# *ALPINE*: Improving Remote Heart Rate Estimation using Contrastive Learning

Lokendra Birla Sneha Shukla Anup Kumar Gupta Puneet Gupta
Indian Institute of Technology Indore
{phd1901201001, phd2101101006, msrphd2105101002, puneet}@iiti.ac.in

## Abstract

*Heart rate (HR) is a crucial physiological indicator of human health and can be used to detect cardiovascular disorders. The traditional HR estimation methods, such as electrocardiograms (ECG) and photoplethysmographs, require skin contact. Due to the increased risk of viral infection from skin contact, these approaches are avoided in the ongoing COVID-19 pandemic. Alternatively, one can use the non-contact HR estimation technique, remote photoplethysmography (rPPG), wherein HR is estimated from the facial videos of a person. Unfortunately, the existing rPPG methods perform poorly in the presence of facial deformations. Recently, there has been a proliferation of deep learning networks for rPPG. However, these networks require large-scale labelled data for better generalization. To alleviate these shortcomings, we propose a method ALPINE, that is,* ***A*** *nove**L** r**P**PG technique for **I**mproving the remote heart rate estimatio**N** using contrastive l**E**arning. ALPINE utilizes the contrastive learning framework during training to address the issue of limited labelled data and introduces diversity in the data samples for better network generalization. Additionally, we introduce a novel hybrid loss comprising contrastive loss, signal-to-noise ratio (SNR) loss and data fidelity loss. Our novel contrastive loss maximizes the similarity between the rPPG information from different facial regions, thereby minimizing the effect of local noise. The SNR loss improves the quality of temporal signals, and the data fidelity loss ensures that the correct rPPG signal is extracted. Our extensive experiments on publicly available datasets demonstrate that the proposed method, ALPINE outperforms the previous well-known rPPG methods.*

## 1. Introduction

The heart is an essential human body organ, and heart rate (HR) can detect cardiovascular diseases. The conventional HR estimation methods, electrocardiograms (ECG) and photoplethysmographs (PPG) require skin contact and perform spuriously in the presence of motion. Thus, these methods are avoided in the ongoing COVID-19 pandemic as skin contact increases the viral infection risk. Alternatively, several methods have been proposed in the literature to perform non-contact HR estimation and overcome the abovementioned limitations. The non-contact HR estimation methods utilize the remote photoplethysmography (rPPG) technique, wherein HR is estimated from a face video of a person by analyzing the color or motion variations present in the video. The rPPG is actively used in chronic disease treatment therapies [16], and cardiac activity monitoring [9]. Apart from the applications in healthcare, rPPG is also used for micro-expression recognition [21], Deepfake detection [11], micro-expression spotting [24], liveness detection [41] and spoof detection [4, 5].

The rPPG is based on the phenomenon that the heart pumps the blood in carotid arteries beneath the facial skin, resulting in color or motion variations in the facial region. Though the variations are latent to the human eye, these can be captured using a camera and processed using signal processing or machine learning techniques to compute the physiological parameters [26]. The facial variation signal is referred to as a temporal signal. Usually, multiple facial regions are used to compute the temporal signals, and these extracted signals are consolidated using blind source separation algorithms to estimate the pulse signal [22]. In this direction, several rPPG methods have been proposed based on signal processing, and deep learning [22, 25, 23, 2, 3, 28, 34].

The existing methods are incompetent for rPPG estimation when the face videos contain facial deformation due to eye-blinking, illumination variations, head movements, and facial expressions. Moreover, the regime of supervised deep learning methods requires extensive labeled data for training the network correctly. Unfortunately, the flexibility of having costly labeled data is absent in our case as it requires enormous data collection and operator intervention [6, 18, 17]. This limitation is mitigated in [18] using self-supervised learning based on contrastive learning (CL). The CL is a self-supervised machine learning technique that allows a network to learn the features from a dataset without utilizing the labels. In this technique, the network learns which features are similar and which are different.

The method proposed in [18] uses temporal augmentation to learn the similarity features between the actual and augmented samples, using a triplet loss. The loss forces the network to learn the similarity between the temporal signal and its augmented version. This leads to incorrect pulse modeling when the temporal signal contains noise; hence resulting in limited network performance.

We propose a novel rPPG-based HR estimation method, *ALPINE*, that is, **A** nove**L** r**P**PG technique for **I**mproving the remote heart rate estimatio**N** using contrastive l**E**arning, which mitigates the limitations mentioned above and improves the HR estimation. It introduces a TCN-based deep learning network that models rPPG information to mitigate the noises from temporal signals using CL. To denoise the temporal signal and predict a clean pulse signal, we propose a novel network that utilizes the temporal signals extracted from the multiple facial regions. These temporal signals enable the network to learn rPPG information across all the regions. Contrastive loss is utilized to learn the similarity between the different temporal signals. It suppresses any facial deformations-based noises that deteriorate the extracted temporal signal. We utilize the SNR loss to reduce the noise present in temporal signals. Finally, we utilize the data fidelity loss to ensure that the denoised signal corresponds to the ground truth. Furthermore, the proposed network is based on TCN, to effectively learns the temporal information of long sequences than its sequential counterparts [1]. Our primary research contributions are:

(1) We propose a novel CL-based network that utilizes the temporal augmentation technique to automatically learns the rPPG information present in different facial regions. (2) We propose a novel hybrid loss function to train the network. It enables the network to automatically learn the similarity between temporal signals extracted from different facial regions, and predicts denoised temporal signals. (3) Our experimental results on publicly available datasets reveal that the proposed method performs better than state-of-the-art rPPG-based methods.

# 2. Literature Survey

## 2.1. Traditional rPPG estimation

The rPPG estimation consists of analyzing the minute variations in the facial videos caused by the pulse motion. The variations are captured by measuring the color variations [37] or the motion variations [2] in the facial video. The information provided by them diminishes in the presence of noise caused by camera artifacts and motion [38], resulting in incorrect HR estimation. Thus, to suppress the noise, several conventional rPPG methods have been proposed. For instance, green color variations are used in [43] for HR estimation as green color is better absorbed by the blood haemoglobin, thereby providing better rPPG information [47]. Independent Component Analysis (ICA) is used for separating the pulse signal from the temporal signals in [37]. Furthermore, color subspace transformation to Chrominance is proposed in [13] to the suppress the illumination variations. Similarly, Eigen decomposition of the temporal signals followed by spatial subspace rotation is used for HR estimation in [44]. These rPPG methods utilized color variations for modeling rPPG information. In contrast, [2] has used facial motion caused by pulse followed by Principle Component Analysis (PCA) for HR estimation. The aforementioned rPPG methods benefit from the handcrafted features for alleviating the effect of noise. However, the noise sources induce varying characteristics that cannot be separated from domain-specific knowledge [3].

