Self-Supervised Representation Learning Framework for Remote Physiological Measurement Using Spatiotemporal Augmentation Loss

Hao Wang, [1] **Euijoon Ahn,** [1,2,3*] **Jinman Kim** [1,2]

[1]School of Computer Science, the University of Sydney, Australia
[2]Telehealth and Technology Centre, Nepean and Blue Mountains Local Health District, Australia
[3]College of Science & Engineering, James Cook University, Australia
hwan7885@uni.sydney.edu.au, euijoon.ahn@jcu.edu.au, jinman.kim@sydney.edu.au

Abstract

Recent advances in supervised deep learning methods are enabling remote measurements of photoplethysmography-based physiological signals using facial videos. The performance of these supervised methods, however, are dependent on the availability of large labelled data. Contrastive learning as a self-supervised method has recently achieved state-of-the-art performances in learning representative data features by maximising mutual information between different augmented views. However, existing data augmentation techniques for contrastive learning are not designed to learn physiological signals from videos and often fail when there are complicated noise and subtle and periodic colour/shape variations between video frames. To address these problems, we present a novel self-supervised spatiotemporal learning framework for remote physiological signal representation learning, where there is a lack of labelled training data. Firstly, we propose a landmark-based spatial augmentation that splits the face into several informative parts based on the Shafer's dichromatic reflection model to characterise subtle skin colour fluctuations. We also formulate a sparsity-based temporal augmentation exploiting Nyquist–Shannon sampling theorem to effectively capture periodic temporal changes by modelling physiological signal features. Furthermore, we introduce a constrained spatiotemporal loss which generates pseudo-labels for augmented video clips. It is used to regulate the training process and handle complicated noise. We evaluated our framework on 3 public datasets and demonstrated superior performances than other self-supervised methods and achieved competitive accuracy compared to the state-of-the-art supervised methods. Code is available at https://github.com/Dylan-H-Wang/SLF-RPM.

Introduction

Physiological signals are critical indicators for human cardiovascular activities such as heart rate (HR), respiration frequency (RF), heart rate variability (HRV) and blood pressure (BP) (Xu, Yu, and Wang 2021). These signals are commonly used to monitor the wellness of patients (Avram et al. 2019). Traditionally, the Electrocardiography (ECG) and Photoplethysmography (PPG) are used to measure physiological signals, and both of them rely on the availability of cuff-based equipment which requires direct contact to human skin. This constrains the application of monitoring and estimation process in an unobtrusive and concomitant way with ubiquitous devices (e.g., smartphone cameras, webcams) (McDuff et al. 2015). In recent years, non-contact video-based remote physiological measurement (RPM) has been of great interest. Remote photoplethysmography (rPPG), using facial videos, has been introduced to overcome the limitation of conventional contact-based measurement approaches. In rPPG, signals are measured based on colour fluctuations on human skin, which are caused by the variations of blood volume during cardiac cycle (De Haan and Jeanne 2013).

Several recent studies using supervised deep learning methods (Yu et al. 2019; Qiu et al. 2019; Niu et al. 2020b; Lee, Chen, and Lee 2020; Li et al. 2018) have shown promising results to remotely estimate physiological signals. They, however, remain problematic because of their dependency on the availability of large-scale labelled training data. The annotation of large-scale data is costly, slow and requires medical equipment. Researchers have employed various approaches to help solve these challenges including transferring knowledge across different domains and fine-tuning those knowledge with a relatively smaller amount of labelled image data. For example, Niu et al. (2020a) used a model that was pre-trained using large labelled natural images (i.e., ImageNet) and fine-tuned this knowledge using rPPG facial videos. Another approach is to create synthetic physiological signals to increase the number of training videos (Niu et al. 2018a; Condrea, Ivan, and Leordeanu 2020). However, such approach is limited by the domain shift between original and synthetic data. An alternative approach is to use self-supervised learning (SSL) to learn and extract image features from unlabelled data. Many recent SSL methods commonly applied the concept of contrastive learning and have achieved state-of-the-art (SOTA) performances in unsupervised image/video representation learning (Ahn et al. 2020; Oord, Li, and Vinyals 2018; Han, Xie, and Zisserman 2019; Qian et al. 2020; Ahn, Feng, and Kim 2021). In these approaches, videos are transformed using standard data augmentation techniques such as frame cropping, resizing, colour jittering and frame re-ordering to produce different views. Invariant video features are then learnt by maximising mutual information between different views in a contrastive manner. These standard augmentation techniques, however, are mainly limited to learn features for ac-

[*]This work was done at the University of Sydney

tion recognition tasks, where large variations among the human anatomy can be modelled, and they are not designed to capture the subtle colour fluctuations on human skin. Many of RPM studies also transformed the video frames (3D) into 2D spatiotemporal map for subsequent 2D convolutional neural networks (CNNs) training (Niu et al. 2020a; Qiu et al. 2019; Lee, Chen, and Lee 2020). However, this transformation potentially neglect nature information contained in original inputs (Yu et al. 2020).

In this paper, we present a new Self-supervised Learning Framework for Remote Physiological Measurement (SLF-RPM). We propose a landmark-based spatial augmentation using Shafer's dichromatic reflection model (Wang et al. 2017) to effectively capture the colour fluctuations on human faces. We also propose a sparsity-based temporal augmentation that characterise periodic colour variations using Nyquist–Shannon sampling theorem (Nyquist 1928) to exploit rPPG signal features. We further formulate a new loss function using the pseudo-labels derived from our augmentations. It regulates the training process of contrastive learning and handles complicated noise. We evaluated our framework by comparing with other SOTA supervised and SSL approaches using 3 public datasets and conducted ablation studies to demonstrate the effectiveness of our SLF-RPM framework.

