Siamese-rPPG Network: Remote Photoplethysmography Signal Estimation from Face Videos

Yun-Yun Tsou
National Tsing Hua University
Hsinchu, Taiwan
tsou0320@gmail.com

Yi-An Lee*
National Tsing Hua University
Hsinchu, Taiwan
s107062576@m107.nthu.edu.tw

Chiou-Ting Hsu
National Tsing Hua University
Hsinchu, Taiwan
cthsu@cs.nthu.edu.tw

Shang-Hung Chang
Linkou Chang Gung Memorial Hospital
Taoyuan, Taiwan
afen.chang@gmail.com

ABSTRACT

Remote photoplethysmography (rPPG) is a contactless method for heart rate (HR) estimation from face videos. In this paper, we propose to estimate rPPG signals directly from input video sequences in an end-to-end manner. We propose a novel Siamese-rPPG network to simultaneously learn the heterogeneous and homogeneous features from two facial regions. Furthermore, to analyze the temporal periodicity of rPPG signals, we construct the network with 3D CNNs and jointly train the two-branch model under the negative Pearson loss function. Experimental results on three benchmark datasets: COHFACE, UBFC, and PURE, show that our method significantly outperforms existing methods with a large margin.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; *Computer vision tasks*; Activity recognition and understanding;

KEYWORDS

siamese network, remote photoplethysmography, heart rate detection, Pearson correlation, region-of-interest

ACM Reference Format:
Yun-Yun Tsou, Yi-An Lee, Chiou-Ting Hsu, and Shang-Hung Chang. 2020. Siamese-rPPG Network: Remote Photoplethysmography Signal Estimation from Face Videos. In *The 35th ACM/SIGAPP Symposium on Applied Computing (SAC '20), March 30-April 3, 2020, Brno, Czech Republic.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3341105.3373905

1 INTRODUCTION

Heart rate (HR) is a major health indicator of human body and has been widely used for the diagnosis of heart diseases (e.g., atrial fibrillation). Electrocardiography (ECG) and Photoplethysmography (PPG) are commonly used to monitor the heart rate through specific contact devices. Although these contact devices provide accurate readings, their application requires specific devices and professional attention and can hardly extend to mass monitoring for a large group of subjects.

Therefore, a number of contactless video-based methods [13, 15, 30] have been developed. Among these methods, remote photoplethysmography (rPPG), which is extracted by analyzing blood volume changes in optical information, has attracted enormous research interests. Nevertheless, without resorting to any contact devices, these video-based methods are inherently vulnerable to environmental interference (e.g., illumination) and subjects' motion (e.g, body and muscular movement) during the recording stage. Recently, with the success of deep learning, several learning-based methods [5, 20, 22, 25] have been developed for remote heart rate detection through rPPG estimation. As rPPG signals relate to chrominance changes reflected on skin, their periodic variation also reflects the periodicity of HR. Therefore, rPPG signals carry diagnostically important information and can be used for heart rate estimation, respiration detection, or even face anti-spoofing [8, 17].

Previous methods [5, 20, 22] usually require quite a few image pre-processing steps, such as facial landmarks detection, region-of-interest (ROI) detection, and format conversion (e.g., conversion of video frames into spatial-temporal maps). However, because these multi-stage methods do not directly estimate rPPG or HR from input videos, their practicability in real-world scenario is doubtful. Moreover, any pre-processing step may unintentionally diminish the subtle chrominance changes in face videos and thus leads to wrong estimation of rPPG signals. Detection of ROIs is also very unstable along a sequence of frames, especially when the target regions are partially or completely occluded during the recoding stage.

Therefore, several efforts have been taken to solve these problems. In [5, 25], the authors designed end-to-end networks with 2D CNNs to estimate the heart rate. However, their performance is far from satisfactory; one possible reason could be that the temporal features are not well characterized with 2D convolutions. In [31], a spatial-temporal network is proposed to measure rPPG signals from face videos. This spatial-temporal network is constructed with

*The first two authors contributed equally to this paper

SAC '20, March 30-April 3, 2020, Brno, Czech Republic

ACM ISBN 978-1-4503-6866-7/20/03…$15.00
https://doi.org/10.1145/3341105.3373905

3D convolutional neural networks and indeed achieve good performance. However, the network requires extensive computation and is impractical for real-world applications.

To tackle the aforementioned problems, in this paper, we propose an end-to-end method to estimate rPPG signals directly from face videos. We use the Siamese network architecture to simultaneously learn rPPG signals from two face regions. Moreover, we construct each branch network with 3D convolutional layers to explicitly characterize the spatial and temporal information of the videos. Each branch network is used to encode one region-of-interest into its correspoding one-dimensional rPPG signals. In the Siamese network, the two branches have the same architecture with shared-weights. Therefore, we can learn their distinctive yet complementary characteristics from different face regions. The learned features are therefore more robust and insensitive to environmental variations. By fusing and jointly train the two branches under the Negative Pearson loss function, the proposed Siamese-rPPG network significantly benefits from the two predictions and achieve state-of-the-art performance on three benchmark datasets: COHFACE [10], PURE [23] and UBFC-RPPG [3], with a great margin.

Our contributions are summarized below:

- We propose an end-to-end method to predict rPPG signals directly from face videos without any pre-processing step.
- We construct the network in terms of 3D CNNs and explicitly model the spatial and temporal characteristics of rPPG signals from face videos.
- With the weight-sharing architecture, the proposed Siamese-rPPG network successfully learns robust, distinctive, and complementary features and achieves state-of-art performance.

