Title

Contactless smartphone camera-based heart rate estimation from facial videos for diverse subject skin tones and scenes using synthetic augmentation

Author
Karinca, Kerim Doruk

Publication Date
2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Contactless smartphone camera-based heart rate estimation from facial videos for diverse subject skin tones and scenes using synthetic augmentation

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

by

Kerim Doruk Karinca

2021

ABSTRACT OF THE THESIS

Contactless smartphone camera-based heart rate estimation
from facial videos for diverse subject skin tones and scenes
using synthetic augmentation

by

Kerim Doruk Karinca
Master of Science in Computer Science
University of California, Los Angeles, 2021
Professor Achuta Kadambi, Chair

The COVID-19 pandemic brought telemedicine applications under the spotlight and the heart rate (HR) is an important clinical vital sign in the evaluation of cardiorespiratory and hemodynamic stability. However, both deep learning- and signal processing-based systems demonstrate biased performance towards dark skin tones in remote HR measurements, as performance measures are restricted to the diversity of dataset subjects. The existing datasets MMSE-HR, AFRL, and UBFC-RPPG contain roughly 10%, 0%, and 5% dark-skinned subjects respectively, leading to poor generalization capability to unseen subjects and lead to unwanted bias toward different demographic groups. We propose a physics-driven algorithm to combat this bias and 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 biasm itigation for all the groups, as

compared with previous work trained with just real samples.

The thesis of Kerim Doruk Karinca is approved.

Omid Salehi-Abari

Majid Sarrafzadeh

Achuta Kadambi, Committee Chair

University of California, Los Angeles

2021

To my parents and sister for their unwavering support . . .

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my advisor Prof. Achuta Kadambi for his support, vision, and insight into this project. I would like thank my committee members, Prof. Majid Sarrafzadeh and Prof. Omid Abari for their time and knowledge.

I am sincerely grateful to my co-authors Pradyumna Chari, Yunhao Ba, Zhen Wang, Krish Kabra, Oyku Deniz Bozkurt, Soumyarup Lahiri, Diplav Srivastava, Kimaya Kulkarni, Tianyuan Chen, Dr. Maxime Cannesson, and Dr. Laleh Jalilian for their sustained endeavor and discipline in carrying this project forward. Thank you, Chinmay Talegaonkar, for being an awesome teammate.

Finally, I am thankful to my family for always believing in me during my journey. As the pandemic made life difficult in every aspect, I cherished their support and being home together.

PREVIOUS PUBLICATIONS

This thesis revises the following publications:

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) [BWK21].

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) [CKK20a].

VITA

2019– Masters Candidate in Computer Science, University of California, Los Angeles (UCLA).

2015–2019 B.S. in Computer Science and Engineering, University of California, Los Angeles (UCLA).

2021 Teaching Assistant, CS 130 (Software Engineering), Winter 2021, UCLA

2020 Teaching Assistant, CS 130 (Software Engineering), Fall 2020, UCLA

2020 Teaching Assistant, CS 31 (Intro to Computer Science), Spring 2020, UCLA

PUBLICATIONS

Scientific Reports F. Ghaderinezhad, H.C. Koydemir, D. Tseng, D. Karinca, K. Liang, A. Ozcan, and S. Tasoglu, "Sensing of electrolytes in urine using a miniaturized paper-based device", August 12, 2020

npj Digital Medicine K. de Haan, H.C. Koydemir, Y. Rivenson, D. Tseng, E. Van Dyne, L.S. Bakic, D. Karinca, K. Liang, M. Ilango, E. Gumustekin, and A. Ozcan, "Automated screening of sickle cells using a smartphone-based microscope and deep learning", May 22, 2020

Scientific Reports H.C. Koydemir, S. Rajpal, E. Gumustekin, D. Karinca, K. Liang, Z. Gorocs, D. Tseng, and A. Ozcan, "Smartphone-based turbidity reader", December 27, 2019

Lab on a Chip Snow, Jonathan W., Hatice Ceylan Koydemir, Doruk Kerim Karinca, Kyle Liangus, Derek Tseng, and Aydogan Ozcan. ”Rapid imaging, detection, and quantification of Nosema ceranae spores in honey bees using mobile phone-based fluorescence microscopy.” January 28, 2019

CHAPTER 1

Introduction and Background

During the pandemic, telehealth consults has increased more than 50-fold for certain groups (e.g., those with chronic diseases) [WBH21] due to the concerns that congregation of people may increase the risks of contraction. However, the remote setting deprives medical professionals of essential diagnostic tools such as vital sign monitors. The assessment of heart rate (HR) in patients with suspected COVID-19 is particularly important, as COVID-19 has been associated with pre-existing cardiovascular disease [NWH20]. Meanwhile, the traditional way of performing this assessment in person using specialized monitoring equipment at a hospital or clinic creates a risk of exposure. Given the clinical relevance of HR in triage decisions, diagnosis, prognosis, and as a criterion for transfer to higher-level medical care, there is a pressing need to develop HR sensing solutions that can facilitate mobile health (mHealth) and telemedicine-based care with remote patient monitoring in order to protect patients and healthcare workers from infectious exposure in a pandemic setting. Outside of a pandemic setting, telemedicine affords the convenience of giving and receiving medical assistance without having to leave one's home.

Presently, HR sensing solutions for telemedicine and remote patient monitoring have relied on the adoption of wearable sensors to make plethysmographic or electrocardiographic measurements [DCC19, LXY20]. Although such wearable technologies have seen major advances in the past decade [KNP13, SMT15], they still require major expenditure on production and distribution of hardware. This expense can create a barrier to adoption of mHealth technologies that affects rural and socioeconomically burdened communities [Saw20].

In contrast to wearable sensors, recent methods have proposed using camera-based hardware present on modern-day smartphones to estimate key vitals including HR. Contact-based methods, where the finger is typically placed overtop the camera module, have already seen widespread applications in major smartphones [PMH19, LWT19]. Despite such methods showing good performance, their long-term practicality for telemedicine video-conferencing visits is potentially limited as the camera module is covered during measurement. This prevents continuous monitoring of patient HR, visual well-being, and collection of other vitals such as respiratory rate and spatial blood perfusion maps.