## 2.2. Deep Learning-based rPPG estimation

Deep learning methods are capable of automatically learning the feature representations from data [19], which can be used for suppressing noise for correct rPPG estimation. In this direction, Convolutional Neural Network (CNN) based VGG15 is used for estimating the HR frequency from the time-frequency representation of the CHROM [13] signals is used in [28]. Spatiotemporal convolution is used [32] for rPPG signal estimation. A time-domain subnet for learning temporal correlations corresponding to the pulse signal is employed for pulse estimation in [29]. A combination of 2D CNN and 1D CNN network is used in [40] for HR estimation. The 2D CNN is used for extracting rPPG information from the facial videos that are fed to the 1D CNN for pulse estimation. Furthermore, normalized frame difference is used by end-to-end rPPG methods. For instance, a two-stream network is used for rPPG-based HR estimation in [10]. The network employs an attention-based appearance stream to identify the facial regions that provide important rPPG information and a CNN-based motion stream to learn the pulse signal features. Spatio-temporal networks are used for pulse estimation in [46]. In essence, they have devised two networks, one utilizing 3D CNN network and the other utilizing a combination of 2D CNN and Long Short Term Memory (LSTM) network for rPPG estimation. These methods comprise deep neural networks with learnable weights that require labeled training data in abundant quantities for correct generalization. Moreover, they fail to provide correct HR estimation for small-scale rPPG datasets [8]. Moreover, the supervised learning-based methods provide incorrect HR estimation for samples having HR outside the HR ranges in the training set [31].

## 2.3. CL and rPPG estimation using CL

The supervised learning-based rPPG methods suffer from a lack of abundant labeled rPPG datasets [8] and the varying data distribution in the training and testing sets of the rPPG datasets [31]. Thus, a meta-learner is used for self-supervised test time weight adaptation in [31]. Moreover, the issue of limited data is prevalent in deep learning-based methods, and CL has been used in this direction for data augmentations. CL methods are widely used in unsupervised visual representation learning [14] like action recognition [33], image captioning [12], and image to image translation [36].

CL-based rPPG methods are merely found in literature as the evaluation in the absence of actual pulse is unreliable [18]. In this direction, a triplet loss is used in [17]. In essence, they have used video resampling to generate positive and negative video samples. The loss function aims to decrease the difference between the pulse estimated for the original video sample and the positive sample while increasing the distance between the pulse estimated for the negative sample and the original video sample. However, this approach is unable to specify whether the information corresponds to the pulse signal or any other source; hence the performance evaluation is unreliable [18]. To mitigate this issue, [18] introduced a saliency sampling layer that helps the rPPG method to identify important facial regions that contribute to rPPG information.

# 3. Proposed Method

The rPPG estimation method, *ALPINE*, is presented in this section. In this method, we divide the input face video into non-overlapping clips of fixed size, and HR is estimated from these input clips. To this end, the proposed method detects the face and divides it into several ROIs. Then we compute the temporal signals corresponding to each ROI. We pass these temporal signals to the Contrastive rPPG Network (CrPPG-Net) to compute the denoised temporal signals. The CrPPG-Net is based on the TCN . Also, it introduces the hybrid loss function during the training in a semi-supervised learning pattern. The loss consisting of three components: contrastive loss, SNR loss, and data fidelity loss. During the time of inference, we consolidate these denoised temporal signals to compute the pulse signal using the blind-source-separation (BSS) based Multi-Kurtosis Optimization method [26]. The proposed method *ALPINE* is depicted using a flow diagram in Fig. 1.

## 3.1. ROI Extraction

In order to localize the facial region with significant rPPG information, we require the discriminatory facial points (known as facial landmark points), which outline the facial regions as facial boundaries. To this end, we use the Deep Alignment Network (DAN) [30]. We utilize the DAN because it provides the landmark points for the forehead region along with the face, whereas well-known facial landmark point extraction networks like CLNF-Openface [48] are unable to provide the forehead region landmark points. Kindly note that we only use the landmark points corresponding to the left cheek, forehead, nose, and right cheek to extract the ROIs, as these regions are least influenced by facial expressions [3]. The eyes and lips regions are avoided as these are easily affected by the deformations, while the chin region skin is avoided as it is occluded in subjects with a beard. We extract these ROIs by estimating the minimum enclosing area rectangle of that region's landmark points, as suggested in [4, 7]. Also, the facial boundary pixels are prone to facial deformations and thereby reduce performance. Hence, we remove the boundary pixels during ROI extraction, as suggested in [24].

