Contactless Pulse Estimation Leveraging Pseudo Labels and Self-Supervision

Zhihua Li, Lijun Yin
Department of Computer Science, Binghamton University
zli191@binghamton.edu, lijun@cs.binghamton.edu

Abstract

Remote photoplethysmography (rPPG) is a promising research area involving non-invasive monitoring of vital signs using cameras. While several supervised methods have been proposed, recent research has focused on contrastive-based self-supervised methods. However, these methods often collapse to learning irrelevant periodicities when dealing with interferences such as head motions, facial dynamics, and video compression. To address this limitation, firstly, we enhance the current self-supervised learning by introducing more reliable and explicit contrastive constraints. Secondly, we propose an innovative learning strategy that seamlessly integrates self-supervised constraints with pseudo-supervisory signals derived from traditional unsupervised methods. This is followed by a corectification technique designed to mitigate the adverse effects of noisy pseudo-labels. Experimental results demonstrate the superiority of our methodology over representative models when applied to small, high-quality datasets such as PURE and UBFC-rPPG. Importantly, on large-scale challenging datasets such as VIPL-HR and V4V, our method, with zero annotation cost, not only significantly surpasses prevailing self-supervised techniques but also showcases remarkable alignment with state-of-the-art supervised methods.

1. Introduction

Remote photoplethysmography (rPPG) is a non-contact method for monitoring human cardiac activity by detecting subtle facial color variations induced by changes in blood volume in skin tissue [16, 26]. Although these cyclic changes are not perceptible to the human eye, they can be captured by video cameras and extracted using computational algorithms [18, 45]. The deployment of rPPG technology has led to various applications, including home health monitoring, fitness training, and face anti-spoofing. Remarkably, consumer cameras that are integrated into smartphones can be used for these applications, making rPPG technology widely accessible [11, 36].

The development of rPPG methods has evolved from hand-crafted features [9, 35, 51] to deep learning methods [22, 32, 56], resulting in significantly improved accuracy due to the powerful feature extraction and representation capabilities of the latter. However, this transition has also led to an increase in the cost of data collection and annotation.

Collecting sufficient data for rPPG can be a challenging task [8], prompting researchers to develop fully self-supervised learning-based methods that can predict pulse signals without annotations. For example, Gideon et al. [13] derived multiple-views contrastive learning utilizing the resampled and recovered videos. Similarly, Contrast-Phys [44] enforced the power spectral density (PSD) of the estimated pulse signal to be as similar as possible for temporally nearby segments and enlarged the distances of PSDs of different videos. Although they opened the door to training with unlabeled data, there exist limitations.

The temporal stability assumption used to acquire positive pairs does not hold if the pulse rate changes instantaneously due to physical or mental activities. Besides, the current contrastive learning [13, 44] enforced the distance between the anchor and the negatives to be further than the positives, which are coarse-grained and unable to provide explicit regularization to the distances. This allows for any videos that are distant from the anchor but not necessarily identical to the corresponding negative samples to satisfy the constraints. The easily breakable assumption of temporal stability and the quite open-ended negative constraints can render the network collapse to learning pulse-irrelevant features, particularly in the absence of supervised constraints.

While some of these methods [13, 44] have shown promising results on small datasets with high-quality and high signal-to-noise ratio (SNR) videos (e.g., videos in PURE [43] and UBFC-rPPG [3]), where the pulse signal is clear enough to be detected using contrastive periodical constraints, their performance can be compromised when dealing with more challenging videos (e.g., VIPL-HR [32] and V4V [39]). In particular, when head motion, varying lighting, and video compression are present, the subtle periodical facial chrominance changes that are crucial for rPPG estimation may be obscured by other types of facial dynamics that conform to the spectral restrictions of self-supervised learning. Therefore, current methods may struggle to converge to learning pulse dynamics solely based on the vulnerable self-supervised contrastive learning, without additional pulse-related supervision. To address the limitations, we propose a spectral ranking loss that imposes more refined constraints on negative pairs to narrow the search space. Our method involves randomly resampling the anchor video at varying frequency ratios to generate negatives and then ranking the frequency peaks of the predicted signals based on their corresponding resampling ratios.

Furthermore, in the context of videos characterized by low SNR, encompassing head motion, variations in lighting, and video compression artifacts, prevalent methodologies frequently encounter a susceptibility to local minima. To alleviate this concern, our devised approach involves constraining the exploration space through the integration of pseudo labels, derived from traditional unsupervised methods [51, 50]. Using these pseudo labels alongside self-supervised learning helps prevent getting stuck in local minima and, as a result, prevents the acquisition of features that are unrelated to the underlying pulse-related dynamics. However, the use of noisy pseudo annotations is a double-edged sword. Although they can help the network avoid superficial local minima and converge to a better solution. Inversely, it causes the network to overfit to misleading information [27, 41]. To mitigate the negative effects of noisy supervision, we proposed to guide self-supervised learning progressively leveraging pseudo labels produced from traditional unsupervised methods. Additionally, inspired by multi-network learning [25, 19, 15], a label rectification technique is invented which involves training two predictors and selectively using the output of one network to supervise the other.

Figure 1 illustrates the rectification process. It begins by generating a decision boundary using a combination of pseudo-supervision and self-supervised learning. Then, we employ a large-deviation trick to select instances that are likely to be wrongly labeled and replace their pseudo-labels with the outputs of a peer network. This corrected subset of instances is then used for supervised training. By iterating this process, our model enters a virtuous circle of self-correction and self-supervision, gradually converging towards the global minimum. We summarize our contributions as follows:

- We develop more reliable and finer-grained self-supervised contrastive constraints.
- We propose a novel strategy for learning rPPG without annotations, using a joint training strategy of pseudo-label supervision and self-supervised learning.
- A label rectification method is proposed to combat the negative effect of incorrect pseudo annotations.

