Remote Heart Rate Measurement from Highly Compressed Facial Videos: an End-to-end Deep Learning Solution with Video Enhancement

Zitong Yu[1*], Wei Peng[1*], Xiaobai Li[1], Xiaopeng Hong[2,4,1], Guoying Zhao[3,1†]
[1]Center for Machine Vision and Signal Analysis, University of Oulu, Finland
[2]Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, PRC
[3]School of Information and Technology, Northwest University, PRC; [4]Peng Cheng Laborotory, China
{zitong.yu, wei.peng, xiaobai.li, xiaopeng.hong, guoying.zhao}@oulu.fi

Abstract

Remote photoplethysmography (rPPG), which aims at measuring heart activities without any contact, has great potential in many applications (e.g., remote healthcare). Existing rPPG approaches rely on analyzing very fine details of facial videos, which are prone to be affected by video compression. Here we propose a two-stage, end-to-end method using hidden rPPG information enhancement and attention networks, which is the first attempt to counter video compression loss and recover rPPG signals from highly compressed videos. The method includes two parts: 1) a Spatio-Temporal Video Enhancement Network (STVEN) for video enhancement, and 2) an rPPG network (rPPGNet) for rPPG signal recovery. The rPPGNet can work on its own for robust rPPG measurement, and the STVEN network can be added and jointly trained to further boost the performance especially on highly compressed videos. Comprehensive experiments are performed on two benchmark datasets to show that, 1) the proposed method not only achieves superior performance on compressed videos with high-quality videos pair, 2) it also generalizes well on novel data with only compressed videos available, which implies the promising potential for real-world applications.

1. Introduction

Electrocardiography (ECG) and Photoplethysmograph (PPG) provide common ways for measuring heart activities. These two types signals are important for healthcare applications since they provide the measurement of both basic average heart rate (HR) and more detailed information like heart rate variability (HRV). However, these signals are mostly measured from skin-contact ECG/BVP sensors, which may cause discomfort and are inconvenient for long-term monitoring. To solve this problem, remote photoplethysmography (rPPG) , which targets to measure heart activity remotely and without any contact, has been developing rapidly in recent years [4, 12, 19, 18, 31, 32, 22].

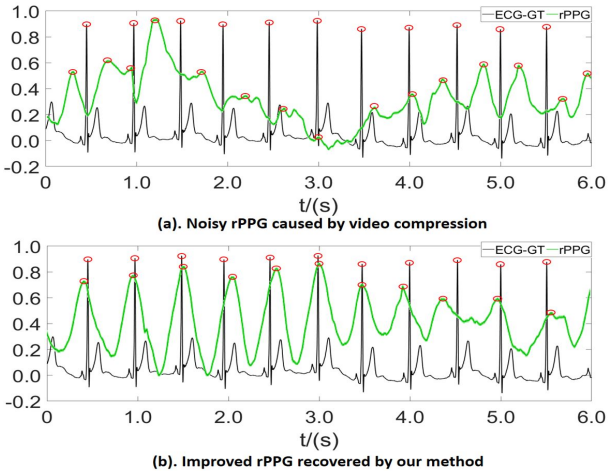

Figure 1. rPPG measurement from highly compressed videos. Due to video compression artifact and rPPG information loss, the rPPG in (a) has very noisy shape and inaccurate peak counts which lead to erroneous heart rate measures, while after video enhancement by STVEN, the rPPG in (b) shows more regular pulse shape with accurate peak locations comparing to the ground truth ECG.

However, most previous rPPG measurement works did not take the influence of video compression into consideration, whereas the fact is that most videos captured by commercial cameras are compressed through different compression codecs with various bitrates. Recently, two works [7, 16] pointed out and demonstrated that the performance of rPPG measurement dropped to various extents when using compressed videos with different bitrates. As shown in Fig. 1(a), rPPG signals measured from highly compressed videos usually suffer from noisy curve shape and inaccurate peak locations due to information loss caused by both intra-frame and inter-frame coding of the video compression process. Video compression is inevitable for remote services considering the convenient storage and transmis-

*Equal contribution † Corresponding author

sion in Internet. Thus it is of great practical value to develop rPPG methods that can work robustly on highly compressed videos. However, no solution has been proposed yet to counter this problem.

To address this problem, we propose a two-stage, end-to-end method using hidden rPPG information enhancement and attention networks, which can counter video compression loss and recover rPPG signals from highly compressed facial videos. Figure 1(b) illustrates the advantages of our method on rPPG measurement from highly compressed videos. Our contributions include:

- To our best knowledge, we provide the first solution for robust rPPG measurement directly from compressed videos, which is an end-to-end framework made up of a video enhancement module STVEN (Spatio-Temporal Video Enhancement Network) and a powerful signal recovery module rPPGNet.
- The rPPGNet, featured with a skin-based attention module and partition constraints, can measure accurately at both HR and HRV levels. Compared with previous works which only output simple HR numbers[17, 25], the proposed rPPGNet produces much richer rPPG signals with curve shapes and peak locations. Moreover, It outperforms state-of-art methods on various video formats of a benchmark dataset even without using the STVEN module.
- The STVEN, which is a video-to-video translation generator aided with fine-grained learning, is the first video compression enhancement network to boost rPPG measurement on highly compressed videos.
- We conduct cross-dataset test and show that the STVEN can generalize well to enhance unseen, highly compressed videos for rPPG measurement, which implies promising potential in real-world applications.

2. Related Work

Remote Photoplethysmography Measurement. In past few years, several traditional methods explored rPPG measurement from videos by analyzing subtle color changes on facial regions of interest (ROI), including blind source separation [19, 18], least mean square [12], majority voting [10] and self-adaptive matrix completion [31]. However, ROI selection in these works were customized or arbitrary, which may cause information loss. Theoretically speaking, all skin pixels can contribute to the rPPG signals recovery. There are other traditional methods which utilized all skin pixels for rPPG measurement, e.g., chrominance-based rPPG (CHROM) [4], projection plane orthogonal to the skin tone (POS) [35], and spatial subspace rotation [36, 34, 13]. All these methods treat each skin pixel with equal contribution, which is against the fact that different skin parts may bear different weights for rPPG recovery.

