UNIVERSITY OF CALIFORNIA

Los Angeles

Improving Skin Color Diversity in Remote-PPG Using Synthetic Subjects

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Electrical and Computer Engineering

by

Oyku Deniz Bozkurt

2022

ABSTRACT OF THE THESIS

Improving Skin Color Diversity in Remote-PPG Using Synthetic Subjects

by

Oyku Deniz Bozkurt
Master of Science in Electrical and Computer Engineering
University of California, Los Angeles, 2022
Professor Achuta Kadambi, Chair

Camera-based remote photoplethysmography (rPPG) provides a non-contact way to measure physiological signals (e.g., heart rate) using facial videos. Recent deep learning architectures have improved the accuracy of such physiological measurement significantly, yet they are restricted by the diversity of the annotated videos. The existing datasets MMSE-HR, AFRL, and UBFC-RPPG contain roughly 10%, 0%, and 5% of dark-skinned subjects respectively. The unbalanced training sets result in a poor generalization capability to unseen subjects and lead to unwanted bias toward different demographic groups. Here we show a first attempt to overcome the lack of dark-skinned subjects by synthetic augmentation. A joint optimization framework is utilized to translate real videos from light-skinned subjects to dark skin tones while retaining their pulsatile signals. In the experiment, our method exhibits around 31% reduction in mean absolute error for the dark-skinned group and 46% improvement on bias mitigation for all the groups, as compared with the previous work trained with just real samples.

The thesis of Oyku Deniz Bozkurt is approved.

Yang Zhang

Anthony Chen

Achuta Kadambi, Committee Chair

University of California, Los Angeles

2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would first like to thank my advisor, Prof. Achuta Kadambi. Throughout my degree, he has inspired me and taught me much of what I know about the science of computational imaging. I would also like to thank my committee members, Prof. Yang Zhang and Prof. Anthony Chen, for their time and support.

Thanks should also go to Yunhao Ba, Zhen Wang, and Doruk Karinca, who have been great coauthors to work with and learn from. Their hard work and dedication to this project have made this thesis possible.

Lastly, I'd like to acknowledge the support of my family and friends. Their encouragement and help were invaluable to me in these past two years.

PREVIOUS PUBLICATIONS

This thesis revises the following publication:

Y. Ba, Z. Wang, D. Karinca, O. D. Bozkurt, and A. Kadambi, "Overcoming difficulty in obtaining dark-skinned subjects for remote-PPG by synthetic augmentation" *arXiv* (2021) [6].

CHAPTER 1

Introduction

Photoplethysmography (PPG) is an optical technique that measures blood volume changes in the skin by detecting the light reflected or transmitted through the skin. It is widely used in medicine due to its low cost and ability to capture vital signs such as oxygen saturation, heart rate, blood pressure and cardiac output. PPG can also assess autonomic function and detect peripheral vascular disease [7]. In clinical settings, a contact device called a pulse oximeter is placed on the finger of the patient to measure PPG signals.

New contactless PPG methods have also been emerging as an alternative to contact-based methods such as pulse oximeters. These contactless methods are called Remote Photoplethysmography (rPPG) and they aim to acquire PPG signals from a camera. In light of the recent COVID-19 pandemic, rPPG is gaining increasing relevance due to its potential to reduce viral transmissions if widely adopted by clinicians. As cameras are ubiquitous in modern electronic devices, rPPG can be used in telemedicine without any equipment set-up needed [8]. Camera-based rPPG techniques have also been used in other applications such as driver monitoring [9] and face anti-spoofing [10].

Algorithms for non-contact rPPG can be roughly classified into three categories: Signal decomposition [11, 2, 12, 13, 14], model-based methods [15, 16, 17], and deep learning methods [18, 3, 19, 20]. Signal decomposition techniques based on Blind Source Separation (BSS) decompose/demix the face videos into different sources utilizing PCA [11] or ICA [2, 12]. For model-based methods, Pulse Blood Vector [15] utilizes the characteristic blood volume signature to weigh different color channels. CHROM [16] first eliminates the specular components and applies color space transforms to linearly combine the chrominance signals. POS [17] modifies this by first projecting the temporally-normalized skin tone onto the plane which is orthogonal to the intensity variation direction and then linearly combining the projected signals. These model-based methods usually use spatially averaged intensity values of skin pixels for pulse extraction, which may achieve sub-optimal results as each pixel can contribute differently to the underlying pulse signals.

More recently, deep learning and convolutional neural networks (CNN) have been more popular due to their expressiveness and flexibility [21, 3, 22, 20, 23, 24, 6, 25]. The relationship between the pulse signal and the color variations on the face captured by cameras is very complex due to various optical effects such as specular reflection. CNNs learn the mapping between the pulse signal and the color variations with end-to-end supervised training on the labeled dataset, thus achieving state-of-the-art performance on vital sign detection. Therefore, the performance of data-driven rPPG networks hinges on the quality of the dataset.

Even though data-driven rPPG networks have exhibited remarkable estimation accuracy for non-contact camera-based sensing [18, 3, 19, 20], there is still a lot of bias in the results of these networks against people with darker skin tones. Unfortunately, this bias is not surprising given other machine learning algorithms that have been shown to discriminate based on race and gender [26, 27]. For the case of rPPG, at least a part of the bias can be attributed to unbalanced datasets with little representation of darker skin tones [28]. Currently, available rPPG datasets MMSE-HR, AFRL, and UBFC-RPPG contain roughly 10%, 0%, and 5% of dark-skinned subjects respectively. Realizing the difficulty of recruiting patients to collect large-scale rPPG datasets in the university setting, synthetic augmentation of facial videos has become an active research topic recently. McDuff et al. [29] use synthetic avatars with ray tracing to reflect the blood volume changes under various configurations. However, as the authors point out, that infrastructure is labor-intensive and requires a significant amount of rendering time for each frame (approximately 20 seconds per frame), which impedes their scalability. Pulse signals can also be incorporated to make the synthetic avatars more lifelike, yet it is difficult for avatar-based methods to generate a balanced dataset due to the lack of dark-skinned avatars [30]. Given this reality, we have set out to directly augment the existing rPPG datasets to create synthetic face videos that can be used to train rPPG networks and reduce dataset bias.

This thesis will first provide a background of PPG technology and discuss notable previous works in rPPG. Then, a novel method for mitigating the effects of dataset bias in rPPG will be presented. This method involves using bio-realistic skin tone translation in order to augment existing datasets with artificial skin tone diversity. Next, the performance of our custom rPPG model will be demonstrated through empirical results and compared against the performance of other state-of-the-art models. Finally, the results of this work will be discussed and future work will be proposed.

1.1 Contributions

To summarize, the contributions of this thesis include:

- We introduce a first attempt to translate facial videos of light-skinned subjects to dark tones while preserving the underlying blood volume variations;
- We demonstrate that our synthetic videos can be directly utilized to improve the performance of the state-of-the-art deep rPPG methods with mitigated bias across different demographic groups;
- We propose a simple yet efficient rPPG estimation model based on 3D convolution operations and show that the proposed model can achieve state-of-the-art performance on various facial videos.

CHAPTER 2

Background

2.1 The working principle of the pulse oximeter

A pulse oximeter is a contact-based sensor that measures light absorption in blood. It is typically placed directly on the finger of a patient in order to read their blood oxygen saturation levels (also called hemoglobin saturation) and pulse rate. The pulse oximeter uses a light source (most commonly an LED) to transmit light through the skin and a photodiode to measure the amount of light received back at the device after traveling through blood.

The design of the pulse oximeter is based on the Beer-Lambert law which relates the attenuation of light through a medium to the properties of that medium. This relationship can be mathematically formulated as:

$$A = \varepsilon l c \tag{2.1}$$

where A represents the absorbance, ε the absorptivity of the medium, l the path length in cm, c the concentration of the solution.

Methemoglobin (MetHb), carboxyhemoglobin (HbCO), hemoglobin (Hb) and oxyhemoglobin (HbO_2) are the only significant light absorbers commonly found in blood [31]. However, pulse oximeters emit two wavelengths (660 nm and 940 nm) and therefore can only account for two separate absorbers. These absorbers are hemoglobin and oxyhemoglobin, which have different absorption spectra at 660 nm (red) and 940 nm (infrared) wavelengths. At 660 nm, HbO_2 absorbs less light than Hb and at 940 nm Hb absorbs less light than HbO_2 [32].

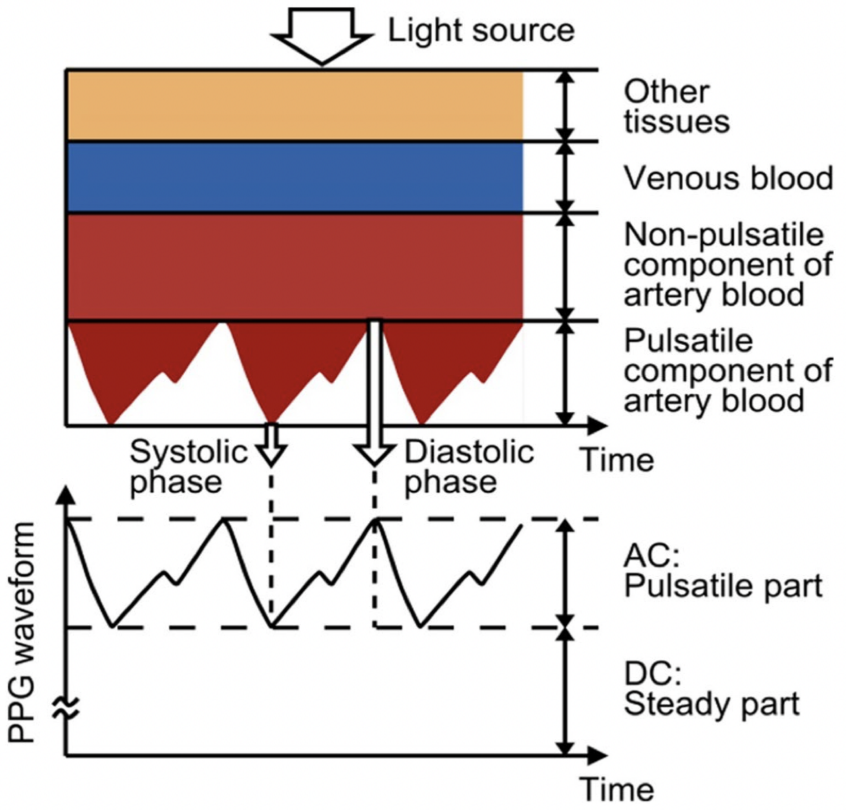

Figure 2.1: **PPG Waveform Generation [1].** Light transmitted through the skin generates a PPG signal on the pulse oximeter's photodiode. The AC component of the PPG signal encodes pulsatile information.