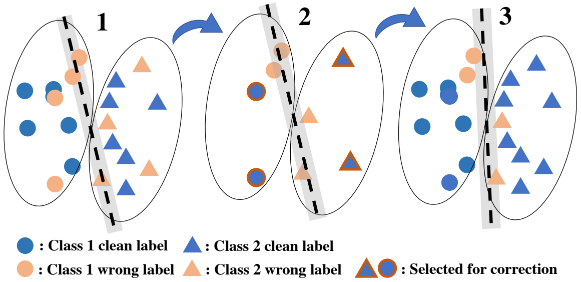

Figure 1. A classification example to illustrate our pseudo-label rectification. **(1)**: draw a decision boundary from all pseudo-labels with weak- and self-supervision; **(2)**: selectively rectify pseudo-labels by confidence scores; **(3)**: draw a new decision boundary using corrected labels. **(4)**: repeat (1)-(3) until convergence.

- In comparison to the state-of-the-art approaches, our annotation-free method consistently attains superior performance across large-scale challenging datasets and compressed videos.

2. Related works

Traditional unsupervised methods: before the emergence of deep learning, traditional methods for rPPG estimation typically involved selecting regions of interest (ROI) and applying techniques such as Independent Component Analysis (ICA) and Blind Source Separation (BSS) to extract the independent volumetric changes in facial blood vessels [35, 34]. To visualize blood flow as it fills the face, eulerian video magnification [53] employed spatial decomposition and temporal filtering to make the pulse signal visible to the naked eye. To tackle the motion problem, the chrominance-based [9] method canceled the specular reflection component using the color difference to emphasize the blood volume pulse signal from motion-induced specular distortions. Spatial subspace rotation (2SR) [51] is the first data-driven method that does not rely on domain knowledge. It measured the skin pixels' eigenspace rotation and filtered the periodical component. POS [50] projected the features to the plane orthogonal to the specular direction, maximizing the changes induced by diffuse reflections. While these traditional methods have shown promise, they are limited by their reliance on domain-specific knowledge and strong assumptions on the light reflection model.

Deep learning methods: Deep learning (DL) enables the extraction of more generalizable representations from videos and corresponding pulse signals. Recent work in this area has focused on spatial-temporal representation learning. For instance, rPPGNet [56] applied 3D convolution for video encoding and recovered the pulse signals from highly compressed videos with a video enhancement module. MTTS-CAN [22] is a temporal shift convolutional attention network that reduces latency and enables real-time prediction on mobile platforms. PhysFormer [57] introduced the first video transformer-based methods for local and global temporal learning using a self-attention mechanism. Some GAN-based approaches [32, 24] pre-processed the video data into 2D MSTmaps and applied CNNs for feature encoding, then used a generative-based feature disentanglement module to reduce subject bias.

Despite their superior performance compared to traditional methods, supervised DL methods demand costly video annotations. Several self-supervised learning methods that rely only on contrastive learning without finetuning have been shown to be effective. For instance, [13] constructed the contrastive triplet using a saliency sampler and frequency resampler, while Contrast-Phys [44] used the features of nearby facial patches to construct positive pairs and the features of other subjects to construct negative pairs. However, the methods solely using coarse-grained contrastive learning are susceptible to noises, easily collapsing to learning pulse-irrelevant dynamics, especially on low SNR videos.

3. Method

The objective of rPPG estimation is to model the mapping ϕ from a video $x \in \mathbf{R}^{t\times w\times h\times c}$ to the corresponding one-dimensional pulse signal $y \in \mathbf{R}^{t\times 1}$, where t, w, h, c represent video frames, width, height, and channels, respectively. Each frame of the video is mapped to a single pulse intensity value, interpreted as facial vessel blood volume.

3.1. Transforming videos to spatial-temporal maps

Since the periodical pulse signals are from the subtle light reflection of the blood vessels, non-skin pixels and facial geometrical characteristics are considered noise compared to skin-chrominance features. Therefore, we chose to transform the raw video to STMap, highlighting the physiological spatial-temporal information of the raw video, which is frequently used in [31, 32, 24]. STMap divides the facial areas into r ROI blocks, and the pixels of each block are averaged separately for each color channel. The results of each frame are concatenated along the temporal dimension, producing a two-dimensional image of size $\mathbf{R}^{t\times r\times c}$, where $c = 3$ denoting RGB channels. Each row of the STMap corresponds to the temporal chrominance dynamics of a specific facial region. Then, we adapt the same backbone encoder and decoder as Dual-gan [24], consisting of basic 2-D convolutions and deconvolutions.

3.2. Pseudo label generation

We use the traditional unsupervised method Spatial Subspace Rotation (2SR) [51] to obtain the pseudo labels. It computes spatial subspaces of skin pixels using eigenvector decomposition and utilizes the temporal rotations of the subspaces to extract pulse signals. Compared to other unsupervised traditional algorithms such as CHROM[9] and POS [50], 2SR does not rely on physiological priors and is empirically found to be robust under complicated conditions. During training, we employ mean squared error (MSE) loss to reduce the distance between model outputs and the pseudo labels from 2SR. Such a weakly-supervised loss is denoted as $\mathcal{L}_{pseudo}$.

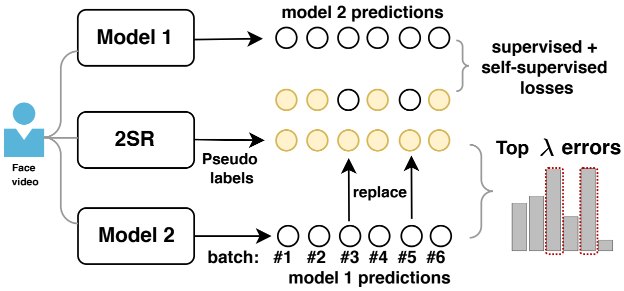

Figure 2. An illustration of the pseudo label co-rectification. The circles denote video-wise pulse signal predictions. Our backbone models are adapted from the backbone of Dual-gan [24]. Here model 2 is employed to correct a subset of pseudo labels and used to supervise the training of model 1. Similarly, model 1 will supervise model 2 in the next epoch. λ=2 is used for demonstration. Instances #3 and #5 have the largest errors, thus are selected for replacements.