## 3.2. Extracting Temporal signal

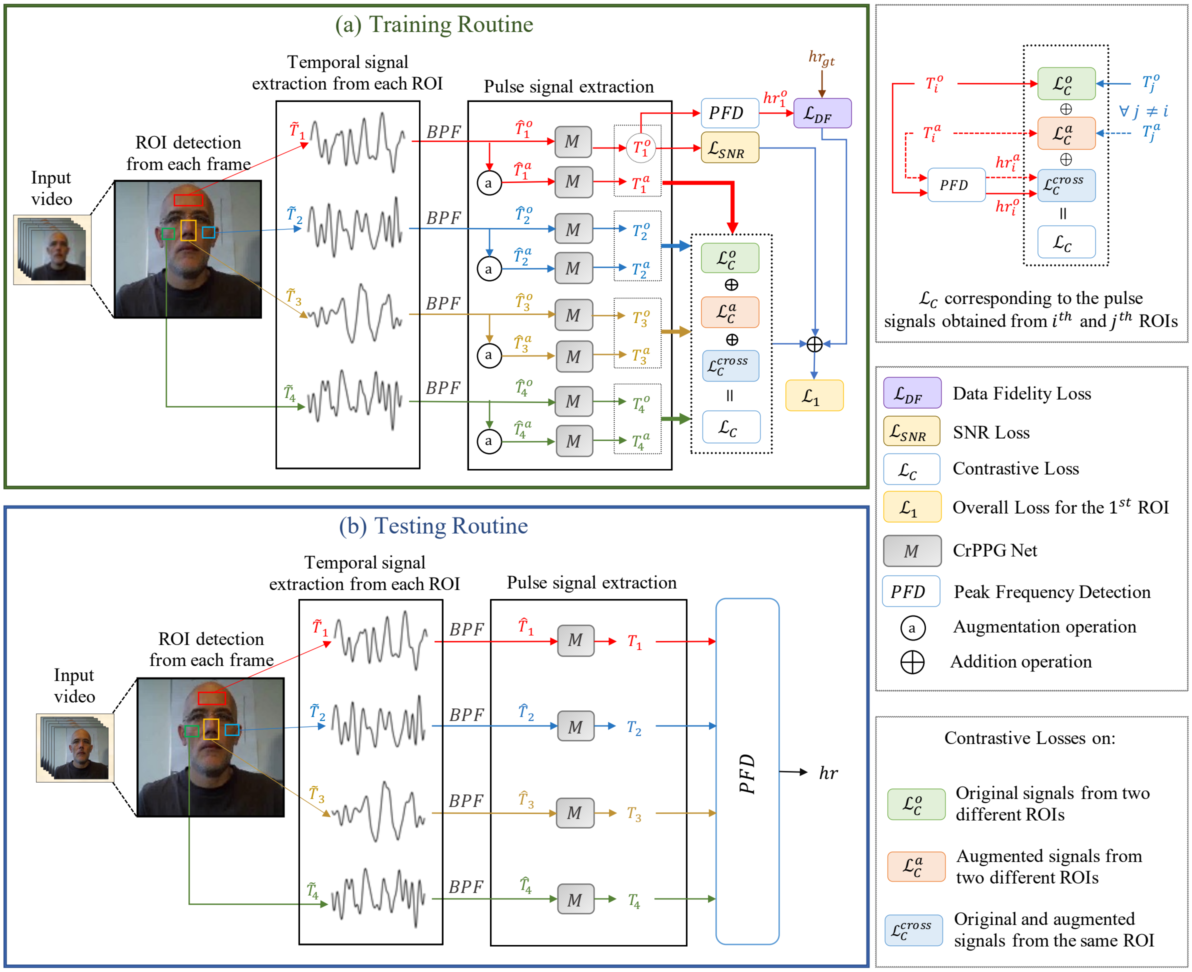
We obtain the temporal signals from each ROI as the average of the intensity values of the green channel across all ROI frames, as described in [3]. Formally, the temporal signal obtained from $i^{th}$ ROI, $\tilde{T}_i$ is given by:

$$\tilde{T}_i = \left( \frac{\sum g_{i,1}}{P_{i,1}}, \frac{\sum g_{i,2}}{P_{i,2}}, \ldots, \frac{\sum g_{i,f}}{P_{i,f}} \right) \quad (1)$$

where $g_{i,k}$ denotes the sum of green intensity values of the pixels belonging to the $i^{th}$ facial ROI of the $k^{th}$ frame, $P_{i,k}$ denotes the total number of pixels in the $i^{th}$ facial ROI of the $k^{th}$ frame and $f$ denotes the total number of frames in the given input video. The temporal signals thus obtained consist of irrelevant signal components, such as noise from facial movements, illumination variations, and other physiological sources [25]. To suppress such spurious components, we pass the obtained signals through a fourth-order Butterworth bandpass filter [45]. The bandpass filter removes any signal components corresponding to frequencies outside the HR range (0.7 Hz to 4.0 Hz) [3]. Further, a detrending filter is applied to remove the non-stationary trends in the temporal signals induced by illumination variations [42]. The filtered temporal signal obtained after the filtering operations is denoted by $\hat{T}_i$.

## 3.3. Data augmentation

There is an uneven distribution in the rPPG datasets where most samples contain HR ranging between 60-90 BPM, whereas the normal human HR ranges between 40–240 BPM. Due to skewed data distribution, the trained network becomes biased to the samples having HR between 60 and 90 BPM. To tackle this issue, temporal scaling-based data augmentation is described in [35], where the videos are scaled up and down to generate samples with HR ranging outside the 60-90 BPM range. Since augmenting the data by interpolating video frames is computationally expensive, we employ temporal signal interpolation. In essence, to obtain a sample with HR $k$ times HR of the original video, the



**Note:**

1. $\hat{T}$, $T$, $hr$ and $BPF$ denote Filtered Signal, Denoised Signal, Heart Rate and Bandpass Filter respectively.
2. The superscript denotes the variable corresponding to either original $(o)$ or augmented $(a)$ Temporal signal and the subscript denotes the ROI index.

Figure 1. The flow diagram of our proposed method *ALPINE*. It shows the (a) Training routine and (b) Testing routine. Kindly note that the loss depicted in the Training routine shows the loss for a particular face ROI. The complete loss is obtained by consolidating the losses for each of the individual ROIs. In the Testing routine, peak frequency detection is employed for obtaining the heart rate from the estimated pulse signal.

temporal signal is scaled by $1/k$ times, using linear interpolation. For instance, to generate temporal signals having the HR half and double of the original HR, $k$ is set to 2 and 1/2, respectively. New samples may have HR outside of 40-240 BPM due to upscaling and downscaling of the temporal signals. We exclude such samples to train the network. Kindly note that the original and augmented temporal signals are denoted by $\hat{T}^o_i$ and $\hat{T}^a_i$, respectively. Furthermore, we employ the augmented signals for contrastive loss-based training (refer section 3.6.1).

