2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

RS-rPPG: Robust Self-Supervised Learning for rPPG

Marko Savic and Guoying Zhao*
Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, FI-90014, Finland

Abstract— **Remote photoplethysmography (rPPG) measures cardiac signals remotely from facial videos, leading to promising applications in telemedicine, face anti-spoofing, emotion analysis, etc. However, recent supervised approaches are limited by data scarcity and current self-supervised rPPG methods struggle to learn physiological features from data recorded in challenging scenarios, which contain overwhelming environmental noise caused by head movements, illumination variations, and recording device changes. We propose a novel contrastive framework that leverages a large set of priors, that enable learning robust and transferable features even from challenging datasets. Ours is the first method to focus on self-supervised learning on challenging data and the first method to use such a large set of priors. The priors include a novel traditional augmentation method, leveraging spatial-temporal maps and self-attention based transformer for SSL. We show that it outperforms current self-supervised methods on four public datasets, especially on the more challenging data where it reaches close to supervised performance. Our code is available at: https://github.com/marukosan93/RS-rPPG**

I. INTRODUCTION

Heart rate (HR), heart rate variation (HRV), respiratory rate and oxygen saturation are important healthcare parameters and emotional cues, since they change accordingly with our well-being and emotional states. Remote photoplethysmography (rPPG) can obtain a signal akin to blood volume pulse (BVP) [5] without any contact, by using ordinary RGB cameras. The rPPG signal is obtained from subtle pixel color variations present in RGB facial videos, relying on the same optical principles as contact photoplethysmography. However, the rPPG signal-to-noise ratio is poor, as the physiological signal has low power when compared to environmental noise that is also captured (caused by head movement, illumination variations, sensor differences, etc.), making robust rPPG very challenging. Early methods relied on hand-crafted features or blind source separation [43], [34], [18], [9], [10], [45], but due to their lack of robustness in scenarios with variable lighting and movement they were surpassed by deep learning approaches. Most deep learning methods employed supervised learning via Convolutional Neural Network (CNN) based models[6], [38], [49], [50], [29], [30], [24], [8] or more recently, self-attention based transformer architectures [22], [35], [51], [15], [36].

Insufficient labeled data is a significant problem in rPPG, as data collection is costly, requires medical devices and presents privacy concerns. Supervised methods lack robustness and generalization capabilities when trained on small datasets with specific noise distributions (from lighting, movement, devices). Solutions to lack of annotated data have been attempted such as data augmentation [31], [48], [41], [17], [1], synthetic data [27], [25], [26] and self-supervised learning (SSL) methods. For generating positive and negative learning samples, several SSL methods exploit intra-data differences in facial videos by utilising spatial-frequency augmentation [13], [47], [52], spatial-temporal augmentation [44], [16] or within video spatial similarity [3]. However, artificial augmentations can introduce additional bias and ignore inter-data information. Leveraging inter-data differences via instance-wise sampling [40], [2] or non-contrastive learning [37] has also proven effective in learning without annotated labels. Nonetheless, there is still a notable gap between self-supervised and the state-of-the-art supervised methods. Moreover, the aforementioned works have been evaluated with controlled environment rPPG datasets (stable illumination and minimal subject movement), and have not proven to be robust to more challenging data. These end-to-end methods all deal with direct input video data, that is polluted with non-physiological noise and propose weak constraints, that cannot hold up in challenging scenarios. For example, Contrast-Phys [40] and SiNC [37] rely on inter-data differences being caused by physiological factors, as different samples likely contain different HR. However, in less uniform and controlled videos, different samples also contain different noise signals that are more pronounced than physiological signals. Thus, focusing only on the inter-data difference on challenging data can lead to learning non-physiological noise present in the data. Furthermore, SSL is unfeasible on a large scale if it can only reliably train from controlled data, as only a small subset of available data can be used, leaving the data scarcity issue unresolved.

In this work, we propose RS-rPPG, a novel contrastive self-supervision framework that is robust to environmental noise, making it applicable to challenging data. RS-rPPG is the first self-supervised method to focus on challenging data and the first to leverage an exhaustive set of seven priors derived from observations about rPPG. The aforementioned priors, most notably include, a novel traditional map augmentation method for learning physiology specific features, leveraging spatial-temporal maps for more robust SSL, and exploiting self-attention based transformer for better temporal modelling. Our method is shown to outperform previous SSL works on four datasets containing both controlled (PURE [39], OBF [19]) and challenging facial videos (MMSE-HR [53], VIPL-HR [28]). We perform ex-

This work was supported by the Research Council of Finland project EmotionAI (grants 336116, 345122) and ICT 2023 project TrustFace (grant 345948), and the University of Oulu & Research Council of Finland Profi 7 (grant 352788). As well, the authors acknowledge CSC – IT Center for Science, Finland, for computational resources.

*Corresponding E-mail: guoying.zhao@oulu.fi

979-8-3503-9494-8/24/$31.00 ©2024 IEEE

tensive intra-dataset, cross/mixed-dataset, demographics and segment length validations, which show that RS-rPPG can learn reliable features from both controlled and challenging data.

II. Related Works

An early study proved the feasibility of measuring rPPG from facial videos by extracting a coarse physiological signal by averaging pixels from the green channel of a physiologically significant facial region [43]. It was followed by numerous traditional methods that relied on handcrafted features and did not need datasets for training. They either relied on optical/physiological considerations expressed through mathematical models like CHROM [9], POS [45], PBV [10], LGI [33] or common blind source separation approaches such as ICA [34] and PCA [18]. These traditional methods were built on assumptions that may not hold in less constrained environments and were surpassed by deep learning methods, for example, CHROM [9] assumes a standardized skin color profile, ICA [34] assumes independence of source signals and PCA [18] assumes their uncorrelation. Among the first deep learning methods were 2DCNN models that extracted the HR from two adiacent frames, such as HR-CNN [38] and DeepPhys [6], that only take spatial information into account. MTTS-CAN [21] extends the DeepPhys architecture by allowing temporal information to be extracted not only from adiacent frames. To overcome the deficiency of temporal information of 2D-CNN methods, as temporal information is crucial for accurate estimation of the quasi-periodical rPPG signal, 3D-CNN were proposed. End-to-end spatial-temporal 3D-CNN models have been used to exploit the temporal information such as PhysNet [49], rPPGNet [50] and Deep-rPPG [20]. The aforementioned methods input a whole facial video block into a 3D-CNN network and extract a temporal signal. Another way to exploit temporal context that can suppress the information that is unrelated to the rPPG signal is computing spatial-temporal maps as input, which act a more compact intermediate representation comprised of coarse signals obtained via averaging. Methods that exploit spatial-temporal maps have shown to have a higher degree of robustness to environmental noise, like RhythmNet [29], CVD [30], Dual-GAN [24] and BVPNet [8]. Due to the success of self-attention based Transformer architectures such as ViT [12] and Swin [23], rPPG methods relying on transformers have been showing promising improvements. Transformer methods such as RADIANT [15], [35] and PhySU-Net [36] are composed of two stages, where intermediary signal embeddings are first obtained and passed to transformer layers to learn global context. EfficientPhys [22] and Physformer [51], on the other hand, use transformer blocks throughout the whole model architecture.