2 BACKGROUND AND RELATED WORK

Photoplethysmography (PPG) is an optically obtained plethysmogram and is used to detect blood volume changes in the microvascular bed of tissue [1]. The technique was first discovered during 1930s [9]. Later on, Verkruysse [26] discovered that the signal can be detected with cameras remotely and thus called this signal as "remote photoplethysmography" (rPPG). The most appealing feature of rPPGs is their retrieval does not require high-end equipment [21, 30]. Nevertheless, rPPG signals are vulnerable to environmental changes and require more efforts to derive robust estimation for real-world applications.

Several methods have been introduced to predict rPPG signals from face videos. In [7], a chrominance-based approach has been proposed to project the RGB channels into a new space. In [28], the authors focused on a subspace of skin-pixels in the spatial domain and used its temporal rotation for rPPG estimation. Because these methods [7][28] are developed based on prior knowledge, their performance heavily depends on whether the data comply with the assumption.

Recently, a number of learning-based methods [20] [25] [24] [6] have been developed for HR estimation. In [20], the authors proposed a spatial-temporal representation and pre-trained the model using a large scale of synthetic rhythm signals. In [25], a two-step CNN network and a new dataset of fitness-themed videos are introduced. In [24], the authors proposed a self-adaptive matrix completion method to dynamically select face regions for HR estimation. In [6], the authors proposed to magnify the subtle color and motion variations for discovering the motion signals.

3 PROPOSED METHOD

In this section, we first describe our motivation and then give an overview of the proposed method. Next, we will detail the proposed Siamese network architecture, which includes twin networks, for learning the rPPG signals from two facial regions through 3D CNNs. Then, we will introduce how we fuse the two branches and define the loss functions for training the network.

3.1 Motivation

Because rPPG signals are estimated from the chrominance changes reflected on human skin, their estimation is highly sensitive to several environmental factors. In [7], the authors mentioned that there exist at least three different sources of noises: (1) camera sensor noises and compression artifact; (2) subject's head movement; and (3) illumination variations. Several existing methods have been proposed to deal with these challenges, by either projecting the RGB channels into a reliable space [7] , by spatial pooling [22], or by temporal normalization [27]. However, these methods either rely on manually designed tricks or require additional pre-processing steps. There is a dearth of research on a unified and robust rPPG estimation.

In this paper, we aim to develop a unified rPPG estimation network and especially focus on the following critical issues. First, to best preserve the subtle chrominance changes in input videos, we include no pre-processing steps but merely use an off-the-shelf face detector to locate two regions of interest (ROIs) as the inputs to our network. Second, to explicitly analyze the temporal characteristics of the input videos, we construct the network with 3D CNNs to directly process the located videos. Third, to mutually learn the characteristic information from two ROIs, we propose a Siamese network on top of two weight-sharing networks. We use the twin networks to learn their corresponding rPPG signals but enforce them to have similar periodic characteristics with the ground truth.

Figure 1 shows the architecture of the proposed Siamese-rPPG network. With the proposed network, we develop an end-to-end training process and directly process an input sequence in the inference stage without any pre-processing steps.

3.2 Siamese-rPPG network

The term "Siamese" is derived from "Siamese twins", which is synonymous for conjoined twins. Hence, siamese neural networks contain two or more branches [4] with identical architecture and also shared weights [11]. Our idea of developing Siamese-rPPG network comes from the observation that, every facial region carries its own appearance and may suffer from different noisy sources, but they should all reflect the same rPPG characteristics. As suggested in [5] [29], forehead and cheek regions usually contain more reliable blood volume pulse (BVP) information. Therefore, as shown in Fig. 1, we propose using the Siamese-rPPG network to simultaneously

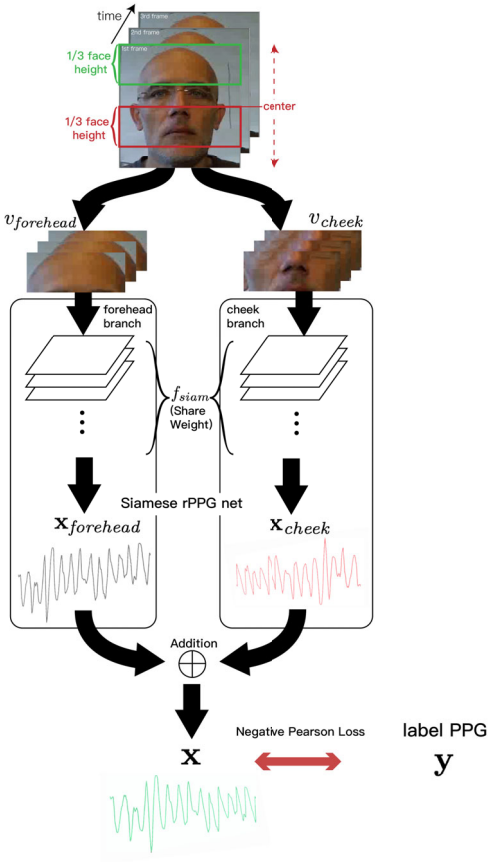


Figure 1: The proposed Siamese-rPPG network

learn from the roughly located forehead and cheek regions. In comparison with using one single network, this network architecture can better learn the homogeneity as well as the heterogeneity of different facial regions. On the other hand, in comparison with using two different models, this weight-sharing architecture has only half the model size and can enforce both branches to learn to approximate the same ground-truth.

Firstly, as different facial regions have heterogeneous visual appearance, the goal of each branch is to learn the characteristic features of its corresponding region. Moreover, unlike other methods such as [20], which needs to pre-process each input video into 2D spatial-temporal maps, we directly use the videos as the input of our network. To explicitly model the periodic variations along the temporal domain, we adopt 3D convolutional layers in our network. The extra dimension is convolved along the temporal dimension to capture the time-dependent information. We therefore construct each branch with six 3D convolutional layers, each is followed by an Avgpool layer. Detail of the network design is summarized in Table. 1. After the six 3D convolutional layers, we use the global average pooling layer to collapse the two spatial dimensions (i.e., width and height) and then use the $1 \times 1 \times 1$ convolutional layer to aggregate the information and to derive the resultant rPPG signal.