More recently, a few deep learning based methods were proposed for average HR estimation, including SynRhythm [17], HR-CNN [25] and DeepPhys [3]. Convolutional neural networks (CNN) were also employed for skin segmentation [2, 28] and then to predict HR from skin regions. These methods were based on spatial 2D CNN, which failed to capture temporal features which are essential for rPPG measurement. Moreover, the skin segmentation task was treated separately from the rPPG recovery task, which lacks the mutual feature sharing between such two highly related tasks.

Video Compression and Its Impact for rPPG. In real-world applications, video compression is widely used because of its great storage capacities with minimal quality degradation. Numerous codecs for video compression have been developed as standards of the Moving Picture Experts Group (MPEG) and International Telecommunication Union Telecommunication Standardization Sector (ITU-T). These include MPEG-2 Part 2/H.262 [8] and the low bitrate standard MPEG-4 Part 2/H.263 [21]. Current-generation standard AVC/H.264 [37] achieves an approximate doubling in encoding efficiency over H.262 and H.263. More recently, next-generation standard HEVC/H.265 [27] utilizes increasingly complex encoding strategies for an approximate doubling in encoding efficiency over H.264.

In the stage of video coding, compression artifacts are inevitable as a result of quantization. Specifically, the existing compression standards drop subtle changes that human eyes cannot see. It does not favor the purpose of rPPG measurement, which mainly relies on subtle changes at invisible level. The impact of video compression on rPPG measurement was not explored until very recently. Three works[7, 16, 24] consistently demonstrated that the compression artifacts do reduce the accuracy of HR estimation. However, these works only tested on small-scale private datasets using traditional methods, and it was unclear whether compression also impacted deep learning based rPPG methods on large dataset. Furthermore, these works just pointed out the problem of compression on rPPG, but no solution has been proposed yet.

Quality Enhancement for Compressed Video. Fueled by the high performance of deep learning, several works introduce it to enhance the quality of compressed videos and get promising results, including ARCNN [5], deep residual denoising neural networks (DnCNN) [39], generative adversarial networks [6] and multi-frame quality enhancement network [38]. However, all of them were designed for solving general compression problems or other tasks like object detection, but not for rPPG measurement. There are two works [15, 40] about rPPG recovery from low quality videos. The [15] focused on frame resolutions but not about video compression and format. The other one [40] tried to address the rPPG issue on compressed videos, but the ap-

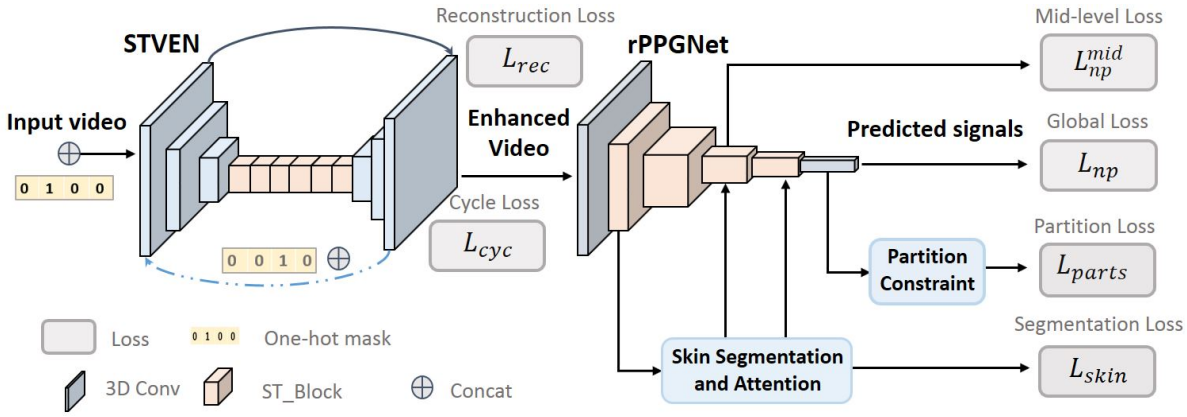

Figure 2. Illustration of the overall framework. There are two models in our framework: video quality enhancement model STVEN (left) and rPPG recovery model rPPGNet (right). Both of them work well by learning with corresponding loss functions. We will also introduce an elaborate joint training, which further improves the rPPG recovery performance.

proach was only on bio-signal processing level AFTER the rPPG was extracted, which has nothing to do with video enhancement. To the best of our knowledge, no video enhancement method has ever been proposed for the problem of rPPG recovery from highly compressed videos.

In order to overcome the above-mentioned drawbacks and fill in the blank, we propose a two-stage, end-to-end deep learning based method for rPPG measurement from highly compressed videos.

3. Methodology

As a two-stage end-to-end method, we will first introduce our video enhancement network STVEN in Section 3.1, then introduce the rPPG signal recovery network rPPGNet in Section 3.2, and at last explain how to jointly train these two parts for boosting performance. The overall framework is shown in Fig 2.

3.1. STVEN

For the sake of enhancing the quality of highly compressed videos, we present a video-to-video generator called Spatio-Temporal Video Enhancement Networks (STVEN), which is shown in the left of Fig.2. Here we perform a fine-grained learning by assuming that compression artifacts from different compression bitrates are with different distributions. As a result, compressed videos are placed into the buckets $[0, 1, 2, ..., \mathcal{C}]$ denoted as $\mathbb{C}$ based on their compression bitrate. Here, 0 and $\mathcal{C}$ represent videos with lowest and highest compression rate, respectively. Let $c_k^\tau = [c_{k1}, c_{k2}, ..., c_{k\tau}]$ be a sequence of the compressed video with length of τ for $k \in \mathbb{C}$. Then our goal is to train a generator $\mathcal{G}$ which can enhance the quality of compressed videos c_k^τ so that the distribution of the video is identical to the one of which $k = 0$, that is original video c_0^τ. Let say the output of generator $\mathcal{G}$ is $\hat{c}_0^\tau = [\hat{c}_{01}, \hat{c}_{02}, ..., \hat{c}_{0\tau}]$. Then the conditional distribution of $\hat{c}_0^\tau$ given input videos c_k^τ and video quality target 0 should be equal to the c_0^τ given input videos c_k^τ and target 0. That is