Related Work

Remote Physiological Measurement

The application of analysing rPPG from camera-captured videos was first proposed by Verkruysse et al. (2008). In early work, many studies have manually designed hand-crafted signal features to characterise the rPPG signals. For example, Poh et al. (2010b; 2010a) used independent component analysis (ICA) with RGB colour sequences to estimate HR signals. Similarly, some methods used chrominance features reflected from the human skin (Wang, Stuijk, and De Haan 2014; Wang et al. 2016). Although these hand-crafted features have shown promising performance, they are required to manually select region of interest (ROI), detect and process skin-pixels signals. This is challenging or even quixotic to be implemented in practical settings. In recent years, deep learning methods based on CNNs (Tulyakov et al. 2016; Hsu, Ambikapathi, and Chen 2017; Niu et al. 2018a) have been developed to overcome such limitations and they have been shown to effectively capture minor colour variations and extract rPPG signals. For example, Niu et al. (2020a) proposed RhythmNet to transform each video clip into a 2D feature map and fed it into a CNN to estimate rPPG signals. Yu et al. (2019) proposed an end-to-end CNN model, i.e, rPPGNet to reconstruct rPPG signals from highly-compressed videos. Lee et al. (2020) applied the concept of meta-learning using 2D CNN coupled with bidirectional long short-term memory to learn spatiotemporal features and enable faster inference adaption. Moreover, Lu et al.(2021) proposed Dual-GAN to model rPPG and noise signals to improve model robustness. The performance of these methods, however, were dependent on the labelled training data. Several unsupervised approaches have been proposed to tackle this problem. Bobbia et al. (2019) introduced an unsupervised skin selection method based on the pulsatility feature to detect rPPG. Similarly, Condrea et al. (2020) proposed an unsupervised LSTM to learn rPPG from synthetic data.

Self-Supervised Video Representations Learning

In recent years, self-supervised video representations learning methods have achieved promising results on action recognition task (Han, Xie, and Zisserman 2019, 2020a; Jing and Tian 2018). For instance, Misra et al.(2016) learnt video representations by classifying shuffled video clips. Similarly, Jenni et al.(2020) generated a set of temporal transformations and constructed a 3D CNN to recognise them. The essential concept with these approaches is to use data augmentations to learn invariant features along spatial and temporal axes. These approaches primarily dealt with object interactions, optical flows, synchronised audios and object tracking. They, however, were not designed to capture the subtle facial colour changes for estimating rPPG signals.

Data Augmentation

Spatial augmentations[1] are commonly used in both supervised and SSL to cover a wider and diverse data distribution, and they have shown to be effective in extracting discriminative image features (Shorten and Khoshgoftaar 2019; Han, Xie, and Zisserman 2020b; Chen et al. 2020b). However, existing RPM studies (Yu, Li, and Zhao 2019; Qiu et al. 2019; Spetlik et al. 2018; Niu et al. 2018b; Chen and McDuff 2018) did not apply spatial augmentations to frames directly, and the most close one is proposed by Niu et al. (2020b) which horizontally and vertically flip the processed *spatiotemporal map* (a representation to compact the 3D video into 2D).

Motion statistics are essential in representing video information and play a key factor on the success of action recognition (Wang et al. 2019). These video dynamics are closely related with temporal information where each action class has a common pattern that can be observed in subsequent frames (Jenni, Meishvili, and Favaro 2020). For instance, many studies (Brattoli et al. 2017; Fernando et al. 2017; Lee et al. 2017; Xu et al. 2019; Kim, Cho, and Kweon 2019) proposed to use image frame re-ordering or optical flows to augment temporal data. These augmentation techniques were, however, not designed to estimate physiological signals in facial videos.

Method

RPM Representation Learning Framework

The overview of our SLF-RPM is illustrated in Fig. 1. Suppose we have randomly sampled N raw videos from the dataset. We first apply our sparsity-based temporal and landmark-based spatial augmentations to generate $2N$ samples in total. Two augmented clips $c_i^{1'}, c_i^{2'}$ originated from

[1] We also consider appearance transformation (such as colour distortion, Gaussian blur) as spatial augmentation.

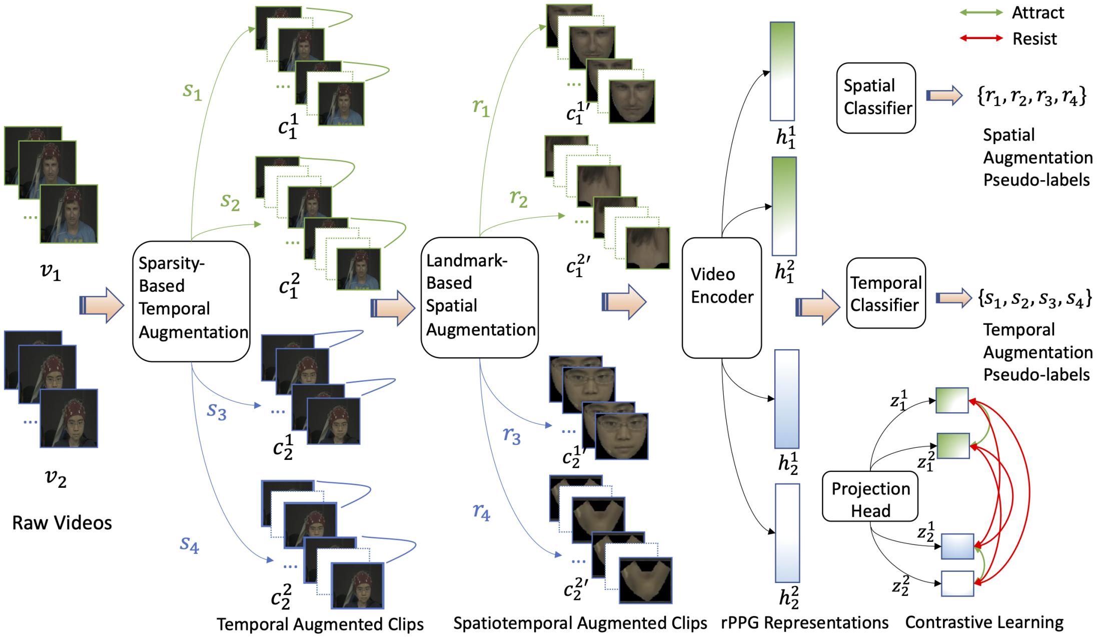

Figure 1: Overview of our SLF-RPM. Raw videos are transformed by the sparsity-based temporal augmentation and landmark-based spatial augmentation to generate different views which are then used in our contrastive learning. Simultaneously, pseudo-labels derived from our augmentations are used to constrain the learning process. Positive samples are denoted by the same subscripts and different superscripts e.g., z_i^1, z_i^2.