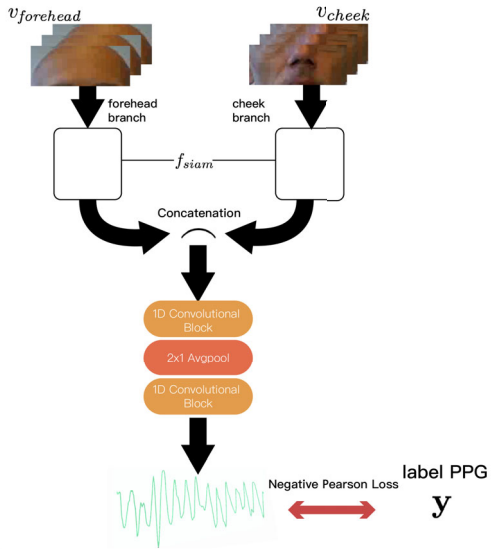


Figure 2: A variation of fusing the two branches, by concatenation of two branches and followed by two convolutional layers and an average pooling.

With the two share-weight branches, we simultaneously learn the model from two different facial regions to capture their homogeneity, and therefore can largely improve the model robustness. Note that, if without the weight-sharing mechanism, we can only train each branch from smaller datasets with less variations. On the other hand, because of the heterogeneous visual characteristics of the two inputs, the proposed network steadily increases the model generalizability. For example, when a subject's forehead region is covered by hair bangs, only few skin area is available to measure the chrominance variation of the forehead region; our model, nevertheless, can refer to the cheek branch to successfully compensate this problem.

Next, after extracting the two rPPG signals independently from each branch, our next goal is to capture their homogeneity so as to jointly train the whole network. One possibility is to concatenate the two rPPG signals and then include additional layers to process the concatenated signals, as shown in Fig.2. However, our preliminary experiments gave unsatisfactory results with this design. We suspect that, as the videos are captured under various variations, some ROIs are occluded and do not provide reliable rPPG estimation for the corresponding branch. Moreover, as we simply use the off-the-shelf face detector to locate the two ROIs, we need to consider the spatial misalignment as another noisy source. Therefore, instead of concatenation, we simply use addition operation to fuse the two branches. With the addition operation, we consider each rPPG signal is independently estimated and, if one is unavailable or unreliable, we can still rely on the other one to derive the result. Moreover, because no additional layers are needed after the addition operation, we have smaller model size than the concatenation fusing. During the inference stage, we can also derive the estimation from one single branch alone, if the other one is unavailable.

Let $v_{forehead} \in \mathbb{R}_{3\times600\times40\times140}$ and $v_{cheek} \in \mathbb{R}_{3\times600\times40\times140}$ denote the located forehead and cheek regions in the input video, where the four dimensions refer to color (RGB channels), time (600 frames), height and width of the resized frames (40×140), respectively. (Note that, the number of 600 frames correspond to 20 secs in the COHFACE and 30 secs in UBFC and PURE datasets. Comparison of other sampling rates will be shown and discussed in Sec. 4.4.4) Let f_{siam} be the proposed Siamese-rPPG network, and θ be the weights of convolutional layers. We then model the Siamese-rPPG network as follows:

$$\mathbf{x}_{cheek} = f_{siam}(v_{cheek}; \theta), \tag{1}$$

$$\mathbf{x}_{forehead} = f_{siam}(v_{forehead}; \theta), \tag{2}$$

$$\mathbf{x} = \mathbf{x}_{cheek} + \mathbf{x}_{forehead}, \tag{3}$$

where $\mathbf{x}_{cheek} \in \mathbb{R}_{600}$ and $\mathbf{x}_{forehead} \in \mathbb{R}_{600}$ are the outputs of the cheek branch and the forehead branch, respectively; and $\mathbf{x} \in \mathbb{R}_{600}$ is their addition and also the final rPPG outcome of the network.

3.3 Loss function

We use the benchmark datasets with ground truth label "Photoplethysmography (PPG) signal" to train our model. PPG is an optically obtained plethysmogram by detecting blood volume changes in the microvascular bed of tissue [1] and is usually obtained through contact devices wore on subjects' fingers. Because both the rPPG (from facial images) and PPG (from devices on fingers) measure the blood volume changes, our objective is to estimate the rPPG as similar as possible to the ground truth PPG signals. Note that, the commonly used matching criteria, such as Mean Square Error (MSE), are inappropriate for these signals. Because the values of PPG and rPPG signals are of diverse ranges, estimation of their exact values is difficult and also unnecessary. Instead, the goal is to derive heart pulse from these wavy signals. Therefore, the estimated rPPG signals are expected to have the same periodic patterns of wave crests and troughs as the ground truth PPG signals. In other words, the goal is to estimate the rPPG signals with the similar "trend" as the PPG ground truth.

