Multi-Task Learning for Simultaneous Video Generation and Remote Photoplethysmography Estimation

Yun-Yun Tsou*, Yi-An Lee*, and Chiou-Ting Hsu

National Tsing Hua University, Hsinchu, Taiwan
tsou0320@gmail.com, andy75527552andy@gmail.com, cthsu@cs.nthu.edu.tw

Abstract. Remote photoplethysmography (rPPG) is a contactless method for estimating physiological signals from facial videos. Without large supervised datasets, learning a robust rPPG estimation model is extremely challenging. Instead of merely focusing on model learning, we believe data augmentation may be of greater importance for this task. In this paper, we propose a novel multi-task learning framework to simultaneously augment training data while learning the rPPG estimation model. We design three networks: rPPG estimation network, Image-to-Video network, and Video-to-Video network, to estimate rPPG signals from face videos, to generate synthetic videos from a source image and a specified rPPG signal, and to generate synthetic videos from a source video and a specified rPPG signal, respectively. Experimental results on three benchmark datasets, COHFACE, UBFC, and PURE, show that our method successfully generates photo-realistic videos and significantly outperforms existing methods with a large margin. [1] [2]

1 Introduction

Heart rate (HR) is a major health indicator of human body and has been widely used to aid diagnosis of heart diseases, such as detection of atrial fibrillation [1, 2]. Existing methods usually rely on specific contact devices to detect Electrocardiography (ECG) or Photoplethysmography (PPG) signals for monitoring the heart rate. Although these contact devices provide accurate readings, they require professional attention to collect the signals and can hardly be used to monitor a large group of subjects at the same time.

A number of contactless video-based methods have been developed [3–5] to support remote estimation of physiological signals. Especially, remote photoplethysmography (rPPG), which aims to analyze the blood volume changes in optical information, has been shown to be able to capture heart rate information [6, 3], to aid detection of atrial fibrillation [1, 2], and can even be extended to applications such as face anti-spoofing [7–10]. However, because visual appearance

[1] The code is publicly available at https://github.com/YiAnLee/Multi-Task-Learning-for-Simultaneous-VideoGeneration-and-Remote-Photoplethysmography-Estimation
[2] The first two authors contributed equally.

is more vulnerable to environmental interference (e.g., illumination) and subjects' motion (e.g, body and muscular movement during the recording stage), many efforts have been devoted to robust estimation of rPPG signals through learning-based methods [3, 4, 11, 2].

Nevertheless, success of learning-based methods heavily relies on large and good supervised datasets; there are unfortunately few datasets available for training robust rPPG or HR estimation. Unlike other video analysis tasks, collection of face videos and their ground truth for rPPG estimation is extremely complex. Voluntary subjects are required to wear specific devices to obtain the ground truth PPG labels. Moreover, if the subjects are hospital patients, not only the data collection takes enormous time, but also the usage of their face videos is highly restricted. Therefore, only a few datasets of face videos with ground-truth PPG signals are publicly available, and these datasets are small-scaled with very limited number of subjects. For example, UBFC-RPPG [12] dataset contains 42 videos from 42 subjects, and PURE dataset [13] contains 60 videos from only 10 subjects. Although in [1] a larger dataset OBF was collected with 200 videos from 100 healthy adults, this dataset cannot be publicly released because of privacy concern. Consequently, performance of existing methods remains unsatisfactory. To tackle the problem of insufficient data, in [14], the authors proposed to pre-train their model on large-scaled data of synthetic rhythm signals. Because these synthetic rhythm signals need to be converted to a spatial-temporal representation before estimating heart rates, its practicability in real-world scenarios is doubtful. Moreover, any pre-processing step may diminish the subtle chrominance changes in face videos and yield inaccurate estimates. For example, previous multi-stage methods usually involve spatial/temporal sampling, and/or conversion from video frames to spatial-temporal maps. The conversion step heavily relies on an accurate and stable ROI selection algorithm (so as to align the same location into the spatial dimension of the map) and also incurs information loss while collapsing the two spatial dimensions into one dimension in the map. There is indeed a dearth of research on resolving this dilemma.

To resolve the aforementioned problems, in this paper, we propose to generate augmented data by synthesizing videos containing specific rPPG signals and to learn the rPPG estimator from both the source videos and the synthetic videos. We formulate two tasks, that is, data augmentation and rPPG estimation, in a multi-task learning framework. Figure 1 illustrates the proposed idea of generating synthetic videos by embedding a target rPPG signal into either a still image or a video sequence. A more detailed framework is given in Figure 2, where the three networks: rPPG network, Image-to-Video network, and Video-to-Video network, are successively trained. We first pre-train the rPPG network using source videos of the benchmark dataset. Next, we train the two video generation networks (i.e., Image-to-Video and Video-to-Video networks) by concatenating them with the pre-trained rPPG network but without updating the parameters of rPPG network. Finally, we fine-tune the pre-trained rPPG network using synthetic videos obtained from the two video generation networks. As mentioned before, the performance of rPPG network highly depends on the quantity and