$$p(\hat{c}_0^\tau | c_k^\tau, 0) = p(c_0^\tau | c_k^\tau, 0). \tag{1}$$

By learning to match the video distributions, our model generates the video sequences with the quality being enhanced. Likewise, in order to make the model more generalizable, the framework is also set to be able to compress the original video with a specific compression bitrate. This means that when our model is fed with video c_0^τ and outputs lower quality target k, the model $\mathcal{G}$ should also be able to generate the video which fits the distribution with the specific compression bitrate k. That is

$$p(\hat{c}_k^\tau | c_0^\tau, k) = p(c_k^\tau | c_0^\tau, k), \tag{2}$$

here $\hat{c}_k^\tau$ is the output of our generator with the inputs c_0^τ and k. Therefore, there will be two parts of the loss function L_{rec} in STVEN: one is the reconstruction loss, for which we introduce a mean squared error (MSE) to deal with the video details, and the other one is the lose for compression reconstruction, here we employ a L1 loss. Then

$$\begin{aligned} L_{rec} =& E_{k\sim\mathbb{C},t}(c_0^\tau(t) - \mathcal{G}(c_k^\tau, 0)(t))^2 \\ &+ E_{k\sim\mathbb{C},t}||c_k^\tau(t) - \mathcal{G}(c_0^\tau, k)(t)|| \end{aligned} \tag{3}$$

Here $t \in [1, \tau]$ is the t-th frame of the output video. In addition, like in [41], we also introduce a cycle-loss for better reconstruction. In this way, we expect our model to satisfy this case: when taking $(\hat{c}_0^\tau)$ of $\mathcal{G}$, which is fed with c_k^τ and the specific compression bitrate label 0, and the compression bitrate label k as its inputs, the following output should match the distribution of the initial input videos. Similarly, we perform the cycle processing for original video. As a result, the cycle loss L_{cyc} in STVEN is

$$\begin{aligned} L_{cyc} =& E_{k\sim\mathbb{C},t}||c_k^\tau(t) - \mathcal{G}(\mathcal{G}(c_k^\tau, 0), k)(t)|| \\ &+ E_{k\sim\mathbb{C},t}||c_0^\tau(t) - \mathcal{G}(\mathcal{G}(c_0^\tau, k), 0)(t)||. \end{aligned} \tag{4}$$

	Layer	Output size	Kernel size
STVEN	Conv_1	$64 \times T \times 128 \times 128$	$3 \times 7 \times 7$
	Conv_2	$128 \times T \times 64 \times 64$	$3 \times 4 \times 4$
	Conv_3	$512 \times \frac{T}{2} \times 32 \times 32$	$4 \times 4 \times 4$
	ST_Block	$512 \times \frac{T}{2} \times 32 \times 32$	$[3 \times 3 \times 3] \times 6$
	DConv_1	$128 \times T \times 64 \times 64$	$4 \times 4 \times 4$
	DConv_2	$64 \times T \times 128 \times 128$	$1 \times 4 \times 4$
	DConv_3	$3 \times T \times 128 \times 128$	$1 \times 7 \times 7$
rPPGNet	Conv_1	$32 \times T \times 64 \times 64$	$1 \times 5 \times 5$
	ST_Block	$64 \times T \times 16 \times 16$	$[3 \times 3 \times 3] \times 4$
	SGAP	$64 \times T \times 1 \times 1$	$1 \times 16 \times 16$
	Conv_2	$1 \times T \times 1 \times 1$	$1 \times 1 \times 1$

Table 1. The architecture of STVEN and rPPGNet. Here "Conv_x" means 3D convolution filters and "DConv_x" denotes 3D transposed convolution filters. "ST_Block" represents spatio-temporal block [30], which is constructed by two sets of cascaded 3D convolution filters with kernel size of $1 \times 3 \times 3$ and $3 \times 1 \times 1$, respectively. Besides, we introduce instance normalization and ReLU into STVEN while batch normalization and ReLU into rPPGNet. "SGAP" is short for spatial global average pooling.

Therefore, the total loss of STVEN L_{STVEN} is the sum of L_{rec} and L_{cyc}. To achieve this goal, we build our model STVEN with a spatial-temporal convolutional neural network. The architecture is composed of two downsampling layers and two upsampling layers at the two ends, with six spatio-temporal blocks in the middle. The details of the architecture is shown in the top of Table. 1.

3.2. rPPGNet

The proposed rPPGNet is composed of a spatio-temporal convolutional network, a skin-based attention module and a partition constraint module. Skin-based attention helps to adaptively selected skin regions, and partition constraint is introduced for learning better rPPG feature representation.

Spatio-Temporal Convolutional Network. Previous works like [4, 35], usually projected spatial pooled RGB into another color space for better representation of the rPPG information. Then temporal context based normalization was used to get rid of irrelevant info (e.g., noise caused by illumination or motion). Here we merge these two steps into one model and propose an end-to-end spatio-temporal convolutional network, which takes T-frame face images with RGB channels as the inputs and outputs rPPG signals directly. The backbone and architecture of rPPGNet is shown in Fig. 2 and Table. 1 respectively.

