Non-Contrastive Unsupervised Learning of Physiological Signals from Video

Jeremy Speth, Nathan Vance, Patrick Flynn, Adam Czajka
University of Notre Dame
{jspeth,nvance1,flynn,aczajka}@nd.edu

Abstract

Subtle periodic signals such as blood volume pulse and respiration can be extracted from RGB video, enabling non-contact health monitoring at low cost. Advancements in remote pulse estimation – or remote photoplethysmography (rPPG) – are currently driven by deep learning solutions. However, modern approaches are trained and evaluated on benchmark datasets with ground truth from contact-PPG sensors. We present the first non-contrastive unsupervised learning framework for signal regression to mitigate the need for labelled video data. With minimal assumptions of periodicity and finite bandwidth, our approach discovers the blood volume pulse directly from unlabelled videos. We find that encouraging sparse power spectra within normal physiological bandlimits and variance over batches of power spectra is sufficient for learning visual features of periodic signals. We perform the first experiments utilizing unlabelled video data not specifically created for rPPG to train robust pulse rate estimators. Given the limited inductive biases and impressive empirical results, the approach is theoretically capable of discovering other periodic signals from video, enabling multiple physiological measurements without the need for ground truth signals.

1. Introduction

Camera-based vitals estimation is a rapidly growing field enabling non-contact health monitoring in a variety of settings [23]. Although many of the signals avoid detection from the human eye, video data in the visible and infrared ranges contain subtle intensity changes caused by physiological oscillations such as blood volume and respiration. Significant remote photoplethysmography (rPPG) research for estimating the cardiac pulse has leveraged supervised deep learning for robust signal extraction [7, 18, 30, 38, 51, 52]. While the number of successful approaches has rapidly increased, the size of benchmark video datasets with simultaneous vitals recordings has remained relatively stagnant.

Robust deep learning-based systems for deployment require training on larger volumes of video data with di-

verse skin tones, lighting, camera sensors, and movement. However, collecting simultaneous video and physiological ground truth with contact-PPG or electrocardiograms (ECG) is challenging for several reasons. First, many hours of high quality videos is an unwieldy volume of data. Second, recording a diverse subject population in conditions representative of real-world activities is difficult to conduct in the lab setting. Finally, synchronizing contact measurements with video is technically challenging, and even contact measurements used for ground truth contain noise.

Fortunately, recent works find that contrastive unsupervised learning for rPPG is a promising solution to the data scarcity problem [13, 42, 45, 50]. We extend this research line into *non-contrastive* unsupervised learning to discover periodic signals in video data. With end-to-end unsupervised learning, collecting more representative training data to learn powerful visual features is much simpler, since only video is required without associated medical information.

In this work, we show that non-contrastive unsupervised learning is especially simple when regressing rPPG signals. We find weak assumptions of periodicity are sufficient for learning the minuscule visual features corresponding to the blood volume pulse from unlabelled face videos. The loss functions can be computed in the frequency domain over batches without the need for pairwise or triplet comparisons. Figure 1 compares the proposed approach with supervised and contrastive unsupervised learning approaches.

This work creates opportunities for scaling deep learning models for camera-based vitals and estimating periodic or quasi-periodic signals from unlabelled data beyond rPPG. Our **novel contributions** are:

1. A general framework for physiological **si**gnal estimation via **n**on-**c**ontrastive unsupervised learning (SiNC) by leveraging periodic signal priors.

2. The first **non-contrastive** unsupervised learning method for camera-based vitals measurement.

3. The first experiments and results of training with **non-rPPG-specific video data** without ground truth vitals.

Source code to replicate this work is available at https://github.com/CVRL/SiNC-rPPG.
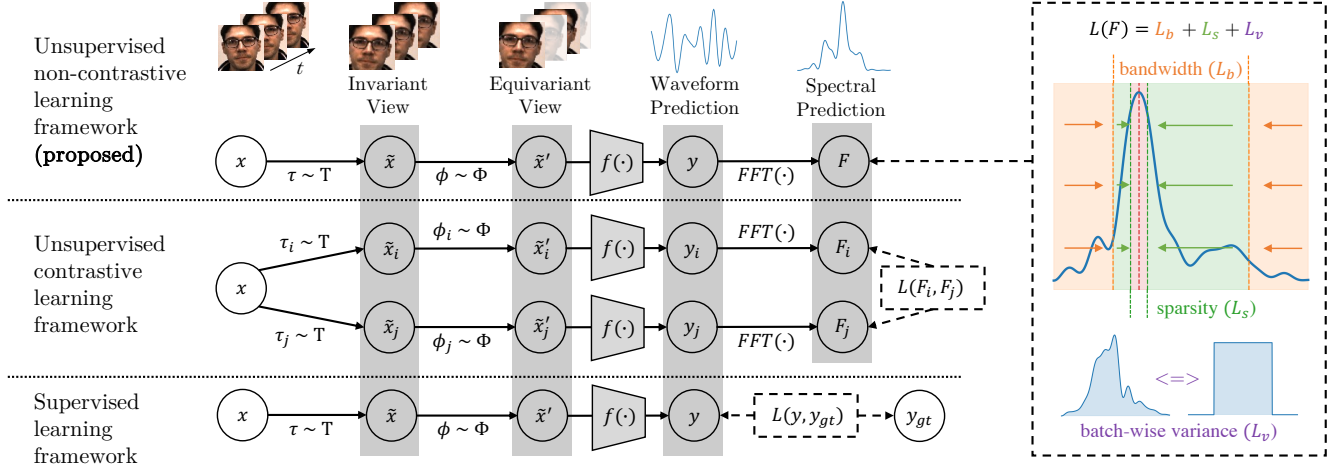
Figure 1. Overview of the SiNC framework for rPPG compared with traditional supervised and unsupervised learning. Supervised and contrastive losses use distance metrics to the ground truth or other samples. Our framework applies the loss directly to the prediction by shaping the frequency spectrum, and encouraging **variance over a batch** of inputs. Power outside of the **bandlimits** is penalized to learn invariances to irrelevant frequencies. Power within the bandlimits is encouraged to be **sparsely** distributed near the peak frequency.