In response to these challenges, contactless methods that use computer vision algorithms and artificial intelligence (AI) tools present an opportunity to remotely extract a blood volume pulse (BVP) signal and corresponding HR estimate from facial videos [PMH19, PMP10a, BDG13, HJ13, PFL18, LBN20, WSH16, KVS15, MSd16, WBS17, CM18, NSH20, YPL19, ANT16, SZC20, TAR16, YLZ19, TKD12, LRK11, HL14, NMV20a, LCZ14, WSH15, VSN08, BMB19, SFC18, PMP11, SND17, PWG20, NMV20b, NMM20, NMV21, TLL18, VCJ19], such as from a video camera feed over the internet. Among these contactless methods, remote photoplethysmography (r-PPG) is the most promising. r-PPG operates by looking for subtle color variations visible on the surface human skin, caused by subdermal light absorption fluctuations from changes in blood volume and content. r-PPG has the potential to respond to additional challenges such as serving as a non-invasive method over physical PPG attachments for patients with sensitive skin, such as preterm infants in neonatal intensive care units [VCJ19], as well as monitoring drivers [NFR18].

Algorithms for non-contact rPPG can be roughly classified into three categories: signal decomposition [LRK11, PMP11, PMP10b, TAR16, WSH16], model-based methods [HL14, HJ13, WBS17, SZC20], and deep learning methods [CM18, McD18, YPL19, YLZ19, NMV20a, RIS19, NSH20, SFC18].

Signal decomposition techniques based on Blind Source Separation (BSS) decompose/demix the face videos into different sources utilizing PCA [LRK11] or ICA [PMP10a, PMP11, PMP10b, TKD12]. For model-based methods, Pulse Blood Vector [HL14] utilizes the characteristic blood volume signature to weight different color channels. CHROM [HJ13] first eliminates the specular components and applies color space transforms to linearly combine the chrominance signals. POS [WBS17] modifies this by first projecting the temporally-normalized skin tone onto the plane which is orthogonal to the intensity variation direction and then linearly combine 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.

While data-driven neural networks have exhibited remarkable estimation accuracy for non-contact camera-based sensing [McD18, YPL19, RIS19, NSH20], there exist several practical constraints towards collecting large-scale data from patients for these deep learning models: (1) demographic biases in society that translate to data (e.g., Innovation happening in some countries/regions may not have access to a diverse dataset as illustrated in Figure 1.1); (2) requirement of medical-grade sensors and necessity of intrusive/semi-intrusive traditional methods for data collection; (3) patient privacy concerns (e.g., OBF dataset [LAS18] is not publicly available due to the licence issue).

Recent study shows that computer vision algorithms have been disadvantaging the underrepresented groups in some applications, such as face recognition [BG18]. Non-contact rPPG estimation is not an exception given the unbalanced and relatively small datasets in the field [NMV20b]. There are very rare subjects with dark skin tones in the existing benchmark datasets. More specifically, MMSE-HR [ZGW16], AFRL [EBM14], and UBFC-RPPG [BMB19] only contain roughly 10%, 0%, and 5% dark-skinned subjects respectively. With the training sets heavily biased towards subjects of light skin tones, the state-of-the-art data-driven rPPG models usually fail to generalize their performance to the underrepresented groups [NMV20b]. This prohibits the clinical deployment of these algorithms, since it is critical for rPPG algorithms to have consistent performance across different demographic groups in the clinical settings.

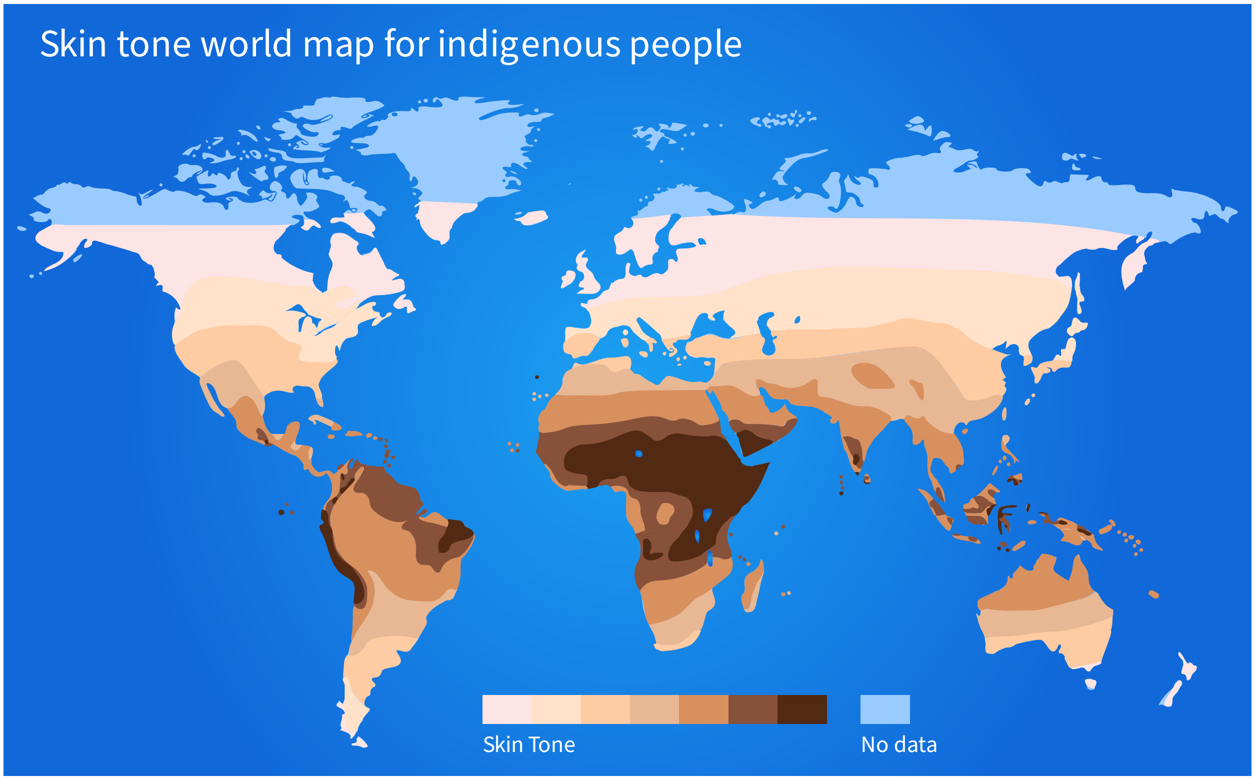


Figure 1.1: **A diverse rPPG dataset with various skin colors may not be accessible for some countries/regions.** This figure illustrates skin color reflectance map predicted from multiple regression for indigenous people. Data from Chaplin, G. [Cha04].

