TRUST: Time-domain Residual Unsupervised Stability Technique for Improved Heart Rate Estimation

Shahzad Ahmad[1] shahzaa@hiof.no
Sania Bano[2] sania.22eez0012@iitrpr.ac.in
Sukalpa Chanda[1] sukalpa@ieee.org
Santosh Kumar Vipparthi[2] skvipparthi@iitrpr.ac.in
Subrahmanyam Murala[3] muralas@tcd.ie
[1] Østfold University College, Norway [2] Indian Institute of Technology Ropar, India [3] Trinity College Dublin, Ireland

Abstract

Camera-based estimation of vital signs is a promising method for non-contact health monitoring, which analyzes minute changes in video data. However, the creation of accurate models for this task is challenging due to the scarcity of datasets that possess synchronized vital sign recordings. Our research enhances an existing non-contrastive unsupervised learning technique for extracting rPPG signals, which does not necessitate ground-truth signals during the training process. We have incorporated new time-domain loss functions and added a feature stabilization block to improve the model's stability and accuracy in detecting low-level features. Additionally, we have devised a metric to evaluate the feature instability in the model's final layer. Our experiments on four public datasets demonstrate that our method surpasses the performance of current state-of-the-art methods. These advancements make our approach a significant breakthrough in the development of scalable deep-learning models for camera-based heart-rate estimation.

1. Introduction

Traditional physiological assessments involve skin-contact sensors, such as photoplethysmography (PPG) and electrocardiography (ECG), often causing discomfort. Physiological parameters, such as heart rate (HR), respiratory frequency (RF), and heart rate variability (HRV), can be extracted from PPG signals and are crucial for healthcare [30, 41] and emotion analysis [17, 29, 45]. However, these methods require specific biomedical equipment such as pulse oximeters, leading to unease and skin irritation. In contrast, remote physiological measurement utilizes a camera to record facial videos, capturing the weak color changes in faces to derive the remote photoplethysmography (rPPG) signal. This method, requiring only off-the-shelf cameras, exhibits substantial promise for applications in remote healthcare [30, 41] and emotion analysis [17, 29, 45].

In previous investigations of rPPG [7, 27, 36, 39], researchers introduced manually crafted features to learn rPPG signals. Subsequently, numerous deep learning-based approaches [5, 6, 13, 14, 16, 23, 23, 25, 33, 44, 44] were proposed, employing supervised methods with diverse network architectures for rPPG signal measurement. Deep learning-based methods have proven to be more robust than traditional handcrafted approaches under specific conditions, such as situations with head movements or diverse video content. However, these methods require extensive video data for diverse scenarios. Furthermore, the simultaneous gathering of video and physiological ground truth, such as contact-PPG or ECG, poses difficulties owing to impractical volumes, difficulties in diverse subject recordings, and technical synchronization challenges. These issues highlight the complexity of obtaining comprehensive datasets for supervised methods, emphasizing the need for alternative approaches, such as unsupervised learning, for more accessible and representative training data. Fortunately, contrastive unsupervised learning proves to be a hopeful solution for addressing the challenge of data scarcity [11, 35, 37, 42].

We explore a non-contrastive unsupervised learning based approach to identify periodic signals in video data. Speth et al. [32] introduced a non-contrastive unsupervised learning framework (SiNC). This novel approach utilizes periodic signal priors for direct physiological signal estimation from unlabeled videos for camera-based physiological measurements. Our investigation reveals that the non-contrastive method yields promising results without relying on the ground truth. However, challenges such as unstable features in low-level spatial resolution significantly affect the quality of the rPPG signal. Therefore, our primary goal is to stabilize features at low spatial resolution levels, thereby enhancing the overall quality of the predicted rPPG signal.

We leverage the SiNC paper, which introduces signal

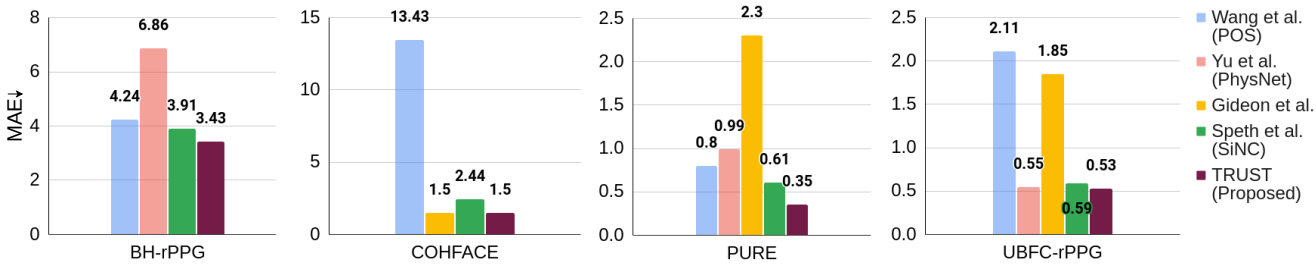

Figure 1. **Effectiveness of TRUST:** TRUST demonstrates superior or competitive performance across four benchmark datasets: BH-rPPG [43], COHFACE [12], PURE [34], and UBFC-rPPG [1]. The evaluation highlights TRUST's robustness and effectiveness in various conditions, consistently outperforming existing methods. The Y-axis represents the Mean Absolute Error (MAE ↓) for heart rate prediction.

filtering in the frequency domain as the loss functions, thereby offering an alternative to the traditional contrastive approaches. However, unstable features at low-level spatial resolutions can affect the predicted signal. In this study, we extend the non-contrastive approach [32] framework by introducing the time-domain loss functions and incorporating dedicated feature stabilization blocks in the architecture. These enhancements stabilize low-level features, ensuring more accurate rPPG signal extraction. To quantitatively measure the effectiveness of our approach, we introduced an entropy-based metric to assess the feature spatial resolution randomness in the low-level feature.

The contributions of this study are as follows.