The more appropriate criterion is to measure the linear correlation through Pearson product-moment correlation coefficient (Pearson correlation):

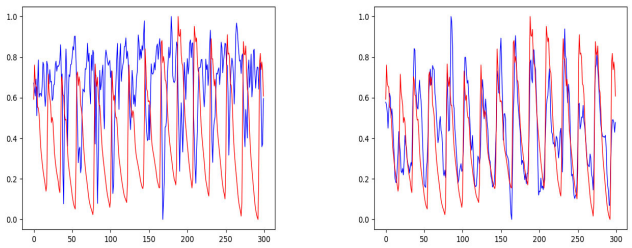
$$r(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{y}_i - \overline{\mathbf{y}})}{\sqrt{\sum_{i=1}^{n}(\mathbf{x_i} - \overline{\mathbf{x}})(\mathbf{x_i} - \overline{\mathbf{x}})}\sqrt{\sum_{i=1}^{n}(\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})}}, \tag{4}$$

where $\mathbf{x}$ and $\mathbf{y}$ are random vectors with sample size n. Pearson correlation $r(\mathbf{x}, \mathbf{y})$ ranges between +1 and -1. The correlation of +1 means total positive linear correlation, 0 means no linear correlation, and -1 means total negative linear correlation. We use two examples to show the differences between good and poor estimations. In Fig. 3(b), the estimated rPPG signal is strongly correlated with the ground truth with $r(\mathbf{x}, \mathbf{y}) = 0.78$; whereas, in Fig 3(a), the estimated result is poorly correlated with the ground truth with $r(\mathbf{x}, \mathbf{y}) = 0.29$.

Therefore, we define our loss function in terms of Negative Pearson by:

$$Loss = 1 - r(\mathbf{x}, \mathbf{y}), \tag{5}$$

where $\mathbf{x}$ and $\mathbf{y}$ are the predicted rPPG and the ground truth PPG signals, respectively. By minimizing the Negative Pearson loss, we train the Siamese-rPPG network to estimate $\mathbf{x}$ with Pearson correlation $r(\mathbf{x}, \mathbf{y})$ to be as close to 1 as possible. That is, the learned rPPGs $\mathbf{x}$ should be linear correlated to the PPG signals $\mathbf{y}$, or to be "in phase"; that is, both are expected to have similar rising and declining patterns.



(a) (b)

Figure 3: The results obtained from the cheek branch alone, when fusing with (a) concatenation (i.e., "concate-cheek" in Table 2) and (b) element-wise addition (i.e., "Proposed-cheek" in Table 2). (Blue: the predicted rPPG signals, red: the ground-truth PPG signals.)

4 EXPERIMENT

We conduct a series of experiments on three benchmark datasets, including COHFACE [10], UBFC-RPPG [3], and PURE [23]. To have a fair comparison with existing methods, we do not evaluate the performance on the estimated rPPG signals. Instead, we follow existing methods and evaluate the performance on the derived heart rate. In this section, we first introduce the datasets and then describe the experimental setting and implementation details. Next, we present our testing results under different experimental settings and show comparison with existing methods.

4.1 Datasets

4.1.1 ***COHFACE dataset.*** COHFACE dataset contains 160 heavily compressed videos from 40 subjects (12 females and 28 males). Each subject contributes four videos of one-minute length: two videos are captured in well lighted condition, and the other two are captured under natural light. The video is recorded by Logitech HD C525 with resolution of 640×480 pixels and frame rate 20 fps. Every subject wore a contact PPG sensor to obtain the blood volume pulse data. The training set contains 24 subjects and the testing set contains 16 subjects.

4.1.2 ***UBFC-RPPG dataset.*** UBFC-RPPG dataset contains 42 videos from 42 subjects, i.e. each subject has only one video. The video is recorded by Logitech C920 HD Pro, with spatial resolution of 640×480 pixels in uncompressed 8-bit RGB format and with the frame rate 30 fps. A CMS50E transmissive pulse oximeter was used to capture PPG data and PPG heart rates. The training set contains 28 subjects and the testing set contains the rest 14 subjects. Since the number of videos in the dataset is relatively small, we pre-train the model on COHFACE dataset and then fine-tune on the UBFC-RPPG dataset.

4.1.3 ***PURE dataset***. PURE dataset contains 60 videos from 10 subjects, and each subject performs six different head motions in front of the camera. The six setups include: (1) sitting still and looking directly at the camera, (2) talking when trying to avoid head movements, (3) slowly moving the head parallel to the camera, (4) moving the head quickly, (5) rotating the head with 20° angle, and (6) rotating the head with 35° angle. The video is recorded by an eco274CVGE camera with resolution of 640 × 480 pixels, frame rate of 30 fps, and with one minute length. Pulox CMS50E finger clip pulse oximeter was used to capture PPG data with a sampling rate of 60 Hz. The training set contains 7 subjects and the testing set contains the rest 3 subjects.

4.2 Implementation setting

We train the network with Nvidia GTX 2080, using Adam optimizer and learning rate of 0.0001. The network is trained for 250 epochs with batch size 2. LeakyReLU activation is used in each convolutional layer. Table 1 summarized the network architecture of the proposed Siamese-rPPG network, which includes approximately 11.80M parameters.

Input Size	Layer Type	Filter Shape
600 x 40 x 140 x 3	Conv 3D (Stride 1/Padding 1)	3 x 3 x 3 x 8
600 x 40 x 140 x 8	AvgPool 3D	1 x 2 x 2
600 x 20 x 70 x 8	Conv 3D (Stride 1/Padding 1)	3 x 3 x 3 x 32
600 x 10 x 35 x 32	AvgPool 3D	1 x 2 x 2
600 x 10 x 35 x 32	Conv 3D (Stride 1/Padding 1)	3 x 3 x 3 x 64
600 x 10 x 35 x 64	AvgPool 3D	1 x 2 x 2
600 x 5 x 17 x 64	Conv 3D (Stride 1/Padding 1)	3 x 3 x 3 x 64
600 x 5 x 17 x 64	AvgPool 3D	1 x 2 x 2
600 x 2 x 8 x 64	Conv 3D (Stride 1/Padding 1)	3 x 3 x 3 x 128
600 x 2 x 8 x 128	AvgPool 3D	1 x 2 x 2
600 x 1 x 4 x 128	Conv 3D (Stride 1/Padding 1)	3 x 3 x 3 x 256
600 x 1 x 4 x 256	Global Average Pooling	-
600 x 1 x 1 x 256	Conv 3D (Stride 1/Padding 0)	1 x 1 x 1 x 1

Table 1: Siamese-rPPG network architecture.