An open challenge in rPPG is the lack of data, as supervised methods tend to not generalise well to samples dissimilar to the training distribution, and varied rPPG data is costly to obtain. To mitigate this issue, other than classic computer-vision augmentation strategies, there have also been efforts to create rPPG specific augmentations. Spatial-temporal augmentations have also been proposed [31], [48] to extend the training set with extra samples containing extremely small or large HR values, by temporally up-sampling and down-sampling videos to achieve this. Augmenting datasets with synthetic videos generated from real data via video-to-video networks was proposed in [41], RErPPGNet [17] and [1]. Particularly, in [1] unwanted bias toward different demographic groups is addressed by augmenting the data with generated darker skinned subjects, as datasets compiled in western academia mostly contain lighter skinned subjects. Another approach is training completely on synthetic data. In [27] the authors construct a dataset of synthetic spatial-temporal maps. Synthetic avatars have also been proposed in [25], where synthetic videos with underlying physiological signals are generated. A large-scale dataset, SCAMPS [26], containing synthetic videos from with diverse subjects and scenarios has been proposed. However, currently synthetic data cannot perfectly imitate real conditions and complex environmental noise. In contrast to the limited amount of labeled rPPG data, facial videos without physiological labels are bountiful.

To overcome data scarcity by leveraging the wast amount of unlabeled real facial video data available, self-supervised learning methods are studied for rPPG. SSL methods can exploit intra-data information from each video sample, for example by utilising spatial and temporal augmentations to produce samples for contrastive learning. In [13], the authors employ a frequency saliency sampler module to generate new negative samples with artificially altered heart rate. SLF-RPM [44] utilises a sparsity based temporal augmentation combined with a landmark based spatial augmentation for negative samples generation. Self-rPPG [16] adopts an approach where negatives are generated by repeating and shuffling frames from the original videos. Simper [47] proposes both periodicity-invariant (crop, resize, reverse, shift) and periodicity-variant (frequency based) augmentations to learn periodic information. ALPINE [3] learns by using the similarity between temporal signals from multiple face regions. In [52] spatial augmentation and learnable frequency augmentations are applied, followed by an aggregation module. Nonetheless, spatial-temporal and frequency artificial augmentations of the data can potentially induce bias into the trained models, and in the aforementioned works inter-data variation is ignored.

Another way to generate training pairs is by using the rich inter-data information present in different video samples. Contrast-Phys [40] employs a 3D-CNN to extract signals that and spatial-temporal instance-wise sampling to learn from inter-data differences from different videos. SimPPG [2] also utilises encoded signals from same video for positive and from different videos for negative sampling. SiNC [37] is a non-contrastive method that does not explicitly define negative pairs, but encourages diverse power spectra over batches of different samples to exploit inter-data differences. Current SSL methods utilise weak self-supervision constraints that may not hold with more challenging data (variable illumination and subject movement). Consequently, they struggle to learn robust features from less controlled data, that is why more robust constraints are necessary to learn physiological features on challenging data.

III. Method

Our framework relies on a large set of priors (**P#**). We will first introduce the motivations followed by their incorporation in our framework. We consider the following concepts, motivated by previous works and observations about rPPG:

- **P1)** Spatial-temporal maps are less subject to noise. This stems from the success of many non-end-to-end methods [32], [29], [24], [8] and traditional methods [43], [9], [45] that proved averaging pixels from regions-of-interest (ROI) can suppress non-physiological noise. Therefore SSL from spatial-temporal maps is less likely to learn physiologically irrelevant noise.
- **P2)** Self-attention based transformers can lead to better temporal modelling, as shown in [22], [35], [51], [15] where the improved long-range spatial-temporal perception from transformers is exploited.
- **P3)** Signals extracted using traditional methods contain more physiological information than raw averaged signals. Traditional methods such as CHROM [9], GREEN [43], POS [45] are motivated by physiological considerations and provide outputs that are closer to the underlying physiological signals compared to the raw averaged signals.
- **P4)** Different facial videos most likely contain different rPPG signals, making inter-data differences rich in information [40], [2], [37]. The frequency characteristics of different input videos will most likely have a different HR peak and spectrum, as they are influenced by many factors that have high variability e.g. subject's resting heart rate, emotional state, respiration, etc. It is highly unlikely to obtain the exact same spectrum from different recordings, even in case where the average HR is similar the rest of the spectrum will not be the exact same.
- **P5)** Due to the spatial redundancy of the imaging sensor, that captures several skin regions, coarse signals obtained from different spatial regions and channels represent the same underlying rPPG signal [46], [3]. Different skin areas on the face are all similarly affected by the cardiac cycle and should results in highly correlated signals. Moreover, the imaging sensor captures multiple channels, that sample the same optical phenomena at different light wavelengths. Therefore, there are both spatial and channel redundancy present in the facial video data.
- **P6)** The rPPG signal is band limited $[0.5, 3]Hz$ corresponding to 30 to 180 heart beats per minute, covering the range of normal heart rates. Frequency components that are outside of this range correspond to external noise caused by environmental factors such as illumination and motion.
- **P7)** The spectrum of rPPG signals is sparse, containing one strong peak at the HR frequency due to the heart's strong periodicity. Focusing on strongly periodic signals with a sparse spectrum can help SSL methods filter out non-physiological noise.

A. Input map generation (*P1, P3*)

Due to their capability to exclude non-physiological noise we utilize MSTmaps [30] as input (**P1**), calculated by averaging pixels from ROIs on the face, we follow the procedure from [30]. Firstly, to extract landmarks we utilise PyFeat[7] with RetinaFace [11] model for face detection and PFLD[14] model for landmark localisation, the landmarks are then stabilised with a 5-point moving average filter. Six informative ROIs are defined within the face (forehead, mouth, upper left cheek, lower left cheek, upper right cheek, lower right cheek), that are joined in $R = 2^6 - 1$ ROI combinations, thus merging global and local information. For each ROI combination R and color channel C a temporal sequence is obtained by averaging the pixels for the whole video. Each of the $C * R$ sequences is then bandpass filtered at $[0.5, 3]Hz$ to reduce interference of non-physiological signal components and is min-max normalised, the R dimension is resized from 63 to 64 for computational ease.