- We propose and develop novel temporal loss functions that operate in the time domain, significantly enhancing the model's ability to capture and maintain smooth, consistent signal transitions. These temporal loss functions complement the existing frequency-domain approaches and provide a more holistic framework for robust rPPG signal extraction for heart rate estimation.
- We update the model architecture by adding feature stabilization blocks at residual connections to efficiently handle robust and stable feature extraction, ensuring accurate estimation of rPPG signals from unlabelled video data.
- We introduce the feature stabilization metric to calculate feature instability.

As shown in Figure 1, TRUST (proposed) demonstrated superior or competitive performance across the four benchmark datasets.

2. Related Work

2.1. Remote Photoplethysmography (rPPG)

Over the past decade, remote pulse estimation approaches have evolved from blind source separation [27,28] to linear color transformations [8, 28, 38, 39] and eventually to the training of supervised deep learning-based models [5,13,14,16,22,23,31,44,46,48]. Although color transformations demonstrate robust generalization, supervised deep learning models exhibit superior accuracy on similar distribution data. Recent developments in deep learning have focused on optimizing architectures to effectively extract resilient spatial and temporal features, especially when dealing with limited datasets. To address data limitations, synthetic datasets such as SCAMPS [18] and UCLA-synthetic [40] have been introduced, offering diverse skin tone coverage. Additionally, unsupervised learning approaches [11, 35, 37, 42] have been explored as an alternative for scenarios with limited data. In the domain of deep learning-based remote photoplethysmography (rPPG) measurement, studies have employed various architectures, including 2D convolutional neural networks (2DCNN) using two consecutive video frames as input has been employed for rPPG measurement [5, 14, 25, 33], other DL methods utilizing spatial-temporal signal maps [16,23,24] and more recently 3DCNN-based methods [11, 44, 44] designed for effective performance on compressed videos [44].

2.2. Unsupervised Learning for rPPG

Current unsupervised rPPG methodologies adopt a contrastive framework [2, 4, 21] that trains the model on pairs of input videos. The aim is to bring predictions for similar videos closer and push those for dissimilar videos apart. Gideon et al. [11] introduced unsupervised rPPG training using a contrastive approach with frequency resampling for negative samples. Their method computes the mean square error between power spectral densities, but the use of negative samples is deemed imprecise. Known resampled frequencies allow for accurate computation, making the repelling estimated spectra less accurate. Yuzhe et al. [42] introduced an alteration to the InfoNCE loss [26] by incorporating a previously ignored resampling factor. This modification adjusts the intended similarity between pairs according to their relative sampling rates. However, their learning framework is not entirely end-to-end unsupervised, requiring fine-tuning with PPG labels after the self-supervised stage. In contrast to [11], both Contrast-Phys [35] and SLF-RPM [37] treat all samples except the anchor as negatives. This assumes that the power spectra differ between subjects or in long windows for the same person. How-

ever, similar to Gideon's method, there are challenges with negative pairs because different people can have the same pulse rate. Therefore, it is common during training to penalize the model for predicting similar frequencies. Yue et al. [47] introduce an innovative self-supervised and contrastive learning-based framework. Notable contributions include a learnable frequency augmentation module for diverse negative sample generation, a local rPPG expert aggregation module enhancing signal estimation from various facial regions, and a suite of frequency-inspired losses for network optimization. In contrast to the traditional contrastive framework, Speth et al. [32] introduced a non-contrastive unsupervised learning framework (SiNC) that utilizes periodic signal priors for direct physiological signal estimation from unlabeled videos. This innovative approach uses signal filtering in the frequency domain as a loss function, offering a novel alternative to conventional contrastive methods.

3. Proposed Method

Our objective was to analyze remote photoplethysmography (rPPG) signals from facial videos without relying on labeled data. Figure 2 illustrates our approach's framework. The core pipeline involves stabilizing features in the low-level spatial resolution of the input video and subsequently predicting the rPPG signal. The model takes an RGB facial video $V \in \mathbb{R}^{C\times W\times H\times T}$ as input, where T represents the number of frames, $W \times H$ denotes the frame dimensions in pixels, and C is the number of channels, model $\mathcal{M}$ that regress a waveform $Y = \mathcal{M}(V), Y \in \mathbb{R}^T$.

We adopted a non-contrastive learning approach inspired by state-of-the-art methods, which exclusively utilizes the model's estimated waveform. By setting specific expectations for the estimated pulse, such as its periodicity and frequency range, unwanted signals outside the preferred frequency range are treated as noise. The loss function acts as a frequency-domain filter to eliminate unwanted frequency components.

Our contribution extends this approach to incorporate time-domain filtering. We introduce time-domain loss functions that discourage the model from retaining irrelevant signals during processing, ensuring robustness to noisy visual features, and maintaining focus on essential information. The following sections will discuss the types of proposed time-domain losses and the adjustments used in our training process.

3.1. Proposed Losses

Unsupervised learning methodologies for periodic signals offer the advantage of imposing strong constraints on the solution space, which is beneficial for physiological signals such as respiration and blood volume pulses owing to their well-defined frequency bounds. We enhance this approach by incorporating time-domain filtering, which stabilizes low-level feature extraction in the model's later layers, thereby improving the accuracy of the predicted rPPG signal.

Our contribution integrates time-domain loss functions with frequency-domain loss functions ($\mathcal{L}_{freq}$) from the SiNC paper [32], including Bandwidth Loss ($\mathcal{L}_b$), Sparsity Loss ($\mathcal{L}_s$), and Variance Loss ($\mathcal{L}_v$). This unified framework combines the strengths of both domains, effectively filtering unwanted signals and addressing the challenges posed by unstable features in video-based rPPG signals. This approach enhances the robustness of the model and simplifies the identification of relevant features for the target signal.

3.1.1 Temporal Variance Loss

The proposed temporal variance loss ($\mathcal{L}_{variance}$) plays a vital role in signal processing. Functioning as a regularization term, it effectively suppresses abrupt temporal variations in the signal and contributes to signal stability and reliability. Utilizing the mean value as a reference, the loss function penalizes deviations, particularly significant changes, promoting a smoother and more consistent signal amplitude. Mathematically, it is defined as

$$\mathcal{L}_{variance} = \lambda_{\text{variance}} \cdot \frac{1}{T}\sum_{i=1}^{T}(Y_i - \bar{Y})^2 \tag{1}$$

where T represents the length of the signal, Y_i denotes the signal value at each time step, and $\bar{Y}$ is the mean value of the signal over the entire sequence. The hyperparameter $\lambda_{\text{variance}}$ controls the strength of regularization.