the same video v_i are then fed into the *video encoder* to obtain corresponding video features h_i^1, h_i^2. Both of them are then mapped by the *projection head* into the space z_i^1, z_i^2 where contrastive loss (Chen et al. 2020a) is applied to maximise mutual information. These two feature vectors are regarded as positive samples and rest of $2(N-1)$ samples within the same mini-batch are considered as negative samples. We then generated pseudo-labels based on applied augmentations of the video v_i and used two classifiers (i.e., single fully-connected (FC) layer) trained on h_i^1, h_i^2 to identify the transformation $\{s_1, r_1\}$ and $\{s_2, r_2\}$.

Video Encoder and Projection Head

We adopted 3D ResNet architecture (Hara, Kataoka, and Satoh 2018) as the video encoder. The 3D architecture allows to learn spatial and temporal information at the same time. Each input was converted into a flatten feature vector h and then fed into the projection head, which is a multi-layer perceptron (MLP) in our experiment, to obtain the final encoded feature vector (i.e., z in Eq. 5). The projection head is removed during the evaluation process, and the feature vector h from the video encoder is used directly as RPM representations to make final predictions.

Preliminary Background: Skin Reflection Model

According to Shafer's dichromatic reflection model (DRM) (Wang et al. 2017), light source has a constant spectral composition with varying intensities and therefore the variation of skin reflections over the time are measured based on body motions (specular variations) and pulse-induced subtle colour changes (diffuse reflection) where only diffuse reflection contains rPPG-related information. Using DRM, we can then define the skin reflection model for the image sequence along the time by

$$C_k(t) = I(t) \cdot (v_s(t) + v_d(t)) + v_n(t) \tag{1}$$

where $C_k(t)$ is the k-th skin pixel of RGB values; $I(t)$ denotes the light intensity level from the light source which is regulated by specular reflection $v_s(t)$ and diffuse reflection $v_d(t)$; $v_n(t)$ is the noise from camera sensor; t is the time step. We can further decompose $v_s(t)$ and $v_d(t)$ by

$$v_s(t) = u_s \cdot (s_0 + s(t)) \tag{2}$$

where u_s is the unit colour vector of the light spectrum, s_0 and $s(t)$ are the stationary and varying parts of specular reflections, i.e., $s(t)$ captures motions.

$$v_d(t) = u_d \cdot d_0 + u_p \cdot p(t) \tag{3}$$

where u_d denotes the unit colour vector of the skin pixel; d_0 refers to the stationary reflection strength; u_p refers to the relative signal strengths in RGB channels; $p(t)$ refers to the rPPG signal. Given the defined notation above, we can rewrite the Eq. 1 using Eq. 2 and 3 by

$$C_k(t) = I(t) \cdot (u_s \cdot (s_0 + s(t)) + u_d \cdot d_0 + u_p \cdot p(t)) + v_n(t) \tag{4}$$

where our aim is to calculate $p(t)$ from $C_k(t)$.

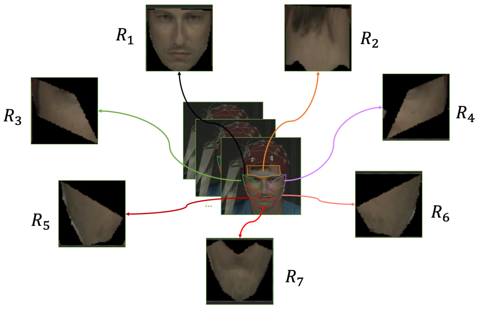

Figure 2: Illustration of landmark-based spatial augmentation. Based on detected facial landmarks, we define 7 ROI areas $\{R_1, R_2, R_3, R_4, R_5, R_6, R_7\}$.

Landmark-Based Spatial Augmentation

We defined two criteria for our landmark-based spatial augmentation: 1) the relationship among colour channels of each pixel needs to be considered so that any pixel value changes (e.g., colour jitter) are not allowed; and 2) the underlying colour variations among different views are similar. From Eq. 4, the rPPG information contained in videos can be maximised by forcing u_d to dominate the equation, and hence the diffuse reflection $v_d(t)$ can be approximated by skin pixel value $C_k(t)$. As such, we selected and cropped each frame into several facial parts according to face landmark locations (Bulat and Tzimiropoulos 2017) as shown in Fig 2. The selection of ROIs considers two factors: 1) the movements of eyes and mouth are more rapid than other parts of face which put more weights on the specular reflection $v_s(t)$ (Li et al. 2014); and 2) facial parts in the same video with similar skin colour $C_k(t)$ should contain similar signals $p(t)$. Theoretically, removal of non-facial areas filters noise from the background which reduces the weight of $v_n(t)$ in Eq. 4. Informative selected ROIs ensure the dominance of diffusion reflections $v_d(t)$ and different ROI sequences from the same facial video contain similar rPPG signals which are used as positive samples (i.e., (z_i, z_j) in Eq. 5) in our contrastive learning. In this paper, we define 7 ROIs including the whole face, forehead, left top cheek, right top cheek, left bottom cheek, right bottom cheek and chin.