Secondly, as augmentation is one of the most important aspects in contrastive SSL, we propose a novel augmentation method. As part of the positive sampling, instead of applying a classic image processing transformation to slightly alter the input (while keeping the a similar distribution) we choose to apply a new transformation with physiological significance. We calculate an augmented spatial-temporal map that we name Tmap, by computing signals with traditional methods. We apply CHROM [9], GREEN [43], POS [45] on the MSTmaps (along the R dimension) and concatenate the resulting signal maps along the channel dimension to form the Tmaps, in which each channel corresponds to the traditional method used for calculating the signals. By using Tmaps as augmentation, we encourage the network to learn similarities between the coarse averaged inputs and the more physiologically relevant traditional signals (**P3**). The MSTmap and Tmap generation procedure is illustrated in Fig. 1.

B. Contrastive learning (*P3, P4, P5*)

To learn from the data itself, without utilising any labels, we exploit both intra-data and inter-data relationships. To fully leverage the underlying physiological information present in each sample we rely on (**P3**) to learn similarities between coarse signals and traditionally augmented signals, and enforce ROI and channel consistency (**P5**) to find common patterns between signals sampled at different spatial locations and channels. Firstly, a mini-batch of MSTmaps $X[b, c, r, t]$ (with b, c, r, t being the dimensions corresponding to batch, channel, ROI and time) is fed into the model F, generating the Anchor $A = F(X)$. Secondly, we input the augmented Tmaps $X'[b, c, r, t]$ into F and generate $F(X')$ that will be used to create the Positives P and Negatives N.

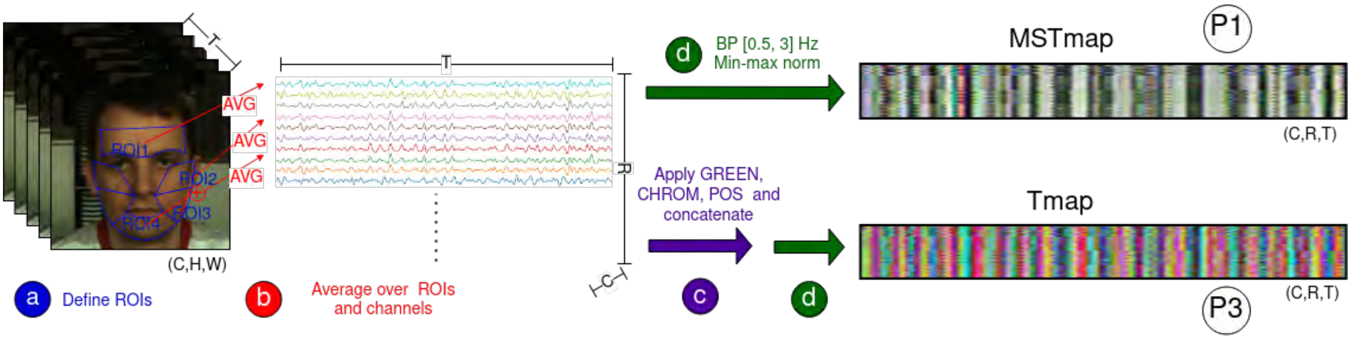

Fig. 1. **Input map generation:** a) ROIs are defined for each frame b) $C \times R$ temporal sequences are extracted by averaging pixels for each channel and ROI combination. c) GREEN, CHROM, POS are applied on the signals and results in $C = 3$ signal maps of $(1, R, T)$, that are concatenated to form the Tmap (C, R, T) d) The sequences are filtered with a pass band of $[0.5, 3]Hz$ and min-max normalized to form the MSTmap or Tmap.

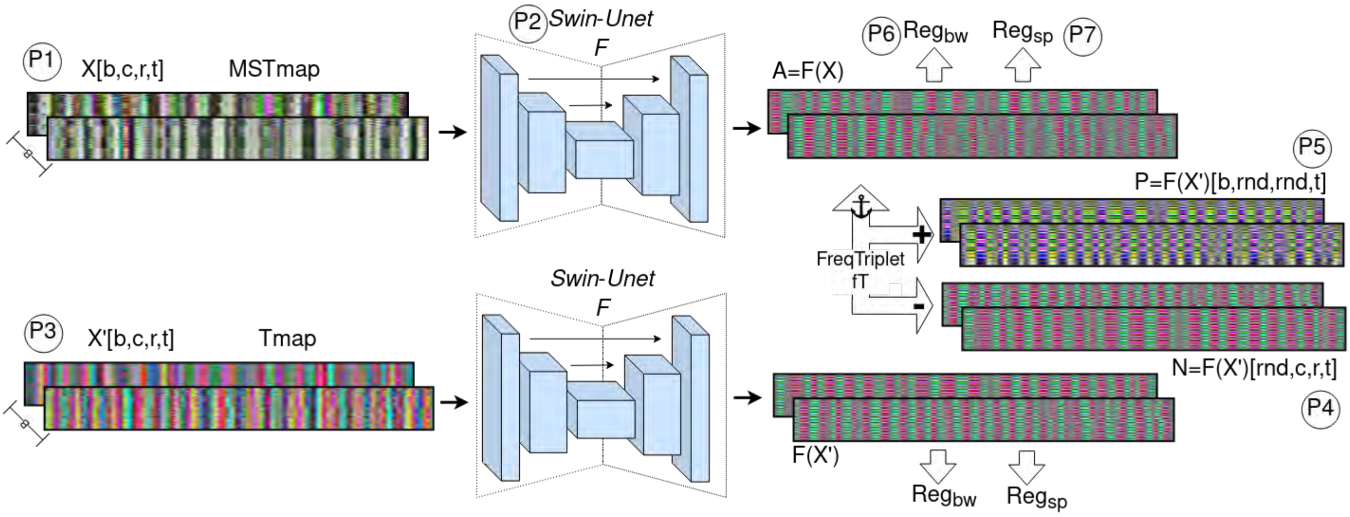

Fig. 2. **Overview:** A mini-batch of the input X (**P1**) and its traditional augmentation X' (**P3**) are fed into F (**P2**) . For our frequency triplet loss learning $F(X)$ is the anchor, the positive (**P5**) is generated by shuffling $F(X')$ along the ROI and channel dimensions and the negative (**P4**) by shuffling $F(X')$ along the batch dimension. $F(X')$ and $F(X)$ are regularized to have limited bandwidth and low normalized spectral power with Reg_{bw} (**P6**) and Reg_{sp} (**P7**).