### 3.4. Contrastive rPPG Network (CrPPG-Net)

Unfortunately, the temporal signals obtained above provide incorrect HR estimation under the influence of noise [3]. Thus, suitable denoising methods are required for correct HR estimation. To this end, we propose a Contrastive rPPG network (CrPPG-Net). The network is based on TCN [1] for learning the correlations among the temporal signals that correspond to the pulse signal. The TCN network effectively learns the long temporal dependencies among the temporal signals for correct pulse signal estimation [20]. The network is trained using a hybrid loss function that comprises contrastive loss, SNR loss, and data fidelity loss functions that force the network to estimate a denoised temporal signal. Finally, the Multi-Kurtosis optimization technique is used to compute the pulse signal from the denoised temporal signals [26]. Kindly note that when we pass the temporal signal $\hat{T}^o_i$ from the CrPPG-Net, the obtained de-

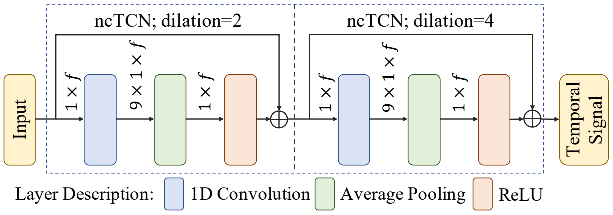


Figure 2. The depiction of our proposed CrPPG-Net.

noised signal is denoted by $T_i^o$.

The TCN-based rPPG network is depicted in Figure 2. It consists of two levels of TCN network stacked sequentially; the output from the TCN network at the first level is fed to the second level TCN network. The first TCN block takes the input signal of dimensions $1 \times f$ and passes it to the 1D convolutional layer with 9 convolutional filters of size $1 \times 9$. The temporal direction of convolution operation utilizes the same padding, and no padding is used in the other direction. The output from the 1D convolutional layer is of size $9 \times 1 \times f$, which is then fed to an average pooling layer that transforms the features into the vector of the same size as the input. Then we introduce non-linearity by using Rectified Linear Unit (ReLU) as the activation layer. A residual connection is provided to the output from the TCN block for performance improvement [15]. The resulting output size $1 \times f$ is passed to the second block. Kindly note that the TCN block at the second level has the same configuration as the first, except for dilations used for the convolutions. A dilation of 2 is used for the first level TCN and 4 for the second level TCN. Please note that TCN can be of two types, casual and non-casual. The casual TCN is used when future time steps are not taken into consideration during the convolution operation, whereas the non-casual TCN utilizes the future information. Since the temporal signals are available for the entire time duration, we have used the non-casual TCN [3]. Further, we also add the skip connection between both the network blocks to improve their performance [15]. In the section 4.2, details of other parameters are discussed.

## 3.5. Estimating Pulse signal and HR

During the time of inference, we consolidate the denoised temporal signals obtained from CrPPG-Net to compute the pulse signal. To this end, we use the Multi-Kurtosis Optimization [25] to compute the pulse. It utilizes periodicity and BSS simultaneously to compute the pulse. Subsequently, we obtain the pulse spectrum by applying the Fast Fourier transform (FFT) to the pulse signal. The maximum amplitude in the pulse spectrum corresponds to the HR frequency. Formally, the HR is given by:

$$\begin{aligned} &hr = freq_{max} \times 60 \\ &\text{where } freq_{max} = \underset{freq}{\text{argmax}}\, PS\left(freq\right) \end{aligned} \tag{2}$$

where $freq_{max}$ is the frequency of pulse spectrum containing maximum amplitude, and $PS\left(k\right)$ denotes the pulse spectrum's amplitude at $k^{th}$ frequency. $PS\left(k\right)$ denotes the pulse spectrum's amplitude at $k^{th}$ frequency. HR is computed by multiplying $freq_{max}$ by 60.

## 3.6. Loss Function

We design a novel hybrid loss function to train the network. The hybrid loss consists of three components: contrastive loss, SNR loss, and data fidelity loss. The contrastive loss, $\mathcal{L}_C(\cdot)$ aims to maximize the normalized correlation between the denoised temporal signals from different facial ROIs. The SNR loss, $\mathcal{L}_{SNR}(\cdot)$, minimizes facial deformation-based noises, thereby improving the quality of the denoised temporal signal. Although these two losses ensure highly correlated and high-quality denoised temporal signals, they do not ensure that the signals correspond to the ground truth. Hence, we use data fidelity loss, $\mathcal{L}_{DF}(\cdot)$, to minimize the error between the predicted temporal signals and ground truth pulse signals. Thus, the proposed hybrid loss enables the network to predict highly correlated signals from different ROIs that are robust to facial deformation-based noises. We discuss each loss component and its contribution in the following sections.

### 3.6.1 Contrastive Loss Function

The contrastive loss $\mathcal{L}_C(\cdot)$ aims at maximizing the similarity among the denoised temporal signals. In this direction, we use three novel contrastive loss functions. The first loss function, $\mathcal{L}_C^o(\cdot)$, ensures a high correlation between the denoised temporal signal obtained from the $i^{th}$ ROI and the denoised temporal signals obtained from other ROIs.

$$\begin{aligned} \mathcal{L}_C^o\left(T_i^o, T_j^o\right) &= 1 - \mathcal{F}_{corr}\left(T_i^o, T_j^o\right) \\ &= 1 - \frac{\sum(T_i^o \times T_j^o)}{\sqrt{\sum(T_i^o)^2 \times \sum(T_j^o)^2}} \end{aligned} \tag{3}$$

where, $T_i^o$ and $T_j^o$, denote the denoised temporal signal obtained from the $i^{th}$ and $j^{th}$ ROIs, respectively; and $\mathcal{F}_{corr}\left(T_i^o, T_j^o\right)$ is the normalised cross correlation between the two signals. Furthermore, for enhancing the generalization of our network over the entire normal HR ranges, we compare augmented temporal signals from the $i^{th}$ ROI with the augmented temporal signals from the other ROIs. Formally, our second contrastive loss, $\mathcal{L}_C^a(\cdot)$, is denoted as:

$$\begin{aligned} \mathcal{L}_C^a\left(T_i^a, T_j^a\right) =& 1 - \mathcal{F}_{corr}\left(T_i^a, T_j^a\right) \\ =& 1 - \frac{\sum(T_i^a \times T_j^a)}{\sqrt{\sum(T_i^a)^2 \times \sum(T_j^a)^2}} \end{aligned} \tag{4}$$

where, $T_i^a$ and $T_j^a$, are obtained by denoising the augmented temporal signals from the $i^{th}$ and $j^{th}$ ROIs, respectively; and $\mathcal{F}_{corr}(\cdot)$ is the normalised cross correlation operation.