3.3. Self-supervised contrastive learning

Positive pairs. Previous methods [13, 44] employ nearby time windows to construct positive pairs, which rely on the temporal stability assumptions. Instead, we randomly shift the timestamps of the selected window x by $\pm f$ ($f \leq t$) frames in the temporal dimension, where $-$ and $+$ denote backward and forward rolling, respectively. It's equivalent to cycle consistency loss [10, 52] when $f = t$ and the video is augmented by reversing the video timestamps. The positive contrastive loss can be described as:

$$\mathcal{L}_p = \mathrm{d}(\mathcal{F}(\phi(x)), \mathcal{F}(\phi(x_{\pm f}))) \tag{1}$$

where d is a distance function using mean squared error (MSE), and $\mathcal{F}$ stands for PSD (power spectral density) transformation following previous works [44, 13]. $x_{\pm f}$ and x represent shifted and original video, respectively.

Negative pairs: Previous study [44] developed the negative loss term by maximizing the dissimilarity of outputs from different videos, which functions only when the model is trained on a small number of videos and each has a distinct pulse rate. Nonetheless, with the escalation in the count of training videos, there is a concurrent rise in the potential occurrence of pairs of two videos coincidentally aligning with the same heart rate. Instead, we employ a negative sample augmentation method modified from [13], which resamples the video and consequently alters the heart rate. The resampling ratio for each instance is selected randomly from a set of values ranging from 0.6 to 1.3. The negative term can be obtained from:

$$\begin{aligned}\mathcal{L}_n = &-\mathrm{d}(\mathcal{F}(\phi(x)), \mathcal{F}(\phi(x_{d\downarrow}))) - \\ &\mathrm{d}(\mathcal{F}(\phi(x)), \mathcal{F}(\phi(x_{u\uparrow})))\end{aligned} \tag{2}$$

where $d \downarrow$ is downsampling and $u \uparrow$ is upsampling.

Furthermore, when a time series signal is upsampled (interpolated), the frequency content of the signal also changes and is shifted to lower frequencies for a fixed fps [29, 48]. That means a down-sampled video leads to an increased HR, and an up-sampled video leads to a decreased HR: $h_{u\uparrow} < h < h_{d\downarrow}$, where h is the heart rate calculated by $h = \mathrm{argmax}(\mathcal{F}(\phi(x)))$. Accordingly, a ranking loss is applied to rank the peaks of PSDs, which can be described as:

$$\mathcal{L}_{rank} = \max(h_{u\uparrow} - h, 0) + \max(h - h_{d\downarrow}, 0) \tag{3}$$

Since the argmax operation is not differentiable, a Gumbel softmax operation is used to convert PSDs into one-hot vectors. h is obtained by multiplying the one-hot vector with the index vector of frequency bins.

3.4. Guided self-supervision with weak-supervision

In addition to the reinvented contrastive learning scheme, more importantly, we seek to tackle the vulnerability of self-supervised learning facing interference from pulse-irrelevant facial dynamics. To this end, we propose to leverage pseudo-label supervision to constrict the optimization path. One straightforward way is to minimize the total loss of the weakly-supervised and self-supervised learning. However, the network likely favors an easier task [7], thus overfit to either the pseudo labels or contrastive constraints [14, 60]. Therefore, we divide the training into two stages. First, we let the learning from pseudo-labels dominate the optimization in early epochs. Then, we gradually raise the ratio of self-supervised learning and employ it to fine-tune the model to search for the optimum leveraging extra information gained from pseudo labels. The pseudo labels, as guidance, pave the way for more trustworthy self-supervised contrastive learning. Therefore, we design the overall loss as follows:

$$\mathcal{L} = (1 - \frac{E(t)}{E_{max}})\mathcal{L}_{pseudo} + \underbrace{\mathcal{L}_p - \mathcal{L}_n + \alpha\mathcal{L}_{rank}}_{\text{self-supervised}} \tag{4}$$

where $E(t)$ is the current epoch number and E_{max} is the total epochs. $\alpha = 0.01$ is a weighting hyperparameter for the spectral ranking loss.

3.5. Co-rectification: collaborative label correction

Weak supervision leveraging pseudo labels boosts the training of clean instances, conversely, this creates a problem of abundant noisy labels which may degrade neural network performance [5]. Intuitively, selecting clean instances out of the noisy ones using a multi-view network yields better performance [25, 19]. However, sample selection is suboptimal due to the high percentage of noisy labels generated from the traditional method, and only selecting the high-SNR instances for training leads the model to underfit. In this work, not only do we select the clean instances, but also rectify the incorrect annotations and then include them for training. Briefly, we propose a novel learning paradigm called “co-rectification” to tackle noisy annotations. Inspired by multi-network learning and sample selection [28, 25, 19, 15], we maintain two networks to reduce the confirmation bias [15]. As illustrated in Figure 2, leveraging the multi-network diversities, the co-rectification composes of two steps: (1) noisy instance selection and (2) updating pseudo-labels with the peer’s output.

Let ϕ_1 and ϕ_2 represent two networks initialized randomly, and let the corresponding outputs be $\tilde{y}_1 = \phi_1(x)$ and $\tilde{y}_2 = \phi_2(x)$, where $\tilde{y}_1$ and $\tilde{y}_2 \in \mathcal{R}^{t\times 1}$. The heart rate calculated from $\tilde{y}_1$ and $\tilde{y}_2$ using frequency peaks are denoted as $\tilde{h}_1$ and $\tilde{h}_2$. We use y_p to represent the pseudo pulse wave calculated from the traditional method 2SR, and the corresponding heart rate is denoted as h_p. To select instances for rectification, here, we employ a large-deviation trick which is inspired by the small-loss method in [40]. Specifically, we first calculate the discrepancy ϵ between the predicted heart rate of the network 1 (i.e., $\tilde{h}_1$) and the pseudo heart rate h_p by taking the absolute value:

$$\epsilon_1^{(b)} = \mathrm{abs}(h_p^{(b)} - \tilde{h}_1^{(b)}) \tag{5}$$

where (b) is the instance index supposing a mini-batch $X = \{x^{(b)}, y_p^{(b)}\}_{b=0}^{B-1}$ is randomly sampled, and B is the batch size. Likewise, $\epsilon_2^{(b)}$ of network 2 is obtained. Subsequently, for each network, the deviation values $\epsilon_1^{(b)}, b \in \{0, ..., B-1\}$ are sorted and the highest λ instance indexes are obtained, which is denoted as $i_1(s), s \in \{0, ..., \lambda - 1\}$, $i_1(s) \in \{0, ..., B-1\}$; vice versa, $i_2(s)$ of network 2 is obtained.