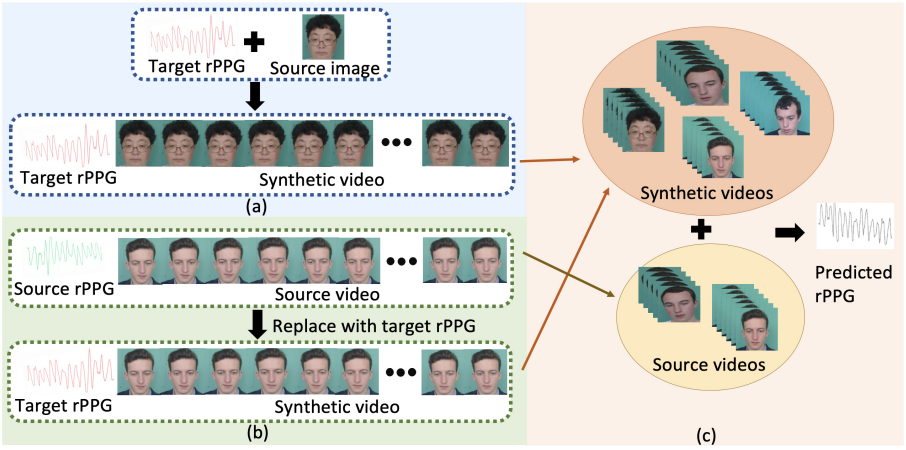

Fig. 1. Illustration of the proposed idea. (a) Generation of a synthetic video from a given source image and a target rPPG signal, (b) generation of a synthetic video from a given source video and a target rPPG signal, and (c) learning the robust rPPG estimator from both the source videos and the synthetic videos.

quality of the training data. In this paper, we aim to create a win-win situation and to reinforce these tasks to mutually support each other. Specifically, because the data augmentation task (i.e., Image-to-Video network and Video-to-Video network) needs to refer to the rPPG network to verify whether the synthesized videos capture the target rPPG signals or not, a robust estimation naturally leads to a better generation performance. On the other hand, with the increased number of synthetics videos, the estimation task is able to learn from various combinations of face videos (e.g., different skin colors, environmental illuminations, and motions) and rPPG signals (e.g., healthy subjects or patients with heart disease) to increase its robustness. Our experimental results on three benchmark datasets: COHFACE [15], PURE [13] and UBFC-RPPG [12], show that we successfully generate photo-realistic videos with different rPPG signals and that the learned rPPG estimator achieves state-of-the-art performance.

Our contributions are summarized below:

- To the best of our knowledge, this is the first work focusing on generating synthetic videos with specific rPPG signals. The augmented dataset will benefit the future study on remote monitoring of human physiological signals.
- We propose a multi-task learning framework to simultaneously learn the data augmentation and the rPPG estimation tasks. These tasks, modeled with three networks, are thus enforced to improve each other to boost the overall performance.
- The proposed method successfully learns rich augmented data and yields robust rPPG estimation. Experimental results show that our method achieves state-of-the-art results.

2 Related Work

Remote Photoplethysmography Estimation has attracted enormous research interests for heart rate estimation [16, 12, 6, 17–21]. Earlier methods focus on designing either feature descriptors or color filters to capture the strongest PPG information from facial videos. For example, in [6], the authors proposed a chrominance-based approach to project RGB channels into a subspace for extracting the rPPG signals. In [21], the authors estimated a spatial subspace of skin-pixels and measured its temporal rotation for rPPG estimation. However, because these methods are developed based on assumed domain knowledge, they may not generalize well to other data not complying with the assumption.

Many learning-based methods [3, 4, 22, 11, 5, 2, 23] have been recently introduced for rPPG or HR estimation. In [2], a 3D CNN-based spatio-temporal network was proposed to measure rPPG signal. In [23], the authors focused on compression artifacts and proposed a two-stage method to recover rPPG signals from highly compressed videos. The other methods [3, 4, 22, 11, 5] mostly focused on improving the estimation accuracy but hardly address the lack of large-scale data issue. Although a larger OBF dataset was introduced in [1], it was not publicly available for experimental comparison. The problem of insufficient training data is still far from being resolved.

Data Augmentation has been widely utilized to overcome the burden of collecting large supervised datasets for training deep neural networks. In addition to traditional augmentation strategies (e.g. horizontal flipping, rotating, cropping), learning to automate the data generation process has been shown to significantly improve object detection and image classification tasks. In [24], the authors proposed a context-based data augmentation for object detection by automatically placing new objects in suitable locations of images. In [25], a data augmentation method was proposed to generates synthetic medical images using Generative Adversarial Networks (GANs) for liver lesion classification.

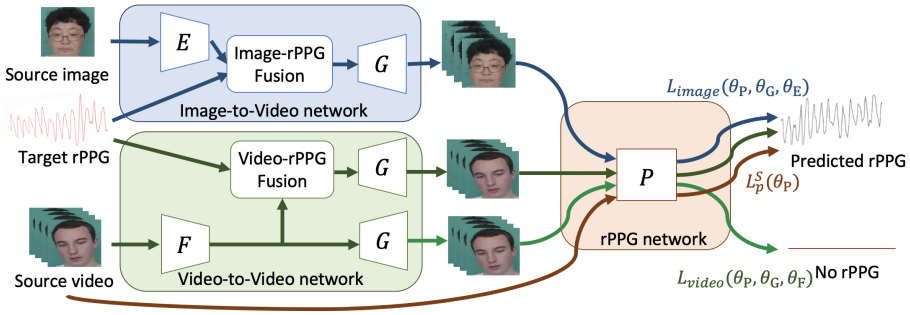

Fig. 2. The proposed multi-task learning framework, which contains three networks: rPPG network, Image-to-Video network, and Video-to-Video network.