Applying the Beer-Lambert law to the light received at the photodiode at these two different wavelengths, the concentration of both hemoglobin and oxyhemoglobin can be measured. Then, the hemoglobin saturation (SaO_2) can be calculated using the ratio between the concentration of hemoglobin and oxyhemoglobin [33]:

$$SaO_2 = \frac{HbO_2}{HbO_2 + Hb} \cdot 100 \tag{2.2}$$

This ratio is calibrated using direct measurements of hemoglobin saturation collected from a group of volunteers and the calibration values are stored inside of the pulse oximeter for later use.

As the pulse oximeter cycles the 660 nm and 940 nm light sources on and off to capture the red light, the infrared light and the ambient light, there is also a photoplethysmographic (PPG) transmission waveform produced [31]. This PPG waveform has a direct current (DC) component and an alternating current (AC) component. While the DC component is dependent on the structure of tissue inside a pulse oximeter, the AC component corresponds to the blood volume changes that happen throughout the cardiac cycle [1]. The fundamental frequency of the AC component can be used to calculate the heart rate (HR).

2.2 Remote PPG

As opposed to the contact-based pulse oximeter, remote PPG techniques only require a video of a subject's facial region to obtain their PPG waveform. Most commonly, the video is taken by a consumer-grade webcam, which captures a mix of the PPG signal and other light sources present in the environment. In this section, different techniques to acquire the underlying PPG signal will be discussed.

2.2.1 Blind Source Separation techniques

2.2.1.1 ICA [2]

Because absorptivity varies with wavelength, each color channel in the RGB sensor captures a different combination of the underlying PPG signals when a face video is recorded by a webcam. The ICA model assumes that these combinations are linear and attempts to find a demixing matrix that can approximate the underlying source signals from the captured signals.

To this end, first, a region of interest (ROI) is determined using the Viola–Jones object detection framework. Then, the video is cropped according to the ROI keeping only the face region in the video. Afterwards, the cropped video is separated into its three channels:

red, green and blue. For each time stamp and for each color channel, the video is spatially averaged to produce a raw 1D signal that varies with time. All three raw signals are detrended and normalized before independent component analysis (ICA) is applied, which returns three separated source signals. The signal whose power spectrum contains the highest peak is then chosen as the underlying PPG signal and smoothed using a five-point moving average filter and bandpass filtered. The passband for the filter is 0.7-4 Hz, which corresponds to the frequency of a normal heart rate.

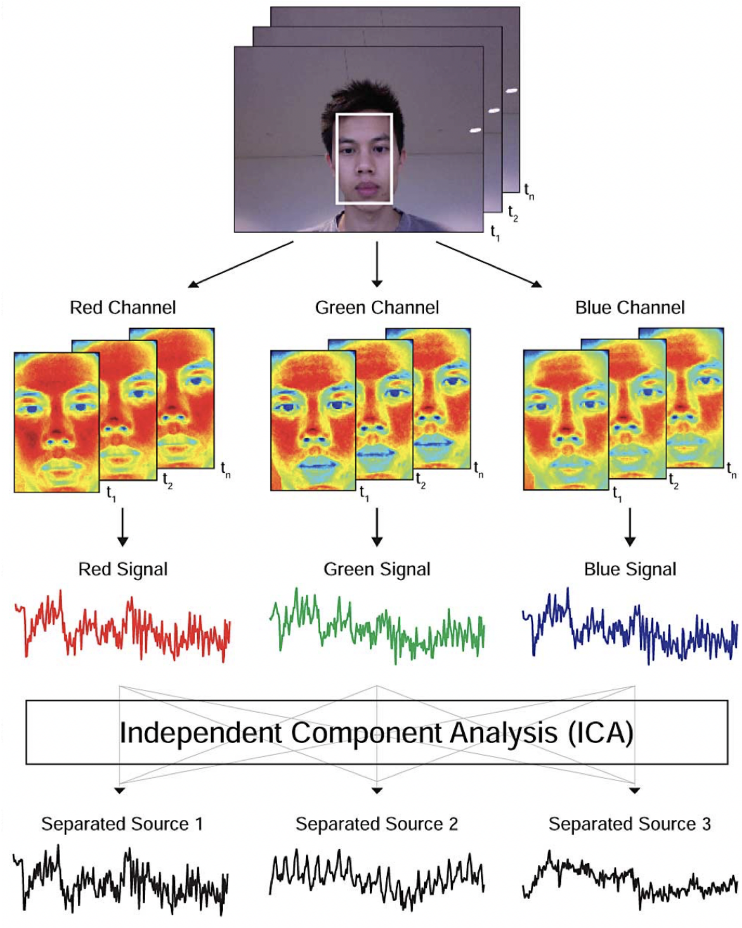

Figure 2.2: **ICA Pipeline [2].** The video is cropped using face detection before it is separated into the R, G, and B channels. Then, each channel is spatially averaged and normalized before applying ICA to get a corresponding source signal which encodes PPG.

2.2.2 Model-based methods

2.2.2.1 CHROM [16]

Similar to ICA, CHROM assumes that blood volume pulse is captured by the camera in the form of light reflected from the face. CHROM models this process for each color channel:

$$C_i = I_{Ci}(\rho_{Cdc} + \rho_{Ci} + s_i), \tag{2.3}$$

where I_{Ci} is the intensity of the light source integrated over the exposure time of the camera in image i for color channel C, ρ_{Cdc} is the stationary part of the reflection coefficient of the skin in color channel C, ρ_{Ci} is the zero-mean time-varying fraction caused by the pulsation of the blood volume and s_i is the specular reflection contribution.

The specular reflection component contains the light that is reflected directly back from the surface and doesn't contain any pulsatile information. Additionally, assuming white light, each color channel is affected by the specular reflection contribution equally. Therefore, CHROM eliminates this specular component by subtracting color channels from each other, which results in two chrominance signals X and Y:

$$X = R - G \tag{2.4}$$

$$Y = 0.5R + 0.5G - B^3 \tag{2.5}$$

Then, a ratio between the normalized chrominance signals (X_n and Y_n) is used to calculate the underlying PPG signal:

$$S = \frac{X_n}{Y_n} - 1 \tag{2.6}$$

Considering non-white illumination and the standard deviation of the chrominance, the coefficients in the chrominance calculation can be further adjusted. For more details, see [16].

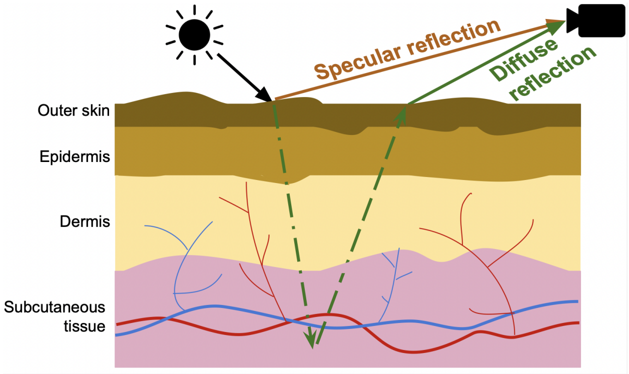

Figure 2.3: **Illustration of the dichromatic skin model.** Light incident on skin produces two reflective components. The specular component is due to the reflection from the skin surface, and the diffuse component is related to the absorption and scattering properties of the skin tissues. CHROM aims to eliminate contributions from the specular component in the RGB video since they do not contain pulsatile information.

2.2.2.2 POS [17]

POS stands for "Plane-Orthogonal-to-Skin" and is yet another rPPG model which estimates a PPG signal from the specular and diffuse reflections of light captured on RGB video. The POS model is similar to CHROM, but alters the order in which the main expected color distortions are reduced using different priors [17]. While CHROM first eliminates specular reflection using chrominance, POS first eliminates intensity variations.

The skin reflection model of POS separates the temporally-normalized and spatially averaged RGB matrix at time t ($\mathbf{C_n}(t)$) into three components: intensity, specular and pulse. Because the intensity component is in the direction of the unit vector $\mathbf{1} = (1, 1, 1)^T$ in this formulation, $\mathbf{C_n}(t)$ is projected onto a plane orthogonal to $\mathbf{1}$ to eliminate intensity variations. This projection can be expressed as:

$$\mathbf{S}(t) = \mathbf{P_p} \cdot \mathbf{C_n}(t), \tag{2.7}$$

where $\mathbf{P_p}$ is a 2x3 projection matrix and $\mathbf{C_n}(t)$ represents the temporally-normalized and spatially averaged RGB matrix at time t.