2. Related Work

2.1. Remote Photoplethysmography (rPPG)

Approaches for remote pulse estimation have shifted over the last decade from blind source separation [34, 35], through linear color transformations [9, 10, 46, 47] to training supervised deep learning-based models [7, 16, 18, 20, 29, 30, 38, 51, 52, 55]. While the color transformations generalize well across datasets, deep learning-based models give better accuracy when tested on data from a similar distribution to the training set. To this end, deep learning research has focused on optimizing architectures for extracting robust spatial and temporal features from the limited datasets.

To get around the data bottleneck, large synthetic datasets have recently been proposed [24, 49]. The SCAMPS dataset [24] contains 2,800 videos of synthetic avatars with corresponding PPG, EKG, respiration, and facial action units. The UCLA-synthetic dataset [49] contains 480 videos, which are shown to improve performance when trained with real data. A strength of synthetic datasets is their ability to cover the broad range of skin tones compared to physical data. Another solution to the lack of physiological data is unsupervised learning, where a large set of videos and periodic priors on the output signal is sufficient [13, 42, 45, 50]. We discuss these methods in Sec. 2.3.

2.2. Unsupervised Learning

Self-supervised learning is progressing for image representation learning, with two main competing classes of approaches: contrastive and non-contrastive (or regularized) learning [3]. Contrastive approaches [5, 6, 27] define criteria for distinguishing whether two or more samples are

the same or different. Non-contrastive methods augment positive pairs, and enforce variance in the predictions over batches to avoid *collapse*, in which the model's embeddings reside in a small subspace of the feature space [3]. Distillation methods only use positive samples and avoid collapse by applying a moving average and stop-gradient operator [8, 14]. Another approaches maximize information content of embeddings [11, 53].

2.3. Unsupervised Learning for rPPG

All existing unsupervised rPPG approaches are contrastive [13, 42, 45, 50], in which the same model is fed pairs of input videos, and the predictions over similar videos are pulled closer, while the predictions from dissimilar videos are repelled. Gideon *et al.* [13] were the first to train a deep learning model without labels for rPPG using the contrastive framework. The core of their approach is frequency resampling to create negative samples. Although spatially similar, a resampled video clip contains a different underlying pulse rate, so the model must learn to attend to the temporal dynamics. For their distance function between pairs, they calculated the mean square error between power spectral densities of the model's waveform predictions. While their approach learns to estimate the pulse, their formulation with negative samples is imprecise. The resampled frequency for the negative sample is known, so the relative translation of the power spectrum from the anchor sample can be directly computed. Thus, rather than repelling the estimated spectra, it is more accurate to penalize differences from the known spectra. Furthermore, resampling close to the original sampling rate causes overlap in the power spectra, so repelling the pair is inaccurate. Yuzhe *et al.* [50]

incorporated the previously ignored resampling factor for a soft modification to the InfoNCE loss [33] that scales the desired similarity between pairs by their relative sampling rate. A downside is that their learning framework is not end-to-end unsupervised and requires fine-tuning with PPG labels after the self-supervised stage.

Differently from [13], Contrast-Phys [42] and SLF-RPM [45] consider all samples different from the anchor to be negatives. This assumes that the power spectra will vary between subjects or sufficiently long windows for the same subject. This runs into similar issues with negative pairs as Gideon's approach. Different subjects may have the same pulse rate, so punishing the model for predicting similar frequencies is common during training. Furthermore, the Fast Fourier Transform (FFT) does not produce perfectly sparse decompositions, resulting in spectral overlap even if the heart rate differs by several beats per minute (bpm). As an example, the last column of the second row in Fig. 2 shows the nulls of the main lobe are nearly 30 bpm apart.

3. Method

We first formulate signal regression from video. A video sample $x_i \in \mathbb{R}^{T \times W \times H \times C}$ sampled from a dataset $\mathcal{D}$ consists of T images of size $W \times H$ pixels across C channels, captured over time. State-of-the-art methods offer models f that regress a waveform $\mathbb{R}^T \ni y_i = f(x_i)$ of the same length as the video. Recently, the task has been effectively modeled end-to-end with the models f being spatiotemporal neural networks [16, 21, 38, 51, 52]. While most previous works are supervised and minimize the loss to a contact pulse measurement, we perform non-contrastive learning using only the model's estimated waveform.

The key realization is that we can place strong priors on the estimated pulse regarding its bandwidth and periodicity. Observed signals outside the desired frequency range are pollutants, so penalizing the model for carrying them through the forward pass results in invariances to such noisy visual features. We find that it is surprisingly easy to impose the desired constraints in the frequency domain. Thus, all waveforms are transformed into their discrete Fourier components with the FFT before computing all losses in our approach, $F = \text{FFT}(y)$. The following sections describe the loss functions and augmentations used during training.

3.1. Losses

One of the advantages of unsupervised learning for periodic signals is that we can constrain the solution space significantly. For physiological signals such as respiration and blood volume pulse, we know the healthy upper and lower bounds of the frequencies. We also desire the extracted signal to be sparse in the frequency domain, and that our model filters out noise signals present in the video. With these constraints, we can greatly simplify the problem of finding good features for the desired signal in the data.

Bandwidth Loss. One of the most powerful constraints we can place on the model is frequency bandlimits. Past unsupervised methods have used the irrelevant power ratio (IPR) as a validation metric [12, 13, 42] for model selection. We find that it is also effective during model training. The IPR penalizes the model for generating signals outside the desired bandlimits. With lower and upper bandlimits of a and b, respectively, our bandwidth loss becomes:

$$L_b = \frac{1}{\sum\limits_{i=-\infty}^{\infty} F_i} \left[\sum_{i=-\infty}^{a} F_i + \sum_{i=b}^{\infty} F_i \right], \quad (1)$$

where F_i is the power in the ith frequency bin of the predicted signal. This simple loss enforces learning of many invariants, such as movement from respiration, talking, or facial expressions which typically occupy low frequencies. In our experiments we specify the limits as $a = 0.\overline{66}$ Hz to $b = 3$ Hz, which corresponds to a common pulse rate range from 40 bpm to 180 bpm. The first column of Fig. 2 shows the result of training exclusively with the bandwidth loss L_b. The last row shows that the model concentrates signal power between the bandlimits.