This thesis approaches this bias problem using two ways: First, we propose a novel r-PPG algorithm that mitigates skin-tone bias to estimate subject HR in a contactless manner using only a smartphone camera. In contrast to prior approaches, this work first establishes a theoretical framework to understand the unique physics that underlies the inconsistency in r-PPG measurement. We establish that the bias is due to imaging noise, and appropriately propose r-PPG denoising methods to alleviate performance losses. To assess the performance of the proposed method, we collect the first remote vital signs detection dataset focused on telemedicine applications that is demographically diverse. As a primary outcome measure, we qualitatively and quantitatively compare the performance of the proposed method against two popular categories of r-PPG algorithmic processing steps across varying skin tones and recording conditions. As a secondary outcome measure, we look at performance gain across skin tones and recording conditions in order to assess how the proposed method bridges the existing performance gaps.

Next, we alleviate the difficulty of recruiting patients to collect large-scale rPPG datasets in the university setting by using synthetic augmentation. Synthetic augmentation of facial videos has become an active research topic recently. McDuff et al. [MHW20] 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 [MN21]. Tsou et al. [TLH20] augment source rPPG videos with other specified rPPG signals. However, their framework is restricted to the face appearance in the original source videos and fails to produce novel videos with dark skin tones.

In contrast to these prior arts, we do a first attempt to directly augment the existing rPPG dataset by translating videos of light-skinned subjects to dark skin tones. This is difficult because the color variations due to blood volume changes are subtle, and the generation network has to be carefully designed to reflect these subtle changes while conducting skin tone translation without accessing real rPPG videos of dark-skinned subjects. However, this technique is rewarding, since it is capable of producing both photo-realistic and physiologically accurate synthetic videos in a fast manner (approximately 0.005 seconds per frame in average for our model) and can assist the development of algorithms and techniques for remote diagnostics and healthcare. In the experiment, our proposed method can reduce around 31% HR estimation error for the dark-skinned group and show 46% improvement on bias mitigation for all the groups, as compared with the existing architecture trained with just real samples.

Yucer et al. [YAA20] introduce a race translation model across various racial domains with a CycleGAN-based architecture. However, their work is not designed to incorporate pulsatile signals. As illustrated in Figure 1.2, this vanilla skin tone translation network [YAA20] merely focuses on the visual appearance, and the pulsatile signals are not preserved. To address this issue, we propose a learning framework that can augment realistic rPPG videos with dark skin tones that are of high fidelity. The framework consists of two interconnected components: (1) a generator to translate light skin tones to dark skin tones and (2) an rPPG estimator named PhysResNet (PRN) to encourage pulsatile signals within the generated videos. The generator is trained to learn both the visual appearance and the subtle color variations with respect to the underlying blood volume variations, and the rPPG network can simultaneously benefit from the generator to generalize its performance in diverse groups. We also demonstrate that our generated synthetic videos can be directly utilized to improve the performance of the state-of-the-art data-driven rPPG estimation methods with reduced bias across different skin color groups.

1.1 Contributions

Here is a summary of contributions described in this work:

1. A GAN-based deep learning architecture for synthetically generating subjects with dark skin tones that incorporates pulsatile signals.
2. A physics-based skin color-agnostic BVP extraction algorithm based on
   (a) RGB-space weighting (as opposed to BVP signal space weighting)
   (b) skin diffuse component weighting.