Finally, to force the network to learn only those correlations that contribute to the pulse signals, we use a third loss term that minimizes the $\ell_2$ distance between the HR obtained from the augmented temporal signal $hr_i^a$ and its supposed HR value. Using the equation 2, we can find the HR value from the augmented temporal signal in terms of the HR from the original temporal signal, which is given as:

$$\mathcal{L}_C^{cross}(hr_i^o, hr_i^a) = \|k \cdot hr_i^o - hr_i^a\|_2^2 \qquad (5)$$

where, $k$ is the scaling factor used to obtain the augmented temporal signal $T_i^a$.

Our complete contrastive loss $\mathcal{L}_C(\cdot)$, is obtained by taking the sum of its three components, that is

$$\begin{aligned}&\mathcal{L}_C\big(T_i^o, T_j^o, T_i^a, T_j^a\big)\\ &= \mathcal{L}_C^o\big(T_i^o, T_j^o\big) + \mathcal{L}_C^m\big(T_i^a, T_j^a\big) + \mathcal{L}_C^{cross}(hr_i^o, hr_i^a) \quad (6)\end{aligned}$$

### 3.6.2 SNR Loss Function

For a high-quality denoised temporal signal, the frequency corresponding to the heart rate should be considerably higher than the noise signals. To this end, we use the Signal-to-Noise Ratio (SNR) loss $\mathcal{L}_{SNR}(\cdot)$, in order to maximize the SNR of the denoised temporal signal obtained from the proposed CrPPG-Net. Mathematically, the SNR loss for a denoised temporal signal $T_o^i$ can be denoted as:

$$\mathcal{L}_{SNR}(T_i^o) = \frac{1}{SNR(T_i^o)} \qquad (7)$$

### 3.6.3 Data Fidelity Loss Function

The above loss functions ensure that correlated and high-quality temporal signals are obtained from the CrPPG-Net. However, they do not ensure that the obtained signals correspond to the ground truth signal. Hence, we use the data fidelity loss function $\mathcal{L}_{DF}(\cdot)$ to force the network to predict the correct pulse signal. To this end, we use the squared $l_2$ distance between the predicted denoised temporal signal $T_o^i$ and the ground truth signal $T_{gt}$. Mathematically, the data fidelity loss is given as:

$$\mathcal{L}_{DF}\left(h_i^o, h_{gt}\right) = \|h_i^o - h_{gt}\|_2^2 \qquad (8)$$

### 3.6.4 Combined Loss Function

The overall loss for the $i^{th}$ ROI of a given video is computed as the sum of: (i) summation of the contrastive loss $\mathcal{L}_C$, computed over the temporal signal obtained from the $i^{th}$ ROI and the temporal signals obtained from all other ROIs $j$, such that $i \neq j$, (ii) the SNR loss $\mathcal{L}_{SNR}$ and (iii) the data fidelity loss $\mathcal{L}_{DF}$. Mathematically, we express the loss for the $i^{th}$ ROI, $\mathcal{L}_i$, as:

$$\begin{aligned}\mathcal{L}_i = \sum_{j \neq i} \mathcal{L}_C\left(T_i^o, T_j^o, T_i^a, T_j^a\right) + \mathcal{L}_{SNR}\left(T_i^o\right)\\ + \mathcal{L}_{DF}\left(h_i^o, h_{gt}\right) \qquad (9)\end{aligned}$$

Finally, the overall hybrid loss for the video is given by:

$$\mathcal{L} = \sum_{i \in \{1,2,\ldots,n\}} \mathcal{L}_i \qquad (10)$$

where $n$ is the total number of extracted ROIs.

# 4. Experimental results

## 4.1. Dataset and Metrics

We have evaluated our proposed method *ALPINE* on the extensively utilized publicly available UBFC-rPPG [6] and COHFACE [27] datasets. The UBFC-rPPG dataset comprises 42 face videos with corresponding ground truth physiological signals recorded from 42 subjects. For comparative analysis, we used the 67% and 33% of the datasets as training, and testing subsets, respectively, similar to [6]. The videos are 2 minutes long with a frame rate of 30 fps. The COHFACE dataset comprises 160 face videos with corresponding ground truth physiological signals recorded from 40 subjects. The duration of videos is 1 minute long with a frame rate of 20 fps. For comparative analysis, we used 60% and 40% of the dataset as training and testing subsets, respectively, similar to the testing protocol as used by [27]. The ground truth of both datasets was acquired using the pulse oximeter during the video recording. Similar to [3], we evaluate our results using mean absolute error (MAE), standard deviation (SD), Pearson's correlation coefficient (r), and root mean squared error (RMSE) between the ground truth HR and estimated HR.

## 4.2. Implementation Details

All experiments were performed on the Intel Xeon Gold (6132) processor. It consists of 192 GB RAM and an NVIDIA V100 GPU server. The CrPPG-Net was trained using the Adam optimizer, with a batch size of 4 and a learning rate of 0.0001, with a maximum number of epochs to 100. We extract 4-second non-overlapping video clips from a given input video to perform the experiments.

## 4.3. Comparative Evaluation

In this subsection, we present a comparative analysis of the various state-of-the-art methods and the proposed method *ALPINE*. For a fair comparative analysis, we have used similar testing protocols as used by [40, 27]. We have presented the comparative analysis in Table 1. The table

Table 1. Performance evaluation of our proposed method and state-of-the-art methods.

| **Methods** | **UBFC-rPPG** | | | | **COHFACE** | | | |
|---|---|---|---|---|---|---|---|---|
| | **SD*** | **MAE*** | **RMSE*** | **r** | **SD*** | **MAE*** | **RMSE*** | **r** |
| Chrominance-rPPG [13] | 5.50 | 4.70 | 6.61 | 0.67 | 10.63 | 7.80 | 12.45 | 0.26 |
| AHRE [22] | 4.95 | 4.20 | 5.78 | 0.61 | 6.38 | 5.72 | 11.52 | 0.31 |
| Fusion-EL [23] | 4.20 | 3.71 | 4.52 | 0.73 | 8.09 | 7.14 | 9.43 | 0.57 |
| RAHR [26] | 4.50 | 3.70 | 4.61 | 0.67 | 10.63 | 7.80 | 12.45 | 0.26 |
| MOMBAT [25] | 3.38 | 3.50 | 4.01 | 0.85 | 6.14 | 5.89 | 7.92 | 0.62 |
| Physnet [46] | 3.85 | 3.63 | 5.29 | 0.94 | 7.9 | 8.59 | 11.60 | 0.36 |
| META-rPPG [31] | 4.50 | 3.70 | 4.61 | 0.67 | 10.63 | 7.80 | 12.45 | 0.26 |
| HR-CNN [40] | 4.15 | 3.82 | 4.92 | 0.71 | 9.23 | 8.10 | 10.78 | 0.29 |
| AND-rPPG [3] | 3.21 | 2.67 | 4.07 | 0.96 | 4.53 | 3.82 | 5.10 | 0.79 |
| CL-rPPG [18] | 4.20 | 4.82 | 3.9 | 0.94 | 4.83 | 4.52 | 5.90 | 0.87 |
| ***ALPINE*** | **3.17** | **2.58** | **4.01** | **0.96** | **4.46** | **3.65** | **5.07** | **0.82** |