Each T length row of the MSTmap contains a coarse signal obtained by averaging a certain channel and ROI, since each of these should contain the same underlying rPPG signal we shuffle $F(X')$ along the channel and ROI dimensions using rnd (a simple operation that returns random indices that are in different positions than the originals), obtaining $P = F(X'[b, rnd, rnd, t])$. We encourage the model to learn similarities between differently positioned (ROI and C) coarse rows (**P5**), and similarities between the MSTmaps and Tmaps (**P3**). With this implementation we effectively confront a large set of signals that all measure the same rPPG signal, but that are either sampled or processed differently.

To learn from the ample information present in inter-data relationships we exploit instance discrimination (**P4**). We implement the negative sampling (**P4**) similarly to the positive. Again, we use the augmented output $F(X')$ and simply shuffle it along the batch dimension using rnd, obtaining $N = F(X'[rnd, c, r, t])$. In this manner, we encourage the model to push apart augmented signals that come from different videos, as they likely correspond to rPPG signals with dissimilar spectra.

To enforce (**P3**, **P4**, **P5**) we design the contrastive frequency based triplet loss fT. For any spatial-temporal map $X(r, t, c)$ containing $R \times C$ temporal signals $x_{r,c}(t)$ of length T, we define $\bar{x}_{r,c}(f)$ as the power-spectral-density (PSD) of $x_{r,c}(t)$. We define a frequency triplets loss fT as in (1), with $f_{min} = 0.5Hz$ and $f_{max} = 3Hz$ being the band limit for rPPG signals (**P6**). Practically, within the relevant frequency band, the loss fT guides F to bring the spectra of positively samples output signals closer to the anchor, and to increase the distance with the negative sample's spectra. The margin m ensures non-negativity of fT and defines a minimum acceptable distance between the anchor-positive and the anchor-negative pair, providing stable convergence as the distance cannot increase indefinitely.

$$fT = max(m + \sum_{f=f_{min}}^{f_{max}} \frac{||\bar{a}_{r,c}(f) - \bar{p}_{r,c}(f)||_1 - ||\bar{a}_{r,c}(f) - \bar{n}_{r,c}(f)||_1}{f_{max} - f_{min}}, 0) \quad (1)$$

C. Network, regularisation and sampling (***P2***,***P3***,***P6***,***P7***)

Inspired by BVPNet [8], we formulate the learning problem as a spatial-temporal map prediction, where F recon-

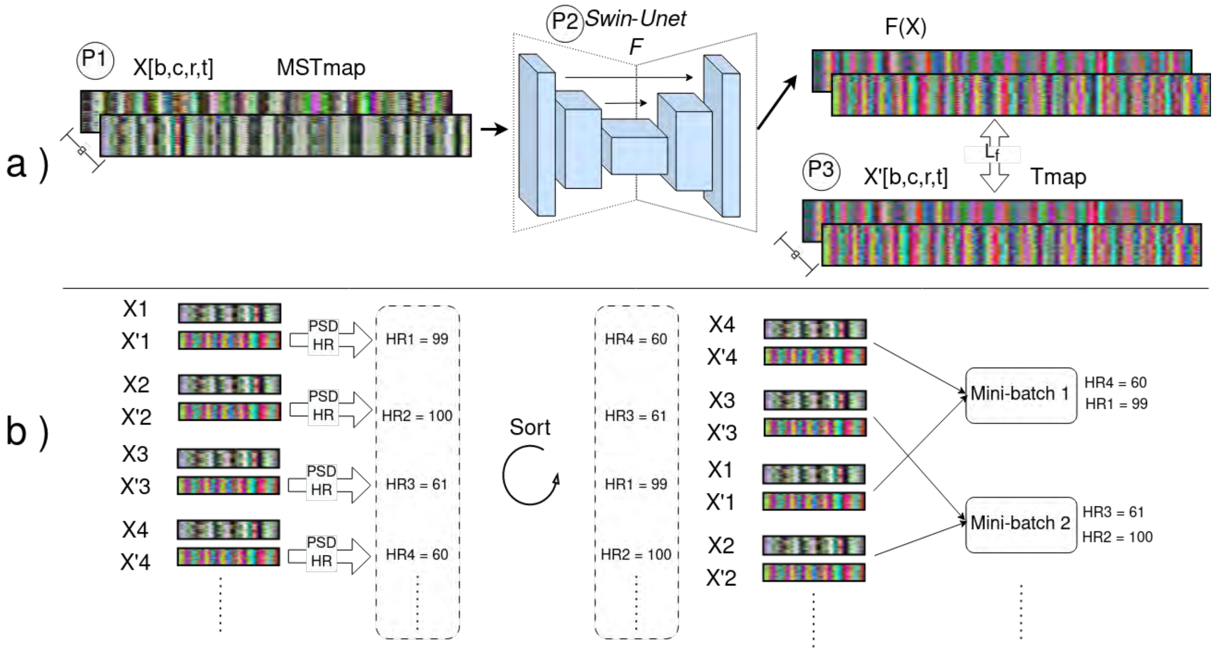

Fig. 3. **a)** The network F is initialised by pre-training it to predict Tmaps from MSTmaps via the frequency based loss L_f (**P3**). **b)** The dataset is ordered by using rough HR predictions calculated via PSD peak from the Tmaps (traditional method signals) resulting in more diverse mini-batche sampling (**P4**).

structs the coarse input signals in accurate physiological signals. Using a Unet type architecture that predicts a signal map (rather than a single signal) allows us to implement the diverse positive/negative sampling by employing standard mini-batches and shuffling along the pertinent dimensions. Given the success of self-attention based transformers [12], [23], we utilize Swin-Unet [4] (**P2**) as the backbone of our framework, with the self-attention mechanism allowing for stronger modeling of temporal sequences. Swin-Unet was designed for medical-image segmentation, but can be used as a general transformer backbone, as it is straightforwardly a general Unet extension of the Swin transformer. This general backbone allows us to take advantage of self-attention for modelling temporal relationships and to implement the priors efficiently on a large set of signals. Moreover, it was shown to handle rPPG data well in PhySU-Net [36], where Swin-Unet was adapted for supervised HR regression. To constrain the network to learn rPPG signals, we add two regularization terms. As the rPPG signal is band limited, Reg_{bw} (2) (**P6**) keeps the model from generating signals outside the HR relevant band as they correspond to irrelevant noise. Essentially, Reg_{bw} tries to minimise the power in the irrelevant band, normalised against the total power.