Because $\mathbf{P_p}$ is a plane that is orthogonal to $\mathbf{1}$ in the normalized RGB space, it is also orthogonal to the temporally-normalized skin tone. The main advantage of projecting the RGB matrix $\mathbf{C_n}(t)$ onto a "Plane-Orthogonal-to-Skin" is that motion-induced intensity variations are usually larger distortions that affect all three RGB channels at once.

POS and CHROM achieve similar performances with POS gaining a slight overall edge in various benchmarks [17]. Most notably, POS has been shown to perform better than CHROM on subjects with diverse skin tones due to its skin-tone-orthogonal projection methodology.

2.2.3 Deep learning based methods

2.2.3.1 PhysNet [3]

PhysNet is an end-to-end spatio-temporal rPPG network that uses a 3D convolutional neural network (3D-CNN) architecture. The input to this network is T frames of RGB video, which is forwarded through the 3D-CNN block. Within the 3D-CNN block, there are multiple convolution and pooling operations as shown in Fig. 2.4. The 3x3x3 convolutions in the 3D-CNN block help extract the semantic rPPG features in both spatial and temporal domain simultaneously [3]. After the 3D-CNN block, the final signal projection is achieved using aggregation.

Because PhysNet aims to not only discover the underlying heart rate but also locate precisely where the peaks in the PPG signal are, the negative Pearson correlation is used as the loss function for this network. The negative Pearson correlation helps maximize the trend similarity and minimize peak location errors [3].

PhysNet was trained on the OBF dataset and tested on both the OBF dataset and MAHNOB-HCI. Through these experiments, it was shown that PhysNet could outperform model-based methods such as CHROM and POS.

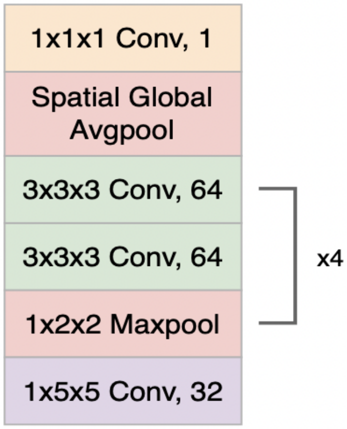

Figure 2.4: **The architecture of the 3D-CNN block used in PhysNet [3].** The input frames go through a convolution filter with a 1x5x5 kernel, followed by a max pool layer and two layers of 3x3x3 convolution which are repeated 4 times. The resulting feature map is then average pooled and passed through a final 1x1x1 convolution.

2.2.3.2 3D-CNN [5]

3D-CNN is a multi-task learning framework that simultaneously generates synthetic videos and estimates PPG signals from video (Note: This is a different 3D-CNN than discussed in the previous section). While being able to augment source rPPG videos by other pulse signals, this framework is restricted to the face appearance in the original source videos and fails to produce novel videos with dark skin tones.

The 3D-CNN multi-task learning framework contains three networks: rPPG network, Image-to-Video network, and Video-to-Video network. The rPPG network estimates PPG signals from input videos, the Image-to-Video network generates a synthetic video given a target rPPG signal and a source image, and the Video-to-Video network takes a source video and imposes a different target rPPG signal onto it. First, the rPPG network is pre-trained using source videos. Then, the Image-to-Video and Video-to-Video networks are trained using the pre-trained rPPG network. During this step, the weights of the rPPG network are frozen. Lastly, the newly generated synthetic videos are used to finetune the rPPG network.

CHAPTER 3

Methods

This chapter presents a novel method for improving the accuracy of deep learning based rPPG models by reducing skin tone bias in existing datasets. Specifically, bio-realistic skin tone translation augments existing datasets by realistically translating the skin tone of some subjects to be darker while retaining PPG signal integrity.

3.1 Bio-realistic skin tone translation

In order to translate real subjects with light skin tones to synthetic subjects with dark skin tones, we utilize two interconnected networks: a video generator G and an rPPG estimator E, as illustrated in Figure 3.1. We next describe the proposed 3D convolutional video generator, the rPPG estimation network, and our joint optimization scheme.

3.1.1 3D convolutional video generator

The goal of our video generator G is to translate frame sequences of real light-skinned subjects to synthetic dark-skinned subjects. We propose a novel 3D convolutional neural network to accomplish this goal. The model consists of an encoder (several convolutional layers), a transformer (6 ResNet Blocks), and finally a decoder (several convolutional layers). The architecture of the generation network can be seen in Fig 3.2.

The generator takes 256 consecutive frames $\mathbf{I}_{light}$ at size 80×80 as the input and generates the corresponding translated frames in the same dimension. Since the paired ground-truth

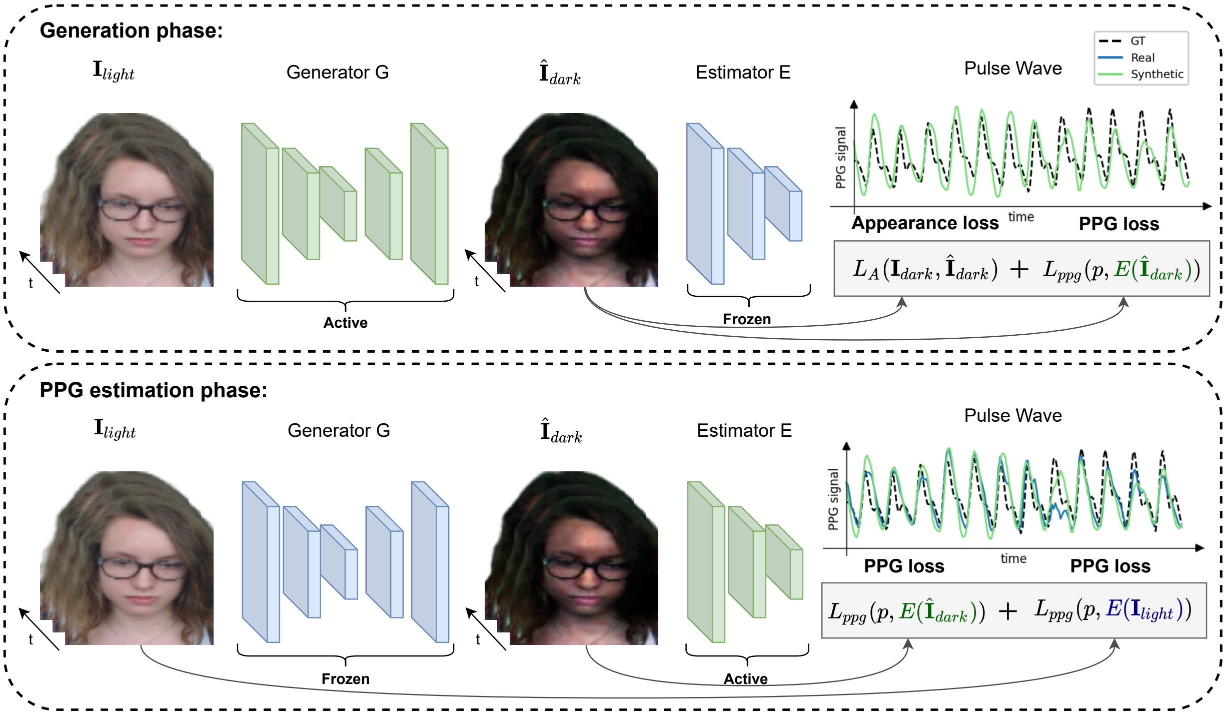

Figure 3.1: **Illustration of the proposed joint optimization framework.** Our framework is capable of translating light-skinned facial videos to dark skin tones while maintaining the original pulsatile signals. With a two-phase weight updating scheme, the rPPG estimation network can benefit from the synthetic dark-skinned videos and gradually learn to conduct inference on dark-skinned subjects without accessing real facial videos with dark skin tones.

translated frames do not exist, we use a race transfer model [4] pretrained on VGGFace2 [34] to generate the pseudo target frames $\mathbf{I}_{dark}$. More specifically, the generator *Caucasian-to-African* in [4] is utilized to translate videos of light-skinned subjects in the existing rPPG dataset to dark skin tones.

The generator is first supervised by the L1 distance between the pseudo target frames $\mathbf{I}_{dark}$ and the generated frames $\hat{\mathbf{I}}_{dark} = G(\mathbf{I}_{light})$ to learn the visual appearance of the synthetic dark-skinned subjects. At this stage, the output frames $\hat{\mathbf{I}}_{dark}$ do not contain pulsatile signal, since the target frames $\mathbf{I}_{dark}$ from [4] are generated in a frame-by-frame manner without temporal pulse correspondence along the time dimension. In the joint optimization section, we describe how to further incorporate the pulsatile signals presented in the original videos $\mathbf{I}_{light}$ into the generated frames.

3.1.2 PRN: rPPG estimator with residual connections

The rPPG estimator is designed to model the BVP temporal information from a sequence of facial frames. Similarly, it takes 256 consecutive frames at size 80×80 as the input, and its output is the corresponding BVP value for each input frame. We build our novel rPPG estimator based on 3D convolution operations. It consists of three consecutive 3D convolutional blocks with residual connections, and an average pooling is performed after each block for the downsampling purpose. The full architecture can be found in Fig 3.3.