3.1.2 Signal Energy Loss

The proposed signal energy loss ($\mathcal{L}_{energy}$) plays a vital role in addressing signal fluctuations. This loss function focuses on identifying and reducing variations in high-energy regions within the signal, which are often caused by factors such as motion. The loss function is mathematically expressed as

$$\mathcal{L}_{energy} = \lambda_{\text{energy}} \cdot \frac{1}{T}\sum_{i=1}^{T} Y_i^2 \tag{2}$$

The loss function focuses on areas with higher energy levels while assisting in the identification and suppression of regions likely affected by fluctuation. This contributes to a more consistent and reliable waveform, thereby enhancing the signal quality.

3.1.3 Derivative Consistency Loss

Our proposed derivative consistency loss function ($\mathcal{L}_{consistency}$) is designed to promote smoothness in the rate of change of a signal. Mathematically expressed as

$$\mathcal{L}_{consistency} = \lambda_{\text{derivative}} \cdot \frac{1}{T}\sum_{i=2}^{T}\left(\frac{dY_i}{dt} - \frac{dY_{i-1}}{dt}\right)^2 \tag{3}$$

where $\frac{dY_i}{dt}$ denotes the derivative of the signal at each time step. The loss function emphasizes the importance of consistent, smooth transitions and removes sharp peaks in the signal. This mechanism helps suppress abrupt changes and fluctuations, contributing to a more stable and reliable signal.

3.2. Combination of Losses

In summary, our training loss functions are the sum of the frequency domain ($\mathcal{L}_{freq}$) [32] and time domain ($\mathcal{L}_{time}$) losses, aiming to improve the model's performance. Frequency domain losses guide the model in capturing sustained frequency characteristics, whereas time domain loss encourages a smooth signal. The combination of these losses in our framework provides a balanced training approach.

$$\mathcal{L}_{time} = \mathcal{L}_{variance} + \mathcal{L}_{energy} + \mathcal{L}_{consistency} \tag{4}$$

$$\mathcal{L}_{freq} = \mathcal{L}_b + \mathcal{L}_s + \mathcal{L}_v \tag{5}$$

where,

$$\mathcal{L}_b = \frac{1}{\sum_{i=-\infty}^{\infty} F_i}\left[\sum_{i=-\infty}^{a} F_i + \sum_{i=b}^{\infty} F_i\right]$$

$$\mathcal{L}_s = \frac{1}{\sum_{i=a}^{b} F_i}\left[\sum_{i=a}^{F^*-\Delta_F} F_i + \sum_{i=F^*+\Delta_F}^{b} F_i\right]$$

$$\mathcal{L}_v = \frac{1}{d}\sum_{i=1}^{d}(CDF_i(Q) - (CDF_i(P))^2$$

$F_i = \text{FFT}(Y)$ be the power in the i-th frequency bin of the predicted signal. Here, $a = 0.66$ Hz and $b = 3$ Hz. F^* represents argmax(F), and Δ_F denotes the frequencies of the spectral peak and the padding around the peak, respectively. CDF refers to the cumulative distribution function.

Our final loss function is defined as

$$\mathcal{L}_{total} = \mathcal{L}_{time} + \mathcal{L}_{freq} \tag{6}$$

3.3. Proposed Model

Our proposed model is inspired by the 3D-CNN structure proposed in [31], emphasizing simplicity and efficiency. This choice is motivated by the need to effectively process 120 consecutive frames in our application, ensuring that the model can handle real-time data while maintaining performance. We excluded the temporal dilations used in [32, 44], focusing on data without considering changes in the temporal dimension. This decision was driven by early experiments, which indicated that temporal dilations could lead to aliasing, potentially limiting the model's bandwidth to specific frequencies. To enhance the stability of low-level feature extraction within the later layers of the model, we incorporated feature stabilization blocks, as illustrated in Figure 2. These blocks are strategically positioned parallel to the B_2, B_3, and B_4 feature extraction blocks and serve as residual connections, ensuring the retention of essential features for improved overall model performance.

3.3.1 Feature Stabilization Block

We introduced feature stabilization blocks (S_i) into our model, as shown in Fig. 2, to enhance the extraction of crucial information in the later layers. These blocks use 3D convolutions of size 1×1 followed by batch normalization, which is designed to stabilize low-level features while retaining high-level information through residual connections.

Starting with an initial video size of 64×64, the model's first feature extraction block B_1 downscales the spatial dimensions to 32×32. Subsequent blocks continue this process until they reach 4×4, whereas the temporal dimension remains consistent. The feature stabilization blocks (S_i), operating parallel to the feature extraction blocks (B_i), stabilize the spatial features at various stages of downsampling.

We demonstrated feature stability at different spatial resolutions using heat maps, as shown in Fig. 3. These maps highlight the stability at low-level resolutions, particularly at 16×16, 8×8, and 4×4, showcasing the method's effectiveness under varying conditions.

3.4. Feature Stabilization Metric

To quantitatively evaluate our approach, we introduced the feature stabilization metric ρ to measure the feature instability in the model's last layer. This metric assesses the performance of the model at low-level feature spatial resolution, providing insight into its interpretability.

The feature stabilization metric ρ is defined as

$$\rho = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{|f|}\sum_{j=1}^{|f|}\sum_{k=1}^{T} E_k\right] \tag{7}$$

where $E_k = -p_k \log_e(p_k)$ and $p_k = \text{softmax}(v_{1\times HW})$.

Equation 7 calculates the spatial entropy across batches (N), features (f), and temporal frames (T). The E_k term captures the entropy of the k-th feature, computed using the softmax probability distribution p_k across the spatial dimensions (height × width).

The calculated values are listed in Table 4.