Sparsity-Based Temporal Augmentation

The main idea of sparsity-based temporal augmentation is motivated by the physical property of signals as shown in Fig. 3, i.e., Nyquist–Shannon sampling theorem (Nyquist 1928) such that a discrete sequence of samples can reconstruct corresponding continuous-time signal if the band-limit B of the signal is less than the sample rate f_s, i.e., $B < \frac{1}{2} f_s$. In our case, the Frames Per Second (FPS) can be treated as sampling rate, and as long as the FPS is bigger than the Nyquist rate (i.e., 2 times of rPPG signal band-limit), rPPG signal can be extracted from the video clip. As such, we can use different strides to augment each video along the time-axis. For example, the larger strides will generate lower FPSs.

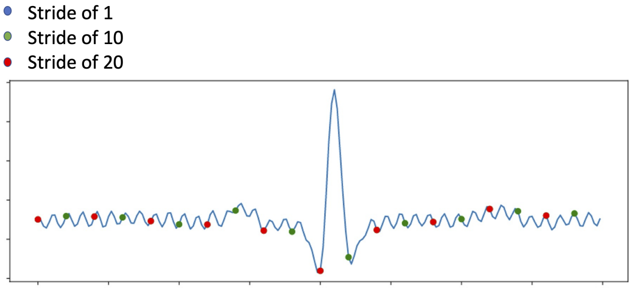

Figure 3: Illustration of signal sparsity. This is an example of 1-second length signal with sampling rate of 256. The *blue* line represents original signal (i.e., stride of 1), the *green* dots represent data points of stride of 10, and the *red* dots represent data points of stride of 20. Note that some red points overlap with green points.

Spatiotemporal Loss with Pseudo-labels

Specific constraints such as inductive bias during model training can help extract representative features. It effectively regulates and generalise the domain information (Argyriou, Evgeniou, and Pontil 2007). As such, we introduce a spatiotemporal loss that regulates the training process of contrastive learning and handles complicated noise. It first generates pseudo-labels for augmented video clips and creates two additional auxiliary classification tasks to predict the pseudo-labels. The pseudo-labels for each of video clip are defined based on the corresponding data augmentations. Suppose that we have a list of ROIs $\{R_1, ..., R_m, ..., R_M\}$ and a list of strides $\{S_1, ..., S_n, ..., S_N\}$. Then, the video clip augmented by R_m and S_n will be labelled as (m, n) which are used as the ground truths of auxiliary classification tasks. According to previous work (Chen et al. 2020a), the mutual information among different views are maximised by

$$\mathcal{L}_{[i,j]} = -\log \frac{\exp\left(sim(z_i, z_j)/\tau\right)}{\sum_{k=0}^{2N} \mathbb{1}_{k \neq i} \exp\left(sim(z_i, z_k)/\tau\right)} \tag{5}$$

where (i, j) is a pair of positive samples, $\mathbb{1}_{k \neq i} \in \{0, 1\}$ is an indicator function which equals to 1 iff $k \neq i$ (i.e., not the feature vector itself) and τ is the temperature hyperparameter. Additionally, we generalise this learning by introducing

$$\mathcal{L}_a(\boldsymbol{y}_a, \hat{\boldsymbol{y}}_a) = -\sum_{i=1}^{\mathrm{C}_M} y_a^i \log(\hat{y}_a^i) \tag{6}$$

where $\boldsymbol{y}_a$ is the predicted ROIs, $\hat{\boldsymbol{y}_a}$ is the target ROIs and C_M is the number of classes, and

$$\mathcal{L}_b(\boldsymbol{y}_b, \hat{\boldsymbol{y}}_b) = -\sum_{i=1}^{\mathrm{C}_N} y_b^i \log(\hat{y}_b^i) \tag{7}$$

where $\boldsymbol{y}_b$ is the predicted strides, $\hat{\boldsymbol{y}_b}$ is the target strides and C_N is the number of classes. Our spatiotemporal loss is the sum of all losses as follows:

$$\mathcal{L}_i = \mathcal{L}_{[i,j]} + \mathcal{L}_a + \mathcal{L}_b \tag{8}$$

Strategy	Method	Dataset	HR (bpm)			
			SD	MAE	RMSE	R
Self-Supervised	DPC	MAHNOB-HCI	11.76	9.16	14.54	-0.35
	MemDPC		10.83	8.23	12.14	0.45
	SeCo		9.48	7.03	10.21	0.67
	SLF-RPM (ours)		4.58	3.60	4.67	0.92
	DPC	VIPL-HR-V2	18.56	14.07	19.20	-0.45
	MemDPC		18.03	13.65	18.12	0.13
	SeCo		16.56	13.32	16.58	0.23
	SLF-RPM (ours)		16.60	12.56	16.59	0.32
	DPC	UBFC-rPPG	11.62	10.60	11.92	-0.32
	MemDPC		12.42	10.85	12.81	0.25
	SeCo		9.74	9.83	10.62	0.58
	SLF-RPM(ours)		9.60	8.39	9.70	0.70
Supervised Learning	DeepPhys (Chen and McDuff 2018)	MAHNOB-HCI	-	4.57	-	-
	STVEN + rPPGNet (Yu et al. 2019)		5.57	4.03	5.93	**0.88**
	AutoHR (Yu et al. 2020)		**4.73**	3.78	5.10	0.86
	Meta-rPPG (Proto+synth) (Lee, Chen, and Lee 2020)		4.90	**3.01**	**3.68**	0.85
	Supervised Baseline (3D ResNet-18)		9.81	7.34	9.76	0.56
	RePSS Team 1 (Li et al. 2020)	VIPL-HR-V2	-	**8.50**	**-**	-
	RePSS Team 5 (Li et al. 2020)		-	12.00	-	-
	Supervised Baseline (3D ResNet-18)		16.69	12.03	16.68	0.37
	POS (Wang et al. 2016)	UBFC-rPPG	10.40	**4.12**	10.5	-
	3D CNN (Bousefsaf, Pruski, and Maaoui 2019)		8.55	5.45	8.64	-
	Meta-rPPG (Proto+synth) (Lee, Chen, and Lee 2020)		**7.12**	5.97	**7.42**	**0.53**
	Supervised Baseline (3D ResNet-18)		9.68	8.08	9.81	**0.53**

Table 1: Linear Evaluation and Supervised Results. The upper section of Table shows the results of SOTA SSL methods and our SLF-RPM on three datasets. The best performing results for each dataset are underlined. The bottom section of Table shows the results of SOTA supervised HR estimation methods and the supervised baseline (i.e., 3D ResNet-18). The best performing results from each dataset are in bold.