3 Proposed method

3.1 Overview

The goal of this paper is two-fold: to augment the training data by synthesizing photo-realistic face videos containing specific rPPG signals, and to leverage the rPPG estimation accuracy by learning from the augmented data. As shown in Figure 2, we propose a multi-task learning framework containing three networks: rPPG network, Image-to-Video network, and Video-to-Video network, to simultaneously fulfill the two tasks.

1. **rPPG network** aims to estimate rPPG signals directly from input face videos.
2. **Image-to-Video network** aims to generate a face video $v_{image} \in \mathbb{R}^{H\times W\times C\times T}$ from a single face image $x_{source} \in \mathbb{R}^{H\times W\times C}$ and a target rPPG signal $y_{target} \in \mathbb{R}^{T}$, where H, W, C denote the height, width, and the number of channels of the source image, and T is the length of the target rPPG signal.
3. **Video-to-Video network** aims to replace the original rPPG signal of a source video $v_{source} \in \mathbb{R}^{H\times W\times C\times T}$ with a target rPPG signal y_{target}. The synthesized video $v_{video} \in \mathbb{R}^{H\times W\times C\times T}$ is expected to look similar to the source video v_{source} but should capture the target rPPG signal y_{target}.

Note that, although it is possible to design and train the three networks independently, their capability will be severely limited by the scale and quality of their individual training data. Below we will detail each network and describe how we formulate these highly correlated problems in a multi-tasking learning framework to strongly reinforce the capability of each network.

3.2 RPPG Network

To estimate rPPG signals from face videos, previous methods [3, 26, 14] usually require image pre-processing steps, such as detection of regions-of-interest (RoIs), conversion of video frames to spatial-temporal maps, etc. However, because any pre-processing step will unavoidably diminish the subtle chrominance changes in face videos, we propose to directly estimate the rPPG signals from face videos in an end-to-end manner without any pre-processing. We develop the rPPG network P in terms of 3D CNN and summarize its network architecture in the supplementary file.

Given a ground truth PPG signal $y' \in \mathbb{R}^{T}$, our goal is to train the rPPG network P to estimate the signal $y \in \mathbb{R}^{T}$ to have the same periodic pattern of wave crests and troughs as y'. An eligible criterion is to measure the linear correlation through Pearson correlation:

$$\rho(y, y') = \frac{(y-\overline{y})^t(y'-\overline{y'})}{\sqrt{(y-\overline{y})^t(y-\overline{y})}\sqrt{(y'-\overline{y'})^t(y'-\overline{y'})}}, \tag{1}$$

where $\overline{y}$ and $\overline{y'}$ are the means of the predicted rPPG signal y and the ground truth PPG signal y', respectively. Hence, we define the loss function of P in terms of Negative Pearson by

$$L_p^S(\theta_P) = 1 - \rho(y_{source}, y'_{source}), \tag{2}$$

where

$$y_{source} = P(v_{source}; \theta_P). \tag{3}$$

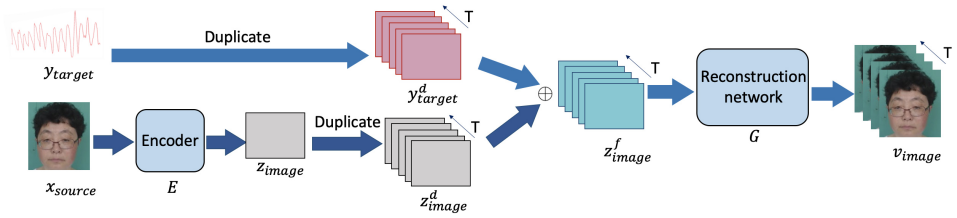

Fig. 3. Image-to-Video Network

3.3 Image-to-Video network

Figure 3 illustrates the proposed Image-to-Video network. Given a source image x_{source}, we first introduce an encoder E to obtain the feature representation of the source image by:

$$z_{image} = E(x_{source}; \theta_E), \tag{4}$$

where $z_{image} \in \mathbb{R}^{H_I \times W_I \times C_I}$; H_I, W_I, and C_I denote the height, width and the number of channels of the image feature, respectively.