Aiming to recover rPPG signals $y \in \mathbb{R}^T$, which should have accurate pulse peak locations compared with the corresponding ground truth ECG signals $y^g \in \mathbb{R}^T$, negative Pearson correlation is used to define the loss function. It can be formulated as

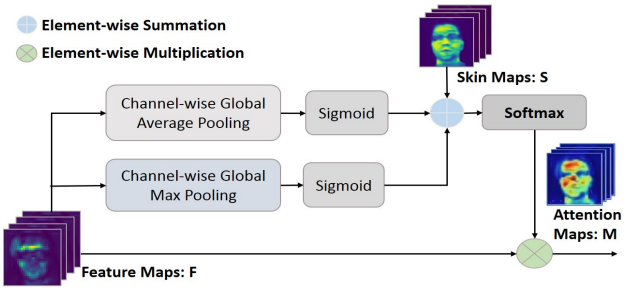

Figure 3. Illustration of the skin-based attention module of the rPPGNet, which is parameter-free. It assigns importance to different locations in accordance with both skin confidence and rPPG feature maps. The softmax operation can be either spatial-wise or spatio-temporal-wise.

$$L_{np} = 1 - \frac{T\sum_{i=1}^{T} y_i y_i^g - \sum_{i=1}^{T} y_i \sum_{i=1}^{T} y_i^g}{\sqrt{(T\sum_{i=1}^{T} y_i^2 - (\sum_{i=1}^{T} y_i)^2)(T\sum_{i=1}^{T} (y_i^g)^2 - (\sum_{i=1}^{T} y_i^g)^2)}}. \quad (5)$$

Unlike Mean Square Error (MSE), our loss is to minimize the linear similarity error instead of the point-wise intensity error. We tried MSE loss in prior test, which achieved much worse performance because the intensity values of signals are irrelevant with our task (i.e., to measure accurate peak locations) and introduces extra noise inevitably.

We also aggregate the mid-level features (outputs of the third ST_Block) into pseudo signals and then constrain them by L_{np}^{mid} for stable convergence. So the basic learning object for recovering rPPG singals is described as

$$L_{rPPG} = \alpha L_{np} + \beta L_{np}^{mid}, \quad (6)$$

where α and β are the weights for balancing the loss.

Skin Segmentation and Attention. Various skin regions have varying density degrees of blood vessels as well as biophysical parameter maps (melanin and haemoglobin), thus contribute at different levels for rPPG signal measurement. So the skin segmentation task is highly related to rPPG signals recovery task. These two tasks can be treated as a multi-task learning problem. Thus we employ a skin segmentation branch after the first ST_Block. The skin segmentation branch projects the shared low-level spatio-temporal features into skin domain, which is implemented by spatial and channel-wise convolutions with residual connections. As there is no ground truth skin map in related rPPG datasets, we generate the binary labels for each frame by adaptive skin segmentation algorithms [29]. With these binary skin labels, the skin segmentation branch is able to predict high quality skin maps $S \in \mathbb{R}^{T \times H \times W}$. Here we adopt binary cross entropy L_{skin} as the loss function.

In order to eliminate the influence of non-skin regions and enhance dominant rPPG features, we construct a skin-based parameter-free attention module which refines the

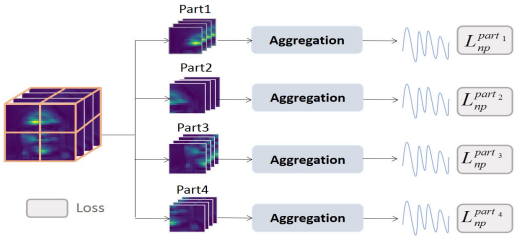

Figure 4. Partition constraints with $N = 4$.

rPPG features by predicted attention maps $M \in \mathbb{R}^{T\times H\times W}$. The module is illustrated in Fig. 3 and the attention maps are computed as

$$M(F,S) = \varsigma(\sigma(AvgPool(F)) + \sigma(MaxPool(F)) + S), \quad (7)$$

where S and F donote the predicted skin maps and rPPG feature maps respectively. σ and ς represent the sigmoid and softmax function respectively.

Partition Constraint. In order to help the model learn more concentrated rPPG features, local partition constraint is introduced. As shown in Fig. 4, the deep features $D \in \mathbb{R}^{C\times T\times H\times W}$ are divided into N uniform spatio-temporal parts $D_i \in \mathbb{R}^{C\times T\times (H/\sqrt{N})\times (W/\sqrt{N})}$, $i \in \{1, 2, ..., N\}$. Afterwards, spatial global average pooling is adopted by each part-level feature for feature aggregation and an independent $1 \times 1 \times 1$ convolution filter is deployed for final signals prediction. The partition loss is described as $L_{parts} = \sum_{i=1}^{N} L_{np}^{part_i}$, where $L_{np}^{part_i}$ is the negative Pearson loss of the i-th part-level feature.

The partition loss can be considered as a dropout [26] for high-level features. It has a regularization effect because each partition loss is independent to each other, thus forcing part features to be powerful enough to recover the rPPG signal. In other words, via the partition constraint, the model can focus more on the rPPG signals instead of interference.

In sum, the loss function of rPPGNet can be written as

$$L_{rPPGNet} = L_{rPPG} + \gamma L_{skin} + \delta L_{parts}, \quad (8)$$

where γ and δ are the weights for balancing the loss.

3.3. Joint Loss Training

When STVEN is trained separately from rPPGNet, the output video cannot guarantee its effectiveness for the latter. Inspired by [14], we design an advanced joint training strategy to ensure that STVEN can enhance the video specifically in favor of rPPG recovery, which boosts the performance of rPPGNet even on highly compressed video.

First, we train the rPPGNet on the high quality videos with the training method described in Section 3.2. Second, we train the STVEN on compressed videos with different bitrates. Finally, we train the cascaded networks, which is illustrated in Fig. 2, with all high-level task model parameters fixed. Therefore, all the following loss functions are designed for the updating of STVEN. Here we employ an application-oriented joint training, where we prefer the end-to-end performance rather than the performance of both stages. In this training strategy, we take away the cycle-loss part since we expect STVEN to recover richer rPPG signals instead of irrelevant information loss during video compression. As a result, we only need to know its target label, and the compression labels of all input videos fed into STVEN can be simply set to 0 as default. This allows the model to be more generalizable since it does not require subjectively compression labeling of input videos, thus can work on novel videos with unclear compression rate. Besides, like [9], we also introduce a perceptual loss L_p for joint training. That is

$$L_p = \frac{1}{T_f W_f H_f} \sum_{t=1}^{T_f} \sum_{i=1}^{W_f} \sum_{j=1}^{H_f} (\phi(c_0^\tau)(t,i,j) - \phi(\mathcal{G}(c_k^T, 0))(t,i,j))^2. \quad (9)$$

Here, ϕ denotes a differentiable function in rPPGNet and the feature maps $\phi(x) \in \mathbb{R}^{T_f\times W_f\times H_f}$. Cost function in Eq. (9) keeps the recovered video and the original video consistent in the feature map space. Besides, we also let STVEN contribute directly to rPPG task by introducing L_{rPPG} as in Eq. (8). In the joint training, we use the rPPG signals recovered from high quality videos as a softer target for the updating of STVEN, and it converges faster and more steadily than using the ECG signals, which might be too far-fetched and challenging as the target for highly compressed videos, as our prior tests proved. In all, the joint cost function L_{joint} for STVEN can be formulated as

$$L_{joint} = L_{rPPGNet} + \varepsilon L_p + \rho L_{STVEN}, \quad (10)$$

here ε and ρ are hyper-parameters.