$$Reg_{bw} = \frac{\frac{1}{f_{min}}\sum_{f=0}^{f_{min}} \bar{x}_{r,c}(f) + \frac{1}{F-f_{max}}\sum_{f=f_{max}}^{F} \bar{x}_{r,c}(f)}{\frac{1}{F}\sum_{f=0}^{F} \bar{x}_{r,c}(f)} \tag{2}$$

To enforce sparsity in the output signal spectrum, we define Reg_{sp} (3) (**P7**), encouraging the model to learn only the salient frequency features likely corresponding to the HR peak frequency, and penalising the model for outputting spectra that do not resemble those of rPPG signals. Instinctively, sparsity is encouraged by minimising the L2 norm of the frequency min-max normalised spectrum, penalising non-sparse frequency distribution.

$$Reg_{sp} = \sqrt{\frac{1}{f_{max}-f_{min}}\sum_{f=f_{min}}^{f_{max}} \left(\frac{\bar{x}_{r,c}(f)-min_f(\bar{x}_{r,c}(f))}{max_f(\bar{x}_{r,c}(f))-min_f(\bar{x}_{r,c}(f))}\right)^2} \tag{3}$$

The final training loss is simply the sum of the triplet loss fT (1) and regularization terms Reg_{bw} (2) and Reg_{sp} (3) averaged over the batch, ROI and channel dimensions. Finally, we define a simple yet effective training strategy. Firstly, we initialize F by pre-training it to predict Tmaps X' from MSTmaps X (**P3**), learning first to predict coarse traditional rPPG signals that it will later refine using contrastive learning. Taking the PSD of the input $\bar{x}_{r,c}(f)$ and of the pseudo-label $\bar{x}'_{r,c}(f)$, the pre-training loss function L_f is defined as the L2-norm of the spectrum difference as shown in 4.

$$L_f = \sqrt{\sum_{r=0}^{R}\sum_{c=0}^{C}\sum_{f=f_{min}}^{f_{max}} (\bar{x}_{r,c}(f) - \bar{x}'_{r,c}(f))^2} \tag{4}$$

With pre-training we condition the network to learn the frequency characteristics of the traditionally augmented signals, making the initial feature representation more physiology specific.

During contrastive learning, to fully exploit the negative sampling strategy (**P4**) it would be optimal to have mini-batches with more diverse spectra as more varied samples are more informative and will result in richer features. Thus, random shuffling of the dataset for mini-batch sampling can lead to having similar samples being used as negative samples. To promote variety in each mini-batch, we first use the Tmaps to get rough HR predictions for each sample, and use these rough predictions to order the dataset so that samples have varied HR in each mini-batch. The pre-training and subsequent mini-batch sampling strategies are shown in Fig. 3.

IV. Experiments

We evaluate RS-rPPG on four rPPG datasets PURE [39], OBF [19], MMSE-HR [53] and VIPL-HR [28], based on the recording environment we classify PURE and OBF as controlled due to stable lighting and minimal subject movement, and MMSE-HR and VIPL-HR as challenging. We perform an intra-dataset evaluation on all four datasets, and show that RS-rPPG outperforms other SSL methods and reaches close to supervised performance, especially on the challenging datasets. We perform an extensive cross dataset evaluation comparing our method directly to Contrast-Phys [40] as it is a recent top performing SSL method, that shares similarities with RS-rPPG, as it is also contrastive and heavily relies on exploiting inter-data differences. We additionally propose a new mixed dataset test by adding data from a challenging dataset to a controlled dataset. Thus, simulating an unknown unlabeled dataset containing non-uniform samples with both controlled and challenging samples. We also provide an ablation study to analyze crucial framework components, perform a demographic based test on the OBF [19] data to evaluate potential skin tone bias and test the framework on shorter length inputs.

A. Experimental Setup

Datasets: 1) Controlled datasets: PURE [39] is a rPPG dataset containing 60 one-minute-long videos from 10 subjects under ambient lighting with small motion tasks (steady, talking, slow translation, fast translation, slow rotation, medium rotation). OBF [19] contains 200 five-minute-long RGB videos recorded from 100 varied subjects, captured with stable lighting and minimal movement of the subjects. It contains videos under resting and elevated HR conditions. 2) Challenging datasets: MMSE-HR [53] has 102 videos of length 20-70s recorded with stable lighting from 40 subjects in emotion elicitation experiments. It contains challenging motions as there are spontaneous facial expressions, and head motions. VIPL-HR [28] contains 2,378 RGB videos of 20-30s length recorded from 108 subjects. It was recorded in a challenging environment with different devices, varied and unstable frame rates, large movements and variable lighting. It contains many sources of environmental noise, making HR estimation very challenging.

Metrics: We follow previous works [13], [40] by using absolute error (MAE), root-mean-square error (RMSE) and Pearson's correlation coefficient (R).

Training: We choose $T = 576$ (19.2s at 30fps) due to VIPL-HR and MMSE containing many videos around 20s long, for computational convenience, and to have a long enough temporal context to benefit from self-attention. The input videos are pre-processed as shown in Fig. 3a, resulting in $[3, 64, 576]$ sized MSTmaps and Tmaps. For the network F, we utilise the official Swin-Unet implementation and change only the following parameters from the default ones to fit our task: img_size=(64, 576), num_classes=3, window_size=4, mlp_ratio=2. Firstly, MSTmap (X) to Tmap (X') pre-training is run for 10 epochs, in this step the augmented Tmaps are used as a pseudo-label for pre-training. Secondly, we sample the mini-batches from the dataset following the strategy shown in Fig. 3b. Thirdly, we perform contrastive training for 30 epochs. Both in pre-training and training, the AdamW optimizer is used with epsilon=1e-8, betas=(0.9, 0.99), lr=5e-5, wd=0.05. We train with batch=4 and m=2. No additional training data augmentation is used, and training samples are 576 frames long with no overlap.

Testing: For fair comparison with other methods in all experiments we use the same validation protocols from previous works, i.e., we use the PURE training-validation split as [24], 3 fold for MMSE-HR [29], 5 fold for VIPL-HR [28] and 10 fold for OBF [49]. All MSTmaps and ground-truth signals are resampled to 30fps. When testing for PURE and OBF 30s segments are evaluated. For MMSE-HR and VIPL-HR experiments we choose to use 20s, as this enables us to use most of the data for evaluation by including shorter videos, providing a more challenging scenario. To calculate the HR prediction, we average the output signals over the ROI and channel dimension and find the highest PSD peak. We only use ground-truth signals for evaluation, as our output signals are taken directly without any labeled re-training necessary. For comparison with Contrast-Phys we faithfully re-implement their method, all the other results come directly from[40] or the corresponding original papers.