Next, we design a fusion method to incorporate the target rPPG signal $y_{target} \in \mathbb{R}^T$ into the feature map z_{image}. Because the dimensions of z_{image} and y_{target} are inconsistent, we cannot directly combine the two signals. We thus pixelwisely duplicate the rPPG signal y_{target} into $y^d_{target} \in \mathbb{R}^{H_I \times W_I \times C_I \times T}$, and temporally duplicate the feature map z_{image} into $z^d_{image} \in \mathbb{R}^{H_I \times W_I \times C_I \times T}$. The resultant y^d_{target} and z^d_{image} are of the same dimension and are then fused through the element-wise addition to obtain the fused feature map $z^f_{image} \in \mathbb{R}^{H_I \times W_I \times C_I \times T}$ by:

$$z^f_{image} = z^d_{image} + y^d_{target}. \tag{5}$$

With this fusion step, we guarantee that all the spatial elements in z^f_{image} reflect the same rPPG characteristics. Then, we design a reconstruction network G to generate synthetic face video $v_{image} \in \mathbb{R}^{H \times W \times C \times T}$:

$$v_{image} = G(z^f_{image}; \theta_G). \tag{6}$$

To ensure the synthetic video v_{image} carries the target rPPG signal, we include the learning of rPPG network P together with the learning of Image-to-Video Network to formulate the loss term. We impose two constraints to define the loss function for Image-to-Video Network. First, we let the rPPG network P guide the encoder E and the reconstruction network G to generate a video v_{image} containing the rPPG signal which is highly correlated with y_{target}:

$$L_p^I(\theta_P, \theta_G, \theta_E) = 1 - \rho(P(G((E(x_{source}; \theta_E)^d + y_{target}^d); \theta_G); \theta_P), y_{target}). \quad (7)$$

Second, as the synthetic video v_{image} should also capture the visual appearance of the input image x_{source}, we define a reconstruction loss in terms of absolute difference by:

$$L_r^I(\theta_G, \theta_E) = \frac{1}{T}\sum_{t=1}^{T} |G((E(x_{source}; \theta_E)^d + y_{target}^d); \theta_G)(t) - x_{source}|. \quad (8)$$

Finally, we define the total loss of Image-to-Video Network as follows,

$$L_{image}(\theta_P, \theta_G, \theta_E) = L_p^I(\theta_P, \theta_G, \theta_E) + \lambda_1 L_r^I(\theta_G, \theta_E), \quad (9)$$

where λ_1 is a hyper-parameter and is empirically set as 0.01 in all our experiments.

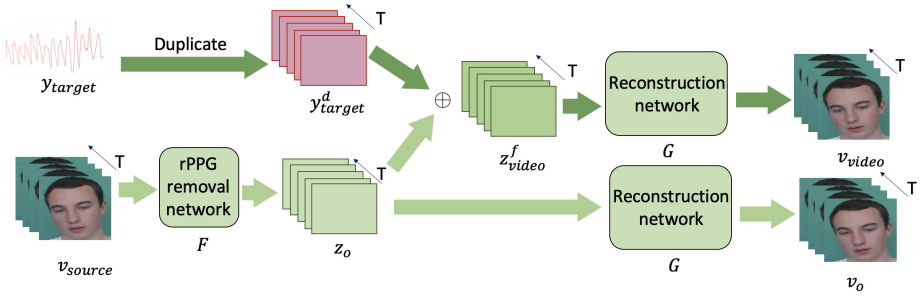

Fig. 4. Video-to-Video network

3.4 Video-to-Video network

Given a source video v_{source} and a target rPPG signal y_{target}, the Video-to-Video network aims to synthesize a target video v_{video} which should be visually similar to the source video but capture the target rPPG signal. Unlike the case in Section 3.3, the source video itself inherently captures its own rPPG signal; thus, we need to erase this rPPG signal before embedding the target signal y_{target}.

As shown in Figure 4, we replace the encoder E in the Image-to-Video network with an rPPG removal network F:

$$z_o = F(v_{source}; \theta_F), \tag{10}$$

where $z_o \in \mathbb{R}^{H_V \times W_V \times C_V \times T}$ is the video feature representation that should contain no rPPG signal; H_V, W_V, and C_V are the height, width and the number of channels, respectively.

Next, we use the same reconstruction network G to generate synthetic video but with additional constraints. Firstly, the reconstructed appearance $v_o \in \mathbb{R}^{H \times W \times C \times T}$ from z_o:

$$v_o = G(z_o; \theta_G), \tag{11}$$

is expected to be visually indistinguishable from the source video v_{source} but should contain no rPPG periodic characteristics. We therefore formulate a reconstruction loss term and a No-rPPG loss term by:

$$L_r^O(\theta_G, \theta_F) = \frac{1}{T}\sum_{t=1}^{T} |G(F(v_{source}; \theta_F); \theta_G)(t) - v_{source}(t)|, \tag{12}$$

and

$$L^O(\theta_P, \theta_G, \theta_F) = Var(P(G(F(v_{source}; \theta_F); \theta_G); \theta_P)), \tag{13}$$

where $Var(\cdot)$ measures the signal variance. Note that, a constant (or zero frequency) signal will have zero variance. Thus, we use $Var(\cdot)$ to quantify the periodicity of the estimated rPPG signal.