Sparsity Loss. The pulse rate is the most common physiological marker associated with the blood volume pulse. Since we are primarily interested in the frequency, we can further improve our model by preventing wideband predictions. This also reveals the true signal we aim to discover by ignoring visual dynamics that are not strongly periodic.

We penalize energy within the bandlimits that is not near the spectral peak:

$$L_s = \frac{1}{\sum\limits_{i=a}^{b} F_i} \left[\sum_{i=a}^{F^*-\Delta_F} F_i + \sum_{i=F^*+\Delta_F}^{b} F_i \right], \quad (2)$$

where $F^* = \text{argmax}(F)$ and Δ_F are the frequencies of the spectral peak and padding around the peak, respectively. For all experiments $\Delta_F = 6$ beats per minute [32]. Figure 2 shows the result of training only with the sparsity loss in the second column. For the whole dataset, the power spectrum is sparsely distributed in the low frequencies, effectively filtering frequencies higher than 1 Hz.

Variance Loss. One of the risks of non-contrastive methods is the model collapsing into trivial solutions and making predictions independently of the input features. In regularized methods such as VICReg [3], a hinge loss on the variance over a batch of predictions is used to enforce diverse outputs. We use a similar strategy to avoid model collapse, but instead spread the variance in power spectral densities towards a uniform distribution over the desired frequency band. Our variance loss processes a uniform prior
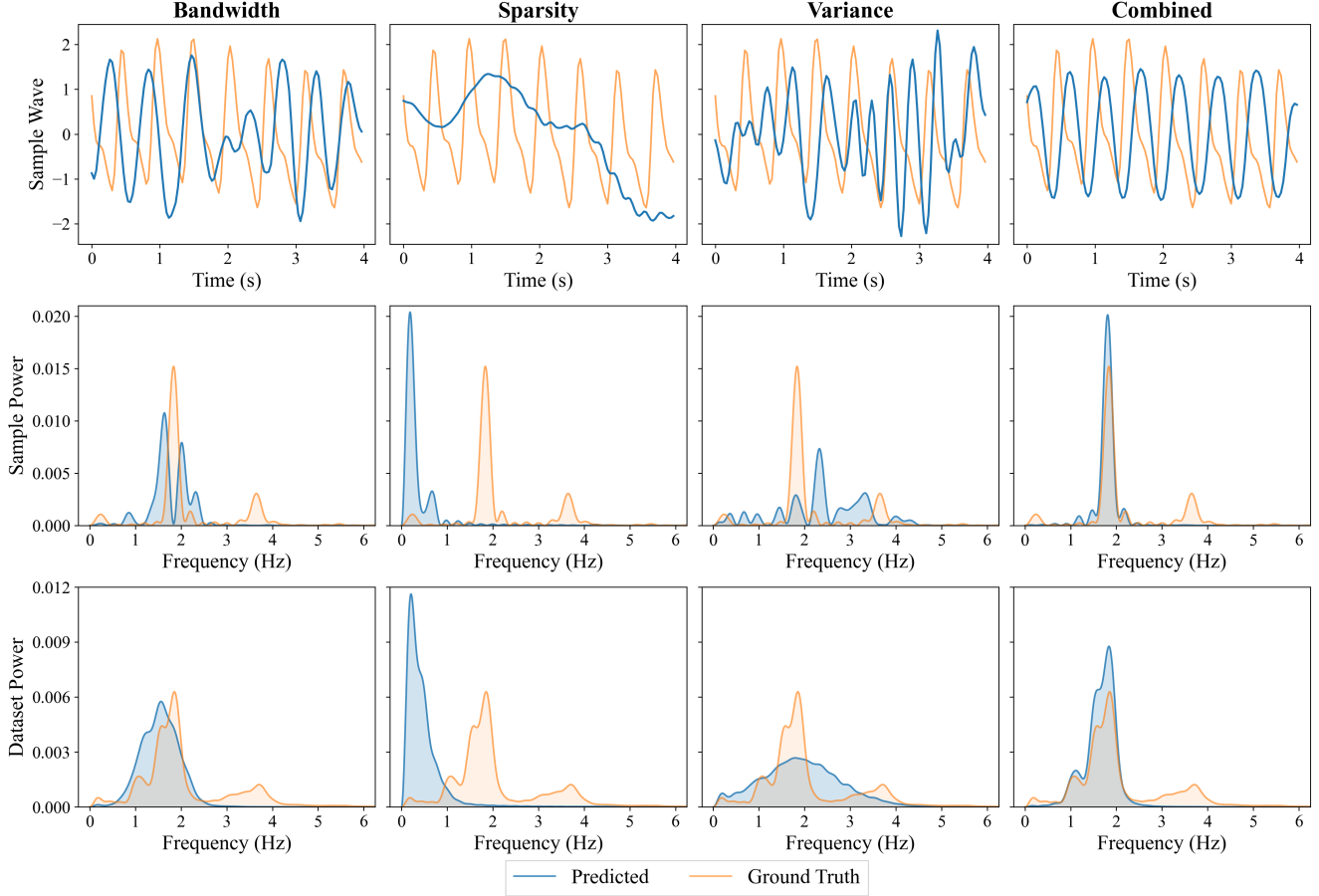
Figure 2. Each column shows predictions from models trained with one or all of the losses for 20 epochs on UBFC-rPPG. The first two rows show a sample in the time and frequency domain, respectively. The last row shows the signal power over the validation set computed by taking the sum of normalized power spectral densities from each sample, then dividing the result by the number of validation samples. The **bandwidth** loss penalizes signal power outside predefined bandlimits (40 to 180 bpm) to constrain the output space. The **sparsity** loss encourages a narrow spectrum containing strong periodicity. The **variance** loss encourages diverse power spectra in a batch, preventing the model from collapsing to a narrow bandwidth. When **combined**, the model estimates periodic signals within the desired bandlimits.

distribution P over d frequencies, and a batch of n spectral densities, $F = [v_1, ..., v_n]$, where each vector is a d-dimensional frequency decomposition of a predicted waveform. We calculate the normalized sum of densities over the batch, Q, and define the variance loss as the squared Wasserstein distance [15] to the uniform prior:

$$L_v = \frac{1}{d} \sum_{i=1}^{d} \left(\mathrm{CDF}_i(Q) - \mathrm{CDF}_i(P) \right)^2, \qquad (3)$$

where CDF is a cumulative distribution function. The third column of Fig. 2 shows the effect of the variance loss during training. For a single sample, wide-band signals containing multiple frequencies are predicted, and the predicted frequencies cover the task's bandwidth. In our experiments we use a batch size of 20 samples. See the suppl. materials for an ablation experiment on the impact of batch size.