4. Experiments

We test the proposed system in four sub-experiments, the first three on OBF [11] dataset and the last one on MAHNOB-HCI [23] dataset. Firstly, we evaluate the rPPGNet on OBF for both average HR and HRV feature measurement. Secondly, we compress OBF videos and explore how video compression influence the rPPG measurement performance. Thirdly, we demonstrate that STVEN can enhance the compressed videos and boost the rPPG measurement performance on OBF. Finally, we cross test the joint system of STVEN and rPPGNet on MAHNOB-HCI, which has only compressed videos, to validate the generalizability of the system.

4.1. Datasets and Settings

Two datasets - **OBF** [11] and **MAHNOB-HCI** [23] are used in our experiments. The OBF is a recently release

Table 2. Performance comparison on OBF. HR is the averaged heart rate within 30 seconds, RF, LF, HF and LF/HF are HRV features that require finer inter-beat-interval measurement of rPPG signals. Smaller RMSE and bigger R values indicate better performance. "rPPGNet_base" denotes the spatio-temporal networks with L_{rPPG} constraint, while "Skin", "Parts" and "Atten" indicate corresponding modules of rPPGNet described in Section 3.2. "rPPGNet (full)" includes all modules of the rPPGNet.

Method	HR(bpm)			RF(Hz)			LF(u.n)			HF(u.n)			LF/HF		
	SD	RMSE	R	SD	RMSE	R	SD	RMSE	R	SD	RMSE	R	SD	RMSE	R
ROI_green [11]	2.159	2.162	0.99	0.078	0.084	0.321	0.22	0.24	0.573	0.22	0.24	0.573	0.819	0.832	0.571
CHROM [4]	2.73	2.733	0.98	0.081	0.081	0.224	0.199	0.206	0.524	0.199	0.206	0.524	0.83	0.863	0.459
POS [35]	1.899	1.906	0.991	0.07	0.07	0.44	0.155	0.158	0.727	0.155	0.158	0.727	0.663	0.679	0.687
rPPGNet_base	2.729	2.772	0.98	0.067	0.067	0.486	0.151	0.153	0.748	0.151	0.153	0.748	0.641	0.649	0.724
rPPGNet_base+Skin	2.548	2.587	0.983	0.067	0.067	0.483	0.145	0.147	0.768	0.145	0.147	0.768	0.616	0.622	0.749
rPPGNet_base+Skin+Parts	2.049	2.087	0.989	0.065	0.065	0.505	0.143	0.144	0.776	0.143	0.144	0.776	0.594	0.604	0.759
rPPGNet_base+Skin+Atten	2.004	2.051	0.989	0.065	0.065	0.515	0.137	0.139	0.79	0.137	0.139	0.79	0.591	0.601	0.76
rPPGNet (full)	**1.756**	**1.8**	**0.992**	**0.064**	**0.064**	**0.53**	**0.133**	**0.135**	**0.804**	**0.133**	**0.135**	**0.804**	**0.58**	**0.589**	**0.773**

dataset for study about remote physiological signal measurement. It contains 200 five-minute-long RGB videos recorded from 100 healthy adults and the corresponding ground truth ECG signals are also provided. The videos are recorded at 60 fps with resolution of 1920x2080, and compressed in MPEG-4 with average bitrate $\approx$ 20000 kb/s (file size $\approx$ 728 MB). The long videos are cut into 30-seconds-long clips for our training and testing. The MAHNOB-HCI dataset is one of the most widely used benchmark for remote HR measurement evaluations. It includes 527 facial videos with corresponding physiological signals from 27 subjects. The videos are recorded with 61 fps with resolution of 780x580, which are compressed in AVC/H.264, average bitrate $\approx$ 4200 kb/s. We use the EXG2 signal as the ground truth ECG in our experimental evaluation. We follow the same routine as in previous works [17, 25, 3] and use 30 seconds (frames 306 to 2135) of each video.

Highly Compressed Videos. Video compression was performed using the latest version of FFmpeg [1]. We used three codecs (MPEG4, x264 and x265) in order to implement the three mainstream compression standards (H.263, H.264 and H.265). In order to demonstrate the effect of STVEN on highly compressed videos (i.e., with small file size and bitrates below 1000 kb/s), we compressed OBF videos into three qualities levels of average bitrate (file size) = 1000 kb/s (36.4 MB),500 kb/s (18.2 MB) and 250 kb/s (9.1 MB). The bitrates (file size) are about 20, 40 and 80 times smaller than those of original videos respectively.

4.2. Implementation Details

Training Setting. For all facial videos, we use the Viola-Jones face detector [33] to detect and crop the coarse face area (see Figure 8 (a)) and remove background. We generate binary skin masks by open source Bob[1] with threshold=0.3 as the ground truth. All face and skin images are normalized to 128x128 and 64x64 respectively.

The proposed method is trained in Nvidia P100 using PyTorch. The length of each video clip is $T = 64$ while videos and ECG signals downsample into 30 fps and 30 Hz respectively. The partition for rPPGNet is $N = 4$. The weights for different losses are set as $\alpha = 1, \beta = 0.5, \gamma = 0.1, \delta = 0.5$. As a part of the input, the compression bitrate label k is represented by an one-hot mask vector. When joint training STVEN with rPPGNet, the loss balance weights $\varepsilon = 1, \rho = 1e-4$. Adam optimizer is used while learning rate is set to $1e{-}4$. We train rPPGNet for 15 epochs and STVEN for 20000 iterations. For the joint training, we fine-tuning STVEN for extra 10 epochs.