Experiment Setup

Datasets We evaluated our RPM framework on HR estimation task by three widely used public datasets: MAHNOB-HCI (Soleymani et al. 2012), VIPL-HR-V2 (Li et al. 2020) and UBFC-rPPG (Bobbia et al. 2019). All videos were re-sampled into 30 FPS and frames were resized into 64×64, and the length of each clip was constrained into 5 seconds.

Metrics The model performance on HR estimation (downstream task of RPM) was measured by comparing metrics of standard deviation (SD), the mean absolute error (MAE), the root mean square error (RMSE) and the Pearson's correlation coefficient (R). During the evaluation process, we adopt subject-exclusive test (Niu et al. 2020a), i.e., subjects in the training set will not appear in the testing set for fair comparisons.

Linear Classification To evaluate the quality of extracted rPPG representations, we followed the common linear classification protocol (Chen et al. 2020a; He et al. 2020) which freezes the weights of self-supervised video encoder layers and trains a subsequent FC layer on the global average pooling features of the video encoder. This evaluation is also used in our ablation studies.

Transfer Learning We firstly pre-trained SLF-RPM without labels, and then fine-tuned the weights of video encoder layers and a linear classifier with labels. We compared our method with 3D ResNet-18 pre-trained using Kinetics-700 (Smaira et al. 2020) and VIPL-HR-V2 (i.e., largest among three datasets) for supervised pre-training.

Augmentation Details Each video was first augmented into two clips using the stride number randomly sampled from the list $\{1, 2, 3, 4, 5\}$, and each augmented clip was constrained to have length of 30-frame. Therefore, the longest clip (i.e., stride of 5) contained 5-second information, while the shortest clip (i.e., stride of 1) had 1-second information. Frames of each clip were then cropped based on a specific ROI by randomly selecting one from 7 predefined facial areas as described in Fig. 2.

We used Adam optimiser (Kingma and Ba 2015) with default settings to update the model weights.

Results and Discussions

Linear Classification Evaluation

The results of linear classification experiment are shown in Table 1. We re-implemented three SOTA video SSL methods including DPC (Han, Xie, and Zisserman 2019), MemDPC (Han, Xie, and Zisserman 2020a) and SeCo (Yao et al. 2020). Our method had the best accuracies relative to other SSL methods on all datasets. SeCo was the closest to our method achieving MAE of 7.03 on MAHNOB-HCI, 13.32 on VIPL-HR-V2 and 9.83 on UBFC-rPPG. We attribute this difference to our augmentation schemes that

Strategy	Method	Pre-training Dataset	Fine-tuning Dataset	HR (bpm)			
				SD	MAE	RMSE	R
Transfer Learning	Supervised-Pre-Training	Kinetics-700	MAHNOB-HCI	11.89	8.98	13.03	-0.07
			VIPL-HR-V2	16.82	12.28	16.86	0.35
			UBFC-rPPG	11.44	9.77	11.79	0.52
		VIPL-HR-V2	MAHNOB-HCI	10.89	8.06	11.90	**0.72**
			UBFC-rPPG	10.87	8.32	10.87	0.61
	SLF-RPM	VIPL-HR-V2	MAHNOB-HCI	**10.19**	**6.23**	**10.35**	0.56
			VIPL-HR-V2	**15.55**	**11.59**	**15.60**	**0.46**
			UBFC-rPPG	**10.19**	**7.35**	**10.53**	**0.63**

Table 2: Transfer Learning Results. This table shows model performances under different pre-training strategies. The transfer abilities of learned representations are evaluated on three datasets. The best performing results from each dataset are in bold.

effectively captured subtle colour fluctuations on facial skin, whereas other SSL methods are limited to learn only from apparent motions of the objects.

Performance comparisons to the supervised methods are also shown in Table 1. Our method on MAHNOB-HCI had a better performance (MAE of 3.60) with a large margin than the Supervised Baseline (MAE of 7.34) and had superior accuracy than previous SOTA Deephys (Chen and McDuff 2018) (MAE of 4.57) and rPPGNet (Yu et al. 2019) (MAE of 4.03). The best performing supervised approach was Meta-rPPG (Lee, Chen, and Lee 2020) (MAE of 3.01) which used unlabelled testing samples for training (i.e., transductive meta-learning). Our method showed a competitive result with Meta-rPPG by using training samples only and achieved higher SD of 4.58 and R of 0.92, suggesting that our method had more consistent predictions and stronger linear correlation between ground truth and predictions. We suggest that our distinctive performance on MAHNOB-HCI over supervised methods is subject to the limitations of the dataset where: 1. number of participants in MAHNOB-HCI is limited (i.e., 27 subjects in total); 2. videos are highly compressed, losing subtle details that human eyes cannot see. This was also reported from other similar work (Yu et al. 2019). Nevertheless, our method implicitly modelled compression-corrupted information patterns and learned robust rPPG signals in a self-supervised manner.

Our SLF-RPM on UBFC-rPPG achieved MAE of 8.39. Supervised methods on UBFC-rPPG had better performances than other SSL methods.This is mainly attributed to less noise on UBFC-rPPG (uncompressed video data), which improved overall supervised learning outcomes. Nevertheless, SLF-RPM reduced the performance gap with Supervised Baseline (MAE of 8.08) to MAE of 0.31.