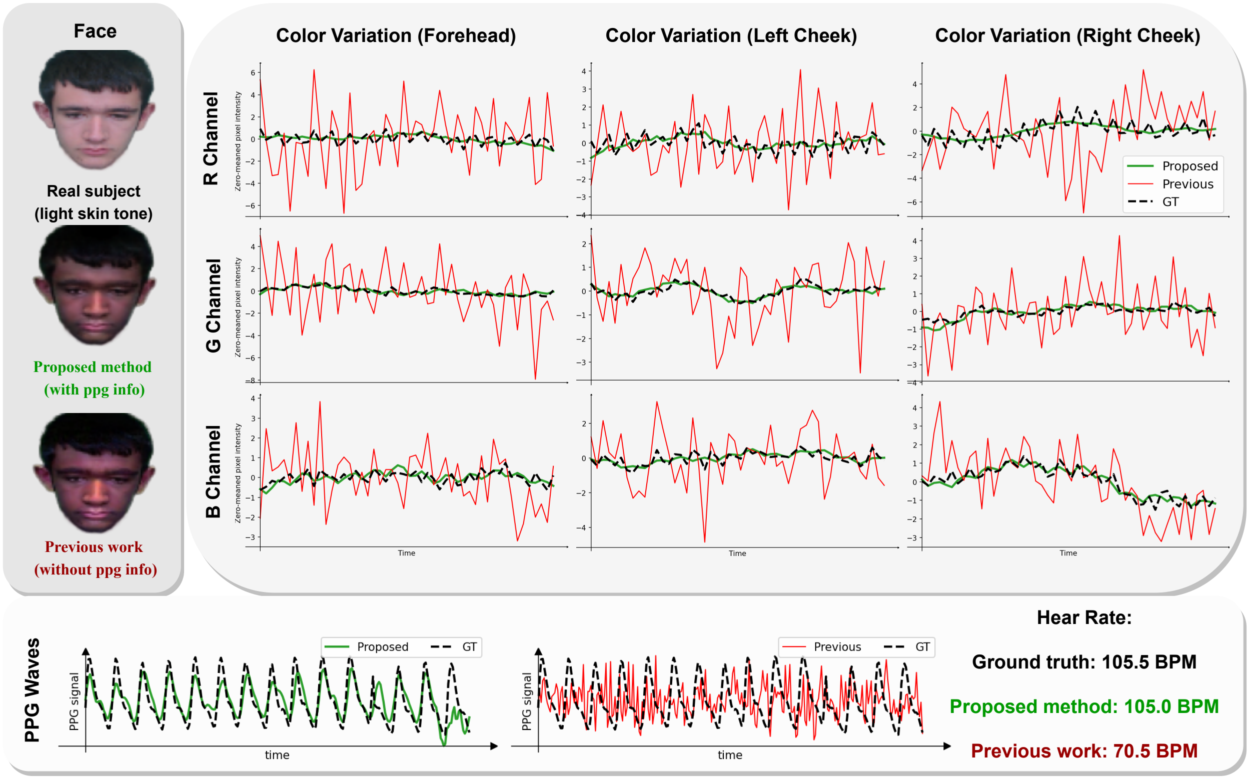


Figure 1.2: **The proposed method successfully incorporates pulsatile signal into the generated videos, while the existing work [YAA20] only focuses on the visual appearance.** For different facial regions, frames generated by the proposed method exhibit similar pixel intensity variations as compared 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 2

Materials and Methods

2.1 Effect of skin tone on PPG signal

In order to develop remote photoplethysmography (r-PPG) denoising and debiasing techniques, we use first principles to derive potential sources of bias and links to statistical noise. Previous work has developed a mathematical model for skin coloration, as a function of melanin content and blood volume fraction [AS17]. We extend this existing coloration model for the new goal of analyzing response of the PPG signal to noise ratio (SNR), in the context of skin tone variation. Let $E(\lambda)$ represent the spectral power distribution of the light source concerned. Let $S_c(\lambda)$ be the spectral sensitivity of the camera in use for color channel c. The model we follow assumes that light from the skin, as seen by the camera, emerges after two transmissions from the epidermis and one reflection from the dermis. That is, $R = T_{ep}^2(\lambda) \cdot R_d(\lambda)$. Using the expressions for $T(\lambda)$ and $R_d(\lambda)$ derived in Alotaibi et al. [AS17], we can evaluate the value of $R(\lambda)$, as a parametric function of f_{mel} (skin melanin fraction), f_{blood} (fraction of blood in the specific skin region) and f_{hg} (fraction of hemoglobin in the blood at the location). Then, the intensity captured in channel c by the camera is given by $\int_\lambda E(\lambda)S_c(\lambda)R(\lambda)d\lambda$. Subsequently, we refer to $R(\lambda)$ as $R(\lambda, f_{mel}, f_{blood}, f_{hg})$ to incorporate all the relevant parameters. To understand the SNR as a function of radiance wavelength, we identify that the PPG signal arises out of temporal variation in the value of f_{blood}. The noise involved is the noise involved in the capture process through the camera. First, we look at the signal strength, while ignoring the effect of imaging noise (analyzed in the next subsection). The PPG signal strength may be given by:

$$L(\lambda) = |R(\lambda, f_{mel}, f_{blood}^{max}, f_{hg}) - R(\lambda, f_{mel}, f_{blood}^{min}, f_{hg})|$$

That is, the strength of the PPG signal is directly related to the radiance change that occurs between the maximum and minimum blood volume fraction in the face. Then, an estimate for the average signal strength for color channel c is given by $M_c(f_{mel}, f_{hg}) = \int_\lambda E(\lambda)S_c(\lambda)L(\lambda)d\lambda$. Since we are interested in analyzing the effect of skin tone, we hold f_{hg} constant and evaluate the above signal strength metric for various reasonable values of f_{mel}. The values of all relevant physiological constants are taken to be as defined in Alotaibi et al. [AS17], i.e. taken to be the average healthy values.

Figure 2.1A. shows the signal strength curves $M_C(f_{mel})$ for different skin melanin fractions, assuming a uniform source spectral power distribution. As is intuitively expected, the signal strength reduces with increasing skin melanin content, for all color channels. The decreasing signal strength leads us to an analysis of imaging noise, which is the major noise phenomenon at play in this case.

2.2 Effect of imaging noise on PPG signal and algorithms

The goal of this subsection is to understand the relationship between imaging noise and r-PPG algorithm estimation. Imaging noise refers to the inherent noise that arises due to the image capture process in a commercial camera. This arises due to various effects related to photon arrival processes, thermal noise in electronics and the quantization noise associated with digitally capturing images [HDF10]. Overall, the entire signal to noise ratio for a pixel of a particular intensity is given by:

$$S(\lambda) = \frac{p}{\sqrt{\frac{p}{g} + (\frac{\sigma_r}{g})^2 + (\sigma_q)^2}}$$

where p is the pixel value (ranging from 0-255), g is the sensor gain (a constant for a given image) and σ_r and σ_q are camera noise parameters (also constant). Plugging in typical values

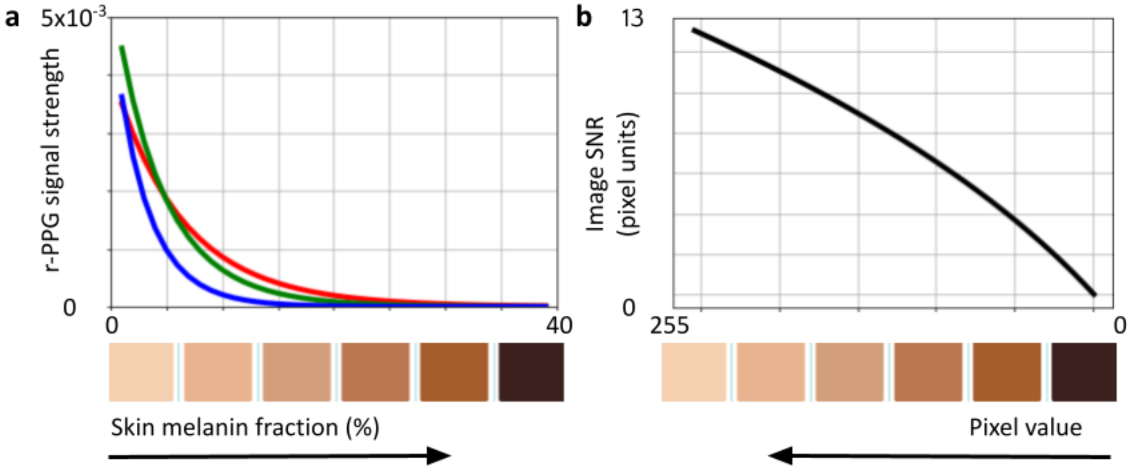


Figure 2.1: **Theoretical analysis links skin melanin fraction to signal characteristics.** **(A)** Plot of signal strength for biophysical PPG signal for Red, Green and Blue channels, varying with skin melanin fraction. **(B)** Plot of signal to noise ratio for a typical camera as a function of decreasing pixel intensity.