* values are in BPM.

demonstrates that the proposed method, *ALPINE* outperforms the existing methods. The methods Chrominance-rPPG [13], and RAHR [26], fail to mitigate the facial expression-based noises present in the temporal signals and hence perform poorly compared to the proposed method. Likewise, the efficacy of methods AHRE [22], Fusion-EL [23], and MOMBAT [25] is also lower than the proposed method because these methods utilize the blind source separation (BSS) techniques directly on the extracted temporal signals. In contrast, *ALPINE* computes denoised temporal signal from a trained network before applying BSS.

*ALPINE* also outperform the deep learning-based methods such as PhysNet [46], HR-CNN [40], and AND-rPPG [3], as these methods fail to rectify erroneous HR. Similarly, the performance of META-rPPG [31] as it uses the LSTM network, which fails to preserve long-term information [1, 20]. In contrast, our proposed method is based on TCN to effectively models and preserves long-term information. Furthermore, our proposed method outperforms CL-rPPG [18] because CL-rPPG provides importance to certain facial regions to compute the pulse signal. Moreover, it utilizes augmented signal from a facial ROI for learning the correlations corresponding to the pulse signal. Such a technique fails when the ROI is affected by local noise due to facial deformations, resulting in the inability of the network to differentiate between pulse signal and noise characteristics. The proposed method effectively mitigates such facial deformations by using multiple facial regions to compute the temporal signals.

### 4.4. Ablation study

This subsection thoroughly analyzes the importance of multiple ROIs, CrPPG-Net, color channels, and loss functions in the proposed method. We modify or change a subpart of the proposed method to perform these experiments and report the results of the ablation study in Table 2. To understand the importance of multiple ROIs, we conducted the experiments *Two-ROIs* and *Three-ROIs*. In *Two-ROIs*, left and right cheek are considered ROIs. Likewise, the left and right cheeks, along with the nose region, are considered ROIs in *Three-ROIs*. Table 2 demonstrates that the proposed method outperforms both these experiments. The reason is that if a large number of ROI are appropriately selected, the facial deformation reduces, thereby increasing performance. Kindly note that if the selected facial region contains facial deformation, the temporal signal will be spurious, leading to incorrect HR predictions. Increasing the number of facial regions and selecting the regions appropriately can mitigate this issue. Hence, we have excluded the facial regions in the experimentation containing frequent motion variations or skin hidden behind the hairs, like the eyes region, lips region, and chin region [3]. The eyes and lip regions are easily affected by deformations, and chin region skin is not visible in subjects with a beard. Additionally, we performed the experiment *RGB* by replacing green color in *ALPINE* with the RGB colors. The results show that our *ALPINE* outperform *RGB* because the green

Table 2. Ablation study of our proposed method. All the values are in BPM.

| | **UBFC-rPPG** | | | **COHFACE** | | |
|---|---|---|---|---|---|---|
| | **SD*** | **MAE*** | **RMSE*** | **SD*** | **MAE*** | **RMSE*** |
| ***ALPINE*** | **3.17** | **2.58** | **4.01** | **4.46** | **3.65** | **5.07** |
| *CO* | 4.64 | 4.05 | 5.86 | 6.21 | 5.29 | 6.92 |
| *CA* | 5.05 | 4.29 | 6.12 | 7.18 | 5.45 | 7.11 |
| *CL* | 4.50 | 3.92 | 5.53 | 6.05 | 5.09 | 6.77 |
| *SNR* | 4.75 | 4.17 | 5.94 | 6.40 | 6.92 | 7.13 |
| *DF* | 5.08 | 4.32 | 5.97 | 7.25 | 7.19 | 7.27 |
| *SNR + CL* | 3.26 | 2.64 | 4.23 | 4.55 | 3.79 | 5.17 |
| *SNR + DF* | 3.40 | 2.82 | 4.49 | 4.47 | 3.82 | 5.37 |
| *DF + CL* | 3.45 | 2.90 | 4.53 | 4.51 | 3.94 | 5.45 |
| *RGB* | 9.94 | 9.20 | 12.01 | 12.28 | 11.68 | 13.72 |
| *Two-ROIs* | 8.34 | 8.00 | 10.11 | 10.92 | 9.23 | 9.73 |
| *Three-ROIs* | 5.29 | 4.08 | 6.32 | 7.12 | 6.15 | 7.68 |
| *LSTM-exp* | 8.32 | 7.72 | 9.38 | 11.05 | 10.05 | 11.29 |

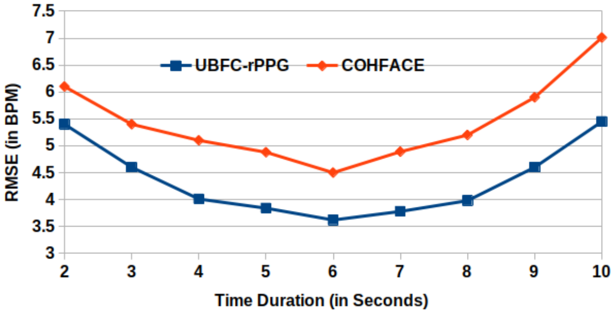


Figure 3. Performance of *ALPINE* for different time duration clips.

color channel provides the strongest PPG signal [25, 43].