The networks tend to learn simple and generalized patterns at the starting epochs [1, 54, 58, 59]. Therefore, they are capable of exploiting partial clean labels to learn generalized discriminative features at the early epochs, and instances with highly deviated predictions are probable to be wrongly-annotated instances. Additionally, because of the memorization nature of DNNs [2], we gradually increase λ by considering the training progress: $\lambda = \frac{E(t)}{E_{max}}\lambda_{max}$. As a result, networks learn general discriminative features in initial epochs with all instances, and as the epoch number goes large, the negative effect of memorizing noisy instances is eliminated as more instances are chosen for rectification.

After the suspect instances are selected, we turn to replace the incorrect supervision signals with the prediction of the same instances from its peer network:

$$\begin{aligned}\phi_1 &: \mathrm{Subst}(y_p^{i_1(s)}/\tilde{y}_2^{i_1(s)})\\ \phi_2 &: \mathrm{Subst}(y_p^{i_2(s)}/\tilde{y}_1^{i_2(s)})\end{aligned} \tag{6}$$

where Subst stands for substitution. Next, the corrected labels are incorporated into supervised training, and the label correction, self-supervision, and weak supervision proceed iteratively until approaching the optimum. The training process is detailed in Algorithm 1.

Algorithm 1: Training code for the model

```
   Input: $x$; networks $\phi_1$ and $\phi_2$; pseudo-labels $y_p$.
1  for e = 1:total_epoch do
2      Sample a batch b = 1:batch_size.
3      Augment samples to form contrastive pairs.
4      Predict $\tilde{y}_1^{(b)} = \phi_1(x^{(b)})$; $\tilde{y}_2 = \phi_2(x^{(b)})$.
5      Calculate self-supervised loss $\mathcal{L}_n$ and $\mathcal{L}_p$.
6      Obtain deviations $\epsilon_1^{(b)}$, $\epsilon_2^{(b)}$ using $y_p^{(b)}$ and Eq (5).
7      if e%2==0 then
8          Obtain indexes $i_1$ from the $\lambda$ highest $\epsilon_1$
             inside the batch.
9          Replace $y_p^{i_1(s)}$ with $\tilde{y}_2^{i_1(s)}$ as Eq (6).
10         Calculate the weakly supervised loss
             $\mathcal{L}_{pseudo}$ using updated pseudo-labels from
             the last step.
11         Update model 1 using Eq (4).
12     else
13         Same as steps 8-11 for model 2.
14     end
15 end
```

4. Experiments

4.1. Datasets and metrics

PURE [43] includes 60 facial videos from ten subjects, which are recorded with controlled head motions (i.e., steady, talking, slow translation, fast translation, and rotations) under natural lighting. We used predefined fold splits for training and testing following [24, 44].

UBFC-rPPG [3] contains uncompressed 42 facial videos from 42 subjects while playing mathematical puzzles. We follow the same test strategy as [31, 32, 24]. Videos in PURE and UBFC-rPPG are of high quality, so the performance difference between existing supervised methods and unsupervised ones is not significant.

VIPL-HR [32] is a challenging large-scale remote physiology measurement database, which contains 2,378 RGB facial videos from 107 subjects. This dataset was collected under complicated and diverse scenarios, which include large head movement and varied ambient lighting. Videos are acquired with multiple types of camera sensors including smartphones and PCs, resulting in inconsistent video frame rates. We follow [24, 57, 44] and use the provided subject-exclusive 5-fold cross-validation protocol.

Vision-for-Vitals (V4V) [39] consists of 179 subjects and 1358 videos. A range of spontaneous emotions was evoked and continuously changing heart rates were observed. The dataset is recorded with significant dynamic head and facial expressions making it a challenging dataset. We follow the same frame-level validation protocol as the challenge.

Evaluation Metrics. Following previous methods [6, 31, 24, 13, 44, 57], we use mean absolute error (MAE), root mean squared error (RMSE) measured in bpm (beats per minute), and Pearson correlation coefficient (PC) to evaluate the model performance for pulse rate prediction.

4.2. Implementation details

For all experiments, we crop the videos into non-overlapping 10 seconds clips and compute the average pulse rate for each clip. Since V4V uses instantaneous frame-wise metrics, we use a shorter length of 5 seconds. On VIPL-HR and V4V, we employ the dataset-provided heart rate values for testing. On UBFC-rPPG and PURE, the heart rate is calculated by locating the highest peak from the PSD of the pulse signal. We use an AdamW [23] optimizer to train our model with a learning rate of 5×10^{-5}, and a batch size of 16 for 80 epochs on one NVIDIA RTX2080 GPU. For each batch, 8 out of 16 instances are selected for rectification ($\lambda_{max} = 8$). Our method is implemented using PyTorch. We use the officially published implementations of Contrast-Phys [44] and Gideon et al. [13] for extra experiments on VIPL-HR and V4V. We follow their configurations and pre-processing steps, and the best results among all training epochs are reported. In particular, as the video frame rate in VIPL-HR is not equal, we resample them to a fixed 30fps before being fed to Contrast-Phys [44].

4.3. Comparisons with state-of-the-art methods

We compare our proposed approach to traditional methods, current state-of-the-art self-supervised, and supervised methods. By checking if blood volume pulse or heart rate annotations are involved during training, the methods are categorized into two: with annotations (**w/ a.**) (i.e., DeepPhys [6], RhythmNet [31], CVD [32], Dual-GAN [24], and [57]) and without annotations (**w/o a.**) (i.e., Contrast-Phys [44], Gideon et al. [13], CHROM [9], 2SR [51], and POS [50]).