Combining All Losses. Summarizing, our training loss function is a simple sum of the aforementioned losses:

$$L = L_b + L_s + L_v. \qquad (4)$$

While one could weight particular components of the loss more than others, we specifically formulated the losses to scale them between 0 and 1. In our experiments, we find that a simple summation without weighting gives good performance. **The combined loss function encourages the model to search over the supported frequencies to discover visual features for a strong periodic signal.** Remarkably, we find that this simple framework is sufficient for learning to regress the blood volume in video, as shown in the last column of Fig. 2.

3.2. Augmentations

Unlike Gideon *et al.*'s [13] approach, which only applies frequency augmentations, we apply several augmentations to both the spatial and temporal dimensions to learn invariances to noisy visual signals. In fact, we found that without augmentations, models did not converge during training (see the supplementary material).

Image Intensity Augmentations. Gaussian noise is added to each pixel with zero mean and a standard deviation of 2 on an image scale from 0 to 255. Each clip is darkened or brightened by adding a constant from a Gaussian distribution with mean 0 and standard deviation of 10.

Spatial Augmentations. We randomly horizontally flip a video clip with 50% probability. The spatial dimension of a clip are randomly square cropped down to between half the original length and the original length. The cropped clip is then linearly interpolated back to the original dimensions.

Temporal Augmentations. With the general assumption that the desired signal is strongly periodic and sparsely represented in the Fourier domain, we randomly flip a video clip along the time dimension with a probability of 50%. Note that the Fourier decomposition of a time-reversed sinusoid is identical to that of the original sinusoid.

Frequency Augmentations. Perhaps the most important augmentation is frequency resampling [13], where the video is linearly interpolated to a different frame rate. This augmentation is particularly interesting for rPPG, because it transforms the video input and target signal equivalently along the time dimension, making it equivariant. Given the aforementioned transformations that are invariant, $\tau(\cdot) \sim \mathcal{T}$, the equivariant frequency resampling operation, $\phi(\cdot) \sim \Phi$, and a model $f(\cdot)$ that infers a waveform from a video we have the following:

$$\phi(f(\tau(x))) = f(\phi(\tau(x))). \tag{5}$$

This is a powerful augmentation, because it allows us to augment the target distribution along with the video input. In our experiments we randomly resample input clips by a factor $c \sim U(0.6, 1.4)$. After applying the resampling augmentation, we scale the bandlimits by c, to avoid penalizing the model if the augmentation pushed the underlying pulse frequency outside of the original bandlimits.

4. Datasets

We use PURE [41], UBFC-rPPG [4], and DDPM [39] as benchmark rPPG datasets for training and testing, and CelebV-HQ dataset [56] and HKBU-MARs [17] for unsupervised training only.

Deception Detection and Physiological Monitoring (DDPM) [39, 43] consists of data from 86 subjects attempting to answer questions deceptively. Interviews were recorded at 90 frames-per-second for more than 10 minutes

on average. Natural conversation and head pose changes make it a difficult and less-constrained rPPG dataset.

PURE [41] is a benchmark rPPG dataset consisting of 10 subjects recorded over 6 sessions. Each session lasted approximately 1 minute, and raw video was recorded at 30 fps. The 6 sessions for each subject consisted of: (1) steady, (2) talking, (3) slow head translation, (4) fast head translation, (5) small and (6) medium head rotations. Pulse rates are at or close to the subject's resting rate.

UBFC-rPPG [4] contains 1-minute long videos from 42 subjects recorded at 30 fps. Subjects played a time-sensitive mathematical game to raise their heart rates, but head motion is limited during the recording.

HKBU 3D Mask Attack with Real World Variations (HKBU-MARs) [17] consists of 12 subjects captured over 6 different lighting configurations with 7 different cameras each, resulting in 504 videos lasting 10 seconds each. The diverse lighting and camera sensors make it a valuable dataset for unsupervised training. We use version 2 of HKBU-MARs, which contains videos with both realistic 3D masks and unmasked subjects.

High-Quality Celebrity Video Dataset (CelebV-HQ) [56] is a set of processed YouTube videos containing 35,666 face videos from over 15,000 identities. The videos vary dramatically in length, lighting, emotion, motion, skin tones, and camera sensors. The greatest challenge in harnessing online videos is their reduced quality due to compression before upload and by the video provider. Compression is a known challenge for rPPG, since the blood volume pulse is so subtle optically [25, 31, 32, 36].

5. Training Details

5.1. Data Preprocessing

To prepare the video clips for the spatiotemporal deep learning models, we first extract 68 face landmarks with OpenFace [2]. We then define a bounding box in each frame with the minimum and maximum (x, y) locations by extending the crop horizontally by 5% to ensure that the cheeks and jaw are present. The top and bottom are extended by 30% and 5% of the bounding box height, respectively, to include the forehead and jaw. We further extend the shorter of the two axes to the length of the other to form a square. The cropped frames are then resized to 64×64 pixels with bicubic interpolation. For faster processing of the massive CelebV-HQ [56] dataset, we instead use MediaPipe Face Mesh [22] for landmarking.

5.2. Model Architectures

We use a 3D-CNN architecture similar to [38] without temporal dilations [51]. We use a temporal kernel width of 5, and replace default zero-padding by repeating the edges. Zero-padding along the time dimension can result

Table 1. **Intra-dataset** pulse rate estimation results. The best results for a given metric/dataset are bolded, and the second-best results are underlined. For a better comparison with results in the literature, this table follows [42]. MAE: Mean Absolute Error; RMSE: Root Mean Square Error; r: Pearson correlation coefficient.