Understanding Why Our Method Works

Instability in low-level spatial features can lead to noisy, predicted waveforms. TRUST addresses this by stabilizing these features, thereby enhancing the rPPG signal quality while preserving the video's temporal dimension. Our model employs time-domain loss functions that are essential for training without ground truth values and stabilizing features in the time domain, which is a novel approach for non-contrastive methods. Our architecture, inspired by the need for efficient processing of 120 consecutive frames, is simple, yet effective. We incorporate novel elements such as residual connections with feature stabilization blocks, which are crucial for stabilizing features and improving the performance of our method.

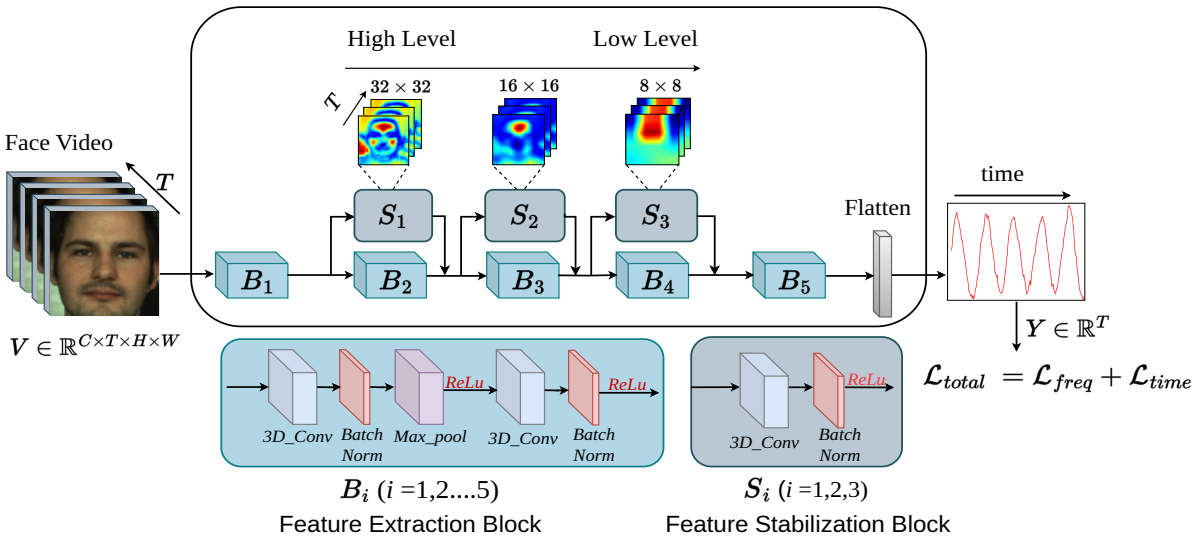

Figure 2. Illustration of the overall TRUST framework. The input video is passed through the model, where each feature extraction block (B_i) performs spatial downsampling while preserving the temporal dimension. Simultaneously, the feature stabilization block (S_i) runs parallel to the feature extraction block to stabilize the spatial features of the downsampled videos. To further enhance stability, the proposed time domain loss functions are employed, effectively securing the waveform in the temporal domain.

Table 1. Intra-dataset Heart Rate (HR) results are presented in the table. An upward arrow (↑) signifies that larger values are better, while a downward arrow (↓) indicates the opposite. The best results are highlighted in bold, and the second-best results are underlined. MAE: Mean Absolute Error; RMSE: Root Mean Square Error; r: Pearson correlation coefficient.

	UBFC-rPPG			**PURE**			**COHFACE**			**BH-rPPG**		
Method	MAE ↓	RMSE↓	r ↑	MAE ↓	RMSE↓	r ↑	MAE ↓	RMSE↓	r ↑	MAE ↓	RMSE↓	r ↑
Traditional Method												
Verkruysse et al. (GREEN) [36]	7.50	14.41	0.62	7.23	17.05	0.69	-	-	-	4.38	8.91	-
Poh et al. (ICA) [27]	5.17	11.76	0.65	3.76	12.60	0.85	-	-	-	4.38	8.91	-
Haan et al. (CHROM) [7]	2.36	9.23	0.87	0.75	2.23	1.00	7.8	12.45	0.26	1.90	4.21	-
Wang et al. (POS) [39]	2.11	9.11	0.87	0.80	4.11	0.98	13.43	17.05	0.24	4.24	14.00	-
Supervised Method												
Špetlık et al. (HR-CNN) (BMVC 18) [33]	-	-	-	1.84	2.37	0.98	10.8	8.1	0.29	-	-	-
Lu et al. (Dual-GAN) (CVPR 21) [16]	**0.44**	**0.67**	**0.99**	0.82	**1.31**	0.99	-	-	-	-	-	-
Speth et al. (RPNet) (21) [31]	0.53 ± 0.01	1.78 ± 0.02	**0.99**	1.15 ± 0.27	5.77 ± 1.25	0.96 ± 0.01	-	-	-	-	-	-
Yu et al. (PhysNet) (19) [44]	0.55 ± 0.03	2.03 ± 0.37	**0.99**	0.99 ± 0.19	5.22 ± 0.93	0.97 ± 0.01	-	-	-	6.86	10.86	-
Deshpande et al. (CVPR 23) [10]	-	-	-	-	-	-	2.92	6.128	0.86	-	-	-
Chen et al (CVPR 23) [3]	-	-	-	-	-	-	-	-	-	**2.042**	**3.142**	**0.959**
Gideon et al. (ICCV 21) [11]	-	-	-	2.1	2.6	0.99	2.5	7.8	0.75	-	-	-
Unsupervised Method												
Gideon et al. (ICCV 21) [11]	1.85	4.28	0.93	2.3	2.9	0.99	**1.5**	4.6	**0.99**	-	-	-
Sun et al. (Contrast-Phys) (ECCV 22) [35]	0.64	1.00	**0.99**	1.00	1.40	0.99	-	-	-	-	-	-
Yue et al. (TPAMI 23) [47]	0.58	0.94	**0.99**	1.23	2.01	0.99	-	-	-	-	-	-
Speth et al. (SiNC)(CVPR 23) [32]	0.59	1.83 ± 0.04	**0.99**	0.61 ± 0.66	1.84 ± 0.40	**1.00**	2.44±0.64	6.02 ± 1.07	0.86 ± 0.05	3.91 ± 0.17	7.32 ± 0.07	0.78 ± 0.02
TRUST (Ours)	0.53 ± 0.03	1.72 ± 0.22	**0.99**	**0.35 ± 0.12**	1.34 ± 0.27	0.99 ± 0.01	**1.50 ± 0.16**	**4.45 ± 0.66**	0.92 ± 0.02	3.43 ± 0.31	6.40 ± 0.56	0.83 ± 0.03

Table 2. Cross-dataset Heart Rate (HR) results are presented in the table. An upward arrow (↑) signifies that larger values are better, while a downward arrow (↓) indicates the opposite.