UBFC and PURE. First, we compare our proposed method with state-of-the-art supervised and unsupervised methods on UBFC-rPPG and PURE, which have minimum interference from facial dynamics and compression. As shown in Table 1, the gap between existing supervised methods and self-supervised methods is small. For exam-

	Method	UBFC			PURE			PURE → UBFC		
		MAE↓	RMSE↓	PC↑	MAE↓	RMSE↓	PC↑	MAE↓	RMSE↓	PC↑
w/ a.	SynRhythm [30]	5.59	6.82	0.72	-	-	-	-	-	-
	PulseGAN [42]	1.19	2.10	0.98	-	-	-	2.09	4.42	0.97
	Dual-GAN [24]	**0.44**	**0.67**	**0.99**	**0.82**	**1.31**	**0.99**	**0.74**	**1.02**	**0.997**
w/o a.	GREEN [49]	7.50	14.41	0.62	-	-	-	8.29	15.82	0.68
	ICA [34]	5.17	11.76	0.65	-	-	-	4.39	11.60	0.82
	CHROM [9]	2.37	4.91	0.89	2.07	9.92	0.99	3.10	6.84	0.93
	2SR [51]	-	-	-	2.44	3.06	0.98	-	-	-
	POS [50]	4.05	8.75	0.78	-	-	-	3.52	8.38	0.90
	Gideon2021 [13]	1.85	4.28	0.93	2.3	2.9	0.99	-	-	-
	Contrast-Phys [44]	0.64	1.00	0.99	1.00	1.40	0.99	-	-	-
	Ours	**0.48**	**0.64**	**0.998**	**0.64**	**1.16**	**0.99**	**0.71**	**1.45**	**0.99**

Table 1. Intra- and cross-dataset pulse rate estimation results on UBFC-rPPG and PURE. The best results are in bold.

	Method	Std↓	MAE↓	RMSE↓	PC↑
w/ a.	I3D [4]	15.9	12.0	15.9	0.07
	SAMC [47]	18.0	15.9	21.0	0.11
	DeepPhy [6]	13.6	11.0	13.8	0.11
	RhythmNet [31]	8.11	5.30	8.14	0.76
	AutoHR [55]	8.48	5.68	8.68	0.72
	CVD [32]	7.92	5.02	7.97	**0.79**
	PhysFormer [57]	**7.74**	**4.97**	**7.79**	0.78
w/o a.	POS [50]	15.3	11.5	17.2	0.30
	CHROM [9]	15.1	11.4	16.9	0.28
	Contrast-Phys [44]	35.9	32.1	36.1	0.04
	Gideon2021 [13]	11.15	9.01	14.02	0.58
	Ours	**7.88**	**5.19**	**8.26**	**0.78**

Table 2. HR estimation results by our method and several state-of-the-art methods on the VIPL-HR dataset

ple, self-supervised Contrast-Phys [44] performs closely to the supervised method Dual-GAN [24]. It aligns with our prior statement that with high SNR videos, solely leveraging self-supervision with spatial-temporal consistences could easily recognize the hidden periodical cues. And in this case, supervised methods with annotations are less advantageous. It can be seen that our method involving the cooperation of self-supervised and weakly-supervised learning suppresses all the existing methods that are trained without annotations. Surprisingly, our method outperforms the most recent supervised approach Dual-GAN [24] on PURE, with an RMSE of 1.16 compared to 1.31.

Cross-dataset testing. Since detecting rPPG requires extracting the subtle blood volume changes underneath the skin, which is naturally vulnerable to external distractions. Verifying cross-dataset performance is an important metric to evaluate the robustness of feature representation learning. The last column of Table 1 presents the model performance by training on PURE and testing on UBFC. Our method achieves the best MAE score and outperforms all other unsupervised methods and most of the supervised methods, such as Pulse-GAN and Contrast-Phys. It suggests that, by compensating the self-supervised constraints with a more specific optimization direction, our approach is capable of discovering more discriminative and generalized pulse information.

VIPL-HR. Previous results indicate that, with high-quality videos, current self-supervised learning captures the underlying pulse signals by enforcing the extraction of hidden cyclicities. However, distractions from environmental lighting, facial muscle activity, and body or camera motions could lower the purity of the desired pulse periodicity, leading the current methods to divergence or converging to noises. As can be seen in Table 2, turning to a more challenging dataset, we found that Contrast-Phys [44], which employs only self-supervision, collapse on VIPL-HR. Despite the fact that apparent periodicity is observed from the predicted signals from Contrast-Phys, the periodicity is likely to be pulse-irrelevant noises. Gideon2021 [13] performs better than existing methods as it utilizes a saliency sampler to magnify the region-of-interest thus manually increasing the SNR.

In contrast, our proposed method greatly outperforms Gideon2021 [13] and all other methods not consuming annotations, and most notably, is close to the most recent supervised method such as PhysFormer [57]. Specifically, we obtained an 8.26 RMSE, exceeding the Gideon2021 [13] by 41% and slightly worse than 7.79 from PhysFormer [57]. It's worth noting that the standard deviation and person correlation of our method ranks the second best among all the existing supervised methods. While our method has zero annotation cost and simpler network structures.

Vision for Vitals. Videos in the V4V dataset present frequent facial expressions and head motions, interfering with the light reflection of blood vessels, thus challenging the current frequency-driven self-supervised methods. In addition, it's the first to utilize frame-wise evaluation metrics, making it one of the most challenging instantaneous datasets. Table 4 compares the performance with the top-ranked methods in the challenge, including [12, 17, 21, 21] and the most recent instantaneous-based method [38], all of which used ground-true annotations for training. Data in Table 4 suggest that our method is superior to the second best and close to the best. Remarkably, ours outperforms all the self-supervised methods by a large margin (20% and 11% improvement in MAE compared to Contrast-Phys [44] and Gideon2021 [13], respectively). The results

Plain self-supervision	Pseudo labels	Gradual guidance	Spectral ranking	Co-rectification	Std↓	MAE↓	RMSE↓	PC↑
√					15.11	23.56	26.44	0.07
√	√				10.95	8.26	12.15	0.57
√	√	√			10.58	7.58	11.10	0.61
√	√	√	√		9.92	7.04	10.38	0.66
√	√		√	√	9.00	6.11	9.14	0.74
√	√	√	√	√	**7.88**	**5.17**	**8.26**	**0.78**

Table 3. Ablation studies on VIPL-HR. **Plain self-supervision:** $\mathcal{L}_p$ and $\mathcal{L}_n$ in Section 3.3; **Pseudo labels:** joint training with a predefined loss weight; **Gradual guidance:** progressive joint learning in Section 3.4; **Spectral ranking:** L_{rank} in Section 3.3; **Co-rectification:** Section 3.5.