To supervise the network, we use a negative Pearson correlation loss between the estimated pulse signals $\hat{p} \in \mathbb{R}^T$ and the ground-truth pulse signals $p \in \mathbb{R}^T$:

$$L_{ppg}(p, \hat{p}) = 1 - \frac{T \sum_{i=1}^{T} p_i \hat{p}_i - \sum_{i=1}^{T} p_i \sum_{i=1}^{T} \hat{p}_i}{\sqrt{\left(T \sum_{i=1}^{T} p_i^2 - \left(\sum_{i=1}^{T} p_i\right)^2\right)\left(T \sum_{i=1}^{T} (\hat{p}_i)^2 - \left(\sum_{i=1}^{T} \hat{p}_i\right)^2\right)}}. \tag{3.1}$$

This negative Pearson correlation loss has shown to be more effective as compared with the point-wise mean squared error (MSE) loss in the previous work [3]. We first train PRN with only real subjects, and this simple yet efficient architecture can already achieve state-of-the-art performance on the existing rPPG datasets. In the next section, we detail how to further incorporate the synthetic subjects into the training process.

3.1.3 Joint optimization

The generator trained with L1 loss in the previous section fails to produce synthetic dark-skinned subjects with desired pulsatile information, and the rPPG estimator trained with only real light-skinned subjects exhibits poor generalization capability on unseen data or data that rarely appears in the training set (i.e., the underrepresented group with dark skin tones). To make use of these two models, we design a joint optimization mechanism to incorporate pulsatile signals into the synthetic videos and improve the generalizability of the rPPG estimator simultaneously.

We use a two-phase weight updating scheme to train the video generator and the rPPG estimator simultaneously. These two phases are alternated within each mini-batch as illustrated in Figure 3.1. In the generation phase, we freeze the weight of the rPPG estimator E, and the generator G is supervised by the following loss function to maintain both the visual appearance and the pulsatile information:

$$L_G(\mathbf{I}_{light}, p) = L_{ppg}(p, E(\hat{\mathbf{I}}_{dark})) + \lambda * L_A(\mathbf{I}_{dark}, \hat{\mathbf{I}}_{dark}), \tag{3.2}$$

$$L_A(\mathbf{I}_{dark}, \hat{\mathbf{I}}_{dark}) = \frac{1}{\sum_i z_i} \sum_i z_i |\mathbf{I}_{dark_i} - \hat{\mathbf{I}}_{dark_i}|, \tag{3.3}$$

$$z_i = \begin{cases} 0 & \text{if } |\mathbf{I}_{dark_i} - \hat{\mathbf{I}}_{dark_i}| < \epsilon \\ 1 & \text{otherwise} \end{cases}, \tag{3.4}$$

where $\hat{\mathbf{I}}_{dark} = G(\mathbf{I}_{light})$ is the generated frame sequence from synthetic dark-skinned subjects, λ is the balance factor, $L_A(\cdot)$ is the visual appearance loss designed based on a threshold L1 loss, and ϵ is the selected threshold. The weighting factor λ is chosen to be 1.0. Directly enforcing a L1 loss between $\mathbf{I}_{dark}$ and $\hat{\mathbf{I}}_{dark}$ causes the generator to struggle between the visual appearance and the pulse information, since the pseudo ground-truth $\mathbf{I}_{dark_i}$ from [4] do not contain the desired BVP variations. Therefore, we relax the appearance loss $L_A(\cdot)$ by a threshold ϵ. The relaxation is based on the observation that the color changes due to BVP variations are subtle in the RGB domain. In our implementation, we select $\epsilon = 0.1$ based on an empirical analysis of the color variations in real videos.

In the rPPG estimation phase, we freeze the weight of the generator G and train the rPPG estimator E with both real and synthetically augmented frame sequences:

$$L_E(\mathbf{I}_{light}, \hat{\mathbf{I}}_{dark}), p) = L_{ppg}(p, E(\hat{\mathbf{I}}_{dark})) + L_{ppg}(p, E(\mathbf{I}_{light})). \tag{3.5}$$

Both real and synthetic subjects are utilized to supervise the rPPG network E while updating its weights. This arrangement allows E to gradually adapt to the synthetic dark-skinned subjects without losing estimation accuracy on real subjects. With this two-phase updating rule, both the generator and the rPPG estimator benefit from each other in an alternate manner. At convergence, the generator G can successfully translate frame sequences from real light-skinned subjects to dark skin tones while maintaining the original BVP variations, and the estimator E can generalize its performance to dark skin tones without using actual real videos from dark-skinned subjects.

3.2 Datasets

UBFC-RPPG [35]: UBFC-RPPG database contains 42 front facial videos from 42 subjects, and the corresponding ground-truth PPG signals are collected from a fingertip pulse oximeter. The videos are recorded at 30 frames per second with a resolution of 640x480 in the uncompressed 8-bit AVI format. Each video is roughly one minute long.

VITAL dataset [36]: Facial videos are recorded at 1920x1080 pixel resolution and 30 frames per second for 60 subjects at room lighting in the highly compressed MP4 format. Each video is roughly 2 minutes long. A Philips IntelliVue MX800 patient monitor is utilized for ground-truth vital sign monitoring. The subject wears a blood pressure cuff, 5-ECG leads, and a finger pulse oximeter, which is connected to the MX800 unit. Diverse skin tones and varied demographic groups are represented in the dataset. We use 58 subjects in the VITAL dataset (subject 26 and subject 40 are left out due to data errors in the collecting process). For the skin types quantified by Fitzpatrick scales [37], there are 5, 16, 14, 11, 5, 7 subjects respectively from I (lightest) to VI (darkest).

3.3 Model training and evaluation

In this section, the training and evaluation pipeline of PRN is described.

3.3.1 Preprocessing

The PRN model takes as input an 80x80 facial video. The facial bounding box for each video is estimated by applying a face detector based on Multitask Cascaded Convolutional Neural Networks (MTCNN) [38] to its first frame, and a square region with 160% width and height of the detected bounding box is cropped and resized to 80×80 using linear interpolation.

3.3.2 Training

The UBFC-RPPG dataset is randomly split into a training set (32 subjects) and a validation set (10 subjects). The training set is used to jointly optimize the generator G and the rPPG estimator E. Models with minimum validation loss are selected for a cross-dataset evaluation on the VITAL videos.

The learning rate for the generator and the rPPG network are 0.0001 and 0.0003 respectively. The learning rates are modified base on a cosine annealing schedule during training [39]. The networks are initialized with Kaiming initialization [40] with a batch size of two and ReLU activation. We use the Adam [41] solver with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The network architectures are implemented with batch normalization [42] in PyTorch [43], and the experiments are conducted on a single NVIDIA Tesla V100 GPU.

3.3.3 Evaluation

In real-world applications, it is common that the test subjects are in a different environment (e.g., illumination conditions) in contrast to the training samples. Therefore, we conduct a cross-dataset evaluation on the VITAL dataset using the models trained on the UBFC-RPPG videos.

After obtaining the estimated pulse waves from each model, we apply a Butterworth filter to the output signals with cut-off frequencies of 0.7 and 2.5 Hz for heart rate estimation. The filtered waves are divided with sliding windows of 30-second length and 1-second stride, and a heart rate is estimated based on the position of the peak frequency for each window. For each subject, four error metrics are calculated and averaged over all windows. The four metrics include MAE, RMSE, PCC between the estimated heart rate and the ground-truth heart rate, and SNR of the estimated PPG waves. The ground-truth HR for UBFC-RPPG is obtained by applying the same procedures as described above to the ground-truth pulse waves, and the ground-truth HR for the VITAL dataset is obtained from the MX800 patient monitor through ECG signals.

Here are the formulas used to calculate MAE, RMSE, PCC, and SNR:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{N} |\mathrm{HR}_i - \mathrm{HR}_i|}{N}, \tag{3.6}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (\mathrm{HR}_i - \mathrm{HR}_i)^2}{N}}, \tag{3.7}$$

$$\mathrm{PCC} = \frac{T\sum_{i=1}^{T} p_i\hat{p}_i - \sum_{i=1}^{T} p_i \sum_{i=1}^{T} \hat{p}_i}{\sqrt{\left(T\sum_{i=1}^{T} p_i^2 - \left(\sum_{i=1}^{T} p_i\right)^2\right)\left(T\sum_{i=1}^{T} (\hat{p}_i)^2 - \left(\sum_{i=1}^{T} \hat{p}_i\right)^2\right)}}, \tag{3.8}$$

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\sum_{f=0.75}^{2.5}\left(U_t(f)\hat{S}(f)\right)^2}{\sum_{f=0.75}^{2.5}\left((1 - U_t(f))\,\hat{S}(f)\right)^2}\right), \tag{3.9}$$

where N is the total number of windows, p is the ground-truth pulse wave, $\hat{p}$ is the estimated pulse signal, $\hat{S}$ is the power spectrum of the pulse signal, f is the frequency in Hz, and $U_t(\cdot)$ is a binary mask. For the heart frequency region from f_{HR} - 0.1 Hz to f_{HR} + 0.1 Hz and its first harmonic region from 2 * f_{HR}- 0.1 Hz to 2 * f_{HR} + 0.1 Hz, $U_t(\cdot)$ is set to be one. For other regions, $U_t(\cdot)$ is set be zero.

Comparison methods: We compare our model with three conventional methods outlined in the Background of this thesis: POS [17], CHROM [16] and ICA [2]. These rPPG baseline methods are implemented based on the publicly available MATLAB toolbox [44], and we follow the procedures in the toolbox to obtain facial pixels of interest, i.e., converting facial frames from RGB to YC_RC_B and identifying skin pixels based on a predefined threshold. We also compare with a data-driven state-of-the-art rPPG algorithm 3D-CNN [5]. It is implemented based on the architecture description as detailed in the original publication.