The performances of all the methods on VIPL-HR-V2 were reduced in comparison to other two datasets. This is because VIPL-HR-V2 has more complex video conditions. Nonetheless, our method (MAE of 12.56) successfully closed the performance gap with the best supervised approach RePSS Team 1 (Li et al. 2020) (MAE of 8.5) and had competitive accuracy with Supervised Baseline (MAE of 12.03).

Transfer Learning

We evaluated the transferable ability of rPPG features extracted by SLF-RPM. The results of the transfer learning are shown in Table 2. Our results indicate that the SLF-RPM improved the estimation of HR when it was used to pre-train CNNs for subsequent supervised fine-tuning. When the available dataset is limited, SLF-RPM can be used as the self-pre-training strategy (i.e., pre-train and fine-tune on the same dataset). This enabled faster adaption for rPPG feature extraction such that MAE was reduced to 11.59 on VIPL-HR-V2 dataset. Moreover, it was most effective when SLF-RPM was applied to pre-train using the largest VIPL-HR-V2 dataset. The performance of Supervised Baseline was improved from MAE of 7.34 to 6.23 on MAHNOB-HCI and from MAE of 8.08 to 7.35 on UBFC-rPPG dataset.We suggest that our self-supervised pre-training acted as an initialisation point for effective supervised fine-tuning, which improved the feature representation of rPPG signals.

In contrast, model pre-trained on Kinetics-700 benchmark, commonly used in action recognition task, degraded HR estimation performance (MAE of 8.98 on MAHNOB-HCI, 12.28 on VIPL-HR-V2 and 9.77 on UBFC-rPPG) compared with the baseline model due to the hard domain shift between tasks. Moreover, we evaluated the model pre-trained with labels using VIPL-HR-V2 and achieved MAE of 8.06 on MAHNOB-HCI and 8.32 on UBFC-rPPG. This demonstrates that our self-supervised pre-training strategy was better at characterising the RPM features than that of supervised pre-training methods.

Ablation Studies

Landmark-based Spatial Augmentation We compared our landmark-based spatial augmentation with 5 other common data augmentation techniques. We applied these techniques to the whole face area (i.e., R_1 of the pre-defined ROI list) and used stride of 1 to avoid temporal augmentation effects. As shown in Table 3, our landmark-based spatial augmentation had the best MAE score of 5.12. Our results show that the *whole face* ROI (R_1) was effective in model training, which improved MAE from 7.98 to 5.12. Among the standard augmentation techniques, *Random Crop and Resize* had the best result with MAE of 6.74. *Random Grayscale* had the worst performance with MAE of 10.66. The combination of all 5 techniques (standard paradigm in simCLR(Chen et al. 2020a)), in fact, reduced the performance giving MAE of 9.54. One possible reason is that although appearance transformations prevent the model from using colour histogram shortcut (Chen et al. 2020a) to distinguish different

Method	HR (bpm) MAE
Random Crop and Resize	6.74
Random Horizontal Flip	6.99
Colour Jitter	8.11
Random Grayscale	10.66
Gaussian Blur	9.81
Combined Above 5 Augmentations	9.54
$\{R_2, R_3, R_4, R_5, R_6, R_7\}$	7.98
$\{R_1, R_3, R_4, R_5, R_6, R_7\}$	6.51
$\{R_1, R_2, R_4, R_5, R_6, R_7\}$	7.03
$\{R_1, R_2, R_3, R_5, R_6, R_7\}$	7.15
$\{R_1, R_2, R_3, R_4, R_6, R_7\}$	6.45
$\{R_1, R_2, R_3, R_4, R_5, R_7\}$	5.82
$\{R_1, R_2, R_3, R_4, R_5, R_6\}$	6.72
$\{R_1, R_2, R_3, R_4, R_5, R_6, R_7\}$	**5.12**

Table 3: Ablation on spatial augmentation. We compared different spatial augmentations and evaluated the impact of each ROI. The best result is in bold.

views, signal-related information contained in colour channels can be also distorted. Another reason could be that geometric transformations cannot guarantee augmented inputs are valuable. They may introduce signal noise by including non-facial areas (Li et al. 2014). Overall, our landmark-based spatial augmentation outperformed standard spatial augmentation techniques by a large margin for the task of HR estimation.

Sparsity-based Temporal Augmentation We evaluated the effectiveness of our sparsity-based temporal augmentation compared to different standard temporal augmentation techniques(Qian et al. 2020; Jenni, Meishvili, and Favaro 2020), and use *whole face* R_1 to avoid spatial augmentation effects. As shown in the bottom part of Table 4, our sparsity-based temporal augmentation follows the Nyquist-Shannon sampling theorem and showed the larger sparsity ranges had under-sampling issue that negatively affect the learning outcome since sparsity range {1,2,3,4,5,6} had worse MAE of 5.92 than the best performing range {1,2,3,4,5} with MAE of 5.28. However, we also noted that we need some enough sparsity (e.g., {1,2,3,4,5}) to have stronger data transformation compared with smaller range {1,2,3,4} achieving MAE of 6.21.

Nevertheless, the proposed sparsity-based temporal augmentation generally outperformed other augmentation techniques. Among these four standard temporal augmentations, *Random Temporal Interval* had the best result with MAE of 5.74 and we attribute this to its ability to characterise the periodic cycle of blood volume changes on facial skin. *Periodic* had the worst MAE score of 7.31 due to the distorted rPPG signals.

Method	HR (bpm) MAE
Random Temporal Interval	5.74
Random Permutation	7.18
Periodic	7.31
Warp	7.22
Sparsity range from {1, 2, 3, 4}	6.21
Sparsity range from {1, 2, 3, 4, 5}	**5.28**
Sparsity range from {1, 2, 3, 4, 5, 6}	5.92

Table 4: Ablation on temporal augmentation. We compared with 4 standard temporal augmentation techniques. The best result is in bold.