Secondly, to embed the target signal y_{target}, we adopt the same duplication and fusing steps of Image-to-Video network by:

$$v_{video} = G(z^f_{video}; \theta_G), \tag{14}$$

where

$$z^f_{video} = z_o + y^d_{target}. \tag{15}$$

Then, we again impose two constraints on the synthetic video v_{video} to ensure that it carries the target rPPG signal y_{target} and also preserves its original visual appearance by:

$$L_p^V(\theta_P, \theta_G, \theta_F) = 1 - \rho(P(G((F(v_{source}; \theta_F) + y^d_{target}); \theta_G); \theta_P), y_{target}), \tag{16}$$

and

$$L_r^V(\theta_G, \theta_F) = \frac{1}{T}\sum_{t=1}^{T} |G((F(v_{source}; \theta_F) + y^d_{target}); \theta_G)(t) - v_{source}(t)|. \tag{17}$$

Finally, we define the total loss of Video-to-Video Network by:

$$\begin{aligned} L_{video}(\theta_P, \theta_G, \theta_F) = {} & L^O(\theta_P, \theta_G, \theta_F) + L_p^V(\theta_P, \theta_G, \theta_F) \\ & + \lambda_2 L_r^O(\theta_G, \theta_F) + \lambda_3 L_r^V(\theta_G, \theta_F), \end{aligned} \tag{18}$$

where λ_2 and λ_3 are hyper-parameters and both are empirically set as 0.01 in our experiments.

3.5 Overall Framework

To sum up, the three networks are designed to mutually support each other. With the rPPG network P, the Image-to-Video network and Video-to-Video network are able to generate synthetic videos containing the target rPPG signals. The reconstruction network G, which is included in both the Image-to-Video and Video-to-Video networks, is constrained to generate photo-realistic synthetic videos. In addition, because the synthetic videos are considered as augmented data, the rPPG network is able to learn from videos with more diversity and with different rPPG characteristics. Our total loss term for the multi-task learning framework is defined by:

$$Loss(\theta_P, \theta_G, \theta_E, \theta_F) = L_p^S(\theta_P) + \alpha L_{image}(\theta_P, \theta_G, \theta_E) + \beta L_{video}(\theta_P, \theta_G, \theta_F), \tag{19}$$

where α and β are coefficients to balance different loss terms and both are set as 0.5 in the experiments.

4 Experiments

4.1 Datasets

We conduct a series of experiments on three benchmark datasets.

COHFACE dataset [15] comprises 160 one-minute-long RGB video sequences of 40 subjects. Each subject contributes four videos: two of them are filmed in well-lighted environment and the other two are filmed under natural light. All the videos are filmed by Logitech HD C525 webcam; the resolution is set to 640×480 and the frame rate is 20 fps. A contact PPG sensor is attached to the subjects to obtain the blood volume pulse signal. The dataset is split into a training set of 24 subjects and a testing set of 16 subjects.

UBFC-RPPG dataset [12] comprises 42 face videos, each belongs to different individuals. The training set consists of 28 subjects and the testing set consists of 14 subjects. All the videos are recorded by Logitech C920 HD Pro, with resolution of 640×480 pixels in uncompressed 8-bit format, and the frame rate is set to 30 fps. CMS50E transmissive pulse oximeter was used to monitor the PPG data and PPG heart rates.

PURE dataset [13] consists of 60 one-minute-long videos from 10 subjects, and each subject was asked to performs six different movements during recording. The six setups are: (1) sitting still and looking directly at the camera, (2) talking but avoiding 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 training set contains 7 subjects and the testing set contains the rest 3 subjects. The videos are recorded using eco274CVGE camera with resolution of 640×480 pixels, and the frame rate is set to 30 fps. Pulox CMS50E finger clip pulse oximeter is adopted to capture PPG data with sampling rate of 60 Hz.

4.2 Implementation Setting

The architectures of the rPPG network P, the encoder E, the reconstruction network G, and the rPPG removal network F in the Video-to-Video Networks are given in the supplementary file. We train the network with Nvidia GTX 2080 for 280 epochs with batch size 4, using Adam optimizer and set the learning rate 0.001. For comparison, experiments using traditional visual data augmentations are also conducted in UBFC and PURE dataset.

Table 1. Ablation study on COHFACE dataset.

Method	MAE	RMSE
Source-150	1.86	3.77
Image-150	1.51	3.29
Video-150	1.54	3.32
All-150	1.33	2.71
Source-200	1.10	2.22
Image-200	1.02	2.12
Video-200	0.98	2.02
All-200	0.93	2.01
Source-256	0.91	2.16
Image-256	0.82	1.81
w/o L_r^O, L^O-256	0.85	1.93
Video-256	0.84	1.76
All-256	0.68	1.65

Table 2. Comparison on COHFACE dataset.

Method	R	MAE	RMSE
2SR [21]	-0.32	20.98	25.84
CHROME [6]	0.26	7.80	12.45
LiCVPR [18]	-0.44	19.98	25.59
HR-CNN [11]	0.29	8.10	10.78
Two stream [5]	0.40	8.09	9.96
Ours-Source	0.68	0.91	2.16
Ours-Image	0.64	0.82	1.81
Ours-Video	0.66	0.84	1.76
Ours-All	0.72	0.68	1.65

4.3 Evaluation Metrics

Because existing methods evaluate the performance on the estimated heart rate instead of on the rPPG signals, to have a fair comparison, we follow [22] to derive the heart rate from the estimated rPPG signals and evaluate the results in terms of the following metrics: (1) Pearson correlation coefficient (R), (2) Mean absolute error (MAE), (3) Root mean square error (RMSE), and (4) Precision at 2.5 and 5 bpm (beats per minute), (5) Peak signal-to-noise ratio (PSNR) and (6) Structural similarity (SSIM).