	Method	MAE↓	RMSE↓	PC↑
w/ a.	Stent et al. [12]	**9.22**	**14.18**	**0.47**
	Hill et al. [17]	9.37	14.59	0.44
	Kossack et al. [20]	10.15	15.38	0.44
	Ouzar et al. [33]	11.60	14.90	0.20
	DeepPhys [6]	14.7	19.7	-
	TS-CAN [22]	13.9	19.2	-
	Revanur et al.[37]	13.0	18.8	-
w/o a.	GREEN [49]	15.45	20.73	0.05
	POS [50]	15.3	21.8	-
	ICA [34]	15.1	20.6	-
	Contrast-Phys [44]	12.0	17.6	0.26
	Gideon2021 [13]$*$	10.70	15.17	0.36
	Ours	**9.57**	**14.55**	**0.46**

Table 4. HR estimation results by our method and state-of-the-art methods on V4V dataset. $*$: results are adapted from the baseline model of the challenge paper [12] where (1) supervised confidence module and (2) training on the test set are excluded.

imply that our approach is able to identify instantaneous heart rate variations, meaning that the abstracted features by the model align well with the pulse volume changes. This attributes to the effective regularization from weak supervision leveraging pseudo labels, which trims the searching space and encourages the learning of pulse-related features.

4.4. Ablation study

Plain self-supervision: As expected, by only utilizing self-supervised training ($\mathcal{L}_p$ in Eq 4) on VIPL-HR, the results (Table 3) are found to be unconvergence or converging to noise. This is consistent with the results of Contrast-Phys [44] on Table 2. Both indicate the limitation of self-supervised methods in capturing cyclicity relating to pulse volume change under video distortions.

Gradual guidance: Section 3.4 proposes to guide self-supervised learning with weak supervision gradually. As shown in Table 3, the performance jumps from 26.44 to 11.10 in RMSE. It reveals that the gradual guidance from pseudo labels can effectively balance the two learning targets and facilitate convergence.

Spectral ranking: By enforcing the ranking of negative augmentations as stated in Section 3.3, RMSE improves from 11.10 to 10.38 and PC increases from 0.61 to 0.66. The reason is that the network is encouraged to focus on intrinsic pulse signals to comply with the spectral distinction.

Co-rectification: In terms of the co-rectification module as described in Section 3.5, it's apparent from the results that co-rectification brings about significant performance boosts over other modules. Moreover, after replacing the progressively increasing weight in Equation 4 with a constant hyperparameter, the pseudo-label rectification still decreases the RMSE from 10.38 to 9.14. The consistent boost in performance indicates that rectifying the noisy labels is an effective approach to reduce the misinformation from noisy annotations, which removes the key barrier of adapting pseudo-labels for supervised training.

4.5. Robustness test under video compression

Video compression is extensively used in practical applications due to its exceptional storage capabilities. Figure 4 compares the performance of our method and two state-of-the-art unsupervised methods (i.e., Contrast-Phys [44] and Gideon2021 [13]) under video compression. Specifically, the videos are compressed with FFmpeg [46] under four rate factors: 6 (low), 10 (medium), 14 (high), and 18 (ultra high). Apparently, as the compression increases, the average RMSE rises accordingly for all methods. Notably, for an ultra-high compression ratio, our method can persist in an average RMSE of 12 compared to 24 and 30 from Gideon2021 [13] and Contrast-Phys [44], respectively. By observing the Std of RMSE from the figure and their statistics, we found that, as more compression is applied, the two competitors inevitably suffer from failing into collapsed solutions and give largely deviated predictions. In contrast, our proposed method continually yields superior output and more importantly with the lowest Std. This attributes to the guidance of pseudo labels obtained from traditional methods and the proposed collaborative rectification approach to solidify the representation learning. It's noteworthy that the occasional collapsing of existing methods raises concerns about deploying them into real applications, while our proposed method can successfully solve this pain point.

4.6. Visualization

Pulse waveforms. How does the model avoid being overfitted to pseudo-labels and proceed to self-correction? To clarify this, we provide four exemplary comparisons between pseudo-labels and rPPGs produced from the training model in Figure 3. It's apparent that pseudo-labels from traditional methods are frequently deformed. In contrast, the network outputs have cleaner periodicity and closer HR

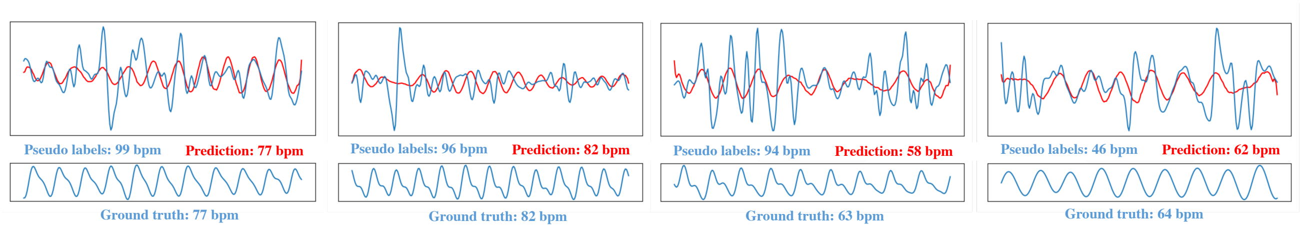

Figure 3. Visualization of model predictions (**top/red**), pseudo labels (**top/blue**), and ground-truths (**bottom**) during training on VIPL-HR.