Method	HR (bpm) MAE
SLF-RPM without pseudo-labels	4.25
SLF-RPM with pseudo-labels	**3.6**

Table 5: Ablation on spatiotemporal loss with pseudo-labels. We compared model performance with and without pseudo-labels integration. The best result is in bold.

Effect of Spatiotemporal Loss with Pseudo-labels To validate the effectiveness of our spatiotemporal loss, we compared the model performance under two different scenarios, i.e., with and without pseudo-labels integration. From Table 5, our model, by additionally assigning two classification tasks, improved the overall performance achieving the MAE of 3.6. We suggest that this is because pseudo-labels enabled better characterisation of complicated ROIs and subtle temporal changes (i.e., noise).

Conclusion

We present a SSL framework for RPM by introducing novel landmark-based spatial augmentation, sparsity-based temporal augmentation and spatiotemporal loss. Our results showed that the SLF-RPM significantly outperformed other SSL methods and achieved a competitive accuracy compared to other supervised methods on HR estimation task. The superior transfer ability of learnt RPM representations using SLF-RPM demonstrates that it can be used as an effective pre-training strategy for many facial video analysis tasks.

Acknowledgements

This work was supported, in part, by the Telehealth and Technology Centre and the ARC.

References

Ahn, E.; Feng, D.; and Kim, J. 2021. A Spatial Guided Self-supervised Clustering Network for Medical Image Segmentation. *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2021).*

Ahn, E.; Kumar, A.; Fulham, M.; Feng, D.; and Kim, J. 2020. Unsupervised Domain Adaptation to Classify Medical Images Using Zero-Bias Convolutional Auto-Encoders and Context-Based Feature Augmentation. *IEEE Transactions on Medical Imaging*, 39: 2385–2394.

Argyriou, A.; Evgeniou, T.; and Pontil, M. 2007. Convex multi-task feature learning. *Machine Learning*, 73: 243–272.

Avram, R.; Tison, G.; Aschbacher, K.; Kuhar, P.; Vittinghoff, E.; Butzner, M.; Runge, R.; Wu, N.; Pletcher, M.; Marcus, G.; and Olgin, J. 2019. Real-world heart rate norms in the Health eHeart study. *NPJ Digital Medicine*, 2.

Bobbia, S.; Macwan, R.; Benezeth, Y.; Mansouri, A.; and Dubois, J. 2019. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognit. Lett.*, 124: 82–90.

Bousefsaf, F.; Pruski, A.; and Maaoui, C. 2019. 3D Convolutional Neural Networks for Remote Pulse Rate Measurement and Mapping from Facial Video. *Applied Sciences*, 9: 4364.

Brattoli, B.; Büchler, U.; Wahl, A.; Schwab, M.; and Ommer, B. 2017. LSTM Self-Supervision for Detailed Behavior Analysis. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3747–3756.

Bulat, A.; and Tzimiropoulos, G. 2017. How Far are We from Solving the 2D & 3D Face Alignment Problem? (and a Dataset of 230,000 3D Facial Landmarks). *2017 IEEE International Conference on Computer Vision (ICCV)*, 1021–1030.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. E. 2020a. A Simple Framework for Contrastive Learning of Visual Representations. *ArXiv*, abs/2002.05709.

Chen, W.; and McDuff, D. 2018. DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks. *ArXiv*, abs/1805.07888.

Chen, X.; Fan, H.; Girshick, R. B.; and He, K. 2020b. Improved Baselines with Momentum Contrastive Learning. *ArXiv*, abs/2003.04297.

Condrea, F.; Ivan, V.-A.; and Leordeanu, M. 2020. In Search of Life: Learning from Synthetic Data to Detect Vital Signs in Videos. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1207–1216.

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

Fernando, B.; Bilen, H.; Gavves, E.; and Gould, S. 2017. Self-Supervised Video Representation Learning with Odd-One-Out Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5729–5738.

Han, T.; Xie, W.; and Zisserman, A. 2019. Video Representation Learning by Dense Predictive Coding. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 1483–1492.

Han, T.; Xie, W.; and Zisserman, A. 2020a. Memory-augmented Dense Predictive Coding for Video Representation Learning. In *ECCV*.

Han, T.; Xie, W.; and Zisserman, A. 2020b. Self-supervised Co-training for Video Representation Learning. *ArXiv*, abs/2010.09709.

Hara, K.; Kataoka, H.; and Satoh, Y. 2018. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6546–6555.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. B. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9726–9735.

Hsu, G.-S.; Ambikapathi, A.; and Chen, M.-S. 2017. Deep learning with time-frequency representation for pulse estimation from facial videos. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, 383–389. IEEE.

Jenni, S.; Meishvili, G.; and Favaro, P. 2020. Video Representation Learning by Recognizing Temporal Transformations. *ArXiv*, abs/2007.10730.

Jing, L.; and Tian, Y. 2018. Self-supervised Spatiotemporal Feature Learning by Video Geometric Transformations. *ArXiv*, abs/1811.11387.

Kim, D.; Cho, D.; and Kweon, I.-S. 2019. Self-Supervised Video Representation Learning with Space-Time Cubic Puzzles. In *AAAI*.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. *CoRR*, abs/1412.6980.

Lee, E.; Chen, E.; and Lee, C. 2020. Meta-rPPG: Remote Heart Rate Estimation Using a Transductive Meta-Learner. In *ECCV*.

Lee, H.-Y.; Huang, J.-B.; Singh, M. K.; and Yang, M.-H. 2017. Unsupervised Representation Learning by Sorting Sequences. *2017 IEEE International Conference on Computer Vision (ICCV)*, 667–676.