Training Dataset	Testing Dataset	Method	MAE (bpm) ↓	r ↑
PURE	UBFC	PhysNet [44]	7.02 ± 3.35	0.60 ± 0.13
	UBFC	Contrast-Pys [35]	10.22 ± 0.38	0.45 ± 0.04
	UBFC	SiNC [32]	6.64 ± 1.76	0.59 ± 0.10
	UBFC	**TRUST**	**3.59 ± 1.19**	**0.78 ± 0.07**
	COHFACE	**TRUST**	**15.84 ± 0.73**	**-0.01 ± 0.02**
	BH-rPPG	**TRUST**	**17.28 ± 0.37**	**-0.03 ± 0.01**
UBFC	PURE	PhysNet [44]	3.81 ± 0.34	0.87 ± 0.02
	PURE	Contrast-Pys [35]	19.61 ± 2.01	0.33 ± 0.06
	PURE	SiNC [32]	4.02 ± 0.06	0.86 ± 0.00
	PURE	**TRUST**	**2.80 ± 0.22**	**0.87 ± 0.01**
	COHFACE	**TRUST**	**10.95 ± 0.36**	**0.17 ± 0.02**
	BH-rPPG	**TRUST**	**17.44 ± 0.68**	**-0.23 ± 0.02**
COHFACE	PURE	**TRUST**	**13.46 ± 0.39**	**0.41 ± 0.01**
	UBFC	**TRUST**	**3.05 ± 0.42**	**0.78 ± 0.05**
	BH-rPPG	**TRUST**	**11.42 ± 0.31**	**-0.07 ± 0.02**

3.5. Augmentations

Inspired by the augmentation techniques from the SiNC paper [32], we applied various transformations in our experiments to enhance our model's resilience to noisy visual signals. Without these augmentations, achieving convergence during training is challenging. We employ several strategies, including Image Intensity Augmentation, which involves introducing Gaussian noise to each pixel and adjusting clip brightness using a Gaussian distribution, thus helping the model handle variations in image intensity. Spatial Augmentation includes random horizontal flipping and cropping of video clips, followed by linear interpolation to restore the original dimensions, aiding the model in generalizing to different spatial orientations and compositions. Temporal Augmentation leverages the assumption of strong periodicity in the Fourier domain by including random flipping along the time dimension and linearly interpolating video frames to a different frame rate, uniformly altering the video input along the time dimension. We also randomly resampled input clips within a specified range to enhance the model's ability to handle temporal variations. These augmentations collectively contribute to improving the model's performance and robustness by introducing variability and ensuring that the model can manage a wide range of noisy and varied inputs.

4. Datasets

We performed our experiments using four publicly available datasets: PURE [34], UBFC-rPPG [1], COHFACE [12], and BH-rPPG [43] for both training and testing.

PURE [34] is a benchmark rPPG dataset with 60 face videos from 10 subjects, recorded over six sessions lasting about a minute each. The sessions involved diverse head motions, such as steady movements, talking, slow and fast head translations, and small and medium head rotations. The video recordings were conducted at a resolution of 640×480 and a frame rate of 30 FPS. Ground-truth PPG signals were obtained using a finger clip pulse oximeter with a sampling rate of 60 Hz.

UBFC-rPPG [1] comprises 42 one-minute face videos from subjects engaged in a math game with time constraints. The dataset includes face videos, each at 640×480 resolution and 30 FPS, featuring PPG signals and heart rates recorded simultaneously.

COHFACE [12] dataset includes 160 one-minute videos featuring 40 subjects, captured in both studio and natural light. Notably, the videos underwent significant compression with MPEG-4 Visual, a factor highlighted by [19] for the potential corruption of rPPG signals. The videos were 640×480 pixels in resolution and had a frame rate of 20 FPS.

BH-rPPG [43] consisting of 105 videos, the dataset includes 35 subjects captured under three lighting conditions: low (8 lux), medium (42.4 lux), and high (104 lux). Videos maintained a resolution of 640 × 480 pixels and a frame rate of 15 FPS. PPG signals were acquired using a CONTEC CMS50E oximeter.

5. Training Details

5.1. Data Preprocessing

For video preprocessing, we utilized RetinaFace [9] to detect and crop faces from each frame, resulting in 64×64 cropped facial images. Instead of cropping faces from each predicted bounding box, we calculated the maximum window among all possible bounding boxes in each frame. This approach helps avoid artificial jerks that can occur when stacking cropped faces from each frame individually. This modification ensures smoother transitions between frames and improves the overall quality of the input images in our model.

5.2. Experimental Setup

We implemented our models using PyTorch and trained them for 200 epochs for each fold. The training was conducted on a Quadro RTX 8000 GPU utilizing a batch size 20. We employed the AdamW optimizer [15] with the learning rate set at 10^{-4}. To ensure consistency, we utilized a clip length of $T = 120$ frames (equivalent to 4 s), adjusting the length of the input signal to achieve a frequency resolution of 0.33 bpm.