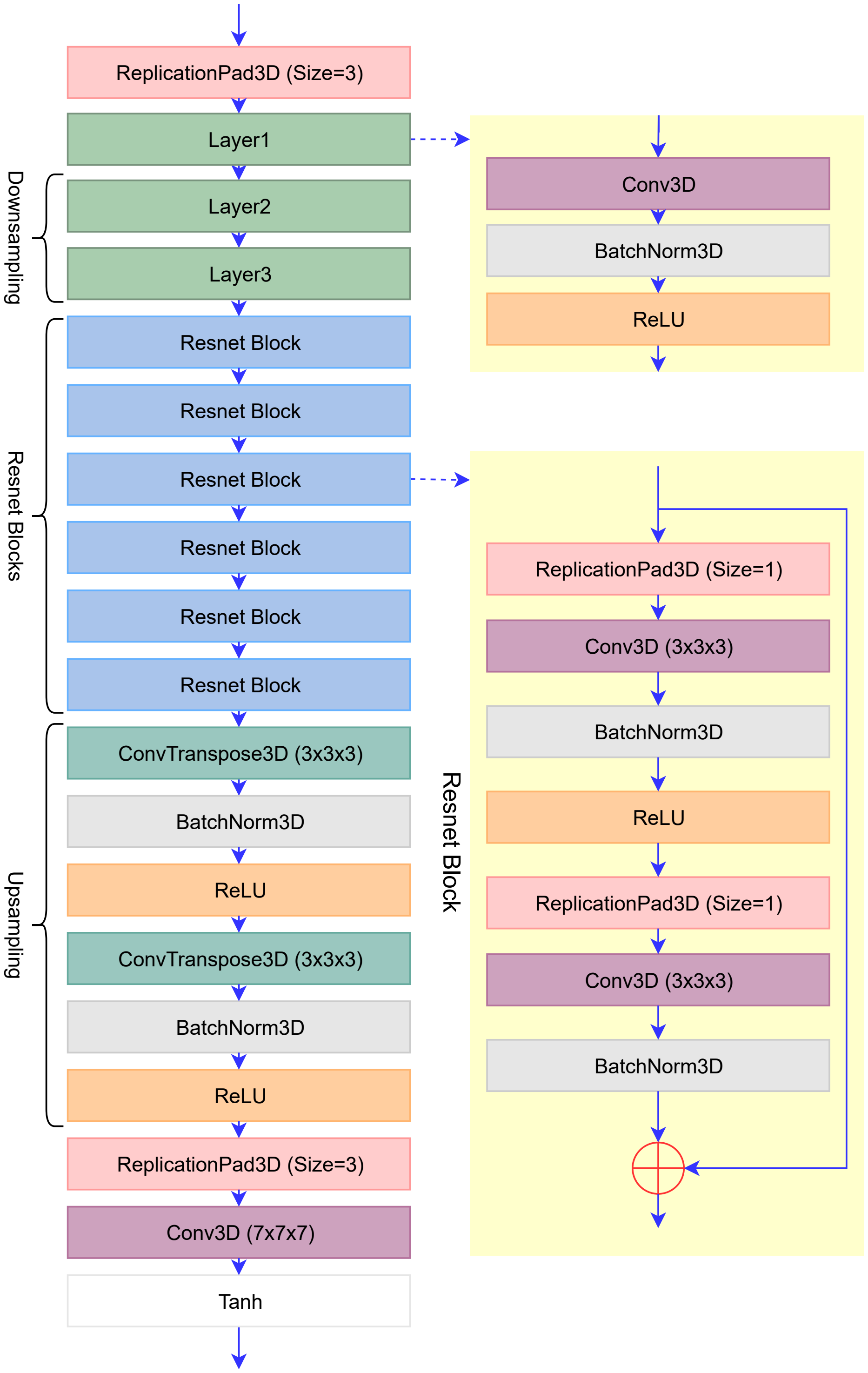

Figure 3.2: **Architecture of the generation network.**

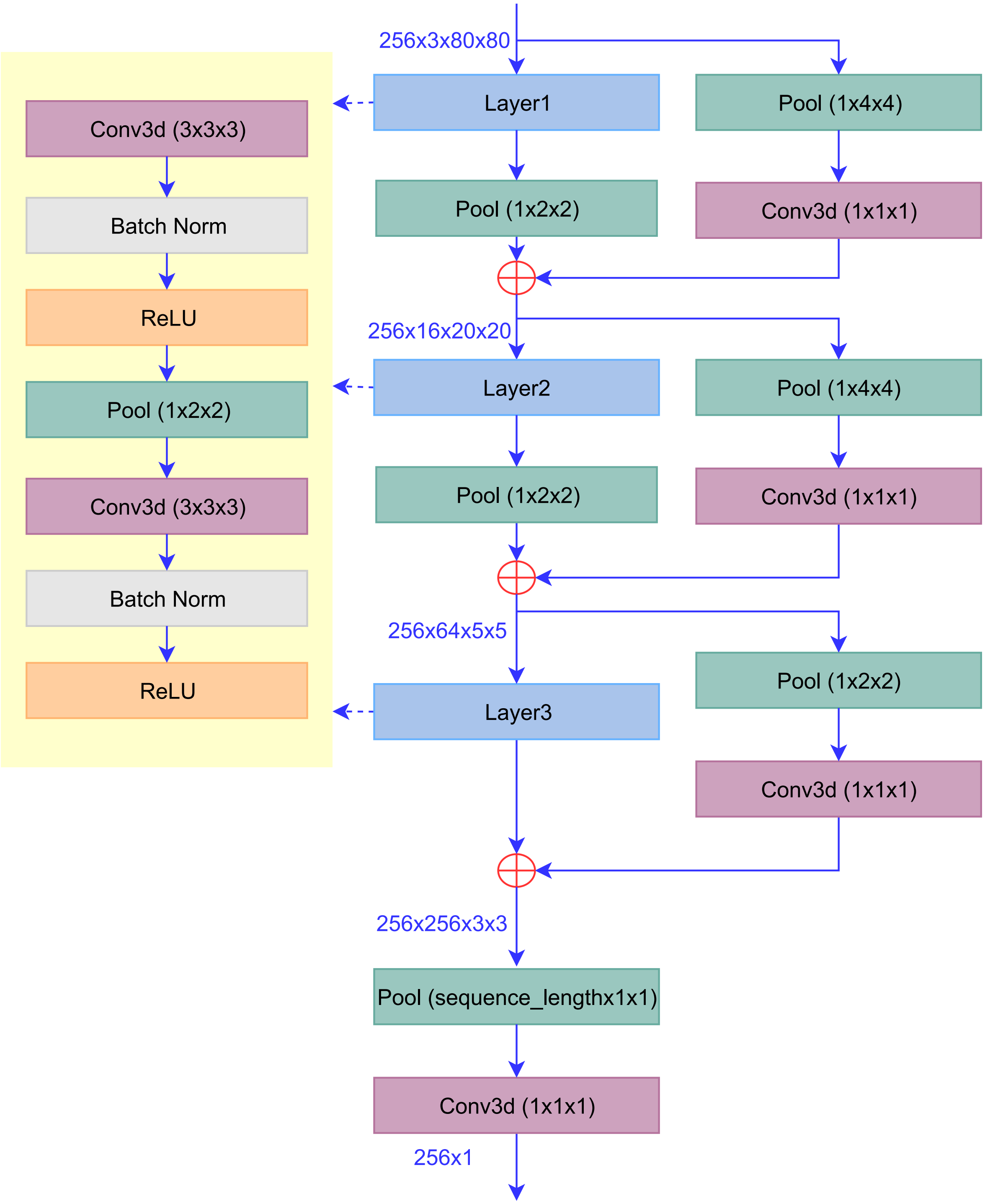

Figure 3.3: **Architecture of the rPPG estimation network.**

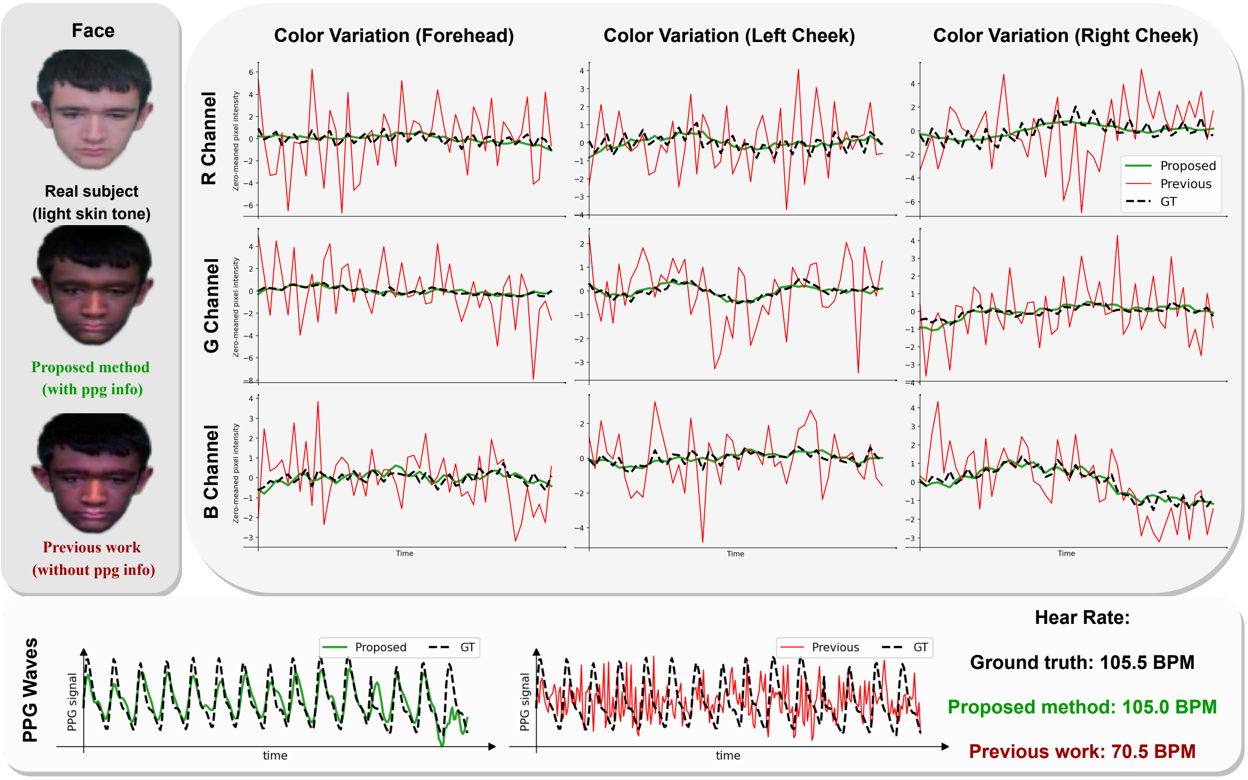

Figure 3.4: **The proposed method successfully incorporates pulsatile signals into the generated videos, while the existing work [4] only focuses on the visual appearance.** For different facial regions, frames generated by the proposed method exhibit similar pixel intensity variations as compared with frames from real videos, while the prior work shows unrealistic RGB variations. As a result, pulsatile signals can be well preserved in our method as opposed to the vanilla skin tone translation.

CHAPTER 4

Results

In this section, we demonstrate the superiority of our proposed method with empirical results on UBFC-RPPG [35] and VITAL [36] for HR estimation using various metrics: mean absolute error (MAE), root mean square error (RMSE), Pearson's correlation coefficient (PCC), and signal-to-noise ratio (SNR). As will be shown, the synthetic videos generated by our model can also improve the performance of the existing data-driven PPG estimation models with reduced bias across different skin tones.

For a qualitative evaluation, some generated frames in the UBFC-RPPG validation set are also illustrated in Figure 4.1. Our generator G can successfully produce photo-realistic videos that reflect the associated underlying blood volume changes. Estimated pulse waves from the real videos and the synthetic videos are both closely aligned with the ground truth. In the frequency domain, power spectrum of the PPG waves is also preserved with a clear peak near the gold-standard HR value.

4.1 Performance of HR estimation on UBFC-RPPG

Performance metrics of different models in the UBFC-RPPG validation set are listed in Table 4.1. We list the HR estimation accuracy of PRN trained with the proposed joint optimization pipeline (referred as PRN augmented), real samples (referred as PRN w/ Real), and synthetic samples (referred as PRN w/ Synth). The synthetic samples are generated by our generator G through translating the real samples in the UBFC-RPPG training set when the joint optimization converges. As a comparison, we also include the performance of a state-of-the-art deep learning model 3D-CNN [5] that is trained with both real and synthetic samples (referred as 3D-CNN w/ Real&Synth), just real samples (referred as 3D-CNN w/ Real), and just synthetic samples (referred as 3D-CNN w/ Synth). Performance of three traditional methods (POS [17], CHROM [16] and ICA [2]) are also provided in the table.

Notably, the proposed PRN architecture has already outperformed other rPPG estimation methods even without synthetic skin color augmentation. More specifically, the proposed PRN has around 31% improvement on MAE and around 14% improvement on RMSE over the state-of-the-art 3D-CNN using real training samples. With the synthetic augmentation, the performance of PRN can be further improved. PRN trained with augmentation achieves 9% improvement on MAE (from 0.75 BPM to 0.68 BPM) as compared with PRN trained with just real samples. This suggests that even for UBFC-RPPG dataset which is overwhelmed by subjects with light skin tones, increasing the diversity of training samples is still able to enhance the performance. This finding is consistent with the recent research [45] that demonstrates a balanced dataset can lead to optimal performance for all the groups.

The joint optimized generator G can be beneficial to other data-driven models as well. For example, we trained 3D-CNN with both real and corresponding synthetic samples from G. As compared with the 3D-CNN model trained with just real samples, the 3D-CNN model trained with both real and synthetic samples exhibits 18% improvement on MAE and 13% improvement on RMSE. This further indicates that our generator has successfully learned to produce both visually-satisfying and BVP-informative facial videos, and these synthetic videos can facilitate the learning progress of the existing data-driven rPPG estimation algorithm without conducting the joint optimization process again to adapt to another new network architecture.

Table 4.1: **Performance of HR estimation on UBFC-RPPG.** Boldface font represents the preferred results.

Method	MAE	RMSE	PCC	SNR
PRN augmented	**0.68**	**1.31**	0.86	5.76
PRN w/ Real	0.75	1.64	0.83	**7.91**
PRN w/ Synth	4.32	6.56	0.54	-1.93
3D-CNN [5] w/ Real&Synth	0.89	1.66	**0.88**	7.74
3D-CNN [5] w/ Real	1.09	1.91	0.84	7.80
3D-CNN [5] w/ Synth	0.95	1.80	0.82	3.48
POS [17]	3.69	5.31	0.75	3.07
CHROM [16]	1.84	3.40	0.77	4.84
ICA [2]	8.28	9.82	0.55	1.45

4.2 Performance of HR estimation on VITAL dataset