Performance Metrics. For evaluating the accuracy of recovered rPPG signals, we follow previous works [11, 17] and report both the average HR and several common HRV features on OBF dataset, and then evaluated several metrics of the average HR measurement on MAHNOB-HCI dataset. Four commonly used HRV features [11, 18] are calculated for evaluation, including respiratory frequency (RF) (in Hz), low frequency (LF), high frequency (HF) and LF/HF (in normalized units, n.u.). Both the recovered rPPGs and their corresponding ground truth ECGs go through the same process of filtering, normalization, and peak detection to obtain the inter-beat-intervals, from which the average HR and HRV features are calculated.

We report the most commonly used metrics for evaluating the performance, which include: the standard deviation (SD), the root mean square error (RMSE), the Pearson correlation coefficient (R), and the mean absolute error (MAE). $\triangle PSNR$ is also employed to evaluate changes of video quality before and after enhancement.

4.3. Results on OBF

OBF has large number of high quality video clips, which is suitable for verifying the robustness of our method in both average HR and HRV levels. We perform subject-independent 10-fold cross validation protocol to evaluate the rPPGNet and STVEN on the OBF dataset. At the testing stage, average HR and HRV features are calculated from

[1] https://gitlab.idiap.ch/bob/bob.ip.skincolorfilter

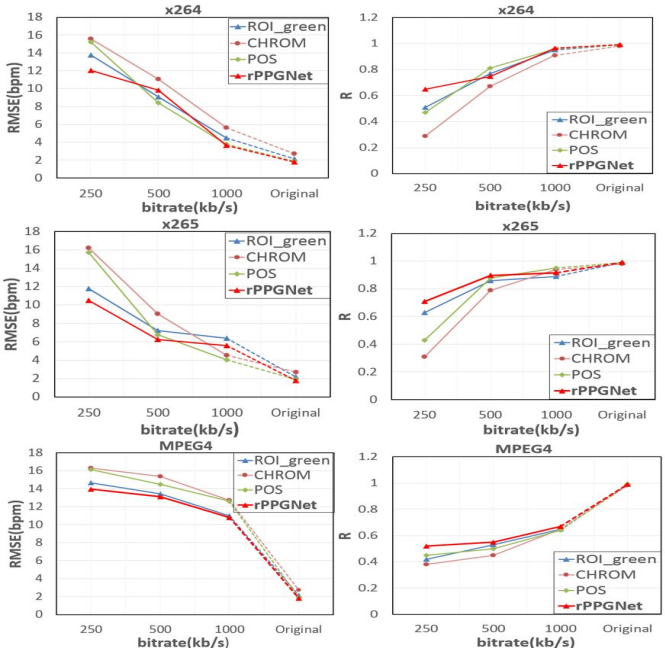

Figure 5. HR measurement on OBF videos at different bitrates: all methods' performance drops with bitrates, while for the same bitrate level, the rPPGNet outperforms other methods.

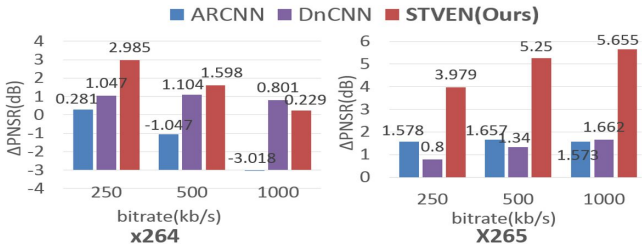

Figure 6. Performance of video quality enhancement networks.

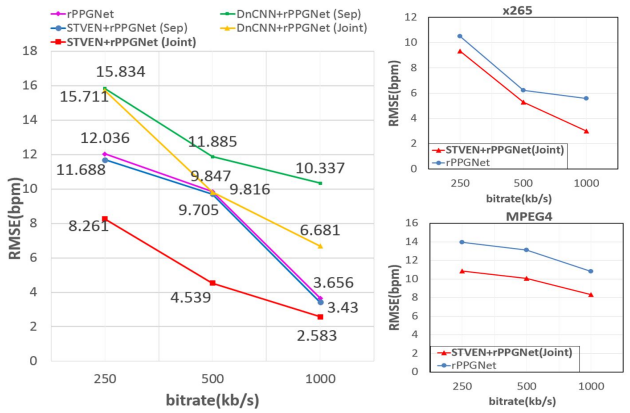

Figure 7. HR measurement using different enhancement methods on highly compressed videos of OBF, left: with x264 codec; right: with x265 and MPEG4 codecs (cross-testing). Smaller RMSE indicates better performance

output rPPG signals of 30 seconds length.

Evaluation of rPPGNet on High Quality Videos. Here, we re-implement several traditional methods [4, 11, 35] on original OBF videos and compare the results in Table. 2. The results show that rPPGNet (full) outperforms other methods for both averaged HR and HRV features. From ablation test results we can conclude that: 1) the skins segmentation module (the fifth row in Table. 2) slightly improves the performance with multi-task learning, which indicates these two tasks may have mutual hidden information. 2) The partition module (sixth row in Table. 2) further improves the performance by helping the model to learn more concentrated features. 3) Skin-based attention teaches the networks where to look and thus improves performance. In our observation, spatial attention with spatial-wise softmax operation works better than spatio-temporal attention, because in the rPPG recovery task the weights for different frames should be very close.

Evaluation of rPPGNet on Highly Compressed Videos. We compressed OBF videos into three bitrates levels (250, 500 and 1000 kb/s) with three codecs (MPEG4, x264 and x265) as described in Section 4.1, so that we have nine groups (3 by 3) of highly compressed videos. We evaluate the rPPGNet together with three other methods on each of the nine groups of videos, using 10-folds cross-validation as before. The results are illustrated in Fig. 5. From the figure we can see that, first, the performance of both traditional methods and rPPGNet drop when bitrate decreases, which is true for all three compression codecs. The observation is consistent with previous findings[16, 24] and proved that compression does impact rPPG measurement. Second, the important result is that when we compare at the same compression condition, rPPGNet can outperform other methods in most cases, especially very low bitrate of 250kb/s. This demonstrate the robustness of rPPGNet. But the accuracy at low bitrates is not satisfactory, and we hope to further improve the performance by video enhancement, i.e., using the proposed STVEN network.