Li, X.; Alikhani, I.; Shi, J.; Seppänen, T.; Junttila, J.; Majamaa-Voltti, K.; Tulppo, M.; and Zhao, G. 2018. The OBF Database: A Large Face Video Database for Remote Physiological Signal Measurement and Atrial Fibrillation Detection. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 242–249.

Li, X.; Chen, J.; Zhao, G.; and Pietikäinen, M. 2014. Remote Heart Rate Measurement from Face Videos under Realistic Situations. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 4264–4271.

Li, X.-B.; Han, H.; Lu, H.; Niu, X.-S.; Yu, Z.; Dantcheva, A.; Zhao, G.; and Shan, S. 2020. The 1st Challenge on Remote Physiological Signal Sensing (RePSS). *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1274–1281.

Lu, H.; Han, H.; and Zhou, S. K. 2021. Dual-GAN: Joint BVP and Noise Modeling for Remote Physiological Measurement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12404–12413.

McDuff, D. J.; Estepp, J. R.; Piasecki, A. M.; and Blackford, E. B. 2015. A survey of remote optical photoplethysmographic imaging methods. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 6398–6404.

Misra, I.; Zitnick, C. L.; and Hebert, M. 2016. Shuffle and Learn: Unsupervised Learning Using Temporal Order Verification. In *ECCV*.

Niu, X.; Han, H.; Shan, S.; and Chen, X. 2018a. SynRhythm: Learning a Deep Heart Rate Estimator from General to Specific. *2018 24th International Conference on Pattern Recognition (ICPR)*, 3580–3585.

Niu, X.; Han, H.; Shan, S.; and Chen, X. 2018b. VIPL-HR: A Multi-modal Database for Pulse Estimation from Less-constrained Face Video. In *ACCV*.

Niu, X.; Shan, S.; Han, H.; and Chen, X. 2020a. RhythmNet: End-to-End Heart Rate Estimation From Face via Spatial-Temporal Representation. *IEEE Transactions on Image Processing*, 29: 2409–2423.

Niu, X.; Yu, Z.; Han, H.; Li, X.; Shan, S.; and Zhao, G. 2020b. Video-based Remote Physiological Measurement via Cross-verified Feature Disentangling. In *ECCV*.

Nyquist, H. 1928. Certain Topics in Telegraph Transmission Theory. *Transactions of the American Institute of Electrical Engineers*, 47(2): 617–644.

Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation Learning with Contrastive Predictive Coding. *ArXiv*, abs/1807.03748.

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

Poh, M.-Z.; McDuff, D. J.; and Picard, R. W. 2010b. Noncontact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 18(10): 10762–10774.

Qian, R.; Meng, T.; Gong, B.; Yang, M.-H.; Wang, H.; Belongie, S. J.; and Cui, Y. 2020. Spatiotemporal Contrastive Video Representation Learning. *arXiv: Computer Vision and Pattern Recognition*.

Qiu, Y.; Liu, Y.; Arteaga-Falconi, J.; Dong, H.; and Saddik, A. E. 2019. EVM-CNN: Real-Time Contactless Heart Rate Estimation From Facial Video. *IEEE Transactions on Multimedia*, 21: 1778–1787.

Shorten, C.; and Khoshgoftaar, T. 2019. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6: 1–48.

Smaira, L.; Carreira, J.; Noland, E.; Clancy, E.; Wu, A.; and Zisserman, A. 2020. A Short Note on the Kinetics-700-2020 Human Action Dataset. *ArXiv*, abs/2010.10864.

Soleymani, M.; Lichtenauer, J.; Pun, T.; and Pantic, M. 2012. A Multimodal Database for Affect Recognition and Implicit Tagging. *IEEE Transactions on Affective Computing*, 3: 42–55.

Spetlik, R.; Franc, V.; Cech, J.; and Matas, J. 2018. Visual Heart Rate Estimation with Convolutional Neural Network. In *BMVC*.

Tulyakov, S.; Alameda-Pineda, X.; Ricci, E.; Yin, L.; Cohn, J. F.; and Sebe, N. 2016. Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2396–2404.

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

Wang, J.; Jiao, J.; Bao, L.; He, S.; Liu, Y.; and Liu, W. 2019. Self-Supervised Spatio-Temporal Representation Learning for Videos by Predicting Motion and Appearance Statistics. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4001–4010.

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

Wang, W.; den Brinker, A. C.; Stuijk, S.; and de Haan, G. 2017. Algorithmic Principles of Remote PPG. *IEEE Transactions on Biomedical Engineering*, 64: 1479–1491.

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

Xu, D.; Xiao, J.; Zhao, Z.; Shao, J.; Xie, D.; and Zhuang, Y. 2019. Self-Supervised Spatiotemporal Learning via Video Clip Order Prediction. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10326–10335.

Xu, Z.; Yu, B.; and Wang, F. 2021. Chapter 4 - Artificial intelligence/machine learning solutions for mobile and wearable devices. In Syed-Abdul, S.; Zhu, X.; and Fernandez-Luque, L., eds., *Digital Health*, 55–77. Elsevier. ISBN 978-0-12-820077-3.

Yao, T.; Zhang, Y.; Qiu, Z.; Pan, Y.; and Mei, T. 2020. SeCo: Exploring Sequence Supervision for Unsupervised Representation Learning. *ArXiv*, abs/2008.00975.

Yu, Z.; Li, X.; Niu, X.; Shi, J.; and Zhao, G. 2020. AutoHR: A Strong End-to-End Baseline for Remote Heart Rate Measurement With Neural Searching. *IEEE Signal Processing Letters*, 27: 1245–1249.

Yu, Z.; Li, X.-B.; and Zhao, G. 2019. Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks. In *BMVC*.

Yu, Z.; Peng, W.; Li, X.-B.; Hong, X.; and Zhao, G. 2019. Remote Heart Rate Measurement From Highly Compressed Facial Videos: An End-to-End Deep Learning Solution With Video Enhancement. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 151–160.