5.3. Evaluation Metrics

We measured the accuracy of the pulse rate using the standard error metrics mean absolute error (MAE), root

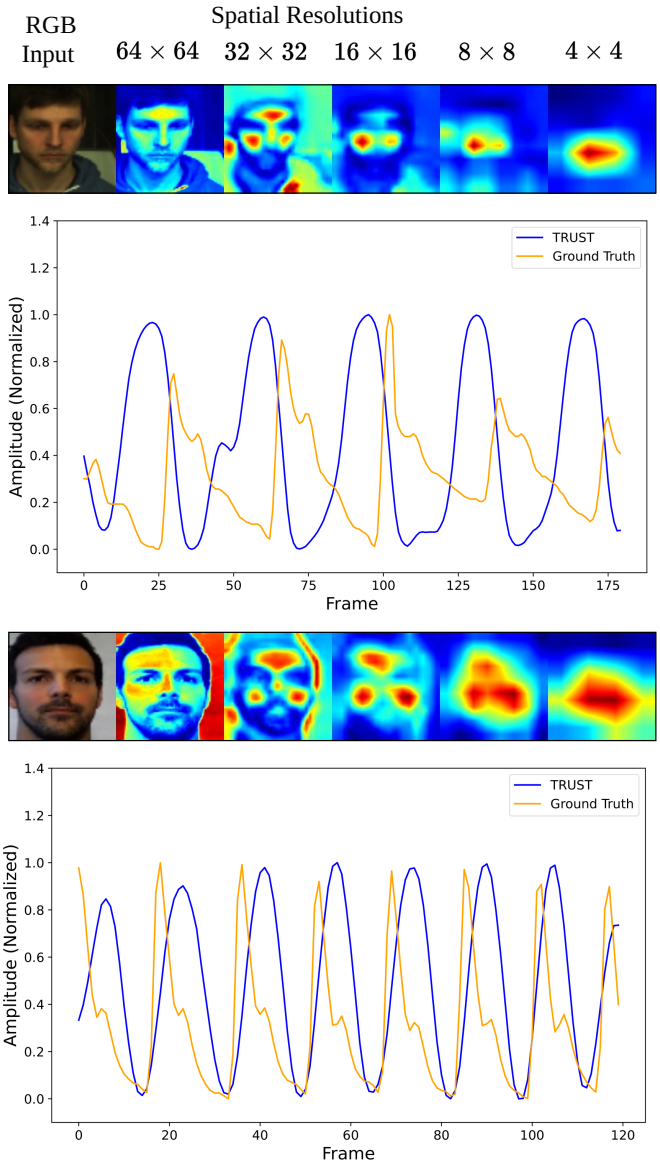

Figure 3. Feature heat maps captured at varying spatial resolution levels reveal stable features. The stability is particularly noticeable in low-level spatial resolutions, showing stable feature extraction from higher to lower resolutions (16×16, 8×8, and 4×4). The wave plot shows smooth wave extraction. The upper sample heat map and wave are from the PURE dataset and the heat map and waveform lower sample are from the COHFACE dataset.

mean squared error (RMSE), and Pearson correlation coefficient (r). Pulse rates are calculated by detecting the highest peak in the frequency spectrum, specifically the highest point within the range of 0.66 Hz to 3 Hz (equivalent to 40 bpm to 180 bpm), employing a 10-second sliding window. We apply the same approach to the ground truth waveforms, maintaining consistency and ensuring a robust evaluation [20].

We employ a 5-fold cross-validation approach for all four PURE, UBFC, COHFACE, and BH-rPPG datasets, adopting the same folds as in [11]. Our methodology uses three folds for training, one for validation, and the remaining one for testing, instead of having separate training and testing partitions. To enhance model robustness, we trained three models with distinct initializations, leading to a total of 15 models trained on all four datasets. The reported results included the mean and standard deviation of the errors.

6. Results & Discussion

6.1. Results on Intra-Dataset Testing

Table 1 presents the intra-dataset results for various methods, including traditional, supervised, and unsupervised approaches. Notably, our proposed model, TRUST, demonstrates superior performance compared to the state-of-the-art unsupervised non-contrastive method SiNC [32] and achieves results comparable to the best-supervised methods. In the PURE and UBFC datasets, TRUST achieved the lowest MAE among all unsupervised methods, and the Pearson correlation coefficient (r) was close to 1, indicating high accuracy and reliability. In the COHFACE dataset, TRUST outperformed all other supervised and unsupervised methods, highlighting its robustness and effectiveness under diverse conditions. For the BH-rPPG dataset, TRUST's performance closely matched that of the best supervised methods. The similarity between the rPPG waveform generated by TRUST and the ground truth rPPG signal, along with the feature heat map, is illustrated in Fig. 3, showcasing the model's ability to extract stable features and smooth waveforms.

6.2. Results on Cross-Dataset Testing

Cross-dataset testing is aimed at assessing the robustness of our approach to variations in lighting, camera sensor, motion, and pulse rate distribution. We present the results in Table 2 which indicate comparable performance between the supervised and unsupervised methods while transferring to diverse datasets. Specifically, training on PURE and testing on UBFC yields favorable outcomes, as does training on UBFC and testing on PURE. However, the performance diminishes when testing on COHFACE and BH-rPPG, which is likely due to disparities in frame rates (20 FPS for COHFACE and 15 FPS for BH-rPPG). Notably, training on COHFACE produces suboptimal results when tested on PURE and BH-rPPG, while achieving satisfactory performance on UBFC. These findings highlight the model's adaptability across datasets, but also underscore the impact of variations in FPS on cross-dataset generalization.

6.3. Performance Analysis Across COHFACE Conditions

We evaluated our model on the COHFACE dataset under various conditions to demonstrate its robustness, including Normal Light, Low Light (natural light), Male, Female,

Table 3. Performance Metrics of SiNC and TRUST Models Under Various Conditions on the COHFACE dataset. In this table, we evaluated our model on the COHFACE dataset under various conditions: Normal Light, Low Light (natural light), Male, Female, Light Skin, Dark Skin, Without Motion, and With Motion to demonstrate robustness in different conditions. We also compared our TRUST model with the SiNC model. The evaluation was conducted across the entire dataset, with samples grouped according to the specified conditions.