Evaluation of rPPGNet with STVEN for Enhancement on Highly Compressed Videos. Firstly, we demonstrate the STVEN does enhance the video quality on general level in terms of $\triangle PSNR$. As shown in Fig. 6, the $\triangle PSNR$ of videos enhanced by STVEN are larger than zero, which indicate quality improvement. We also compared the STVEN to two other enhancement networks (ARCNN[5] and DnCNN[39]) and STVEN achieved even larger $\triangle PSNR$ than the other two methods.

Then we cascade STVEN with rPPGNet for verifying that the video enhancement model can boost performance of rPPGNet for HR measurement. We compare the performance of two enhancement networks (STVEN vs. DnCNN[39]) with two training strategies (separate training vs. joint training) on x264 compressed videos. Separate training means that the enhancement networks are pretrained on highly compressed videos and the rPPGNet was

Table 3. Results of average HR measurement on MAHNOB-HCI.

Method	HR_{SD} (bpm)	HR_{MAE} (bpm)	HR_{RMSE} (bpm)	HR_R
Poh2011 [18]	13.5	-	13.6	0.36
CHROM [4]	-	13.49	22.36	0.21
Li2014 [12]	6.88	-	7.62	0.81
SAMC [31]	5.81	4.96	6.23	0.83
SynRhythm [17]	10.88	-	11.08	-
HR-CNN [25]	-	7.25	9.24	0.51
DeepPhys [3]	-	4.57	-	-
rPPGNet	7.82	5.51	7.82	0.78
STVEN+rPPGNet	**5.57**	**4.03**	**5.93**	**0.88**

pre-trained on high quality original videos, while joint training fine tunes the results of the two separate training with joint loss of the two tasks. The results in Fig. 7(left) shows that: for rPPG recovery and HR measurement on highly compressed videos, 1) STVEN helps to boost the performance of rPPGNet while DnCNN does not; and 2)joint training works better than separate training. It is surprising that STVEN boosts rPPGNet while DnCNN[39] suppresses rPPGNet in both separate training and joint training modes, which may be caused by the excellent spatio-temporal structure with fine-grained learning in STVEN and the limitation of the single-frame model of DnCNN. The generalization ability of STVEN-rPPGNet is shown in Fig. 7(right), in which the joint system trained on x264 videos was cross-tested on MPEG4 and x265 videos. Due to the quality and rPPG information enhancement by STVEN, rPPGNet is able to measure more accurate HR from untrained videos with MPEG4 and x265 compression.

4.4. Results on MAHNOB-HCI

In order to verify the generalization of our method, we evaluate our methods on the MAHNOB-HCI dataset. MAHNOB-HCI is the most widely used dataset in HR measurement and the video samples are challenging because of the high compression rate and spontaneous motions, e.g., facial expressions. Subject-independent 9-fold cross validation protocol (3 subjects in a fold, totally 27 subjects) is adopted. As there are no original high quality videos available, the STVEN is trained with x264 highly compressed videos on OBF firstly and then cascades with the rPPGNet trained on MAHNOB-HCI for testing. Compared to the state-of-the-art methods in Table. 3, our rPPGNet outperforms the deep learning based methods [17, 25] in subject-independent protocol. With the help of video enhancement with richer rPPG information via STVEN, our two-stage method (STVEN+rPPGNet) surpasses all other methods. It indicates that STVEN can cross-boost the performance even when high-quality videos ground truth are not available.

4.5. Visualization and Discussion.

In Fig. 8, we visualize an example to show the interpretability of our STVEN+rPPGNet method. The predicted attention map from rPPGNet Fig. 8(c) focuses on the skin regions with strongest rPPG information (e.g., forehead and cheeks), which is in accordance with the priori knowledge mentioned in [32]. As shown in Fig. 8(b), the STVEN enhanced face image seems to have richer rPPG information and stronger pulsatile flows in similar skin regions, which indicates the consistency of Fig. 8(c).

Figure 8. Visualization of model output images. (a) face image in compressed video; (b) STVEN enhanced face image; (c) rPPGNet predicted attention map.

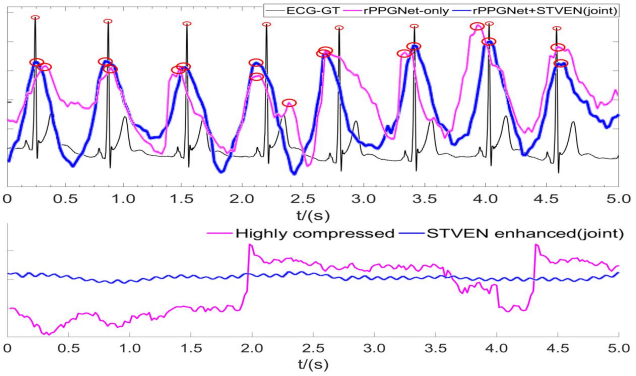

Figure 9. Predicted rPPG signals (top) and corresponding video PSNR curves (bottom).

We also plot the rPPGNet recovered rPPG signals on highly compressed videos with and without STVEN. As shown in Fig. 9(top), benefited from the enhancement from STVEN, the predicted signals are with more accurate IBIs. Besides, Fig. 9(bottom) shows less objective quality (PSNR) fluctuation of the highly compressed videos with STVEN enhancement, which seems to help recover smoother and robust rPPG signals.

5. Conclusions and Future Work

In this paper, we proposed an end-to-end deep learning based method for rPPG signals recovery from highly compressed videos. The STVEN is used to enhance the videos, and the rPPGNet is cascaded to recover rPPG signals for further measurement. In future, we will try using compression related metrics like PSNR-HVS-M [20] to constrain the enhancement model STVEN. Moreover, we will also explore ways of building a novel metric for evaluating the video quality specially for the purpose of rPPG recovery.