Cross-dataset verification can provide more visibility on the generalization capability of models trained on UBFC-RPPG videos. In order to demonstrate the improved HR estimation accuracy on the VITAL dataset using the proposed method, we report MAE, RMSE, PCC, and SNR of various models trained with real and synthetic samples in Table 4.2. Since the VITAL dataset contains testing subjects of diverse skin tones with the associated Fitzpatrick scale labels (F1-6), we group the subjects into three categories, i.e., F1-2 (light skin color), F3-4 (medium skin color), and F5-6 (dark skin color), to measure performance across different demographic groups.

PRN trained with the joint optimization pipeline exhibits significant improvement across these metrics as compared with PRN trained with just real samples. More precisely, there is 1.01 BPM reduction on MAE and 1.33 BPM reduction on RMSE for the light skin color group, 1.72 BPM reduction on MAE and 2.01 BPM reduction on RMSE for the medium skin color group, and 2.22 BPM reduction on MAE and 2.5 BPM reduction on RMSE for the dark skin color group. For all the methods, it is observed that the error of light skin tone group is generally lower than other groups. This is probably due to the melanin concentration of the light-skinned subjects is the least, and more light can be reflected to the camera. However, it should also be noted that models trained by both real and synthetic data have a relatively smaller performance difference among the three groups. For the dark skin color groups, PRN trained with synthetic data shows lower estimation errors as compared with real data, and the errors are reversed for the light skin color group. This validates the fact that data-driven rPPG estimation models are heavily impacted by the skin color distribution of training samples, and it is critical to create a diverse and balanced training set for generalizability and real-world deployment of rPPG algorithms.

To assess the cross-dataset generalization capability of synthetic videos, we also evaluate 3D-CNN trained on real and synthetic samples from UBFC-RPPG on the VITAL dataset. Similar improvement can be observed in the 3D-CNN model, where 3D-CNN trained with both real and synthetic samples outperforms the model trained on only real or only synthetic samples. This supports that our synthetic videos can accurately reflect subtle color variations due to blood volume changes and can serve as a bio-realistic augmentation to the real samples.

POS [17], CHROM [16] and ICA [2] show relatively large HR estimation errors as compared with the data-driven models, where their MAEs on the light skin color group is usually larger than 4 BPM. Their MAEs are even higher for other groups. Unlike the end-to-end rPPG estimation networks, these conventional methods usually require preprocessing steps which may diminish the subtle color changes on the face and degrade the performance. Besides, these models need to average the pixel intensities over the skin region, and this might be a sub-optimal solution since skin pixels at different facial regions can contribute differently to the pulse signals.

The cross-dataset experiment indicates that the improvement of our proposed framework is more substantial as compared with intra-dataset evaluation where all the samples are obtained within the same environment. This suggests that synthetic videos can provide more significant benefit by diversifying the training samples when there exist some data distribution shifts between real training and testing videos. This finding is also consistent with the observation for ray-tracing based augmentation method [29]. Synthetic augmentation techniques thus become particularly effective for cross-domain learning and can improve the generalization capability of HR estimation for real-world applications.

4.3 Bias mitigation

It is critical for an algorithm to have consistent performance across different demographic groups in real-world medical deployment. To quantify the performance gap for each group, we use the standard deviation of MAE and RMSE for each Fitzpatrick scale as the measurement. This measurement has also been used in some prior work [29, 4]. The standard deviation for each method in the VITAL dataset is illustrated in Figure 4.2, together with a sample portrait for each skin scale from F1 to F6. Conventional POS method exhibits large variation (MAE: 2.66 BPM, RMSE: 3.19 BPM) across different Fitzpatrick scales, while the jointly optimized PRN shows the lowest bias (MAE: 1.53 BPM, RMSE: 1.80 BPM) as compared with all the conventional methods. In contrast to PRN trained with just real samples (MAE: 2.03 BPM), the augmented training offers a 25% improvement of bias mitigation among different groups while simultaneously improving the overall performance of all the groups. This suggests our joint training framework can provide a more desired trade-off between performance and bias. For 3D-CNN, the standard deviations for MAE and RMSE are also reduced by adding the synthetic samples into the training set. We attribute this improvement to the more diverse and balanced dataset augmented by our generator.

Table 4.2: **The proposed method shows an improved HR estimation accuracy on the VITAL dataset.** Boldface font denotes the preferred results.