for the constants, Figure 2.1B. shows the trend for the SNR as a function of pixel value. The SNR is smaller for lower pixel values (corresponding to darker skin or shadowed regions) as compared to higher pixel values (corresponding to brighter skin or lit up regions). These observations, coupled with the observations from the previous subsection, allow us to make the following inferences:

1. Imaging noise creates skin tone bias (and lighting bias): The performance gap across skin tones, as well as across lighting differences, can be understood in terms of imaging noise. Darker skin regions have lower signal strength that manifest as lower pixel value changes in the video. This results in poorer SNRs. Note that this inference also holds true for shadowed regions, thereby extending this analysis towards understanding lighting bias.

2. Imaging noise and specular reflections degrade the r-PPG signal: Imaging noise, coupled with specular highlights due to lighting, are the major contributing factors to

signal degradation. The corruption due to imaging noise depends on signal intensity. The corruption due to specular highlights depends on lighting conditions- regions with strong specular highlights have relatively lower PPG signal information. Combating the highlighted biases in existing r-PPG would therefore involve a principled approach towards reduction of the above highlighted imaging noise and specular highlight removal. Note that specular highlight removal, in addition to reducing lighting related biases, also indirectly affects skin tone bias: darker skin subjects are worse affected by these interferences, since the intensity difference between the signal and the highlight is much more.

3. Denoising to be done before signal inference: This noise removal must be carried out in the combination step (shown in Figure 2.2) as opposed to after signal aggregation.

2.3 Analysis of existing methods

In order to understand the origin of the performance bias, for the first time, we theoretically analyzed the r-PPG measurement process and the role of imaging noise using biophysical first principles in the previous sections. We note three key observations: (i) Imaging noise creates skin tone bias (and lighting bias), (ii) imaging noise and specular reflections degrade the r-PPG signal, and (iii) denoising is to be done before signal inference. This sets the stage for understanding how existing algorithms improve the noise performance in the combination step. The most straightforward approach is to simply average all face pixels in a frame to arrive at time samples of the RGB signal. We refer to this as facial aggregation [PMP10a, HJ13, WSH16, WBS17, LRK11, HL14].

To improve upon this, previous approaches have sought to modify this averaging process. We describe the best performing result amongst these on the Vital-sign Imaging for Telemedicine AppLications (VITAL) dataset we collected (We explain in Section 2.7 how the data collection study is designed). The face is gridded into smaller rectangular regions.

Pixels within each region are averaged to arrive at individual time series for each region. Each of these gridded temporal signals is passed through the inference step, to obtain the corresponding blood volume signal estimate. Approaches use measures such as SNR at peak frequency of this blood volume signal to characterize the 'goodness' of each signal [HJ13, PFL18, LBN20, KVS15, BMB19], with higher weights being assigned for better signals. In this paper we use the two harmonic SNR estimate, which was found to be more robust. That is, for a signal s (frequency domain S) with a HR p, the SNR at the HR frequency is given by:

$$SNR = \frac{\int_{p-w}^{p+w} |S(f)|^2 df + \int_{2(p-w)}^{2(p+w)} |S(f)|^2 df}{\int_{-\infty}^{\infty} |S(f)|^2 df - \int_{p-w}^{p+w} |S(f)|^2 df - \int_{2(p-w)}^{2(p+w)} |S(f)|^2 df}$$

where w is the peak window size for estimation (for this work's experiments, we use $w = 0.1$Hz). This resultant signal is passed to the HR step. We call this method SNR weighting [PFL18, LBN20, KVS15, BMB19]. Finally, these weights are used to average the blood volume signals together.

A few key issues arise with the SNR weighting method. Firstly, we empirically observe that the weight maps from previous methods (based on region-based SNR estimates) have the tendency to be sparse, especially for darker skin tones. Therefore, the expected improvements due to weighted averaging are lost to noise corruption for darker skin tone subjects since much lesser signal is being aggregated. This poorer denoising for darker skin subjects results in worse SNRs, thereby degrading performance. Datasets on which these previous methods were tested were not as diverse across skin tones: these performance caveats were therefore missed. Secondly, the previous method of SNR weighting may also fall prey to specular highlights. With these, the signal contains no information of the pulsatile signal, which gets buried in the light from the source. This is a considerable factor when looking at scene conditions, such as camera angle, lighting direction, lighting color and intensity, as well as skin tone. Previous weighting approaches do not explicitly take this into account and use the gridded weighting method to implicitly combat these highlights. However, since the nature of this gridding itself degrades for darker skin tones, we observe that specular effects must be directly addressed. Finally, the SNR weighting performs denoising after signal inference, as opposed to before. Given that the inference method (CHROM) is non-linear, such a weighting regime may not be the most optimal.

2.4 Novel algorithmic modifications

Having identified the reasons for poor performance of existing methods, we propose novelties to be incorporated in the combination step, that look to achieve a performance gain in a manner that is fair across skin tones. We focus our novelties to this step since the origin of the performance bias is the image SNR. In order to move towards debiasing, it is critical that major modifications are applied during combination, so that the effect of noise during inference is minimized. This also allows for the proposed modifications to be applied independent of the inference algorithm, thereby making the modifications more generally applicable.

Specifically, we propose two major novelties: (i) weighting in RGB space, rather than blood volume signal space and (ii) skin diffuse component weighting.