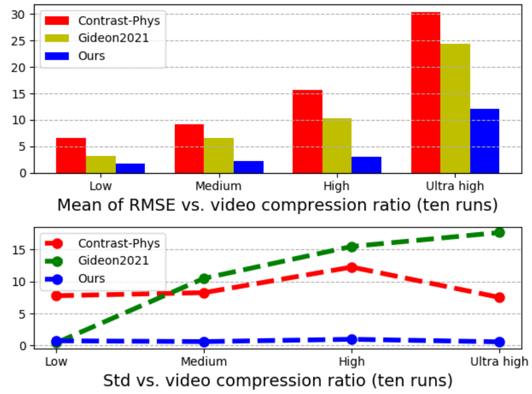

Figure 4. Robustness analysis in terms of video compression ratios on UBFC-rPPG. We plot the average and Std of RMSE from ten times run by altering random seeds.

predictions to ground truths. Particularly, we find that the network tends to align with certain cycles of circulations (e.g., Figure 4-1 right and 4-4 middle), especially when the waveforms of pseudo labels are less distorted. The model then leverages the selected cycles as the anchor and extends the periodicity to the rest of the signal. This observation implies that the network selectively learns from pseudo-labels and manages to reject the misinformation, in line with what we expect.

The reason for this is twofold. First, the network starts with fitting as much as possible to the pseudo labels in the early epochs, then self-supervised contrastive learning is involved in searching for better minima. Such a fit-and-escape strategy encourages the network to specify anchors from the pseudo-labels and extend the characteristics to the distorted segment with contrastive learning. Particularly, the joint training scheme that gradually decreases the weight of the weak supervision (Section 3.4) contributes to this transition. Second, the label co-rectification procedure reinforces the merit of joint self-supervised and weekly-supervised learning. Because once the network escapes from overfitting to the noisy labels, it's crucial to rectify the pseudo labels and move one step further toward optimum.

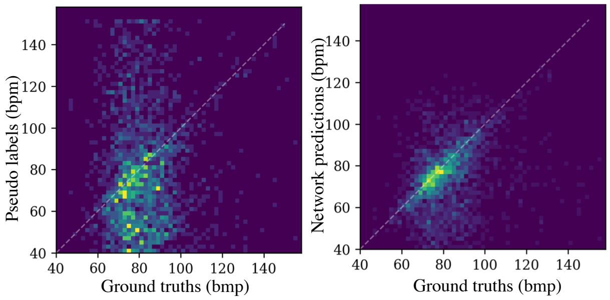

Figure 5. The diagonal position is when the predicted HR is exactly the same as ground truths. The network **(right)** yields more accurate HR predictions than the pseudo labels **(left)**, providing the basis for rectification when training. Here we only plot the samples that are selected for rectification.

Feasibility analysis of co-rectification. To demonstrate that the peer model's output provides more accurate supervision signals than the pseudo labels in the training phase (without co-rectification), we use a scatter plot to visualize the predictions and the corresponding ground-truth pulse rate in Figure 5 with the instances selected for rectification. The left shows that the pseudo labels of the selected instances are randomly spread in the entire plane, meaning that the pseudo labels are far away from the ground truths. In contrast, the outputs of the peer network are clustered around the diagonal line, showing the model outputs are close to true labels. This evidence sets the stage for the following label rectification process. On the other hand, it suggests that utilizing a large-deviation criterion to select wrongly-labeled samples is efficacious.

Conclusion

This paper studies the task of remote photoplethysmography estimation without label annotations. While current methods have shown promise in utilizing unlabeled data for training, they have limitations. The current models are vulnerable to noisy facial dynamics and video quality, collapsing to learning pulse-unrelated periodicity. To address these issues, this paper proposes several strategies to narrow the searching space of contrastive learning, thereby helping the model to escape local minimums induced by noise. We demonstrate that our method outperforms both state-of-the-art unsupervised methods and most supervised methods, particularly on challenging datasets like VIPL-HR and V4V. Our findings suggest a new direction for learning without annotations in the rPPG field.

Acknowledgement The material is based on the work supported in part by the NSF under grant CNS-1629898 and the Center of Imaging, Acoustics, and Perception Science (CIAPS) of the Research Foundation of Binghamton University.

References

[1] Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8:1–74, 2021. 4

[2] Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *ICML*, pages 233–242. PMLR, 2017. 4

[3] 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. 1, 5

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 6

[5] Pengfei Chen, Ben Ben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and utilizing deep neural networks trained with noisy labels. In *ICML*, pages 1062–1070. PMLR, 2019. 4

[6] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *ECCV*, pages 349–365, 2018. 5, 6, 7

[7] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *ICML*, pages 794–803. PMLR, 2018. 4

[8] Ananyananda Dasari, Sakthi Kumar Arul Prakash, László A Jeni, and Conrad S Tucker. Evaluation of biases in remote photoplethysmography methods. *NPJ digital medicine*, 4(1):91, 2021. 1

[9] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013. 1, 2, 3, 5, 6

[10] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *CVPR*, pages 1801–1810, 2019. 3

[11] Víctor Ferrer-Mileo, Federico Guede-Fernandez, Mireya Fernández-Chimeno, Juan Ramos-Castro, and Miguel A García-González. Accuracy of heart rate variability estimation by photoplethysmography using an smartphone: Processing optimization and fiducial point selection. In *EMBC*, pages 5700–5703. IEEE, 2015. 1

[12] John Gideon and Simon Stent. Estimating heart rate from unlabelled video. In *ICCV*, pages 2743–2749, 2021. 6, 7

[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*, pages 3995–4004, 2021. 1, 3, 5, 6, 7

[14] Rick Groenendijk, Sezer Karaoglu, Theo Gevers, and Thomas Mensink. Multi-loss weighting with coefficient of variations. In *WACV*, pages 1469–1478, 2021. 4

[15] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *NeurIPS*, 31, 2018. 2, 4

[16] Alrick B Hertzman. Photoelectric plethysmography of the fingers and toes in man. *Proceedings of the society for experimental biology and medicine*, 37(3):529–534, 1937. 1

[17] Brian L Hill, Xin Liu, and Daniel McDuff. Beat-to-beat cardiac pulse rate measurement from video. In *ICCV*, pages 2739–2742, 2021. 6, 7

[18] Markus Huelsbusch and Vladimir Blazek. Contactless mapping of rhythmical phenomena in tissue perfusion using ppgi. In *Medical Imaging 2002: Physiology and Function from Multidimensional Images*, volume 4683, pages 110–117. SPIE, 2002. 1

[19] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, pages 2304–2313. PMLR, 2018. 2, 4

[20] Benjamin Kossack, Eric Wisotzky, Peter Eisert, Sebastian P Schraven, Brigitta Globke, and Anna Hilsmann. Perfusion assessment via local remote photoplethysmography (rppg). In *CVPR*, pages 2192–2201, 2022. 7