To understand the importance of each component of the proposed hybrid loss function, we perform different experiments by modifying the hybrid loss of the proposed method. To analyze the importance of each loss, we perform the experiments *CO*, *CA*, *SNR*, *DF*, and *CL* by using $\mathcal{L}_C^o$ loss, $\mathcal{L}_C^a$ loss, $\mathcal{L}_{SNR}$ loss, $\mathcal{L}_{DF}$ loss, and $\mathcal{L}_C$ loss respectively. The results are reported in Table 2. Kindly note that out of all these settings, *CL* provides us with the best results. This behaviour can be attributed to the fact that $\mathcal{L}_C$ loss forces the network to increase the similarity of both the original and augmented signals. In contrast, the $\mathcal{L}_C^o$ and $\mathcal{L}_C^a$ work solely on the original and augmented signals, respectively. Whereas $\mathcal{L}_{SNR}$ and $\mathcal{L}_{DF}$ loss work only on the original signals and do not take the augmented signals into consideration. Furthermore, to better understand the behaviour of the losses, we performed the experiments *SNR+CL*, *SNR+DF*, and *DF+CL* by taking combinations of two losses at the same time. The experiment *SNR+CL* is formed by utilizing $\mathcal{L}_{SNR}$ and $\mathcal{L}_C$ losses. Likewise, *SNR+DF* utilizes the $\mathcal{L}_{SNR}$ and $\mathcal{L}_{DF}$ losses, and *DF+CL* utilizes the $\mathcal{L}_{DF}$ and $\mathcal{L}_C$ losses. The table indicates that when we used contrastive and MSNR loss simultaneously, the performance is improved because both the losses mitigate the facial deformation-based noises from the temporal signals by learning and mitigating the similarity features and noises from the temporal signals. It is observed from the table that these experiments do not perform better than the proposed method because our proposed hybrid loss learns the relationship between the temporal signals of different regions, minimizes the noise and forces the network to predict signals similar to the ground truth. Furthermore, the experiment *LSTM-exp* is performed by replacing the CrPPG-Net in the proposed method with LSTM network. It indicates that TCN can model long sequences of temporal information better than the sequential networks [1].

In the proposed method, the clip size is set to 4 seconds. For rigorous analysis, we evaluate the performance of the proposed method for different time duration clips, and the corresponding results are shown in Figure 3. The figure depicts that if we decrease the clip duration, the proposed method's performance reduces due to the reduction of rPPG information in short-duration clips. In contrast, if we increase the clip duration from 4 to 6 seconds, performance increases because long-duration clips provide more rPPG information. However, if clip duration increases by more than 6 seconds, performance decreases because it reduces the number of clips used to train the network. Nevertheless, we have selected the 4-second duration clips for our experimentation because it is a widely-accepted practice [39, 29].

Our novel hybrid contrastive loss $\mathcal{L}_C(\cdot)$ automatically learns the similarity between temporal signals extracted from different facial regions for denoising the temporal signals. It improves the performance and enables the proposed method to outperform state-of-the-art methods, but unfortunately, it is incompetent in several cases. While investigating the failure cases, we observe that our method provides the incorrect HR when the input clip contains significant deformations in all ROIs. In such cases, our method also learns noises along with pulse signals.

## 5. Conclusion

Conventional rPPG methods cannot handle facial deformation-based noises, and deep learning-based rPPG methods require large-scale labelled data to train the network. The proposed method, *ALPINE*, has mitigated these limitations by utilizing CL. The CL has equipped our novel network to learn the similarity between the temporal signals of multiple ROIs without employing the labelled data. Such learning has facilitated the denoising of temporal signals and introduced diversity in the data samples for better network generalization. Our network has been trained by combining multiple loss functions, which used the contrastive loss to compute the similarity between the temporal signals, SNR loss and data fidelity loss. The experiments conducted on the publically available UBFC-rPPG and COHFACE datasets, revealed that the proposed method outperforms the state-of-the-art methods. Further, it has demonstrated that our novel network can mitigate the issue of facial deformations, and the network could be trained using small-scale labelled data. Also, it indicates that the performance can be improved when SNR and data fidelity losses are combined with our proposed contrastive loss. In future work, we will explore the possibility of utilizing transformer-based networks in unsupervised settings for rPPG computation.

**Acknowledgement:** This research is supported partially by SERB, DST and project number is SRG/2020/001383. The work of Anup Kumar Gupta is partially supported by Prime Minister's Research Fellowship (PMRF), the Ministry of Education, Government of India (2101306).

## References

[1] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[2] Guha Balakrishnan, Fredo Durand, and John Guttag. Detecting pulse from head motions in video. In *Computer Vision and Pattern Recognition*, pages 3430–3437, 2013.

[3] Lokendra Birla and Puneet Gupta. AND-rPPG: A novel denoising-rPPG network for improving remote heart rate estimation. *Computers in Biology and Medicine*, page 105146, 2021.

[4] Lokendra Birla and Puneet Gupta. PATRON: Exploring respiratory signal derived from non-contact face videos for face anti-spoofing. *Expert Systems with Applications*, page 115883, 2021.

[5] Lokendra Birla, Puneet Gupta, and Shravan Kumar. SUNRISE: Improving 3D mask Face Anti-spoofing for Short Videos using Pre-emptive Split and Merge. *IEEE Transactions on Dependable and Secure Computing*, 2022.

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

[7] Sylvain Boltz, Eric Debreuve, and Michel Barlaud. High-dimensional statistical measure for region-of-interest tracking. *IEEE Transactions on Image Processing*, 18(6):1266–1283, 2009.

[8] Frédéric Bousefsaf, Alain Pruski, and Choubeila Maaoui. 3D convolutional neural networks for remote pulse rate measurement and mapping from facial video. *Applied Sciences*, 9(20):4364, 2019.

[9] Christoph Bruser, Christoph Hoog Antink, Tobias Wartzek, Marian Walter, and Steffen Leonhardt. Ambient and unobtrusive cardiorespiratory monitoring techniques. *IEEE Reviews in Biomedical Engineering*, 8:30–43, 2015.

[10] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *European Conference on Computer Vision*, pages 349–365, 2018.

[11] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. FakeCatcher: Detection of synthetic portrait videos using biological signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020.

[12] Bo Dai and Dahua Lin. Contrastive learning for image captioning. *Advances in Neural Information Processing Systems*, 30, 2017.

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

[14] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *International Conference on Computer Vision*, pages 1422–1430, 2015.