1. RGB-space weighting: Existing spatial averaging methods estimate weights for each grid region, based on the blood volume signal quality [PFL18, LBN20, KVS15, BMB19]. Instead of using these estimated weights to average the blood volume signals, as done in previous methods, we propose using these weights to average in RGB space. As a result, we obtain one consolidated SNR weighted RGB signal, which is again passed through the inference step to obtain the final blood volume signal. The motivation for this can be understood in the context of noise. Averaging the RGB signal results in a less noisy signal passing through the inference step, enabling the inference method to provide better estimates, as compared to when noisier signals are passed through the method, to be averaged later. If the inference method is non-linear (such as CHROM),

a pre-weighting would lead to additional noise performance gain.

2. Skin diffuse component weighting: An image can be split into two constituent components: the diffuse component, that arises out of transmission and reflection through the skin, and a specular component, that arises from mirror-like surface reflections. Since the diffuse component contains the signal of interest for us, we utilize gridded diffuse components as additional weights. For each frame, the diffuse component is estimated [YWA10]. It is then gridded and averaged across the grid dimensions and time, in order to arrive at weights for each grid element. The diffuse weights play two key roles in improving bias in performance as well as overall performance: first, they can remove specular affected regions from the average explicitly. Second, they combat the sparsity issue observed in traditional SNR weights, since the diffuse component is continuous and non-sparse. The SNR weights and the novel diffuse weights are multiplied together and renormalized to arrive at the final spatial weights for the gridded video.

The overall pipeline, therefore, involves using the novel weights together, to arrive at efficiently weighted RGB signals. These are averaged together and passed through the estimation step and HR step. This pipeline is visually highlighted as such in Figure 2.2.

2.5 Bio-realistic skin tone translation

The second line of contribution of this thesis is a generative deep learning model for creating subjects with dark skin tones that contain rPPG data. 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 2.3. We next describe the proposed 3D convolutional video generator, the rPPG estimation network, and our joint optimization scheme.

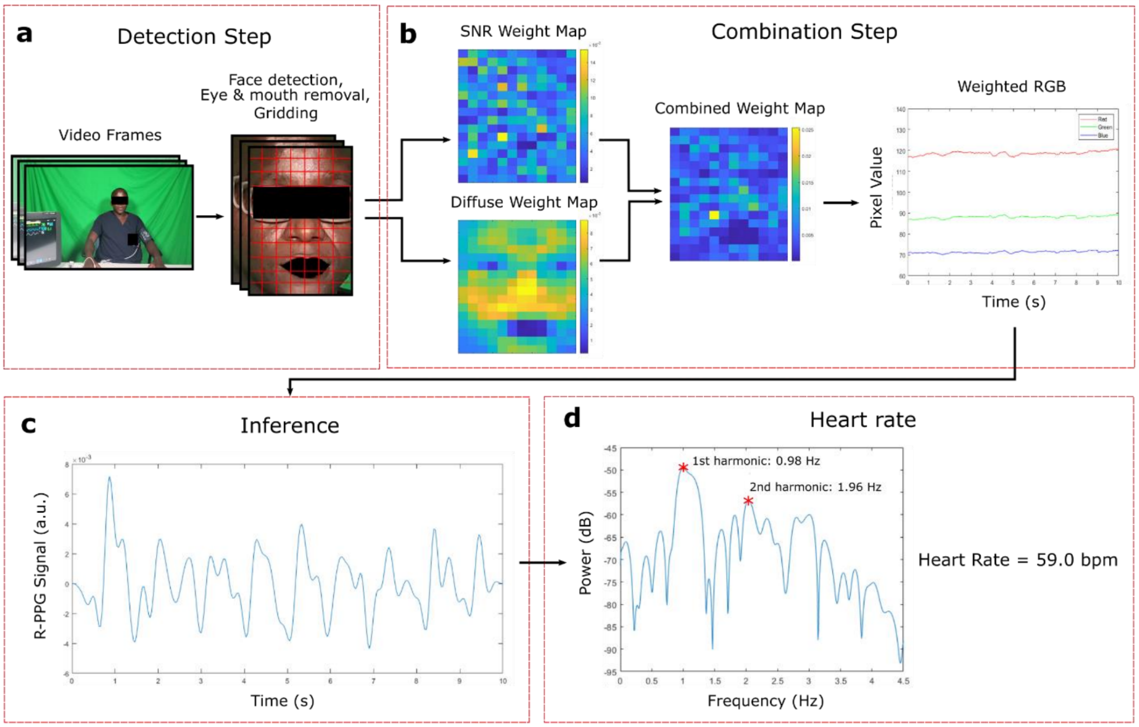


Figure 2.2: **The proposed heart rate estimation algorithm consists of four steps.** The proposed novelty in the combination step of the pipeline incorporates skin diffuse information weighting, in addition to SNR weighting in RGB space, to achieve robust r-PPG performance across skin tones. Written consent was obtained from the subject for using their image in the publication.