	Normal Light			**Low Light**			**Male (28 subjects)**			**Female (12 subjets)**			**Light Skin**			**Dark Skin(2 subjects)**			**Without motion**			**With motion**		
Method	MAE ↓	RMSE ↓	r ↑	MAE ↓	RMSE ↓	r ↑	MAE ↓	RMSE ↓	r ↑	MAE ↓	RMSE ↓	r ↑	MAE ↓	RMSE ↓	r ↑	MAE ↓	RMSE ↓	r ↑	MAE ↓	RMSE ↓	r ↑	MAE ↓	RMSE ↓	r ↑
SiNC [32]	2.22	6.12	0.86	2.71	6.85	0.85	2.23	5.48	0.87	2.98	8.24	0.77	2.12	5.34	0.88	6.79	13.97	0.43	2.59	6.87	0.83	2.34	6.11	0.85
TRUST	1.09	3.84	0.95	1.76	5.21	0.90	1.28	3.58	0.95	1.78	6.27	0.87	1.11	3.21	0.96	5.52	12.38	0.55	1.46	4.75	0.92	1.39	4.38	0.93

Table 4. Effect of varying loss function combinations on the feature stabilization block. Evaluation of metrics through an ablation study An upward arrow (↑) signifies that larger values are better, while a downward arrow (↓) indicates the opposite.

Feature Stabilization Block	$\mathcal{L}_{variance}$	$\mathcal{L}_{energy}$	$\mathcal{L}_{smoothness}$	MAE↓	RMSE↓	r ↑	ρ ↓
✓	✓	✗	✗	3.17 ± 0.24	7.90 ± 0.35	0.76 ± 0.07	2.7
✓	✗	✓	✗	3.15 ± 0.31	7.89 ± 0.86	0.76 ± 0.05	2.5
✓	✗	✗	✓	2.91 ± 0.48	7.44 ± 0.22	0.79 ± 0.08	2.3
✓	✓	✓	✗	1.91 ± 0.12	6.73 ± 0.31	0.81 ± 0.01	1.9
✓	✗	✓	✓	1.86 ± 0.19	6.15 ± 0.07	0.86 ± 0.02	1.8
✓	✓	✗	✓	2.12 ± 0.42	6.32 ± 0.10	0.80 ± 0.04	2.0
✗	✓	✓	✓	2.31 ± 0.27	6.35 ± 0.03	0.82 ± 0.03	2.1
✓	✓	✓	✓	**1.50 ± 0.16**	**4.45 ± 0.66**	**0.92 ± 0.02**	**1.7**

Light Skin, Dark Skin, Without Motion, and With Motion. In this evaluation, we compared our TRUST model with the SiNC model. The evaluation was conducted across the entire dataset by grouping the samples according to the specified conditions. Performance metrics, including the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Pearson correlation coefficient (r), were measured for each condition. The results, summarized in Table 3, show that TRUST demonstrates superior or competitive performance compared to SiNC, highlighting its robustness for heart rate prediction across different conditions.

6.4. Ablation Study on Losses

We performed an ablation study to explore the contributions of different loss-function components and a feature extraction block. The results, presented in Table 4 for the COHFACE dataset, reveal insights into the behavior of the model. Initially, the evaluation of a single component from the loss function yielded poor results. Gradual improvements were observed when the two loss components were incorporated. Surprisingly, considering all three loss components without the feature extraction block led to diminished performance. However, combining all three loss components with the feature extraction block resulted in a remarkable enhancement of the model's overall performance. Notably, the feature extraction block emerged as a crucial component for uncovering the rPPG signal, highlighting its significance in the model's learning process, as the losses alone did not autonomously capture the desired signal.

We assessed the stability of spatial features through the feature stabilization metric (ρ), calculated specifically on the 4×4 spatial resolution in the final layers of our model. The results presented in Table 4 indicate a strong correlation between the MAE values and the metric. When the MAE reflects poor performance, the metric also indicates instability in the features. In contrast, when the model attains a low MAE, indicating strong performance, the metric aligns, emphasizing the stability of the features at that layer. This metric serves as a valuable indicator of feature reliability.

7. Conclusion

Our study introduces TRUST, an advanced non-contrastive unsupervised learning framework. In contrast to existing methods, TRUST focuses on stabilizing features at low-level spatial resolution through the incorporation of time-domain loss functions and a feature stabilization block. We also introduce the feature stabilization metric to measure the feature instability in the last layer of the model. This innovative approach significantly enhances the overall quality of feature extraction. Notably, TRUST outperforms conventional unsupervised methods, signifying substantial progress in obtaining accurate signals without reliance on labeled data. Our contributions underscore the potential of TRUST in advancing the field of non-contact health monitoring, offering a more robust and effective solution for vital sign estimation through non-contrastive unsupervised learning.

Acknowledgment

This research was conducted with support from the Eureka Eurostars funding program under Project number 55, with the project acronym BabySensor-Pre.

References

[1] 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. 2, 6

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 2

[3] Shutao Chen, Sui Kei Ho, Jing Wei Chin, Kin Ho Luo, Tsz Tai Chan, Richard HY So, and Kwan Long Wong. Deep learning-based image enhancement for robust remote photoplethysmography in various illumination scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6076–6084, 2023. 5

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[5] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the european conference on computer vision (ECCV)*, pages 349–365, 2018. 1, 2

[6] Tamal Chowdhury, Sukalpa Chanda, Saumik Bhattacharya, Soma Biswas, and Umapada Pal. Contact-less heart rate detection in low light videos. In Christian Wallraven, Qingshan Liu, and Hajime Nagahara, editors, *Pattern Recognition - 6th Asian Conference, ACPR 2021, Jeju Island, South Korea, November 9-12, 2021, Revised Selected Papers, Part I*, volume 13188 of *Lecture Notes in Computer Science*, pages 77–91. Springer, 2021. 1

[7] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013. 1, 5

[8] Gerard De Haan and Arno Van Leest. Improved motion robustness of remote-ppg by using the blood volume pulse signature. *Physiological measurement*, 35(9):1913, 2014. 2

[9] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019. 6

[10] Yogesh Deshpande, Surendrabikram Thapa, Abhijit Sarkar, and A Lynn Abbott. Camera-based recovery of cardiovascular signals from unconstrained face videos using an attention network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5974–5983, 2023. 5

[11] John Gideon and Simon Stent. The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3995–4004, 2021. 1, 2, 5, 7

[12] G Heusch, A Anjos, and S Marcel. A reproducible study on remote heart rate measurement. arxiv 2017. *arXiv preprint arXiv:1709.00962*. 2, 6