Method	F1-2		F3-4		F5-6		Overall	
	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓
PRN augmented	2.37	3.13	**2.95**	**3.82**	**4.39**	**5.98**	**3.04**	**4.01**
PRN w/ Real	3.38	4.46	4.67	5.83	6.61	8.48	4.60	5.88
PRN w/ Synth	4.27	6.01	4.52	6.18	5.64	8.33	4.66	6.57
3D-CNN [5] w/ Real&Synth	**2.32**	**3.11**	3.18	4.09	5.45	7.07	3.34	4.35
3D-CNN [5] w/ Real	3.31	4.64	5.86	6.78	7.07	8.89	5.19	6.44
3D-CNN [5] w/ Synth	3.88	5.23	4.68	6.07	7.81	9.88	5.04	6.56
POS [17]	4.97	6.28	5.36	6.86	7.25	9.74	5.69	7.25
CHROM [16]	6.51	8.92	5.01	6.38	7.83	14.56	6.14	8.99
ICA [2]	7.65	9.66	7.14	8.40	5.75	7.31	7.04	8.63
	F1-2		F3-4		F5-6		Overall	
	PCC↑	SNR↑	PCC↑	SNR↑	PCC↑	SNR↑	PCC↑	SNR↑
PRN augmented	0.40	3.45	0.63	**5.73**	**0.30**	**-3.38**	**0.48**	**3.02**
PRN (w/ Real)	0.36	0.32	0.50	0.03	0.08	-7.00	0.36	-1.32
PRN (w/ Synth)	0.29	-0.64	0.42	-0.44	0.11	-6.35	0.31	-1.74
3D-CNN [5] (w/ Real&Synth)	**0.42**	**3.96**	**0.65**	5.21	0.17	-4.84	0.47	2.68
3D-CNN [5] (w/ Real)	0.30	-0.61	0.48	-1.26	0.11	-8.26	0.34	-2.47
3D-CNN [5] (w/ Synth)	0.07	-2.04	0.38	-1.34	0.10	-6.38	0.21	-2.64
POS [17]	0.26	-2.22	0.42	-1.04	0.27	-5.59	0.33	-2.41
CHROM [16]	0.15	-2.14	0.46	-1.11	-0.10	-5.53	0.23	-2.40
ICA [2]	0.24	-2.06	0.32	-1.73	0.06	-5.04	0.23	-2.53

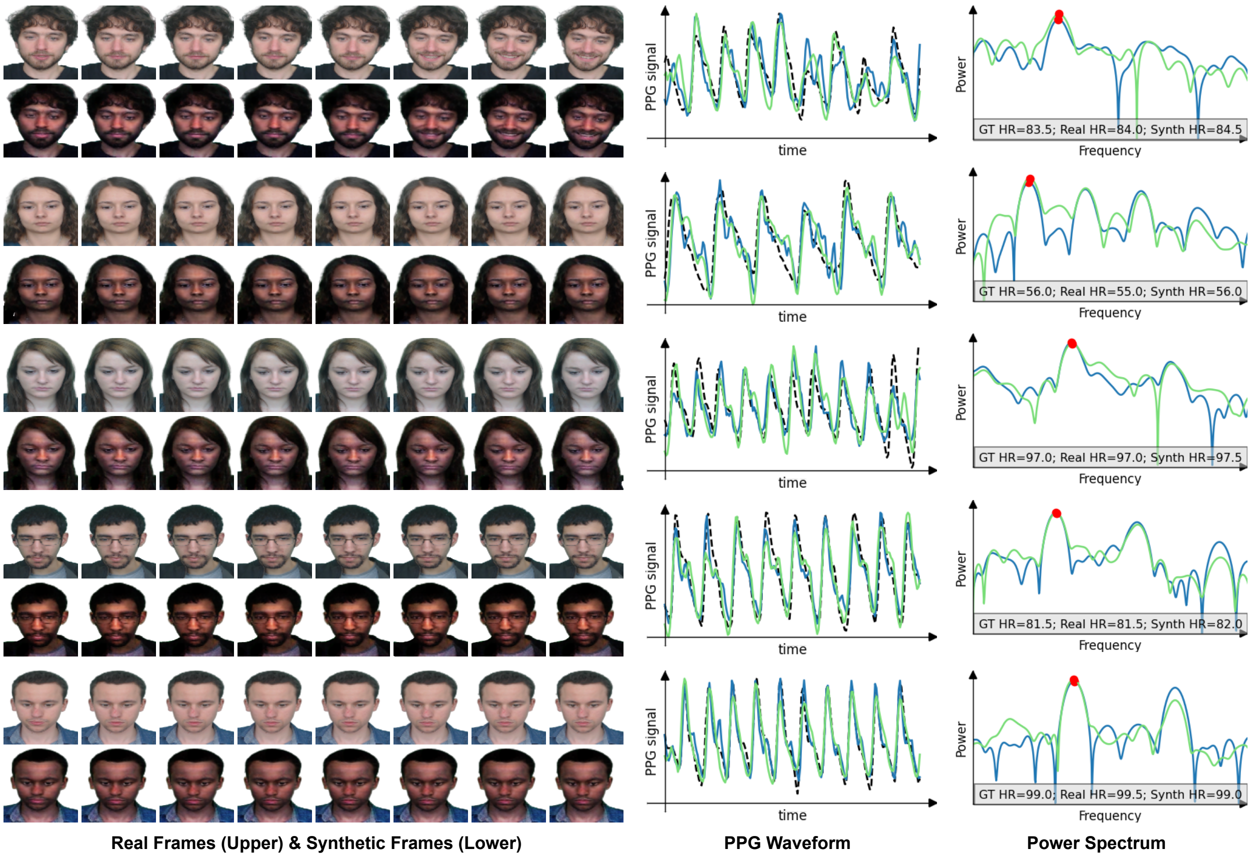

Figure 4.1: **Illustration of real frames and the corresponding synthetic frames in the UBFC-RPPG dataset.** Our proposed framework has successfully incorporated pulsatile signals when translating the skin color. The estimated pulse waves from PRN exhibit high correlation to the ground-truth waves, and the heart rates are preserved in the frequency domain.

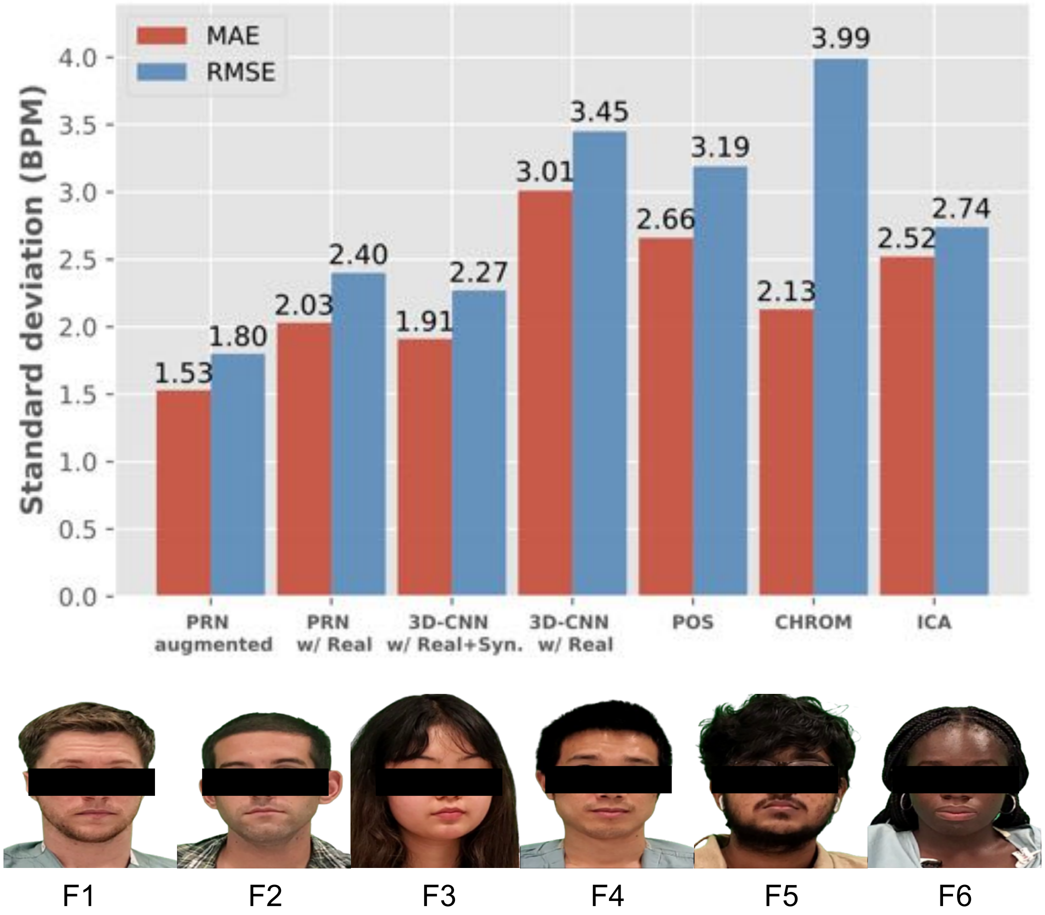

Figure 4.2: **Synthetic dark-skinned videos can help to reduce bias in HR estimation.** The augmented PRN and the 3D-CNN [5] trained on both real and synthetic videos show a reduced standard deviation on MAE and RMSE across Fitzpatrick scales F1-6 in the VITAL dataset.