4.2.1 ***ROI extraction***. During the training stage, we randomly sample 600 consecutive frames from one video as an input to the proposed Siamese-rPPG network. Given an input video of 600 frames, we only use the Dlib [12] detector once to locate the face region at the first frame and do not re-locate the region for the following 599 frames. We then define the forehead ROI as the upper 1/3 of the detected face, and the cheek ROI as the area right below the center of the detected face region. These two ROIs are resized into 40×140 and fed into the proposed network for training, as illustrated in Fig. 1. Although this simple setting does not give accurate locations of ROIs, we found it works pretty well and very efficient. One major reason is that, most face detector does not give stable detection under natural light setting and will thus result in misaligned ROIs. In addition, we found that the subjects usually have small or insignificant motion within the interval of 600 frames. Therefore, there is no need to detect the face region for every frame.

Method	R	MAE	RMSE
w/o siamese-cheek	0.48	1.66	2.70
w/o siamese-forehead	0.59	1.90	3.04
w/o siamese-whole face	0.59	2.18	3.81
MSE loss-cheek	0.64	1.62	2.47
MSE loss-forehead	0.55	2.60	3.82
MSE loss-all	0.62	0.73	1.35
concate-cheek	0.53	3.51	5.52
concate-forehead	0.59	1.05	2.00
concate-all	0.61	1.04	1.88
mul-cheek	0.45	3.39	5.11
mul-forehead	0.49	1.81	2.72
mul-all	0.78	0.80	1.40
Proposed-cheek-256	0.40	3.19	9.09
Proposed-forehead-256	0.52	2.23	7.74
Proposed-all-256	0.60	1.30	4.72
Proposed-cheek-400	0.51	2.69	4.86
Proposed-forehead-400	0.49	2.22	3.53
Proposed-all-400	0.68	1.05	1.76
Proposed-cheek	0.61	1.52	2.50
Proposed-forehead	0.65	1.56	2.47
Proposed-all	0.73	0.70	1.29

Table 2: Ablation study on COHFACE dataset.

In the testing stage, we also use the Dlib [12] detector once at the first frame and then directly estimate rPPG signals from the input videos.

4.2.2 ***Heart rate estimation***. To evaluate the performance in terms of heart rate, we need to derive HR from the ground truth PPG and the estimated rPPG signals. As heart rate is determined by the periodicity of PPG and rPPG, we simply use Fast Fourier transformation to convert the signals from time domain into frequency domain. Under normal condition, the heartbeat frequency for adults is about 40-180 beats per minute (bpm). Therefore, the frequency components less than 40 beats per minutes can be considered as noises. Because the components with frequency larger than 180 are of small magnitude, we simply remove the components smaller than 40 bmp and determine the HR as the frequency with the strongest magnitude.

4.3 Evaluation metrics

To evaluate the performance of our model, we use the following metrics:

(1) Pearson correlation coefficient (R) represents the correlation between the heart rate calculated from the rPPG signal and from the ground-truth PPG signal.

$$R = \frac{\sum_{i=1}^{N}(H_i - \overline{H})(H_i' - \overline{H'})}{\sqrt{\sum_{i=1}^{N}(H_i - \overline{H})^2}\sqrt{\sum_{i=1}^{N}(H_i' - \overline{H'})^2}}, \tag{6}$$

(2) Mean absolute error (MAE)

$$MAE = \frac{\sum_{i=1}^{N} | H_i - H_i' |}{N}, \quad (7)$$

(3) Root mean square error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (H_i - H_i')^2}{N}}, \quad (8)$$

(4) Precision at 2.5 or 5 bpm

$$Precision_t = \frac{P_t}{N}, \quad (9)$$

where H_i and H_i' denote the heart rates estimated from the predicted rPPG signal and from the ground-truth PPG signal, respectively; $\overline{H}$ denotes the mean of heart rate, and P_t denotes the number of subjects whose MAE values are under t bpm (t =2.5 or 5).

4.4 Ablation study

Table 2 summarizes the ablation study of our method under different settings on the COHFACE dataset. Note that, "-cheek" and "-forehead" refer to the output of their corresponding branch before the fusing step, "-all" refer to the fused rPPG signal, and "-256" and "-400" are two other settings of frame numbers in one input video (the default setting is 600 frames).

4.4.1 ***Evaluation of the siamese network.*** We first evaluate the effectiveness of the Siamese network. When we use only one single branch of network to process one facial region, including cheek, forehead, and the whole face region (i.e., w/o Siamese-cheek, w/o Siamese-forehead, and w/o Siamese-whole face in Table 2), the Pearson correlation coefficient (R) is about 0.48, 0.59, and 0.59. With the proposed Siamese-rPPG network, we significantly improve the performance (R) to 0.61, 0.65, and 0.73. The improvement can also be seen from the reduced MAE and RMSE. These results verify that, by training under the proposed Siamese network, we successfully learn the homogeneity and heterogeneity of two ROIs and achieve significant improvement.

4.4.2 ***Evaluation of the loss term.*** Next, if we replace the Negative Pearson loss with Mean Square Error loss in Equation (5), the model will enforce the amplitude of the estimated rPPG signals to be as close as possible to the ground truth PPG signals. Although this loss term does not reflect our goal of learning the periodic trend and is also much difficult for the model to converge, our results show that the Siamese network (i.e., MSE loss-all) still outperforms one single branch network (i.e., MSE loss-cheek and MSE loss-forehead).