Types	Methods	UBFC-rPPG			PURE			DDPM		
		MAE (bpm)	RMSE (bpm)	r	MAE (bpm)	RMSE (bpm)	r	MAE (bpm)	RMSE (bpm)	r
Traditional	GREEN [44]	7.50	14.41	0.62	7.23	17.05	0.69	32.79	43.09	0.04
	ICA [34]	5.17	11.76	0.65	3.76	12.60	0.85	22.22	35.77	0.40
	CHROM [9]	2.36	9.23	0.87	<u>0.75</u>	2.23	**1.00**	13.48	28.53	0.56
	2SR [48]	6.90	18.50	0.65	2.44	3.06	0.98	22.08	39.73	0.22
	POS [47]	2.11	9.11	0.87	0.80	4.11	0.98	9.03	23.07	0.70
Supervised	HR-CNN [40]	-	-	-	1.84	2.37	0.98	-	-	-
	SynRhythm [29]	5.59	6.82	0.72	-	-	-	-	-	-
	PulseGAN [37]	1.19	2.10	<u>0.98</u>	-	-	-	-	-	-
	Dual-GAN [21]	**0.44**	**0.67**	**0.99**	0.82	**1.31**	<u>0.99</u>	-	-	-
	RPNet [38]†	<u>0.53 ± 0.01</u>	1.78 ± 0.02	**0.99 ± 0.00**	1.15 ± 0.27	5.77 ± 1.25	0.96 ± 0.02	**3.46 ± 0.24**	**12.47 ± 0.68**	**0.91 ± 0.01**
	PhysNet [51]†	0.55 ± 0.03	2.03 ± 0.37	**0.99 ± 0.00**	0.99 ± 0.19	5.22 ± 0.93	0.97 ± 0.01	<u>3.96 ± 0.76</u>	13.57 ± 1.74	<u>0.89 ± 0.03</u>
Unsupervised	Gideon2021 [13]	1.85	4.28	0.93	2.3	2.9	<u>0.99</u>	-	-	-
	SLF-RPM [45]*	8.39	9.70	0.70	-	-	-	-	-	-
	SimPer [50]*	4.24	-	-	3.89	-	-	-	-	-
	Contrast-Phys [42]	0.64	<u>1.00</u>	**0.99**	1.00	<u>1.40</u>	0.99	9.70 ± 2.90	25.02 ± 6.01	0.58 ± 0.19
	SiNC (ours)	0.59 ± 0.00	1.83 ± 0.04	**0.99 ± 0.00**	**0.61 ± 0.06**	1.84 ± 0.40	**1.00 ± 0.00**	5.87 ± 0.11	17.44 ± 0.16	0.81 ± 0.00

* Some methods in the unsupervised row require fine-tuning on labeled data with a linear classifier.

† Some supervised methods were trained with identical data augmentations to SiNC for fair comparison.

in edge effects that add artificial frequencies to the predictions. Early experiments showed that temporal dilations caused aliasing and reduced the bandwidth of the model to specific frequencies. Our losses and framework may be applied to any task and architecture with dense predictions along one or more dimensions. However, popular rPPG architectures such as DeepPhys [7] and MTTS-CAN [18] may be ill-suited for the approach, since they consume very few frames, and the number of time points should be large enough to give sufficient frequency resolution with the FFT. In our experiments, we use the AdamW [19] optimizer with a learning rate of 0.0001. We use a clip length of $T = 120$ frames (4 seconds), and we set the input signal's length to achieve a frequency resolution of $0.\overline{33}$ bpm.

5.3. Supervised Training

To properly compare our approach to its supervised counterpart we use the same model architecture and train it with the commonly used negative Pearson loss between the predicted waveform and the contact sensor ground truth [51]. During training we apply all of the same augmentations except time reversal. Models are trained for 200 epochs on PURE and UBFC-rPPG, and for 40 epochs on DDPM. The model from the epoch with the lowest loss on the validation set is selected for testing.

5.4. Unsupervised Training

Unsupervised models are trained for the same number of epochs as the supervised setting for both PURE and UBFC-rPPG, but we train for an additional 40 epochs on DDPM,

since this dataset is considerably more difficult. We set the batch size to 20 samples during training. Contrary to previous unsupervised approaches [13, 42], we leverage validation sets for model selection by selecting the model with the lowest combined bandwidth and sparsity losses. The creation of the dataset splits is described in the next section.

5.5. Evaluation

Pulse rates are computed as the highest spectral peak between 0.66 Hz and 3 Hz (equivalent to 40 bpm to 180 bpm) over a 10-second sliding window. The same procedure is applied to the ground truth waveforms for a reliable evaluation [26]. We apply common error metrics such as mean absolute error (MAE), root mean square error (RMSE), and Pearson correlation coefficient (r) between the pulse rates.

We perform 5-fold cross validation for both PURE and UBFC with the same folds as [13], and use the predefined dataset splits from DDPM [38]. Differently from [13], we use 3 of the folds for training, 1 for validation, and the remaining for testing rather than only training and testing partitions. We train 3 models with different initializations, resulting in 15 models trained on PURE and UBFC each, and 3 models trained on DDPM. We present the mean and standard deviation of the errors in the results.

6. Results

6.1. Within-Dataset Testing

Table 1 shows the results for models trained and tested on subject-disjoint partitions from the same datasets. For

PURE and UBFC we achieve MAE lower than 1 bpm, performing better or on par with all traditional and supervised learning approaches. For PURE, our approach gives the lowest MAE and a Pearson r of nearly 1. Performance drops on DDPM due to the overall difficulty of the dataset. SiNC outperforms contrastive approaches, only being surpassed by supervised deep learning models.

In comparison to other unsupervised methods, Contrast-Phys [42] gives the most competitive performance on all but DDPM. Note that our approach gives the lowest MAE on all datasets, but has higher RMSE. We believe this is due their use of harmonic removal as a post-processing step when estimating the pulse rate, which is not described in [42], but can be found in their publicly available code.