4.4 Ablation Study

To show the effectiveness of our proposed method, we design several ablative settings on the COHFACE dataset and summarize the results in Table 1. "Source-", "Image-", "Video-", "w/o L_r^O, L^O-", "All-" refer to different combinations of the

tasks described in Figure 2, that is, (1) the single task P trained from the source data, (2) training of Image-to-Video network and P, (3) training of Video-to-Video network and P, (4) the proposed multi-tasking framework but without loss terms L_r^O, L^O, (5) the proposed multi-task framework, respectively. "-150", "-200" and "-256" refer to different lengths of T in terms of frame numbers.

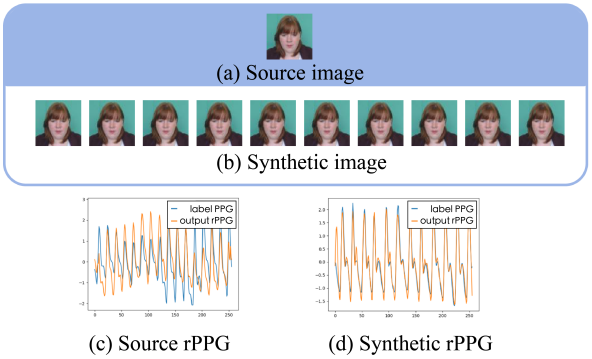

Fig. 5. A visualized example of Image-to-Video network. (a) A source image from the UBFC-RPPG dataset; (b) The synthetic video with target rPPG; (c) The source rPPG label (blue) and the predicted rPPG (orange); (d) The target rPPG (blue) and the predicted rPPG (orange).

As shown in Table 1, all the settings yield better performance with increased T. These results suggest that observing a longer duration of video frames is essential to derive stable rPPG periodicity and robust HR estimation. In addition, for each setting of T, "Image-", "Video-" and "All-" all have improved performance over "Source-". These results verify that the data augmentation and rPPG estimation tasks indeed promote each other and substantially boost the overall performance.

Figures 5 and 6 give two examples synthesized by Image-to-Video network and Video-to-Video network, respectively. In both cases, the generated videos in Figs. 5(b) and 6(b) are visually indistinguishable from the source data, and the estimated rPPG signals in Figs. 5(c) and 6(d) are very accurately aligned with the ground truth. In addition, Fig. 6(e), where the estimated rPPG signal from the rPPG-removed video becomes a flat signal (with the average variance of 3.6×10^{-4}), demonstrates that the proposed removal network F successfully erases the rPPG information from the source video. The target rPPG signals and their estimated results in Figs. 5(d) and 6(f) also demonstrate that the proposed framework successfully embeds the target signals into the synthesized videos.

We also evaluate the visual quality of synthetic videos in terms of PSNR and SSIM on COHFACE dataset. As shown in Table 3, both the Image-to-Video network and the Video-to-Video network generate synthetic videos with

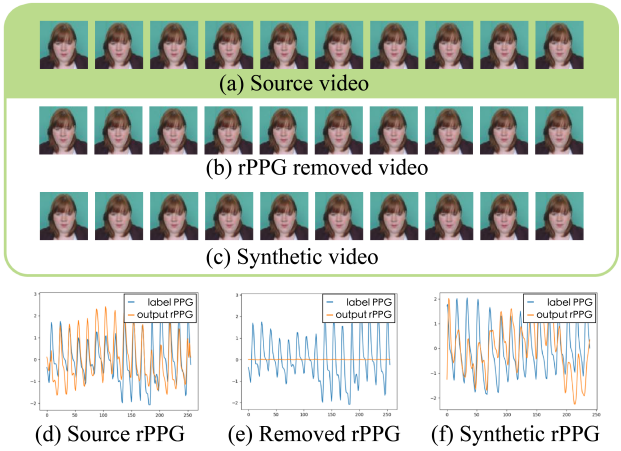

Fig. 6. A visualized example of Video-to-Video network. (a) Source video from the UBFC-RPPG dataset; (b) rPPG removed video after F; (c) The synthesized videos; (d) The source PPG label (blue) and the predicted rPPG (orange); (e) The source PPG label (blue) and the estimated rPPG from rPPG-removed video (orange); (f) The target rPPG (blue) and the predicted rPPG (orange).

high PSNR and SSIM and show that our method successfully generates photo-realistic videos visually indistinguishable from source data.

4.5 Results and Comparison

Table 2 shows the comparison with existing methods on the COHFACE dataset. The first three methods, i.e., 2SR [21], CHROME [6], and LiCVPR [18] are not learning-based methods; thus, there is a performance gap between them and the other two learning-based methods HR-CNN [11] and Two stream [5]. As to the proposed method "Ours-", we use the same settings "-source", "-image", "-video" and "-all" as mentioned in Table 1 with $T = 256$. The result of "Ours-Source" shows that, even without data augmentation, the proposed rPPG network P alone has already outperformed all these existing methods with a large margin. We believe there are two main reasons. First, P is an end-to-end network which directly processes the input video without any pre-processing step; hence, there involves no information loss in comparison with other multi-stage methods. Second, the 3D CNN architecture in P effectively captures the temporally periodic characteristics of rPPG signals in face video. Once we further include the data augmentation task, the proposed method "Ours-all" achieves the best performance with correlation coefficient (R) 0.72.