[13] Eugene Lee, Evan Chen, and Chen-Yi Lee. Meta-rppg: Remote heart rate estimation using a transductive meta-learner. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 392–409. Springer, 2020. 1, 2

[14] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems*, 33:19400–19411, 2020. 1, 2

[15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[16] Hao Lu, Hu Han, and S Kevin Zhou. Dual-gan: Joint bvp and noise modeling for remote physiological measurement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12404–12413, 2021. 1, 2, 5

[17] Daniel McDuff, Sarah Gontarek, and Rosalind Picard. Remote measurement of cognitive stress via heart rate variability. In *2014 36th annual international conference of the IEEE engineering in medicine and biology society*, pages 2957–2960. IEEE, 2014. 1

[18] Daniel McDuff, Miah Wander, Xin Liu, Brian Hill, Javier Hernandez, Jonathan Lester, and Tadas Baltrusaitis. Scamps: Synthetics for camera measurement of physiological signals. *Advances in Neural Information Processing Systems*, 35:3744–3757, 2022. 2

[19] Daniel J McDuff, Ethan B Blackford, and Justin R Estepp. The impact of video compression on remote cardiac pulse measurement using imaging photoplethysmography. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 63–70. IEEE, 2017. 6

[20] Yuriy Mironenko, Konstantin Kalinin, Mikhail Kopeliovich, and Mikhail Petrushan. Remote photoplethysmography: Rarely considered factors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 296–297, 2020. 7

[21] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6707–6717, 2020. 2

[22] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V 14*, pages 562–576. Springer, 2019. 2

[23] 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. 1, 2

[24] Xuesong Niu, Zitong Yu, Hu Han, Xiaobai Li, Shiguang Shan, and Guoying Zhao. Video-based remote physiological measurement via cross-verified feature disentangling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 295–310. Springer, 2020. 2

[25] Ewa M Nowara, Daniel McDuff, and Ashok Veeraraghavan. The benefit of distraction: Denoising camera-based physiological measurements using inverse attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4955–4964, 2021. 1, 2

[26] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[27] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, 58(1):7–11, 2010. 1, 2, 5

[28] 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(10):10762–10774, 2010. 2

[29] Rita Meziati Sabour, Yannick Benezeth, Pierre De Oliveira, Julien Chappe, and Fan Yang. Ubfc-phys: A multimodal database for psychophysiological studies of social stress. *IEEE Transactions on Affective Computing*, 2021. 1

[30] Jingang Shi, Iman Alikhani, Xiaobai Li, Zitong Yu, Tapio Seppänen, and Guoying Zhao. Atrial fibrillation detection from face videos by fusing subtle variations. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(8):2781–2795, 2019. 1

[31] Jeremy Speth, Nathan Vance, Patrick Flynn, Kevin Bowyer, and Adam Czajka. Unifying frame rate and temporal dilations for improved remote pulse detection. *Computer Vision and Image Understanding*, 210:103246, 2021. 2, 4, 5

[32] Jeremy Speth, Nathan Vance, Patrick Flynn, and Adam Czajka. Non-contrastive unsupervised learning of physiological signals from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14464–14474, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[33] Radim Špetlík, Vojtech Franc, and Jirí Matas. Visual heart rate estimation with convolutional neural network. In *Proceedings of the british machine vision conference, Newcastle, UK*, pages 3–6, 2018. 1, 2, 5

[34] Ronny Stricker, Steffen Müller, and Horst-Michael Gross. Non-contact video-based pulse rate measurement on a mobile service robot. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 1056–1062. IEEE, 2014. 2, 6

[35] Zhaodong Sun and Xiaobai Li. Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast. In *European Conference on Computer Vision*, pages 492–510. Springer, 2022. 1, 2, 5, 6

[36] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008. 1, 5

[37] Hao Wang, Euijoon Ahn, and Jinman Kim. Self-supervised representation learning framework for remote physiological measurement using spatiotemporal augmentation loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2431–2439, 2022. 1, 2

[38] Wenjin Wang, Albertus C Den Brinker, and Gerard De Haan. Single-element remote-ppg. *IEEE Transactions on Biomedical Engineering*, 66(7):2032–2043, 2018. 2

[39] Wenjin Wang, Albertus C Den Brinker, Sander Stuijk, and Gerard De Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2016. 1, 2, 5

[40] Zhen Wang, Yunhao Ba, Pradyumna Chari, Oyku Deniz Bozkurt, Gianna Brown, Parth Patwa, Niranjan Vaddi, Laleh Jalilian, and Achuta Kadambi. Synthetic generation of face videos with plethysmograph physiology. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20587–20596, 2022. 2

[41] Bryan P Yan, William HS Lai, Christy KY Chan, Stephen Chun-Hin Chan, Lok-Hei Chan, Ka-Ming Lam, Ho-Wang Lau, Chak-Ming Ng, Lok-Yin Tai, Kin-Wai Yip, 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. 1

[42] Yuzhe Yang, Xin Liu, Jiang Wu, Silviu Borac, Dina Katabi, Ming-Zher Poh, and Daniel McDuff. Simper: Simple self-supervised learning of periodic targets. *arXiv preprint arXiv:2210.03115*, 2022. 1, 2

[43] Ze Yang, Haofei Wang, and Feng Lu. Assessment of deep learning-based heart rate estimation using remote photoplethysmography under different illuminations. *IEEE Transactions on Human-Machine Systems*, 52(6):1236–1246, 2022. 2, 6

[44] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. *arXiv preprint arXiv:1905.02419*, 2019. 1, 2, 4, 5, 6

[45] 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. 1

[46] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip HS Torr, and Guoying Zhao. Physformer: Facial video-based physiological measurement with temporal difference transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4186–4196, 2022. 2

[47] Zijie Yue, Miaojing Shi, and Shuai Ding. Facial video-based remote physiological measurement via self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3, 5

[48] Yu Zhao, Bochao Zou, Fan Yang, Lin Lu, Abdelkader Nasreddine Belkacem, and Chao Chen. Video-based physiological measurement using 3d central difference convolution attention network. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–6. IEEE, 2021. 2