4.4.3 ***Evaluation of different fusion architectures.*** We also evaluate the performance of using different fusing designs in the Siamese network. In Table 2, "concate" refers to the concatenation fusion of Fig.2 (with two convolutional layers and one pooling layer), and "mul" refers to using element-wise multiplication to fuse the two branches. The results show that, both "mul" (element-wise multiplication) and our proposed "element-wise addition" outperform "concate". Fig 3 also gives an example to show the difference between "concate" and the proposed method. Note that, although "mul-all" achieve good performance, its single branch results (i.e., "mul-cheek" and "mul-forehead") perform much worse than the proposed method. We believe it is because element-wise multiplication tends to be dominated by one single branch than element-wise addition. Therefore, with the proposed method, even one single branch alone (i.e., Proposed-cheek or Proposed-forehead) is capable to estimate reliable results. By combining two face regions, we increase the Pearson correlation to 0.73 and decrease the mean average error to 0.70. These results verify the superiority of the proposed network as well as the proposed loss function.

4.4.4 ***Evaluation of sampling different numbers of frames.*** We also show the results when we only sample 256 or 400 frames as the input to the proposed network (i.e., "-256" and "-400"). Both settings give poorer performance than our default setting (i.e., 600 frames). These results suggest that, to derive temporally periodic patterns, we need to involve sufficient number of frames.

4.5 Experimental results and comparison

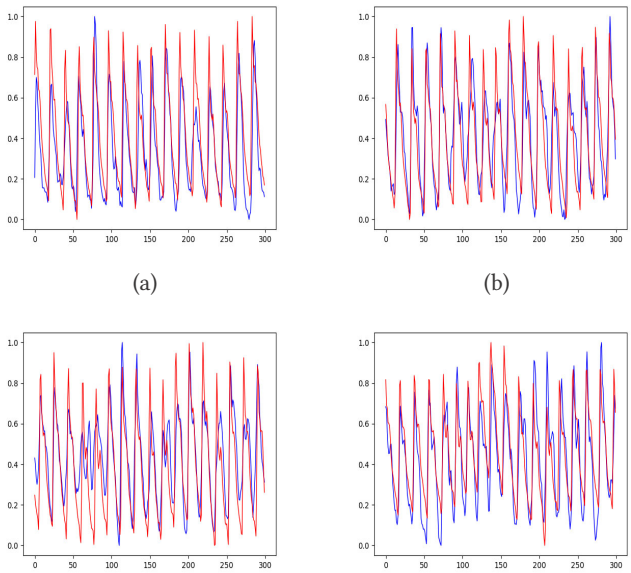
Method	R	MAE	RMSE
2SR [28]	-0.32	20.98	25.84
CHROME [7]	0.26	7.80	12.45
LiCVPR [16]	-0.44	19.98	25.59
HR-CNN [25]	0.29	8.10	10.78
Two stream [29]	0.40	8.09	9.96
Ours	0.73	0.70	1.29

Table 3: Comparison on COHFACE dataset

Table 3 shows the comparison with existing methods on the COHFACE dataset. 2SR [28], CHROME [7], and LiCVPR [16] are traditional methods; HR-CNN [25] and Two stream [29] are data-driven, learning-based methods. The results show that our method significantly outperforms these methods in terms of the three metrics. We also show the rPPG signals estimated under different scenarios for comparison. Fig 4 (a) and (b) are the cases captured under controlled scenario and show that the estimated signals highly overlap with the ground-truth signals. On the other hand, Fig 4 (c) and (d) are the more challenging cases captured under natural scenario. Although the estimated results are not as good as in the controlled scenario, the results also show well-aligned wave trough and crest between the estimated and ground-truth signals.

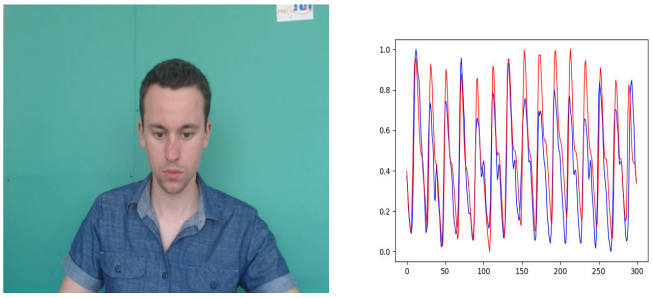
Table 4 shows the results and comparison on the UBFC-RPPG dataset. Our method makes a significant breakthrough and decreases both MAE and RMSE to be less than 1. We even obtain the precision 1.0 at both 2.5bpm and 5pm; in other words, there is no subjects' MAE is smaller than 2.5 bpm. One example is given in Fig 5, where our predicted rPPG signal is highly correlated with the ground-truth PPG signal.

Table 5 shows our results and comparison on the original PURE dataset, which contains lossless PNG files. In this case, because the uncompressed videos have better quality, there is a significant performance gap between this dataset and COHFACE dataset. We also follow the settings [25] to compress the PURE dataset into MPEG-4 format and conduct experiments on the compressed videos.



(a) (b) (c) (d)

Figure 4: Estimated rPPG signals of the same subject in COHFACE dataset under: (a)(b) controlled scenario, and (c)(d) natural scenario. (Blue: the predicted rPPG signals, red: the ground-truth PPG signals.)



(a) (b)

Figure 5: An example from UBFC-RPPG dataset: (a) the original video frame, and (b) the estimated results. (Blue: the predicted rPPG signals, red: the ground-truth PPG signals.)