Table 4 shows the results on the original PURE dataset containing lossless PNG files. Because the uncompressed PNG files have better visual quality, most of the methods achieve a high correlation coefficient (R) larger than 0.9. Once we follow the settings in [11] to compress the PURE dataset into MPEG-4 and ex-

Table 3. Quality evaluation of synthetic videos on COHFACE dataset.

	PSNR	SSIM
Image-to-Video	33.38	0.993
Video-to-Video	34.07	0.994
rPPG removed video	37.71	0.996

periment on the compressed videos, the results in Table 5 show that our method is least sensitive to compression artifacts and significantly outperforms previous methods.

Moreover, in order to show that the proposed video generation largely enriches the training data than the simple data augmentation, we also conduct the experiments "Ours-Trad.Aug" using traditional augmentation, including random rotation, brightness, and saturation, for comparison. In Table 4 and Table 5, "Ours-Trad.Aug" shows no improvement over "Ours-Source" in both cases. Note that, because the training and testing videos are recorded in similar lighting environments with the same device, traditional data augmentation provides little information than the original training set. Instead, the proposed video generation method is able to generate videos with a variety of rPPG signals and is particularly advantageous for creating new benchmark datasets for rPPG estimation task.

Table 6 shows the results and comparison on the UBFC-RPPG dataset. The setting "Ours-Source" again outperforms the other methods. The precision 1.0 at 2.5 bpm indicates that there is no subjects' MAE larger than 2.5 bpm and verifies the robustness of the proposed method. As to the traditional augmentation, although "Ours-Trad.Aug." shows little improvement by decreasing the MAE and RMSE to 0.63 and 2.08, the proposed method "Ours-All" further decreases MAE and RMSE to 0.47 and 2.09, respectively. These results again show that our augmentation method effectively enriches the variety of dataset and largely improves the robustness of model training.

In addition, we conduct cross-dataset experiments to evaluate the generalization of the proposed method. As shown in Table 7, we show the results when training the network on PURE dataset but directly testing the model on UBFC-RPPG dataset and vice versa. Note that, this cross-dataset HR estimation inevitably leads to degraded performance, especially when the training dataset is of smaller-scale and less diversity than the testing dataset. Therefore, we have increased MAE from 0.40 to 4.24 and RMSE from 1.07 to 6.44 when testing on PURE dataset but training on UBFC-RPPG dataset. The main cause of this performance degradation comes from that PURE dataset contains different head movements whereas UBFC-RPPG dataset has none; therefore, the model trained on UBFC-RPPG is unable to adapt to different poses and head movements of PURE testing data. On the other hand, when testing on UBFC-RPPG dataset,

Table 4. Comparison on PURE dataset.

Method	R	MAE	RMSE
2SR [21]	0.98	2.44	3.06
CHROME [6]	0.99	2.07	2.50
LiCVPR [18]	-0.38	28.22	30.96
HR-CNN [11]	0.98	1.84	2.37
Ours-Source	0.88	0.44	1.16
Ours-Trad.Aug.	0.79	1.06	2.12
Ours-All	0.92	0.40	1.07

Table 5. Comparison on PURE dataset (MPEG-4 visual).

Method	R	MAE	RMSE
2SR [21]	0.43	5.78	12.81
CHROME [6]	0.55	6.29	11.36
LiCVPR [18]	-0.42	28.39	31.10
HR-CNN [11]	0.7	8.72	11.00
Two stream [5]	0.42	9.81	11.81
Ours-Source	0.86	0.79	1.76
Ours-Trad.Aug.	0.70	1.19	2.61
Ours-All	0.87	0.75	1.69

we only have a slight increase in MAE and RMSE, and our result (i.e., MAE 1.06) still outperforms existing methods and shows good generalization of the proposed model.

Table 6. Comparison on UBFC-RPPG dataset.

Method	MAE	RMSE	2.5 bpm	5 bpm
PVM [20]	4.47	-	0.71	0.81
MODEL [17]	3.99	5.55	0.75	0.87
SKIN-TISSUE [12]	-	2.39	0.89	0.83
MAICA [19]	3.34	-	0.72	0.88
BIC [16]	1.21	2.41	0.951	0.975
Ours-Source	0.73	2.38	1.0	1.0
Ours-Trad.Aug.	0.63	2.08	1.0	1.0
Ours-All	0.47	2.09	1.0	1.0

Table 7. Comparison of cross-dataset estimation.

Training-Testing	MAE	RMSE
PURE-PURE	0.40	1.07
UBFC-PURE	4.24	6.44
UBFC-UBFC	0.47	2.09
PURE-UBFC	1.06	2.70

5 Conclusions

To the best of our knowledge, this is the first work targeting generating synthetic face videos with specific rPPG signals. We study the impact of data augmentation and propose a novel multi-task learning method to simultaneously accomplish the data augmentation and the rPPG estimation tasks. By generating photo-realistic videos, we successfully augment the existing small-scale datasets with enriched characteristics and yield robust rPPG estimation. Our experimental results verify the effectiveness of the proposed method and show its great potential for promoting contactless estimation of human physiological signals.

