal Processing Magazine **38**(6), 50–58 (2021)
58. Yu, Z., Peng, W., Li, X., Hong, X., Zhao, G.: Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 151–160 (2019)
59. Zhang, Z., Girard, J.M., Wu, Y., Zhang, X., Liu, P., Ciftci, U., Canavan, S., Reale, M., Horowitz, A., Yang, H., et al.: Multimodal spontaneous emotion corpus for human behavior analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3438–3446 (2016)