As shown in Table 6, our method is insensitive to the compression artifacts and still outperforms all the existing methods.

To evaluate the generalizability of our proposed method, we perform cross-dataset experiments on UBFC-RPPG and PURE datasets and show the results in Table 7. That is, we train the network on one dataset but directly test the model on the other dataset. As shown in the first part of Table 7, when testing on UBFC-RPPG dataset, we obtain degraded performance if using the model trained on PURE dataset instead of the same UBFC-RPPG dataset. This degradation is not unexpected, because the two datasets behave differently. Nonetheless, the result still outperforms existing methods and verify the generalizability of our network. On the other hand, in the second part of Table 7, when testing on PURE dataset, we have better cross-dataset performance than the original same-dataset setting. We believe this result comes from the fact that UBFC-RPPG covers wider variations on poses and illuminations and thus is more suitable to train a robust model than the PURE dataset.

Method	MAE	RMSE	2.5 bpm	5 bpm
PVM [19]	4.47	-	0.71	0.81
MODEL [14]	3.99	5.55	0.75	0.87
SKIN-TISSUE [3]	-	2.39	0.89	0.83
MAICA [18]	3.34	-	0.72	0.88
BIC [2]	1.21	2.41	0.951	0.975
Ours-cheek	0.43	0.89	1.0	1.0
Ours-forehead	0.73	1.40	1.0	1.0
Ours-all	0.48	0.97	1.0	1.0

Table 4: Comparison on UBFC-RPPG dataset

Method	R	MAE	RMSE
2SR [28]	0.98	2.44	3.06
CHROME [7]	0.99	2.07	2.50
LiCVPR [16]	-0.38	28.22	30.96
HR-CNN [25]	0.98	1.84	2.37
Ours-cheek	0.73	0.96	3.97
Ours-forehead	0.76	0.96	2.32
Ours-all	0.83	0.51	1.56

Table 5: Comparison on PURE dataset

Method	R	MAE	RMSE
2SR [28]	0.43	5.78	12.81
CHROME [7]	0.55	6.29	11.36
LiCVPR [16]	-0.42	28.39	31.10
HR-CNN [25]	0.7	8.72	11.00
Two stream [29]	0.42	9.81	11.81
Ours-cheek	0.57	3.55	12.12
Ours-forehead	0.56	3.04	9.32
Ours-all	0.85	0.63	2.70

Table 6: Comparison on PURE dataset (MPEG-4 visual)

5 CONCLUSIONS

In this paper, we propose a novel Siamese-rPPG network to estimate remote photoplethysmography (rPPG) from facial videos. We construct 3D CNNs to model the spatial and temporal characteristics of rPPG signals from two facial regions. Without involving any pre-processing step, the proposed Siamese-rPPG network performs

Training->Testing	MAE	RMSE
UBFC->UBFC	0.48	0.97
PURE->UBFC	1.29	8.73
PURE->PURE	0.63	2.70
UBFC->PURE	0.63	2.51

Table 7: Comparison of cross-dataset estimation

in an end-to-end manner. Moreover, the weight-sharing mechanism greatly leverages the capability of learning robust, distinctive, and complementary features from multiple facial regions. Our experimental results on three benchmark datasets verify the effectiveness of our proposed method and significantly outperform all the existing methods.4.2.1.

6 ACKNOWLEDGEMENT

This work was funded in part by Qualcomm through a Taiwan University Research Collaboration Project, and was also supported in part by Ministry of Science and Technology (MOST) under grant MOST 108-2221-E-007 -065 -MY3.

REFERENCES

[1] John Allen. 2007. Photoplethysmography and Its Application in Clinical Physiological Measurement. *Physiological Measurement* 28, 3 (Feb 2007), R1–R39. https://doi.org/10.1088/0967-3334/28/3/R01

[2] Yannick Benezeth, Serge Bobbia, Keisuke Nakamura, Randy Gomez, and Julien Dubois. 2019. Probabilistic Signal Quality Metric for Reduced Complexity Unsupervised Remote Photoplethysmography. 1–5. https://doi.org/10.1109/ISMICT.2019.8744004

[3] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. 2017. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters* (10 2017). https://doi.org/10.1016/j.patrec.2017.10.017

[4] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. Signature Verification Using a "Siamese" Time Delay Neural Network. In *Proceedings of the 6th International Conference on Neural Information Processing Systems (NIPS'93)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 737–744. http://dl.acm.org/citation.cfm?id=2987189.2987282

[5] Weixuan Chen and Daniel McDuff. 2018. DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks. In *Computer Vision – ECCV 2018*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer International Publishing, Cham, 356–373.

[6] Weixuan Chen and Daniel J. McDuff. 2018. DeepMag: Source Specific Motion Magnification Using Gradient Ascent. *CoRR* abs/1808.03338 (2018). arXiv:1808.03338 http://arxiv.org/abs/1808.03338

[7] Gerard de Haan and Vincent Jeanne. 2013. Robust Pulse Rate From Chrominance-Based rPPG. *IEEE Transactions on Biomedical Engineering* 60, 10 (Oct 2013), 2878–2886. https://doi.org/10.1109/TBME.2013.2266196

[8] Javier Hernandez-Ortega, Julian Fierrez, Aythami Morales, and Pedro Tome. 2018. Time Analysis of Pulse-Based Face Anti-Spoofing in Visible and NIR. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.

[9] A. B. Hertzmann. 1937. Observations on the Finger Volume Pulse Recorded Photo-electrically. *American Journal of Physiology* 119 (1937), 334–335.