B. Results

To the best of our knowledge, RS-rPPG is the first SSL method focused on challenging rPPG data, as lack of robustness is a serious limitation of current SSL methods.

Intra-dataset: In Table I we show the intra-dataset evaluation. We rank the datasets based on their challenge level with PURE [39] (small, good lighting, small movements, resting HR) and OBF [19] (large, varied subjects, good lighting, small movements, elevated HR) being considered controlled datasets and MMSE-HR [53] (small, challenging movement) and VIPL-HR [28] (large, unstable fps, different devices, challenging lighting and movement) considered challenging. We show obvious improvement on OBF and PURE, with overall better performance than other SSL and comparable to supervised. On the more challenging data, RS-rPPG vastly outperforms other SSL methods, and obtains comparable results to supervised learning methods on MMSE-HR, and slightly lower than supervised on VIPL-HR dataset with an RMSE of 10.5.

Cross/mixed-dataset: In Table II we perform an extensive cross and mixed validation with all four datasets and directly compare with Contrast-Phys [40]. RS-rPPG learns physiologically relevant features even from the challenging data, with some larger RMSE in cross testing due to dataset differences,

TABLE I

Intra-dataset evaluation on PURE [39], OBF [19], MMSE-HR [53] and VIPL-HR [28], best per type underlined

Type	Method	PURE [39]			OBF [19]			MMSE-HR [53]			VIPL-HR [28]		
		MAE↓ (bpm)	RMSE↓ (bpm)	R↑	MAE↓ (bpm)	RMSE↓ (bpm)	R↑	MAE↓ (bpm)	RMSE↓ (bpm)	R↑	MAE↓ (bpm)	RMSE↓ (bpm)	R↑
Traditional	GREEN [43]	-	-	-	-	2.16	0.99	-	-	-	15.9	21.0	0.11
	CHROM [9]	2.07	9.92	0.99	-	2.73	0.98	-	14.0	0.55	11.4	16.9	0.28
	POS [45]	-	-	-	-	1.91	0.99	-	-	-	11.5	17.2	0.30
Supervised	Physnet [49]	2.10	2.60	0.99	-	1.81	0.99	-	-	-	10.8	14.9	0.20
	RhythmNet [29]	-	-	-	-	-	-	-	5.03	0.86	5.30	8.14	0.76
	Dual-GAN [24]	0.82	1.31	0.99	-	-	-	-	-	-	4.93	7.68	0.81
	BVPNet [8]	-	-	-	-	-	-	-	7.47	0.79	5.34	7.85	0.70
	Physformer [51]	-	-	-	-	0.804	0.99	-	-	-	4.97	7.79	0.78
Self-supervised	Gideon2021 [13]	2.30	2.90	0.99	2.83	7.88	0.82	-	-	-	-	-	-
	Contrast-Phys [40]	0.69	1.10	0.99	0.42	1.34	0.98	1.78	6.03	0.86	17.8	22.8	0.17
	Yue2023 [52]	1.23	2.01	0.99	-	-	-	-	-	-	-	-	-
	SiNC [37]	0.61	1.84	0.99	-	-	-	-	-	-	-	-	-
	RS-rPPG (Ours)	0.29	0.59	0.99	0.63	1.32	0.99	1.10	2.34	0.98	5.98	10.5	0.56

TABLE II

Cross and mixed dataset evaluation on PURE [39], OBF [19], MMSE-HR [53] and VIPL-HR [28], best underlined

Test Data	Train Data	RS-rPPG (Ours)			Contrast-Phys [40]		
		MAE↓ (bpm)	RMSE↓ (bpm)	R↑	MAE↓ (bpm)	RMSE↓ (bpm)	R↑
VIPL	MMSE	5.10	9.07	0.71	17.8	24.5	0.21
	OBF	7.94	13.1	0.58	6.95	11.9	0.50
	PURE	9.45	15.1	0.71	16.6	26.9	0.16
MMSE	VIPL	3.42	8.54	0.81	33.3	35.9	0.01
	OBF	1.50	3.27	0.97	5.09	12.3	0.51
	PURE	3.21	9.74	0.78	27.8	40.3	-0.23
OBF	VIPL	1.92	6.08	0.87	17.8	25.3	0.16
	MMSE	0.38	1.01	0.99	16.9	25.3	0.04
	PURE	1.03	3.39	0.95	1.29	4.09	0.94
PURE	VIPL	0.99	1.54	0.99	14.5	19.7	0.26
	MMSE	0.42	0.95	0.99	21.4	31.2	0.01
	OBF	0.65	1.23	0.99	0.57	0.90	0.99
PURE	PURE+VIPL	0.44	0.72	0.99	13.5	19.1	0.19
	PURE+MMSE	0.30	0.68	0.99	0.74	1.07	0.99
OBF	OBF+VIPL	0.63	1.39	0.99	13.6	22.7	0.11
	OBF+MMSE	0.64	1.45	0.99	0.43	1.43	0.98

TABLE III

Ablation study on PURE [39] and VIPL-HR [28]

	PURE			VIPL-HR		
	MAE↓ (bpm)	RMSE↓ (bpm)	R↑	MAE↓ (bpm)	RMSE↓ (bpm)	R↑
RS-rPPG	0.29	0.59	0.99	5.97	10.5	0.56
w BVPNet (P2)	1.58	5.68	0.92	6.91	11.3	0.48
w/o Tmap (P3)	0.25	0.46	0.99	34.5	38.6	0.28
w/o MSTmap to Tmap Pre-train (P3)	1.77	7.55	0.84	39.7	42.5	-0.03
w/o ROI Channel sampling (P5)	0.34	0.80	0.99	7.30	12.0	0.43
w/o Regbw (P6)	0.32	0.79	0.99	6.18	10.58	0.56
w/o Regsp (P7)	0.61	2.02	0.99	8.53	14.3	0.50
w/o ordering Fig. 3b (P4)	1.72	7.99	0.82	7.15	12.3	0.37

as samples from the testing data are not well represented in the training data. Contrast-Phys [40] obtains good performance only when trained on the large and controlled OBF dataset, training on MMSE and VIPL-HR leads it to learn non-physiological noise due to its weak self-supervision. RS-rPPG retains similar performance when mixing samples from challenging datasets into the controlled datasets, while Contrast-Phys becomes unreliable when samples from the challenging VIPL-HR are added to the training data.

Ablation: We perform an ablation study to show the effectiveness of several crucial framework components in Table III. In the ablation experiment related to (**P2**), we simply replace the Swin-Unet [4] backbone with the convolutional alternative BVPNet [8]. For the other experiments we remove the components related to the specific parts of the method (and the corresponding priors). As can be seen , their contribution is much more noticeable with the challenging VIPL-HR data, as without robust learning constraints, the framework fails to learn physiological features from challenging data. In particular the novel traditional augmentation (**P3**) is a crucial for training, as removing it leads to the network learning non-rPPG features on challenging data.