[21] Benjamin Kossack, Eric Wisotzky, Anna Hilsmann, and Peter Eisert. Automatic region-based heart rate measurement using remote photoplethysmography. In *ICCV*, pages 2755–2759, 2021. 6

[22] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *NeurIPS*, 33:19400–19411, 2020. 1, 2, 7

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[24] Hao Lu, Hu Han, and S Kevin Zhou. Dual-gan: Joint bvp and noise modeling for remote physiological measurement. In *CVPR*, pages 12404–12413, 2021. 3, 5, 6

[25] Eran Malach and Shai Shalev-Shwartz. Decoupling" when to update" from" how to update". *NeurIPS*, 30, 2017. 2, 4

[26] Daniel McDuff. Camera measurement of physiological vital signs. *ACM Computing Surveys*, 55(9):1–40, 2023. 1

[27] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *NeurIPS*, 26, 2013. 2

[28] Jingchao Ni, Shiyu Chang, Xiao Liu, Wei Cheng, Haifeng Chen, Dongkuan Xu, and Xiang Zhang. Co-regularized deep multi-network embedding. In *Proceedings of the 2018 world wide web conference*, pages 469–478, 2018. 4

[29] Jan Nilsson, Marcela Panizza, and Mark Hallett. Principles of digital sampling of a physiologic signal. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 89(5):349–358, 1993. 4

[30] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Synrhythm: Learning a deep heart rate estimator from general to specific. In *ICPR*, pages 3580–3585. IEEE, 2018. 6

[31] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Transactions on Image Processing*, 29:2409–2423, 2019. 3, 5, 6

[32] Xuesong Niu, Zitong Yu, Hu Han, Xiaobai Li, Shiguang Shan, and Guoying Zhao. Video-based remote physiological measurement via cross-verified feature disentangling. In *ECCV*, pages 295–310. Springer, 2020. 1, 3, 5, 6

[33] Yassine Ouzar, Djamaleddine Djeldjli, Frédéric Bousefsaf, and Choubeila Maaoui. Lcoms lab's approach to the vision for vitals (v4v) challenge. In *ICCV*, pages 2750–2754, 2021. 7

[34] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, 58(1):7–11, 2010. 2, 6, 7

[35] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 18(10):10762–10774, 2010. 1, 2

[36] Donghao Qiao, Farhana Zulkernine, Raihan Masroor, Roshaan Rasool, and Nauman Jaffar. Measuring heart rate and heart rate variability with smartphone camera. In *MDM*, pages 248–249. IEEE, 2021. 1

[37] Ambareesh Revanur, Ananyananda Dasari, Conrad S Tucker, and László A Jeni. Instantaneous physiological estimation using video transformers. In *Multimodal AI in healthcare: A paradigm shift in health intelligence*, pages 307–319. Springer, 2022. 7

[38] Ambareesh Revanur, Ananyananda Dasari, Conrad S Tucker, and László A Jeni. Instantaneous physiological estimation using video transformers. In *Multimodal AI in Healthcare*, pages 307–319. Springer, 2023. 6

[39] Ambareesh Revanur, Zhihua Li, Umur A Ciftci, Lijun Yin, and László A Jeni. The first vision for vitals (v4v) challenge for non-contact video-based physiological estimation. In *ICCV*, pages 2760–2767, 2021. 1, 5

[40] Yanyao Shen and Sujay Sanghavi. Learning with bad training data via iterative trimmed loss minimization. In *ICML*, pages 5739–5748. PMLR, 2019. 4

[41] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 2

[42] 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

[43] Ronny Stricker, Steffen Müller, and Horst-Michael Gross. Non-contact video-based pulse rate measurement on a mobile service robot. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 1056–1062. IEEE, 2014. 1, 5

[44] Zhaodong Sun and Xiaobai Li. Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast. In *ECCV*, pages 492–510. Springer, 2022. 1, 3, 5, 6, 7

[45] Chihiro Takano and Yuji Ohta. Heart rate measurement based on a time-lapse image. *Medical engineering & physics*, 29(8):853–857, 2007. 1

[46] Suramya Tomar. Converting video formats with ffmpeg. *Linux Journal*, 2006(146):10, 2006. 7

[47] Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, Lijun Yin, Jeffrey F Cohn, and Nicu Sebe. Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In *CVPR*, pages 2396–2404, 2016. 6

[48] Parishwad P Vaidyanathan. *Multirate systems and filter banks*. Pearson Education India, 2006. 4

[49] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008. 6, 7

[50] Wenjin Wang, Albertus C Den Brinker, Sander Stuijk, and Gerard De Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2016. 2, 3, 5, 6, 7

[51] 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, 2015. 1, 2, 3, 5, 6

[52] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *CVPR*, pages 2566–2576, 2019. 3

[53] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM transactions on graphics (TOG)*, 31(4):1–8, 2012. 2

[54] Quanming Yao, Hansi Yang, Bo Han, Gang Niu, and James Tin-Yau Kwok. Searching to exploit memorization effect in learning with noisy labels. In *ICML*, pages 10789–10798. PMLR, 2020. 4

[55] Zitong Yu, Xiaobai Li, Xuesong Niu, Jingang Shi, and Guoying Zhao. Autohr: A strong end-to-end baseline for remote heart rate measurement with neural searching. *IEEE Signal Processing Letters*, 27:1245–1249, 2020. 6

[56] Zitong Yu, Wei Peng, Xiaobai Li, Xiaopeng Hong, and Guoying Zhao. Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement. In *ICCV*, pages 151–160, 2019. 1, 2

[57] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip HS Torr, and Guoying Zhao. Physformer: facial video-based physiological measurement with temporal difference transformer. In *CVPR*, pages 4186–4196, 2022. 3, 5, 6

[58] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Michael C Mozer, and Yoram Singer. Identity crisis: Memorization and generalization under extreme overparameterization. *arXiv preprint arXiv:1902.04698*, 2019. 4

[59] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021. 4

[60] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Rongrong Ji. Pyramidal person re-identification via multi-loss dynamic training. In *CVPR*, pages 8514–8522, 2019. 4