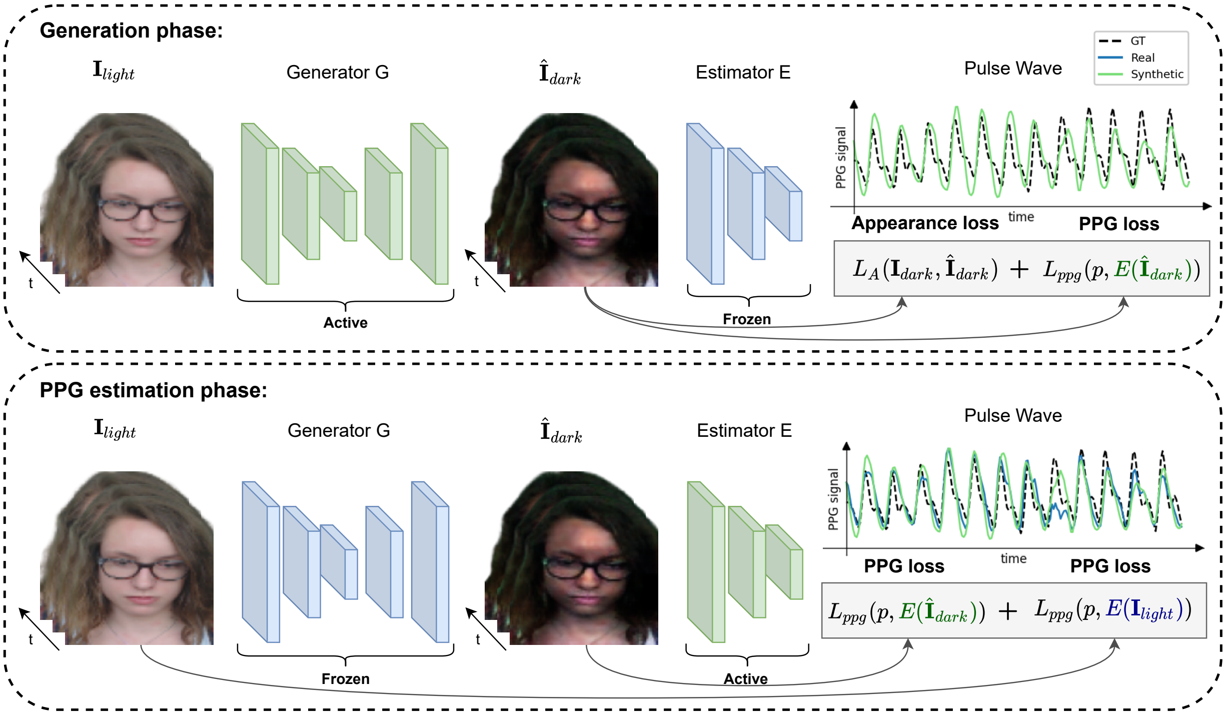
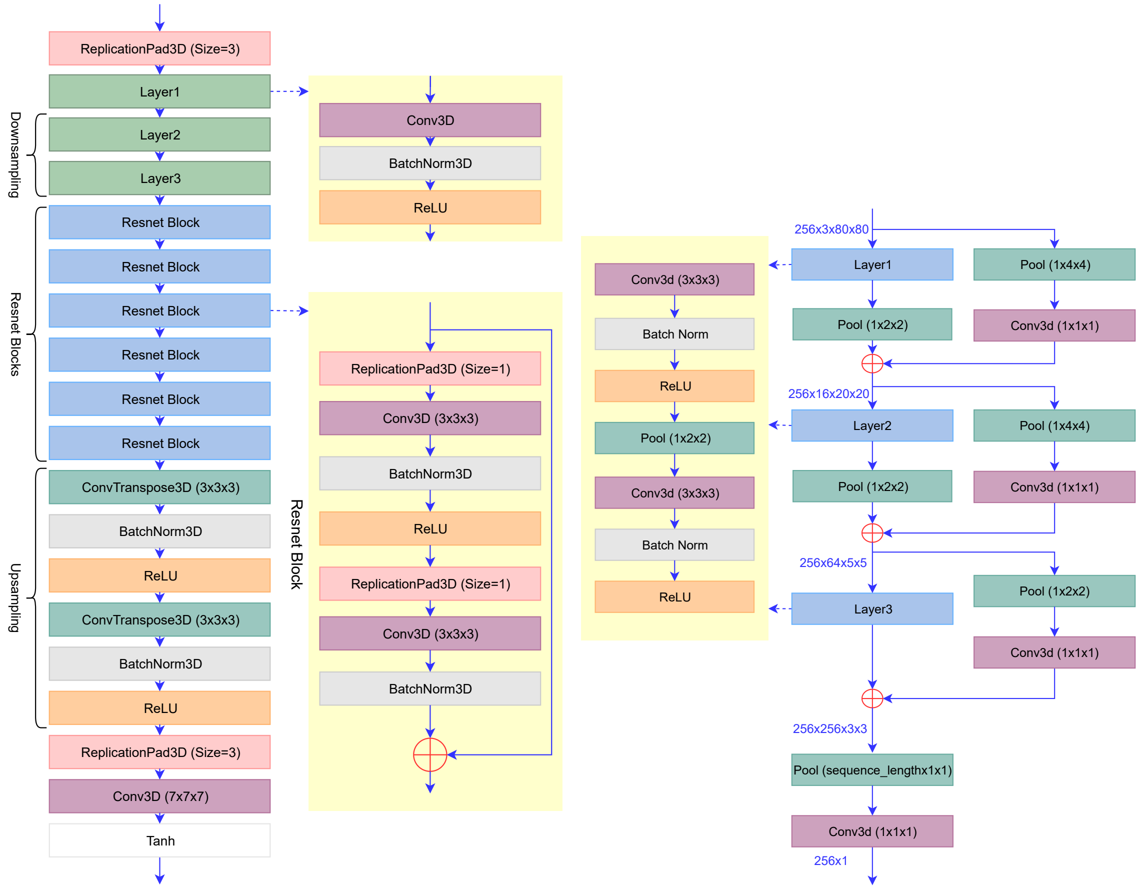


Figure 2.3: **Illustration of the proposed joint optimization framework.** Our framework is capable of translating light-skinned facial videos to dark skin tones while maintaining the original pulsatile signals. With a two-phase weight updating scheme, the rPPG estimation network can benefit from the synthetic dark-skinned videos and gradually learn to inference on dark-skinned subjects without accessing real facial videos with dark skin tones.

2.5.1 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 block diagram of the generator is illustrated in Figure 2.4A. We adapt the architecture from the image translation networks in CycleGAN [ZPI17] and make the operation of 2D convolution to 3D convolution. The generator model consists of an encoder (several



(A) **Architecture of the generation network.**

(B) **Architecture of the rPPG estimation network.**

Figure 2.4: **Block diagrams of the networks used.**

convolutional layers), a transformer (6 ResNet Blocks), and a decoder (several convolutional layers).

The diagram of the rPPG estimation network is shown in Figure 2.4B. 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 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 translated frames do not exist, we use a race transfer model [YAA20] pretrained on VGGFace2 [CSX18] to generate the pseudo target frames $\mathbf{I}_{dark}$. Specifically, the generator *Caucasian-to-African* in [YAA20] 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 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 [YAA20] are generated in a frame-by-frame manner without temporal pulse correspondence along the time dimension. In the joint optimization part, we describe how to further incorporate the pulsatile signals presented in the original videos $\mathbf{I}_{light}$ into the generated frames.