Demographics: As rPPG methods have been shown to be biased towards demographic groups, especially with darker skin tones [1], we conduct a demographics based testing on the OBF [19] dataset. We choose the OBF dataset, as it contains 100 subjects of varied ethnicities and skin tone, we use the annotations provided by the dataset authors that categorise the subjects based on skin tone. The subjects are categorised in three groups of: 31 subjects of lighter skin tone (group 1), 41 subjects with middle-range skin tone (group 2) and 28 subjects of darker skin tone (group 3). We perform cross testing by training on two groups and evaluating on the remaining one, in this way simulating a lack of skin-tone diversity in the training data. In Table IV we first show two traditional methods evaluated on the demographics groups, there is a noticeable difference with the third group that performs significantly worse on both methods with a relative RMSE increase of 222% (CHROM [9]) and 423% (POS [45]) when changing the testing group from 1 to 3. We

TABLE IV

DEMOGRAPHIC CROSS TESTING ON OBF [19] FOR THREE SKIN TONE TESTING GROUPS, WITH RELATIVE INCREASE FROM GROUP 1 TO 3

	Test Group	MAE↓ (bpm)	RMSE↓ (bpm)	R↑
CHROM [9]	1	0.545	1.007	0.998
	2	0.612	1.559	0.992
	3	1.154	3.248	0.979
	1→3	+112%	+222%	-1.90%
POS [45]	1	0.392	0.726	0.999
	2	0.431	1.060	0.999
	3	1.098	3.795	0.970
	1→3	+180%	+423%	-2.90%
Contrast-Phys [40]	1	0.235	0.665	0.999
	2	0.284	0.742	0.998
	3	0.564	1.982	0.992
	1→3	+140%	+198%	-0.70%
RS-rPPG (Ours)	1	0.444	0.801	0.998
	2	0.390	0.806	0.998
	3	0.881	1.913	0.992
	1→3	+98.4%	+139%	-0.60%

TABLE V

SHORTER INPUTS INFERENCE ON PURE [39] AND VIPL-HR [28]

	PURE			VIPL-HR		
	MAE↓ (bpm)	RMSE↓ (bpm)	R↑	MAE↓ (bpm)	RMSE↓ (bpm)	R↑
RS-rPPG	0.29	0.59	0.99	5.97	10.5	0.56
Inference T=256	0.79	3.58	0.97	7.07	11.7	0.47
Inference T=128	1.52	3.20	0.98	8.11	12.7	0.40
Inference T=64	4.78	7.97	0.86	12.0	17.1	0.26

then evaluate Contrast-Phys [40] and RS-rPPG, and notice that compared to traditional methods, they exhibit lower relative increase in RMSE with 198% and 139% respectively, with RS-rPPG showing the most demographically balanced performance. Intuitively, we conclude that both SSL methods do not significantly amplify the demographics bias and that the higher error is most likely due to the poorer SNR ratio for group 3, as the signal strength is highly correlated to the skin tone.

Shorter inputs: We evaluate how RS-rPPG would scale down to predict shorter length clips, as certain applications could require prediction from segments much shorter than 20-30s. We perform this evaluation by varying the T length of non overlapping input segments, and achieve this by padding the input MSTmap to fit the base model's length of 576. In Tab. V we see that performance naturally decreases with smaller T, as the task becomes more challenging. The performance decrease is gradual and consistent with the higher difficulty of the task, with a sharper decline at 64 frames ($\approx 2s$) as the observed time-frame is much shorter and very challenging.

Visualisation: In Fig. 4 we visualize the output of RS-rPPG compared to GREEN [43] and Contrast-Phys [40]. With minimal environmental noise (PURE) even the raw signal from GREEN [43] gives an accurate prediction, but for a more challenging sample (VIPL-HR) the most prominent raw signal peak corresponds to lower frequency noise, which GREEN and Contras-Phys erroneously predict as the HR peak. Due to its strong self-supervision, RS-rPPG ignores the false peak and gives a more accurate output. In Fig. 5, we visualise the features of the network F with t-SNE [42], and compare supervised learning with RS-rPPG self-supervised. The feature distribution between RS-rPPG and standard supervised learning on the same backbone are indistinguishable, meaning that RS-rPPG can obtain the same high quality features as supervised learning but without the need for labels.

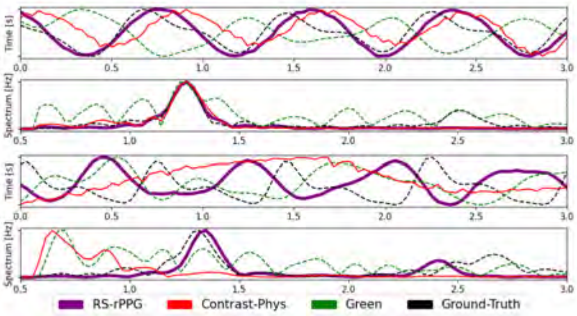

Fig. 4. rPPG extracted using RS-rPPG, Contrast-Phys [40] and Green [43].

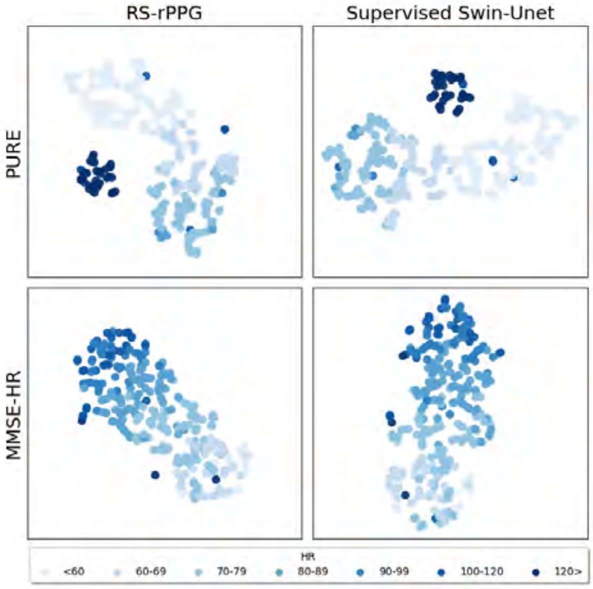

Fig. 5. t-SNE [42] feature visualisation of supervised and RS-rPPG self-supervised learning on the same backbone Swin-Unet backbone F

V. CONCLUSION

We propose RS-rPPG, a robust contrastive self-supervised method that, in contrast to current SSL methods, can reliably learn physiological features even from challenging data. RS-rPPG is carefully constructed on a large set of priors that enable strong self-supervision and greatly outperforms current SSL on challenging data with close to supervised performance. Future work can include investigating unlabeled data from non-rPPG datasets and semi-supervised learning.