[10] Guillaume Heusch, André Anjos, and Sébastien Marcel. 2017. A Reproducible Study on Remote Heart Rate Measurement. *CoRR* abs/1709.00962 (2017). arXiv:1709.00962 http://arxiv.org/abs/1709.00962

[11] Branislav HollÃďnder. 2018. Siamese Networks: Algorithm, Applications And PyTorch Implementation. https://becominghuman.ai/siamese-networks-algorithm-applications-and-pytorch-implementation-4ffa3304c18 Last accessed on Aug 29, 2019.

[12] Davis E. King. 2009. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* 10 (2009), 1755–1758.

[13] Georg Lempe, Sebastian Zaunseder, Tom Wirthgen, Stephan Zipser, and Hagen Malberg. 2013. ROI Selection for Remote Photoplethysmography. In *Bildverarbeitung für die Medizin 2013*. Springer Berlin Heidelberg, Berlin, Heidelberg, 99–103.

[14] Peixi Li, Keisuke Nakamura Yannick Benezeth, Randy Gomez, and Fan Yang. 2019. Model-based Region of Interest Segmentation for Remote Photoplethysmography. In *14th International Conference on Computer Vision Theory and Applications*. 383–388.

[15] Xiaobai Li, Iman Alikhani, Jingang Shi, Tapio Seppanen, Juhani Junttila, Kirsi Majamaa-Voltti, Mikko Tulppo, and Guoying Zhao. 2018. 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)*. 242–249.

[16] Xiaobai Li, Jie Chen, Guoying Zhao, and Matti PietikÃďinen. 2014. Remote Heart Rate Measurement from Face Videos under Realistic Situations. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 4264–4271. https://doi.org/10.1109/CVPR.2014.543

[17] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. 2018. Learning Deep Models for Face Anti-Spoofing: Binary or Auxiliary Supervision. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 389–398.

[18] Richard Macwan, Yannick Benezeth, and Alamin Mansouri. 2019. Heart rate estimation using remote photoplethysmography with multi-objective optimization. *Biomedical Signal Processing and Control* 49 (03 2019), 24–33. https://doi.org/10.1016/j.bspc.2018.10.012

[19] Richard Macwan, Serge Bobbia, Yannick Benezeth, Julien Dubois, and Alamin Mansouri. 2018. Periodic Variance Maximization Using Generalized Eigenvalue Decomposition Applied to Remote Photoplethysmography Estimation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 1413–14138. https://doi.org/10.1109/CVPRW.2018.00181

[20] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. 2018. SynRhythm: Learning a Deep Heart Rate Estimator from General to Specific. In *2018 24th International Conference on Pattern Recognition (ICPR)*. 3580–3585.

[21] Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard. 2010. Non-contact, Automated Cardiac Pulse Measurements Using Video Imaging and Blind Source Separation. *Optics Express* 18, 10 (2010), 10762–10774.

[22] Ying Qiu, Yang Liu, Juan Arteaga-Falconi, Haiwei Dong, and Abdulmotaleb E. Saddik. 2019. EVM-CNN: Real-Time Contactless Heart Rate Estimation From Facial Video. *IEEE Transactions on Multimedia* 21, 7 (July 2019), 1778–1787. https://doi.org/10.1109/TMM.2018.2883866

[23] Ronny Stricker, Steffen MÄijller, and Horst-Michael Gross. 2014. Non-contact video-based pulse rate measurement on a mobile service robot. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication* 2014, 1056–1062. https://doi.org/10.1109/ROMAN.2014.6926392

[24] Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, Lijun Yin, Jeffrey F. Cohn, and Nicu Sebe. 2016. Self-Adaptive Matrix Completion for Heart Rate Estimation from Face Videos under Realistic Conditions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2396–2404. https://doi.org/10.1109/CVPR.2016.263

[25] Radim Špetlík, Vojtěch Franc, Jan Čech, and Jiří Matas. 2018. Visual Heart Rate Estimation with Convolutional Neural Network. In *Proceedings of British Machine Vision Conference*.

[26] Wim Verkruysse, Lars O. Svaasand, and J S. Nelson. 2008. Remote Plethysmographic Imaging Using Ambient Light. *Optics Express* 16, 26 (2008), 21434–21445.

[27] Wenjin Wang, Albertus C. den Brinker, Sander Stuijk, and Gerard de Haan. 2017. Algorithmic Principles of Remote PPG. *IEEE Transactions on Biomedical Engineering* 64, 7 (July 2017), 1479–1491. https://doi.org/10.1109/TBME.2016.2609282

[28] Wenjin Wang, Sander Stuijk, and Gerard de Haan. 2016. A Novel Algorithm for Remote Photoplethysmography: Spatial Subspace Rotation. *IEEE Transactions on Biomedical Engineering* 63, 9 (Sep. 2016), 1974–1984. https://doi.org/10.1109/TBME.2015.2508602

[29] Zhi-Kuan Wang, Ying Kao, and Chiou-Ting Hsu. 2019. Vision-based Heart Rate Estimation via a Two-stream CNN. In *2019 IEEE International Conference on Image Processing (ICIP)*. 3327–3331. https://doi.org/10.1109/ICIP.2019.8803649

[30] Sun Yu, Sijung Hu, Vicente Azorin-Peris, Jonathon A. Chambers, Yisheng Zhu, and Stephen E. Greenwald. 2011. Motion-compensated noncontact imaging photoplethysmography to monitor cardiorespiratory status during exercise. *Journal of biomedical optics* 16 (07 2011), 077010. https://doi.org/10.1117/1.3602852

[31] Zitong Yu, Xiaobai Li, and Guoying Zhao. 2019. Recovering remote Photoplethysmograph Signal from Facial videos Using Spatio-Temporal Convolutional Networks. *CoRR* abs/1905.02419 (2019). arXiv:1905.02419 http://arxiv.org/abs/1905.02419