[15] Michal Drozdzal, Eugene Vorontsov, Gabriel Chartrand, Samuel Kadoury, and Chris Pal. The importance of skip connections in biomedical image segmentation. In *Deep Learning and Data Labeling for Medical Applications*, pages 179–187. Springer, 2016.

[16] Haoyuan Gao, Xiaopei Wu, Jidong Geng, and Yang Lv. Remote heart rate estimation by signal quality attention network. In *Computer Vision and Pattern Recognition*, pages 2122–2129, 2022.

[17] John Gideon and Simon Stent. Estimating heart rate from unlabelled video. In *International Conference on Computer Vision*, pages 2743–2749, 2021.

[18] John Gideon and Simon Stent. The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video. In *International Conference on Computer Vision*, pages 3995–4004, 2021.

[19] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[20] Anup Kumar Gupta, Puneet Gupta, and Esa Rahtu. FATALRead-Fooling visual speech recognition models. *Applied Intelligence*, 52(8):9001–9016, 2022.

[21] Puneet Gupta. MERASTC: Micro-expression recognition using effective feature encodings and 2D convolutional neural network. *IEEE Transactions on Affective Computing*, 2021.

[22] Puneet Gupta, Brojeshwar Bhowmick, and Arpan Pal. Accurate heart-rate estimation from face videos using quality-based fusion. In *International Conference on Image Processing*, pages 4132–4136, 2017.

[23] Puneet Gupta, Brojeshwar Bhowmick, and Arpan Pal. Serial fusion of Eulerian and Lagrangian approaches for accurate heart-rate estimation using face videos. In *Engineering in Medicine and Biology Society*, pages 2834–2837, 2017.

[24] Puneet Gupta, Brojeshwar Bhowmick, and Arpan Pal. Exploring the feasibility of face video based instantaneous heart-rate for micro-expression spotting. In *Computer Vision and Pattern Recognition Workshops*, pages 1316–1323, 2018.

[25] Puneet Gupta, Brojeshwar Bhowmick, and Arpan Pal. MOMBAT: heart rate monitoring from face video using pulse modeling and Bayesian tracking. *Computers in Biology and Medicine*, 121:103813, 2020.

[26] Puneet Gupta, Brojeshwar Bhowmik, and Arpan Pal. Robust adaptive heart-rate monitoring using face videos. In *Winter Conference on Applications of Computer Vision*, pages 530–538, 2018.

[27] Guillaume Heusch, Andre Anjos, and Sebastien Marcel. A reproducible study on remote heart rate measurement. *arXiv preprint arXiv:1709.00962*, 2017.

[28] Gee-Sern Hsu, ArulMurugan Ambikapathi, and Ming-Shiang Chen. Deep learning with time-frequency representation for pulse estimation from facial videos. In *International Joint Conference on Biometrics*, pages 383–389, 2017.

[29] Min Hu, Fei Qian, Dong Guo, Xiaohua Wang, Lei He, and Fuji Ren. ETA-rPPGNet: Effective time-domain attention network for remote heart rate measurement. *IEEE Transactions on Instrumentation and Measurement*, 70:1–12, 2021.

[30] Marek Kowalski, Jacek Naruniec, and Tomasz Trzcinski. Deep alignment network: A convolutional neural network

for robust face alignment. In *Computer Vision and Pattern Recognition Workshops*, pages 88–97, 2017.

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

[32] Si-Qi Liu and Pong C Yuen. A general remote photoplethysmography estimator with spatiotemporal convolutional network. In *Automatic Face and Gesture Recognition*, pages 481–488, 2020.

[33] Guillaume Lorre, Jaonary Rabarisoa, Astrid Orcesi, Samia Ainouz, and Stephane Canu. Temporal contrastive pretraining for video action recognition. In *Winter conference on Applications of Computer Vision*, pages 662–670, 2020.

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

[35] Xuesong Niu, Xingyuan Zhao, Hu Han, Abhijit Das, Antitza Dantcheva, Shiguang Shan, and Xilin Chen. Robust remote heart rate estimation from face utilizing Spatial-temporal attention. In *Automatic Face and Gesture Recognition*, pages 1–8. IEEE, 2019.

[36] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pages 319–345. Springer, 2020.

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

[38] Jaromir Przybyło. A deep learning approach for remote heart rate estimation. *Biomedical Signal Processing and Control*, 74:103457, 2022.

[39] Ying Qiu, Yang Liu, Juan Arteaga-Falconi, Haiwei Dong, and Abdulmotaleb El Saddik. EVM-CNN: Real-time contactless heart rate estimation from facial video. *IEEE Transactions on Multimedia*, 21:1778–1787, 2018.

[40] Radim Spetlik, Vojtech Franc, and Jiri Matas. Visual heart rate estimation with convolutional neural network. In *British Machine Vision Conference*, pages 3–6, 2018.

[41] Jan Spooren, Davy Preuveneers, and Wouter Joosen. $PPG^2$ live: Using dual PPG for active authentication and liveness detection. In *International Conference on Biometrics*, pages 1–6, 2019.

[42] Mika P Tarvainen, Perttu O Ranta-Aho, and Pasi A Karjalainen. An advanced detrending method with application to HRV analysis. *IEEE Transactions on Biomedical Engineering*, 49:172–175, 2002.

[43] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics Express*, 16:21434–21445, 2008.

[44] Wenjin Wang, Sander Stuijk, and Gerard De Haan. A novel algorithm for remote photoplethysmography: Spatial subspace rotation. *IEEE Transactions on Biomedical Engineering*, 63:1974–1984, 2015.

[45] M Yang, J Liu, Y Xiao, and H Liao. 14.4 nW fourth-order bandpass filter for biomedical applications. *Electronics Letters*, 46:973–974, 2010.

[46] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In *British Machine Vision Conference*, page 277, 2019.

[47] Zitong Yu, Xiaobai Li, and Guoying Zhao. Facial-video-based physiological signal measurement: Recent advances and affective applications. *IEEE Signal Processing Magazine*, 38(6):50–58, 2021.

[48] Amir Zadeh, Yao Chong Lim, Tadas Baltrusaitis, and Louis-Philippe Morency. Convolutional Experts Constrained Local Model for 3D Facial Landmark Detection. In *International Conference on Computer Vision Workshops*, pages 2519–2528, 2017.