References

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

[2] S. Bhattachrjee, H. Li, J. Xia, and W. Xu. Simppg: Self-supervised photoplethysmography-based heart-rate estimation via similarity-enhanced instance discrimination. *Smart Health*, 28:100396, 2023.

[3] L. Birla, S. Shukla, A. K. Gupta, and P. Gupta. Alpine: Improving remote heart rate estimation using contrastive learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5029–5038, 2023.

[4] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

[5] A. Challoner and C. Ramsay. A photoelectric plethysmograph for the measurement of cutaneous blood flow. *Physics in Medicine & Biology*, 19(3):317, 1974.

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

[7] J. H. Cheong, T. Xie, S. Byrne, and L. J. Chang. Py-feat: Python facial expression analysis toolbox. *CoRR*, abs/2104.03509, 2021.

[8] A. Das, H. Lu, H. Han, A. Dantcheva, S. Shan, and X. Chen. Bvpnet: Video-to-bvp signal prediction for remote heart rate estimation. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE, 2021.

[9] G. De Haan and V. Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013.

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

[11] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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

[14] X. Guo, S. Li, J. Yu, J. Zhang, J. Ma, L. Ma, W. Liu, and H. Ling. Pfld: A practical facial landmark detector. *arXiv preprint arXiv:1902.10859*, 2019.

[15] A. K. Gupta, R. Kumar, L. Birla, and P. Gupta. Radiant: Better rppg estimation using signal embeddings and transformer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4976–4986, 2023.

[16] Z. Hasan, A. Z. M. Faridee, M. Ahmed, and N. Roy. Self-rppg: Learning the optical & physiological mechanics of remote photoplethysmography with self-supervision. In *2022 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pages 46–56. IEEE, 2022.

[17] C.-J. Hsieh, W.-H. Chung, and C.-T. Hsu. Augmentation of rppg benchmark datasets: Learning to remove and embed rppg signals via double cycle consistent learning from unpaired facial videos. In *European Conference on Computer Vision*, pages 372–387. Springer, 2022.

[18] M. Lewandowska, J. Rumiński, T. Kocejko, and J. Nowak. Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity. In *2011 federated conference on computer science and information systems (FedCSIS)*, pages 405–410. IEEE, 2011.

[19] X. Li, I. Alikhani, J. Shi, T. Seppanen, J. Junttila, K. Majamaa-Voltti, M. Tulppo, and G. Zhao. The obf database: A large face video database for remote physiological signal measurement and atrial fibrillation detection. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 242–249. IEEE, 2018.

[20] S.-Q. Liu and P. C. Yuen. A general remote photoplethysmography estimator with spatiotemporal convolutional network. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 481–488. IEEE, 2020.

[21] X. Liu, J. Fromm, S. Patel, and D. McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems*, 33:19400–19411, 2020.

[22] X. Liu, B. Hill, Z. Jiang, S. Patel, and D. McDuff. Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5008–5017, 2023.

[23] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[24] H. Lu, H. Han, and S. K. Zhou. Dual-gan: Joint bvp and noise modeling for remote physiological measurement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12404–12413, 2021.

[25] D. McDuff, J. Hernandez, E. Wood, X. Liu, and T. Baltrusaitis. Advancing non-contact vital sign measurement using synthetic avatars. *arXiv preprint arXiv:2010.12949*, 2020.

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

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

[28] X. Niu, H. Han, S. Shan, and X. 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.

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

[30] X. Niu, Z. Yu, H. Han, X. Li, S. Shan, and G. 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.

[31] X. Niu, X. Zhao, H. Han, A. Das, A. Dantcheva, S. Shan, and X. Chen. Robust remote heart rate estimation from face utilizing spatial-temporal attention. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pages 1–8. IEEE, 2019.

[32] X. Niu, X. Zhao, H. Han, A. Das, A. Dantcheva, S. Shan, and X. Chen. Robust remote heart rate estimation from face utilizing spatial-temporal attention. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8, 2019.

[33] C. S. Pilz, S. Zaunseder, J. Krajewski, and V. Blazek. Local group invariance for heart rate estimation from face videos in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1254–1262, 2018.

[34] M.-Z. Poh, D. J. McDuff, and R. W. Picard. Advancements in non-contact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, 58(1):7–11, 2010.

[35] A. Revanur, A. Dasari, C. S. Tucker, and L. A. Jeni. Instantaneous physiological estimation using video transformers. In *Multimodal AI in healthcare: A paradigm shift in health intelligence*, pages 307–319. Springer, 2022.

[36] M. Savic and G. Zhao. Physu-net: Long temporal context transformer for rppg with self-supervised pre-training, 2024.

[37] J. Speth, N. Vance, P. Flynn, and A. Czajka. Non-contrastive unsupervised learning of physiological signals from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14464–14474, 2023.

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

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

[40] Z. Sun and X. Li. Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel,*

October 23–27, 2022, Proceedings, Part XII, pages 492–510. Springer, 2022.

[41] Y.-Y. Tsou, Y.-A. Lee, and C.-T. Hsu. Multi-task learning for simultaneous video generation and remote photoplethysmography estimation. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[42] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[43] W. Verkruysse, L. O. Svaasand, and J. S. Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008.

[44] H. Wang, E. Ahn, and J. 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.

[45] W. Wang, A. C. den Brinker, S. Stuijk, and G. De Haan. Algorithmic principles of remote PPG. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2016.

[46] W. Wang, S. Stuijk, and G. De Haan. Exploiting spatial redundancy of image sensor for motion robust rppg. *IEEE transactions on Biomedical Engineering*, 62(2):415–425, 2014.

[47] Y. Yang, X. Liu, J. Wu, S. Borac, D. Katabi, M.-Z. Poh, and D. McDuff. Simper: Simple self-supervised learning of periodic targets. In *International Conference on Learning Representations*, 2023.

[48] Z. Yu, X. Li, X. Niu, J. Shi, and G. Zhao. Autohr: A strong end-to-end baseline for remote heart rate measurement with neural searching. *IEEE Signal Processing Letters*, 27:1245–1249, 2020.

[49] Z. Yu, X. Li, and G. Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In *British Machine Vision Conference*, 2019.

[50] Z. Yu, W. Peng, X. Li, X. Hong, and G. Zhao. Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 151–160, 2019.

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

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

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