6.2. Cross-Dataset Testing

Table 2. **Cross-dataset** pulse rate estimation performance. The top 3 training datasets are common rPPG benchmarks, while HKBU was not designed for rPPG and has no pulse ground truth. Note that testing on UBFC after training on DDPM performs well, since their frequency distributions are similar, while PURE's pulse rates tend to be much lower.

Training Dataset	Testing Dataset	Method	MAE (bpm)	r
DDPM	UBFC	SiNC	0.88 ± 0.25	0.98 ± 0.01
	UBFC	Contrast-Phys	1.14 ± 0.38	0.96 ± 0.03
	UBFC	PhysNet	1.11 ± 0.42	0.95 ± 0.02
	PURE	SiNC	3.12 ± 1.07	0.88 ± 0.05
	PURE	Contrast-Phys	13.02 ± 6.12	0.19 ± 0.59
	PURE	PhysNet	1.46 ± 0.34	0.95 ± 0.02
UBFC	DDPM	SiNC	18.53 ± 0.36	0.38 ± 0.01
	DDPM	Contrast-Phys	22.93 ± 1.02	0.18 ± 0.04
	DDPM	PhysNet	18.58 ± 0.12	0.40 ± 0.00
	PURE	SiNC	4.02 ± 0.06	0.86 ± 0.00
	PURE	Contrast-Phys	19.61 ± 2.01	0.33 ± 0.06
	PURE	PhysNet	3.81 ± 0.34	0.87 ± 0.02
PURE	UBFC	SiNC	6.64 ± 1.76	0.59 ± 0.10
	UBFC	Contrast-Phys	10.22 ± 0.38	0.45 ± 0.04
	UBFC	PhysNet	7.02 ± 3.35	0.60 ± 0.13
	DDPM	SiNC	24.92 ± 0.65	0.20 ± 0.00
	DDPM	Contrast-Phys	29.63 ± 0.48	0.03 ± 0.02
	DDPM	PhysNet	28.03 ± 2.20	0.13 ± 0.05
HKBU (non-rPPG)	UBFC	SiNC	1.08 ± 0.03	0.95 ± 0.00
	PURE	SiNC	2.43 ± 0.20	0.90 ± 0.02
	DDPM	SiNC	20.34 ± 0.25	0.19 ± 0.02

We perform cross-dataset testing to analyze whether the approach is robust to changes to the lighting, camera sensor, pulse rate distribution, and motion. Table 2 shows the results for SiNC and supervised training with the same architecture. We find that the performance is similar for the supervised and unsupervised approaches when transferring to different data sources. Training on PURE exclusively gives relatively poor results when transferring to

UBFC-rPPG and DDPM, due to the low pulse rate variability within PURE samples and lack of movement. Training on DDPM gives the best results overall, since the dataset is the largest and captures larger subjects' movements compared to other datasets.

6.3. Training with CelebV-HQ Videos

Given the abundance of face videos publicly available online, we trained a model on the CelebV-HQ dataset [56]. After processing the available videos with MediaPipe and resampling to 30 fps, our unlabeled dataset consisted of 34,029 videos. We trained the model for 23 epochs and manually stopped training due to a plateau in the validation loss. Unfortunately, we found that the model could not converge to the true blood volume pulse. We attribute the failure to poor video quality from compression. Although the videos were downloaded with the highest available quality, they have likely been compressed, removing the pulse signal entirely. See the supplementary material for an assessment of rPPG quality using POS [47] on CelebV-HQ.

6.4. Training with HKBU-MARs Videos

The HKBU-MARs dataset [17] was designed for face presentation attack detection, but we trained models on the "real" video sessions in the dataset. The bottom rows in Table 2 show the results for training on HKBU-MARs, then testing on the benchmark rPPG datasets. Training on HKBU-MARs gives better results when testing on UBFC-rPPG and PURE than all training sets except DDPM, which is an order of magnitude larger. **To our knowledge, this is the first succesful experiment showing that non-rPPG videos can be used to train robust rPPG models, even if they do not have ground-truth pulse labels.**

6.5. Ablation Study on Losses

Table 3. Ablation study on the loss functions used during training. Results are shown for models trained and tested on UBFC-rPPG.

Loss	MAE (bpm)	RMSE (bpm)	r
L_b	3.08 ± 1.69	8.08 ± 3.61	0.87 ± 0.08
L_s	45.50 ± 1.22	50.04 ± 0.94	-0.04 ± 0.08
L_v	22.89 ± 2.83	31.51 ± 2.36	0.22 ± 0.09
$L_s + L_v$	51.24 ± 5.36	57.80 ± 7.39	-0.04 ± 0.09
$L_b + L_s$	9.99 ± 2.55	17.14 ± 2.36	0.51 ± 0.14
$L_b + L_v$	4.18 ± 2.88	8.90 ± 5.24	0.82 ± 0.14
$L_b + L_s + L_v$	$\mathbf{0.59 \pm 0.00}$	$\mathbf{1.83 \pm 0.04}$	$\mathbf{0.99 \pm 0.00}$

We trained models using all combinations of loss components to analyze their contributions. Table 3 shows the results for training and testing on UBFC-rPPG. The bandwidth loss is the most critical for discovering the true blood volume pulse, while the sparsity and variance losses do not

UBFC-rPPG

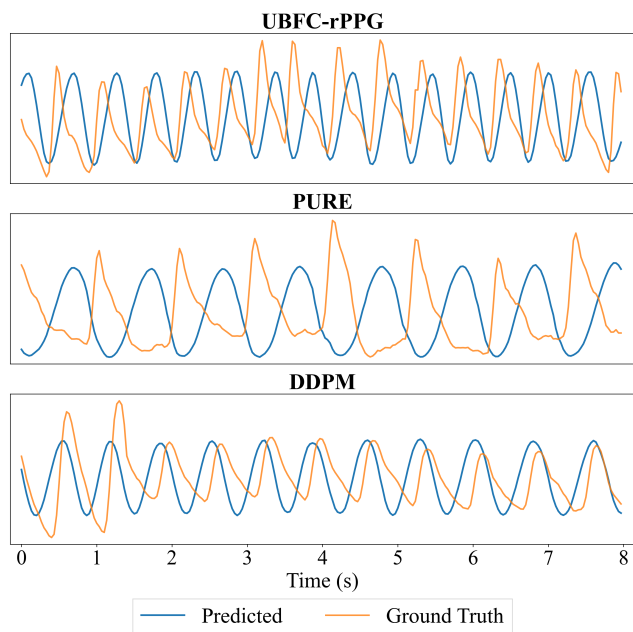**PURE**

DDPM

Time (s)

— Predicted   — Ground Truth

Figure 3. Within-dataset waveform predictions on all baseline datasets from end-to-end unsupervised models over an 8-second window. The model predictions are remarkably periodic without any form of filtering. Note that phase is not considered during training, so each model learns its own phase shift.

learn the desired signal by themselves. Surprisingly, combining the bandwidth loss with just one of the sparsity or variance losses gives worse performance than just the bandwidth loss. However, when combining all three components, the model achieves impressive results.