CHAPTER 5

Conclusions

In this thesis, we showed a first attempt to overcome the lack of dark-skinned subjects in existing rPPG datasets using synthetic augmentation. Our method involves translating subjects with light skin in existing datasets to subjects with dark skin while keeping the rPPG signals intact. For this purpose, we used a joint optimization framework composed of a generation phase and a PPG estimation phase. During the generation phase, we froze the weights of the rPPG estimator and trained a novel 3D convolutional neural network to translate skin tone. During the PPG estimation phase, we froze the weights of the generator. Using this approach, we were able to take advantage of the strengths of both architectures to create bio-realistic synthetic subjects.

The proposed jointly optimized rPPG estimator can outperform the existing state-of-the-art methods with reduced estimation bias across different demographic groups. More specifically, PRN trained with augmentation has around 31% reduction in MAE for the dark-skinned group along with 46% improvement on bias mitigation in the VITAL dataset, as compared with 3D-CNN [5] trained with just real samples. Our generated synthetic videos maintain both photo-realistic and bio-realistic features, and can also be beneficial to other data-driven models. For example. 3D-CNN trained with both real and synthetic samples exhibits 18% improvement on MAE compared to 3D-CNN trained with only real samples.

Our current pipeline focuses on skin color translation, and therefore all the remaining factors (e.g., pulse signals, body motion, and other facial attributes) are directly copied from the original videos. To improve the generalization capabilities of rPPG networks and to maximize the benefit of synthetic augmentation, it is also critical to extend the generation framework to incorporate arbitrary facial attributes and pulse waves. Besides, it should also be noted that the generated frames are limited by a fixed resolution at 80×80. Future work may produce solutions to generate frames at arbitrary pixel resolution to fit the requirements of various subsequent rPPG estimation models without frame size interpolation.

Finally, we acknowledge that video synthesis, such as deepfakes, has raised public concerns [46]. Over half a decade, these 'fake' videos generated by deep learning have been used for face manipulation, and the malicious usage has drawn a lot of social attention. We demonstrate a positive example that these bio-realistic 'fake' videos can also be utilized for the purpose of social good. Our synthetic videos are capable of reducing both HR estimation error and bias for rPPG models and further facilitate the development of remote healthcare. We hope our framework can act as a tool to address some social issues in existing medical applications.

REFERENCES

[1] T. Tamura, “Current progress of photoplethysmography and spo2 for health monitoring,” *Biomedical engineering letters*, vol. 9, no. 1, pp. 21–36, 2019.

[2] M.-Z. Poh, D. J. McDuff, and R. W. Picard, “Advancements in noncontact, multiparameter physiological measurements using a webcam,” *IEEE transactions on biomedical engineering*, vol. 58, no. 1, pp. 7–11, 2010.

[3] Z. Yu, W. Peng, X. Li, X. Hong, and G. Zhao, “Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 151–160.

[4] S. Yucer, S. Akçay, N. Al-Moubayed, and T. P. Breckon, “Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 18–19.

[5] Y.-Y. Tsou, Y.-A. Lee, and C.-T. Hsu, “Multi-task learning for simultaneous video generation and remote photoplethysmography estimation,” in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[6] Y. Ba, Z. Wang, K. D. Karinca, O. D. Bozkurt, and A. Kadambi, “Overcoming difficulty in obtaining dark-skinned subjects for remote-ppg by synthetic augmentation,” *arXiv preprint arXiv:2106.06007*, 2021.

[7] J. Allen, “Photoplethysmography and its application in clinical physiological measurement,” *Physiological measurement*, vol. 28, no. 3, p. R1, 2007.

[8] E. Allado, M. Poussel, A. Moussu, V. Saunier, Y. Bernard, E. Albuisson, and B. Chenuel, “Innovative measurement of routine physiological variables (heart rate, respiratory rate and oxygen saturation) using a remote photoplethysmography imaging system: A prospective comparative trial protocol,” *BMJ open*, vol. 11, no. 8, p. e047896, 2021.

[9] E. M. Nowara, T. K. Marks, H. Mansour, and A. Veeraraghavan, “Sparseppg: Towards driver monitoring using camera-based vital signs estimation in near-infrared,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 1353–135 309.

[10] S. Liu, P. C. Yuen, S. Zhang, and G. Zhao, “3d mask face anti-spoofing with remote photoplethysmography,” in *European Conference on Computer Vision*. Springer, 2016, pp. 85–100.

[11] M. Lewandowska, J. Rumiński, T. Kocejko, and J. Nowak, "Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity," in *2011 federated conference on computer science and information systems (FedCSIS)*. IEEE, 2011, pp. 405–410.

[12] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation." *Optics express*, vol. 18, no. 10, pp. 10 762–10 774, 2010.

[13] S. Tulyakov, X. Alameda-Pineda, E. Ricci, L. Yin, J. F. Cohn, and N. Sebe, "Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2396–2404.

[14] W. Wang, S. Stuijk, and G. De Haan, "A novel algorithm for remote photoplethysmography: Spatial subspace rotation," *IEEE transactions on biomedical engineering*, vol. 63, no. 9, pp. 1974–1984, 2015.

[15] G. De Haan and A. Van Leest, "Improved motion robustness of remote-ppg by using the blood volume pulse signature," *Physiological measurement*, vol. 35, no. 9, p. 1913, 2014.

[16] G. De Haan and V. Jeanne, "Robust pulse rate from chrominance-based rppg," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 10, pp. 2878–2886, 2013.

[17] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, "Algorithmic principles of remote ppg," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1479–1491, 2016.

[18] D. McDuff, "Deep super resolution for recovering physiological information from videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1367–1374.

[19] A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven, "Deep ppg: Large-scale heart rate estimation with convolutional neural networks," *Sensors*, vol. 19, no. 14, p. 3079, 2019.

[20] X. Niu, S. Shan, H. Han, and X. Chen, "Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation," *IEEE Transactions on Image Processing*, vol. 29, pp. 2409–2423, 2019.

[21] W. Chen and D. McDuff, "Deepphys: Video-based physiological measurement using convolutional attention networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 349–365.

[22] X. Liu, J. Fromm, S. Patel, and D. McDuff, “Multi-task temporal shift attention networks for on-device contactless vitals measurement,” *arXiv preprint arXiv:2006.03790*, 2020.

[23] X. Niu, Z. Yu, H. Han, X. Li, S. Shan, and G. Zhao, “Video-based remote physiological measurement via cross-verified feature disentangling,” in *European Conference on Computer Vision*. Springer, 2020, pp. 295–310.

[24] H. Lu, H. Han, and S. K. Zhou, “Dual-gan: Joint bvp and noise modeling for remote physiological measurement,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 404–12 413.

[25] E. M. Nowara, D. McDuff, and A. Veeraraghavan, “The benefit of distraction: Denoising camera-based physiological measurements using inverse attention,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4955–4964.

[26] J. Buolamwini and T. Gebru, “Gender shades: Intersectional accuracy disparities in commercial gender classification,” in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81. PMLR, 23–24 Feb 2018, pp. 77–91. [Online]. Available: https://proceedings.mlr.press/v81/buolamwini18a.html

[27] A. Kadambi, “Achieving fairness in medical devices,” *Science*, vol. 372, no. 6537, pp. 30–31, 2021.

[28] E. M. Nowara, D. McDuff, and A. Veeraraghavan, “A meta-analysis of the impact of skin tone and gender on non-contact photoplethysmography measurements,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 284–285.

[29] D. McDuff, J. Hernandez, E. Wood, X. Liu, and T. Baltrusaitis, “Advancing non-contact vital sign measurement using synthetic avatars,” *arXiv preprint arXiv:2010.12949*, 2020.

[30] D. McDuff and E. Nowara, “"warm bodies": A post-processing technique for animating dynamic blood flow on photos and avatars,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, 2021.

[31] M. W. Wukitsch, M. T. Petterson, D. R. Tobler, and J. A. Pologe, “Pulse oximetry: analysis of theory, technology, and practice,” *Journal of clinical monitoring*, vol. 4, no. 4, pp. 290–301, 1988.

[32] A. Jubran, “Pulse oximetry,” *Critical care*, vol. 3, no. 2, pp. 1–7, 1999.

[33] K. K. Tremper, “Pulse oximetry,” *Chest*, vol. 95, no. 4, pp. 713–715, 1989.

[34] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, “Vggface2: A dataset for recognising faces across pose and age,” in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[35] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois, “Unsupervised skin tissue segmentation for remote photoplethysmography,” *Pattern Recognition Letters*, vol. 124, pp. 82–90, 2019.

[36] P. Chari, K. Kabra, D. Karinca, S. Lahiri, D. Srivastava, K. Kulkarni, T. Chen, M. Cannesson, L. Jalilian, and A. Kadambi, “Diverse r-ppg: Camera-based heart rate estimation for diverse subject skin-tones and scenes,” *arXiv preprint arXiv:2010.12769*, 2020.

[37] T. B. Fitzpatrick, “The validity and practicality of sun-reactive skin types i through vi,” *Archives of dermatology*, vol. 124, no. 6, pp. 869–871, 1988.

[38] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” *IEEE signal processing letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[39] I. Loshchilov and F. Hutter, “SGDR: stochastic gradient descent with warm restarts,” in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=Skq89Scxx

[40] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[41] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[42] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[43] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[44] D. McDuff and E. Blackford, “iphys: An open non-contact imaging-based physiological measurement toolbox,” in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 6521–6524.

[45] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante, “Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis,” *Proceedings of the National Academy of Sciences*, vol. 117, no. 23, pp. 12 592–12 594, 2020.

[46] Y. Mirsky and W. Lee, “The creation and detection of deepfakes: A survey,” *ACM Computing Surveys (CSUR)*, vol. 54, no. 1, pp. 1–41, 2021.