References

1. Li, X., Alikhani, I., Shi, J., Seppanen, T., Junttila, J., Majamaa-Voltti, K., Tulppo, M., Zhao, G.: 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). (2018) 242–249
2. Yu, Z., Li, X., Zhao, G.: Recovering remote photoplethysmograph signal from facial videos using spatio-temporal convolutional networks. CoRR **abs/1905.02419** (2019)
3. Chen, W., McDuff, D.: Deepphys: Video-based physiological measurement using convolutional attention networks. In Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., eds.: The European Conference on Computer Vision (ECCV), Cham, Springer International Publishing (2018) 356–373
4. Chen, W., McDuff, D.J.: Deepmag: Source specific motion magnification using gradient ascent. CoRR **abs/1808.03338** (2018)
5. Wang, Z.K., Kao, Y., Hsu, C.T.: Vision-based Heart Rate Estimation via a Two-stream CNN. In: 2019 IEEE International Conference on Image Processing (ICIP). (2019) 3327–3331
6. de Haan, G., Jeanne, V.: Robust pulse rate from chrominance-based rppg. IEEE Transactions on Biomedical Engineering **60** (2013) 2878–2886
7. Hernandez-Ortega, J., Fierrez, J., Morales, A., Tome, P.: Time Analysis of Pulse-Based Face Anti-Spoofing in Visible and NIR. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. (2018)
8. Liu, Y., Jourabloo, A., Liu, X.: Learning Deep Models for Face Anti-Spoofing: Binary or Auxiliary Supervision. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. (2018) 389–398
9. Liu, S., Lan, X., Yuen, P.C.: Remote photoplethysmography correspondence feature for 3d mask face presentation attack detection. (2018)
10. Liu, S., Yuen, P.C., Zhang, S., Zhao, G.: 3d mask face anti-spoofing with remote photoplethysmography. Volume 9911. (2016) 85–100
11. Špetlík, R., Franc, V., Čech, J., Matas, J.: Visual Heart Rate Estimation with Convolutional Neural Network. In: Proceedings of British Machine Vision Conference. (2018)
12. Bobbia, S., Macwan, R., Benezeth, Y., Mansouri, A., Dubois, J.: Unsupervised skin tissue segmentation for remote photoplethysmography. Pattern Recognition Letters (2017)
13. Stricker, R., Müller, S., Gross, H.M.: Non-contact video-based pulse rate measurement on a mobile service robot. Volume 2014. (2014) 1056–1062
14. Niu, X., Han, H., Shan, S., Chen, X.: Synrhythm: Learning a deep heart rate estimator from general to specific. In: 2018 24th International Conference on Pattern Recognition (ICPR). (2018) 3580–3585
15. Heusch, G., Anjos, A., Marcel, S.: A reproducible study on remote heart rate measurement. CoRR **abs/1709.00962** (2017)
16. Benezeth, Y., Bobbia, S., Nakamura, K., Gomez, R., Dubois, J.: Probabilistic signal quality metric for reduced complexity unsupervised remote photoplethysmography. (2019) 1–5
17. Li, P., Yannick Benezeth, K.N., Gomez, R., Yang, F.: Model-based region of interest segmentation for remote photoplethysmography. In: 14th International Conference on Computer Vision Theory and Applications. (2019) 383–388

18. Li, X., Chen, J., Zhao, G., Pietikäinen, M.: Remote heart rate measurement from face videos under realistic situations. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. (2014) 4264–4271
19. Macwan, R., Benezeth, Y., Mansouri, A.: Heart rate estimation using remote photoplethysmography with multi-objective optimization. Biomedical Signal Processing and Control **49** (2019) 24–33
20. Macwan, R., Bobbia, S., Benezeth, Y., Dubois, J., Mansouri, A.: 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). (2018) 1413–14138
21. Wang, W., Stuijk, S., de Haan, G.: A novel algorithm for remote photoplethysmography: Spatial subspace rotation. IEEE Transactions on Biomedical Engineering **63** (2016) 1974–1984
22. Tsou, Y.Y., Lee, Y.A., Hsu, C.T., Chang, S.H.: Siamese-rppg network: Remote photoplethysmography signal estimation from face video. In: The 35th ACM/SIGAPP Symposium on Applied Computing (SAC'20). (2020)
23. Yu, Z., Peng, W., Li, X., Hong, X., Zhao, G.: Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement. In: International Conference on Computer Vision (ICCV). (2019)
24. Dvornik, N., Mairal, J., Schmid, C.: Modeling visual context is key to augmenting object detection datasets. In: The European Conference on Computer Vision (ECCV). (2018)
25. Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: Synthetic data augmentation using gan for improved liver lesion classification. (2018) 289–293
26. Qiu, Y., Liu, Y., Arteaga-Falconi, J., Dong, H., Saddik, A.E.: Evm-cnn: Real-time contactless heart rate estimation from facial video. IEEE Transactions on Multimedia **21** (2019) 1778–1787