7. Discussion

7.1. Improvements over Supervised Learning

It is initially surprising that unsupervised training leads to similar or improved rPPG estimation models compared to those trained in a supervised manner. However, there are several potential benefits to unsupervised training. From a hardware perspective, one of the difficulties in supervised training is aligning the contact pulse waveform with the video frames [54]. The pulse sensor and camera may have a time lag, effectively giving the model an out-of-phase target at training time. Unsupervised training gives the model freedom to learn the phase directly from the video. The contact-PPG signal is also sensitive to motion and may be noisy. Since motion may co-occur at the face and fingertip, the contact signal may misguide the model towards visual features for which they should be invariant.

From a physiological perspective, the pulse observed optically at the fingertip with a contact sensor has a different phase than that of the face, since blood propagates along a

different path before reaching the peripheral microvasculature, making alignment nearly impossible without shifting the targets to rPPG estimates from existing methods [38]. Additionally, the morphological shape of the contact-PPG waveform depends on numerous factors such as the wavelength of light (and corresponding tissue penetration depth), external pressure from the oximeter clip, and vasodilation at the measurement site [1, 28]. This indicates that the morphology and phase of the target PPG waveform is likely different from the observed rPPG waveform.

7.2. Why Does It Work?

The success of SiNC depends on specific properties of the data, model, and how the two interact. Limited model capacity is actually required, since it forces learning features that generalize across inputs. By constraining the model's predictions to have specific periodic properties, the limited-capacity model must find a general set of features to produce a signal that exists in most or all of the training samples, which is the blood volume pulse in our datasets.

As a beneficial side-effect, the model intrinsically learns to ignore common noise factors such as illumination, rigid motion, non-rigid motion (*e.g.* talking, smiling, etc.), and sensor noise, since they may preside outside the predefined bandlimits or with uniform power spectra. Even if noise exhibits periodic tendencies within the bandlimits for some samples, those features would produce poor signals on other samples. Therefore, end-to-end unsupervised approaches like SiNC are well-suited for periodic problems.

8. Conclusions

We proposed a novel non-contrastive learning approach for end-to-end unsupervised signal regression, with specific experiments on blood volume pulse estimation from face videos. This SiNC framework effectively learns powerful visual features with only loose frequency constraints. We demonstrated this by training accurate rPPG models using non-rPPG data and our simple loss functions. Given the subtlety of the rPPG signal, we believe our work can be extended to other signal regression tasks in the domain of remote vitals estimation.

References

[1] Pierre Abraham, Mélissa Bourgeau, Maïte Camo, Anne Humeau-Heurtier, Sylvain Durand, Pascal Rousseau, and Guillaume Mahe. Effect of skin temperature on skin endothelial function assessment. *Microvascular Research*, 88:56–60, 2013. 8

[2] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency. Openface 2.0: Facial behavior analysis toolkit. In *IEEE International Conference on Automatic Face Gesture Recognition (FG)*, pages 59–66, 2018. 5

[3] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR*, 2022. 2, 3

[4] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124:82–90, 2019. 5

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 2020-Decem(NeurIPS):1–13, 2020. arXiv: 2006.09882. 2

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *37th International Conference on Machine Learning, ICML 2020*, pages 1575–1585, 2020. arXiv: 2002.05709 ISBN: 9781713821120. 2

[7] Weixuan Chen and Daniel McDuff. DeepPhys: Video-based physiological measurement using convolutional attention networks. In *European Conference on Computer Vision (ECCV)*, pages 356–373, 2018. 1, 2, 6

[8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 15745–15753, 2021. arXiv: 2011.10566 ISBN: 9781665445092. 2

[9] G. De Haan and V. Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013. 2, 6

[10] Gerard De Haan and Arno Van Leest. Improved motion robustness of remote-PPG by using the blood volume pulse signature. *Physiological Measurement*, 35(9):1913–1926, 2014. 2

[11] Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *International Conference on Machine Learning*, pages 3015–3024. PMLR, 2021. 2

[12] John Gideon and Simon Stent. Estimating heart rate from unlabelled video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2743–2749, October 2021. 3

[13] John Gideon and Simon Stent. The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3995–4004, October 2021. 1, 2, 3, 5, 6

[14] Jean Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 2020-Decem(NeurIPS):1–14, 2020. arXiv: 2006.07733. 2

[15] Le Hou, Chen-Ping Yu, and Dimitris Samaras. Squared Earth Mover's Distance-based Loss for Training Deep Neural Networks, Apr. 2017. arXiv:1611.05916 [cs]. 4

[16] Eugene Lee, Evan Chen, and Chen-Yi Lee. Meta-rppg: Remote heart rate estimation using a transductive meta-learner. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3

[17] Siqi Liu, Baoyao Yang, Pong C. Yuen, and Guoying Zhao. A 3d mask face anti-spoofing database with real world variations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2016. 5, 7

[18] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19400–19411. Curran Associates, Inc., 2020. 1, 2, 6

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6

[20] Hao Lu, Hu Han, and S Kevin Zhou. Dual-GAN : Joint BVP and Noise Modeling for Remote Physiological Measurement. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12404–12413, 2021. 2

[21] Hao Lu, Hu Han, and S Kevin Zhou. Dual-gan: Joint bvp and noise modeling for remote physiological measurement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12404–12413, 2021. 3, 6