Aknowledgement This work was supported by the National Natural Science Foundation of China (No. 61772419), Tekes Fidipro Program (No. 1849/31/2015), Business Finland Project (No. 3116/31/2017), Academy of Finland, and Infotech Oulu.

References

[1] Fabrice Bellard and M. Niedermayer. Ffmpeg. [online]. available: http://ffmpeg.org. 6

[2] Sitthichok Chaichulee, Mauricio Villarroel, Joao Jorge, Carlos Arteta, Gabrielle Green, Kenny McCormick, Andrew Zisserman, and Lionel Tarassenko. Multi-task convolutional neural network for patient detection and skin segmentation in continuous non-contact vital sign monitoring. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 266–272. IEEE, 2017. 2

[3] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 349–365. 2018. 2, 6, 8

[4] 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, 4, 6, 7, 8

[5] Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang. Compression artifacts reduction by a deep convolutional network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 576–584, 2015. 2, 7

[6] Leonardo Galteri, Lorenzo Seidenari, Marco Bertini, and Alberto Del Bimbo. Deep generative adversarial compression artifact removal. In *ICCV, 2017*. 2

[7] Sebastian Hanfland and Michael Paul. Video format dependency of ppgi signals. In *Proceedings of the International Conference on Electrical Engineering*, 2016. 1, 2

[8] ITU-T. Rec. h.262 - information technology - generic coding of moving pictures and associated audio information: Video. *International Telecommunication Union Telecommunication Standardization Sector (ITU-T), Tech. Rep.*, 1995. 2

[9] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 5

[10] Antony Lam and Yoshinori Kuno. Robust heart rate measurement from video using select random patches. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3640–3648, 2015. 2

[11] Xiaobai Li, Iman Alikhani, Jingang Shi, Tapio Seppanen, Juhani Junttila, Kirsi Majamaa-Voltti, Mikko Tulppo, and Guoying Zhao. The obf database: A large face video database for remote physiological signal measurement and atrial fibrillation detection. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 242–249. IEEE, 2018. 5, 6, 7

[12] Xiaobai Li, Jie Chen, Guoying Zhao, and Matti Pietikainen. Remote heart rate measurement from face videos under realistic situations. pages 4264–4271, 2014. 1, 2, 8

[13] Xiaobai Li, Xiaopeng Hong, Antti Moilanen, Xiaohua Huang, Tomas Pfister, Guoying Zhao, and Matti Pietikainen. Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. *IEEE Transactions on Affective Computing*, 2017. 2

[14] Ding Liu, Bihan Wen, Xianming Liu, Zhangyang Wang, and Thomas S Huang. When image denoising meets high-level vision tasks: A deep learning approach. In *IJCAI, 2018*. 5

[15] Daniel McDuff. Deep super resolution for recovering physiological information from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1367–1374, 2018. 2

[16] Daniel J McDuff, Ethan B Blackford, and Justin R Estepp. The impact of video compression on remote cardiac pulse measurement using imaging photoplethysmography. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 63–70. IEEE, 2017. 1, 2, 7

[17] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Synrhythm: Learning a deep heart rate estimator from general to specific. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3580–3585. 2018. 2, 6, 8

[18] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, 58(1):7–11, 2010. 1, 2, 6, 8

[19] 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

[20] Nikolay Ponomarenko, Flavia Silvestri, Karen Egiazarian, Marco Carli, Jaakko Astola, and Vladimir Lukin. On between-coefficient contrast masking of dct basis functions. In *Proceedings of the third international workshop on video processing and quality metrics*, volume 4, 2007. 8

[21] Atul Puri and Alexandros Eleftheriadis. Mpeg-4: An object-based multimedia coding standard supporting mobile applications. *Mobile Networks and Applications*, 3(1):5–32, 1998. 2

[22] Jingang Shi, Iman Alikhani, Xiaobai Li, Zitong Yu, Tapio Seppänen, and Guoying Zhao. Atrial fibrillation detection from face videos by fusing subtle variations. *IEEE Transactions on Circuits and Systems for Video Technology*, DOI 10.1109/TCSVT.2019.2926632, 2019. 1

[23] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2012. 5

[24] Radim Spetlík, Jan Cech, and Jiri Matas. Non-contact reflectance photoplethysmography: Progress, limitations, and myths. In *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*, pages 702–709. IEEE, 2018. 2, 7

[25] Radim Špetlík, Vojtech Franc, and Jirí Matas. Visual heart rate estimation with convolutional neural network. In *BMVC, 2018*. 2, 6, 8

[26] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 5

[27] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video

coding (hevc) standard. *IEEE Transactions on circuits and systems for video technology*, 22(12):1649–1668, 2012. 2

[28] Chuanxiang Tang, Jiwu Lu, and Jie Liu. Non-contact heart rate monitoring by combining convolutional neural network skin detection and remote photoplethysmography via a low-cost camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1309–1315, 2018. 2

[29] Michael J Taylor and Tim Morris. Adaptive skin segmentation via feature-based face detection. In *Real-Time Image and Video Processing 2014*, volume 9139, page 91390P. International Society for Optics and Photonics, 2014. 4

[30] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 4

[31] 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. pages 2396–2404, 2016. 1, 2, 8

[32] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008. 1, 8

[33] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *null*, page 511. IEEE, 2001. 6

[34] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3360–3367. Citeseer, 2010. 2

[35] 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, 2017. 2, 4, 6, 7

[36] 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. 2

[37] Thomas Wiegand, Gary J Sullivan, Gisle Bjontegaard, and Ajay Luthra. Overview of the h. 264/avc video coding standard. *IEEE Transactions on circuits and systems for video technology*, 13(7):560–576, 2003. 2

[38] Ren Yang, Mai Xu, Zulin Wang, and Tianyi Li. Multi-frame quality enhancement for compressed video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6664–6673, 2018. 2

[39] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017. 2, 7, 8

[40] Changchen Zhao, Chun-Liang Lin, Weihai Chen, and Zhengguo Li. A novel framework for remote photoplethysmography pulse extraction on compressed videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1299–1308, 2018. 2

[41] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 3