[22] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 5

[23] Daniel McDuff. Camera measurement of physiological vital signs. *ACM Comput. Surv.*, aug 2022. Just Accepted. 1

[24] Daniel McDuff, Miah Wander, Xin Liu, Brian L. Hill, Javier Hernandez, Jonathan Lester, and Tadas Baltrusaitis. SCAMPS: Synthetics for camera measurement of physiological signals. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 2

[25] Daniel J. McDuff, Ethan B. Blackford, and Justin R. Estepp. The Impact of Video Compression on Remote Cardiac Pulse Measurement Using Imaging Photoplethysmography. *IEEE International Conference on Automatic Face and Gesture Recognition Workshops (FG)*, pages 63–70, 2017. 5

[26] Yuriy Mironenko, Konstantin Kalinin, Mikhail Kopeliovich, and Mikhail Petrushan. Remote photoplethysmography:

Rarely considered factors. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1197–1206, June 2020. 6

[27] Ishan Misra and Laurens van der Maaten. Self-Supervised Learning of Pretext-Invariant Representations. *CVPR*, 2020. 2

[28] Andreia Moco, Sander Stuijk, Gerard De Haan, Ruikang K. Wang, and Wim Verkruysse. A Model for Waveform Dissimilarities in Dual-Depth Reflectance-PPG. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, 2018-July:5125–5130, 2018. ISBN: 9781538636466. 8

[29] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video. In *Asian Conference on Computer Vision (ACCV)*, pages 562–576, 2018. 2, 6

[30] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. RhythmNet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Transactions on Image Processing*, 29:2409–2423, 2020. 1, 2

[31] Ewa Nowara and Daniel McDuff. Combating the Impact of Video Compression on Non-Contact Vital Sign Measurement Using Supervised Learning. In *IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 1706–1712, Oct. 2019. ISSN: 2473-9944. 5

[32] Ewa M. Nowara, Daniel McDuff, and Ashok Veeraraghavan. Systematic analysis of video-based pulse measurement from compressed videos. *Biomedical Optics Express*, 12(1):494, 2021. 3, 5

[33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018. 3

[34] M. Poh, D. J. McDuff, and R. W. Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Transactions on Biomedical Engineering*, 58(1):7–11, 2011. 2, 6

[35] Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Opt. Express*, 18(10):10762–10774, May 2010. 2

[36] Michal Rapczynski, Philipp Werner, and Ayoub Al-Hamadi. Effects of video encoding on camera-based heart rate estimation. *IEEE Transactions on Biomedical Engineering*, 66(12):3360–3370, 2019. Publisher: IEEE. 5

[37] Rencheng Song, Huan Chen, Juan Cheng, Chang Li, Yu Liu, and Xun Chen. Pulsegan: Learning to generate realistic pulse waveforms in remote photoplethysmography. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1373–1384, 2021. 6

[38] Jeremy Speth, Nathan Vance, Adam Czajka, Kevin Bowyer, and Patrick Flynn. Unifying frame rate and temporal dilations for improved remote pulse detection. *Computer Vision and Image Understanding (CVIU)*, pages 1056–1062, 2021. 1, 2, 3, 5, 6, 8

[39] Jeremy Speth, Nathan Vance, Adam Czajka, Kevin Bowyer, Diane Wright, and Patrick Flynn. Deception detection and remote physiological monitoring: A dataset and baseline ex-

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

[41] Ronny Stricker, Steffen Muller, and Horst Michael Gross. Non-contact video-based pulse rate measurement on a mobile service robot. *IEEE International Symposium on Robot and Human Interactive Communication*, pages 1056–1062, 2014. 5

[42] Zhaodong Sun and Xiaobai Li. Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast. *European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 3, 6, 7

[43] Nathan Vance, Jeremy Speth, Siamul Khan, Adam Czajka, Kevin W. Bowyer, Diane Wright, and Patrick Flynn. Deception detection and remote physiological monitoring: A dataset and baseline experimental results. *IEEE Transactions on Biometrics, Behavior, and Identity Science (TBIOM)*, pages 1–1, 2022. 5

[44] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Opt. Express*, 16(26):21434–21445, Dec 2008. 6

[45] Hao Wang, Euijoon Ahn, and Jinman Kim. Self-Supervised Representation Learning Framework for Remote Physiological Measurement Using Spatiotemporal Augmentation Loss. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(2):2431–2439, June 2022. 1, 2, 3, 6

[46] Wenjin Wang, Albertus C. Den Brinker, and Gerard De Haan. Single-Element Remote-PPG. *IEEE Transactions on Biomedical Engineering*, 66(7):2032–2043, 2019. 2

[47] 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, 2017. 2, 6, 7

[48] Wenjin Wang, Sander Stuijk, and Gerard De Haan. A Novel Algorithm for Remote Photoplethysmography: Spatial Subspace Rotation. *IEEE Transactions on Biomedical Engineering*, 63(9):1974–1984, 2016. 6

[49] Zhen Wang, Yunhao Ba, Pradyumna Chari, Oyku Deniz Bozkurt, Gianna Brown, Parth Patwa, Niranjan Vaddi, Laleh Jalilian, and Achuta Kadambi. Synthetic generation of face videos with plethysmograph physiology. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20555–20564, 2022. 2

[50] Yuzhe Yang, Xin Liu, Jiang Wu, Silviu Borac, Dina Katabi, Ming-Zher Poh, and Daniel McDuff. Simper: Simple self-supervised learning of periodic targets, 2022. 1, 2, 6

[51] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 1–12, 2019. 1, 2, 3, 5, 6

[52] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip H.S. Torr, and Guoying Zhao. Physformer: Facial video-based physiological measurement with temporal difference transformer. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR), pages 4186–4196, June 2022. 1, 2, 3

[53] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. 2021. arXiv: 2103.03230. 2

[54] Qi Zhan, Wenjin Wang, and Gerard de Haan. Analysis of CNN-based remote-PPG to understand limitations and sensitivities. *Biomedical Optics Express*, 11(3):1268–1283, 2020. 8

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

[56] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. CelebV-HQ: A large-scale video facial attributes dataset. In *ECCV*, 2022. 5, 7