2.5.2 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.

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 $p \in \mathbb{R}^T$:

$$L_{ppg}(p, \hat{p}) = 1 - \frac{T\sum_{i=1}^{T} p_i\hat{p}_i - \sum_{i=1}^{T} p_i \sum_{i=1}^{T} \hat{p}_i}{\sqrt{\left(T\sum_{i=1}^{T} p_i^2 - \left(\sum_{i=1}^{T} p_i\right)^2\right)\left(T\sum_{i=1}^{T} (\hat{p}_i)^2 - \left(\sum_{i=1}^{T} \hat{p}_i\right)^2\right)}} \tag{2.1}$$

This negative Pearson correlation loss has shown to be more effective as compared with the point-wise MSE loss in the previous work [YPL19]. 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 part, we detail how to further incorporate the synthetic subjects into the training process.

To evaluate the HR estimation against the gold-standard ground truth, we use the following four metrics: mean absolute error (MAE), root mean square error (RMSE), Pearson's correlation coefficient (PCC), and signal-to-noise ratio (SNR):

$$\text{MAE} = \frac{\sum_{i=1}^{N} |\text{HR}_i - \text{HR}_i|}{N}, \tag{2.2}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (\text{HR}_i - \text{HR}_i)^2}{N}}, \tag{2.3}$$

$$\text{PCC} = \frac{T\sum_{i=1}^{T} p_i\hat{p}_i - \sum_{i=1}^{T} p_i \sum_{i=1}^{T} \hat{p}_i}{\sqrt{\left(T\sum_{i=1}^{T} p_i^2 - \left(\sum_{i=1}^{T} p_i\right)^2\right)\left(T\sum_{i=1}^{T} (\hat{p}_i)^2 - \left(\sum_{i=1}^{T} \hat{p}_i\right)^2\right)}}, \tag{2.4}$$

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

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.

2.5.3 Joint optimization

The generator trained with L1 loss in the previous part 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 deploy a joint optimization mechanism to incorporate pulsatile signals into the synthetic videos and improve the generalizability of the rPPG estimator simultaneously.

We deploy 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 2.3. 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:

$$L_G(\mathbf{I}_{light}, p) = L_{ppg}(p, E(\hat{\mathbf{I}}_{dark})) + \lambda * L_A(\mathbf{I}_{dark}, \hat{\mathbf{I}}_{dark}), \tag{2.6}$$

$$L_A(\mathbf{I}_{dark}, \hat{\mathbf{I}}_{dark}) = \frac{1}{\sum_i z_i} \sum_i z_i |\mathbf{I}_{dark_i} - \hat{\mathbf{I}}_{dark_i}|, \tag{2.7}$$

$$z_i = \begin{cases} 0 & \text{if } |\mathbf{I}_{dark_i} - \hat{\mathbf{I}}_{dark_i}| < \epsilon \\ 1 & \text{otherwise} \end{cases}, \tag{2.8}$$

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 [YAA20] 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:

$$L_E(\mathbf{I}_{light}, \hat{\mathbf{I}}_{dark}), p) = L_{ppg}(p, E(\hat{\mathbf{I}}_{dark})) + L_{ppg}(p, E(\mathbf{I}_{light})). \tag{2.9}$$

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 a frame sequence from a real light-skinned subject to dark skin tone 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.

2.6 Generating synthetic subjects with dark skin tones

We demonstrate the superiority of our proposed method with empirical results on UBFC-RPPG [BMB19] and VITAL [CKK20b] 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). The synthetic videos generated by our model can also further improve the performance of the existing data-driven PPG estimation model with reduced bias across different skin tones.

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. As the majority of subjects in current benchmark datasets are light-skinned, we translate the light-skinned subjects in UBFC dataset to dark-skinned subjects while retaining the pulse signals. Some generated frames in the UBFC-RPPG validation set are illustrated in Figure 2.5. 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.

2.7 Human study design

We also collect the Vital-sign Imaging for Telemedicine AppLications (VITAL) dataset, which features facial videos and vital sign data of a set of subjects with diverse skin tones. The human study protocol was approved by the UCLA Institutional Review Board (IRB#20-001025-AM-00001), and participants provided written informed consent to take part in the study. Figure 2.6 shows the data collection setup. Each subject is made to sit on a height-adjustable chair, in the field of view of two cell-phone cameras (with different view angles): one camera (Samsung Galaxy S10) is perfectly front-on, while the other (Samsung Galaxy A51) is directly in front of the face, at a dip (lower) of 15 degrees. The front-on camera is placed approximately 130 cm from the subject, and the lower camera at a dip is approximately 90 cm from the subject. The height of the chair is chosen so that the subject is centered in the front-on frame. The controlled lights are set up on either side of the front-on camera, with a baseline of 100 centimeters between them.

We record subjects using these cameras under four different scene conditions: (1) controlled lighting at 5600K ("cool" lighting) with the subject remaining stationary, (2) controlled lighting at 3200K ("warm" lighting) with the subject remaining stationary, (3) ambient room lighting (distributed white LED lighting) with the subject remaining stationary, and (4) ambient room lighting with the subject speaking. Controlled lighting is enabled by a pair of professional bi410 color LED photography lights (Neewer Bi-Color 480 LED). The controlled lighting recording conditions were enabled with the room lights off, allowing for fine-tuned control over the illumination spectral properties. As incorporating controlled lighting only enables a front-facing illumination angle, two recording conditions in ambient room lighting were captured where the subject was lit more completely from several angles. The final recording condition involved variations in the subject, including talking, natural head movements, and facial expressions. Each scene recording session lasts for 2 minutes, for a total of 16 minutes of video footage across 8 videos.

During data collection, volunteers are fitted with standard anesthesiology cardiopulmonary monitors: pulse oximeter (Red DCI, Masimo), blood pressure cuff (Comfort Care, Philips), and 5-lead electrocardiogram (Philips IntelliVue). To collect vital sign data, we utilize the Philips IntelliVue MX800 patient monitor to perform real time monitoring of four vital signs- HR, respiratory rate, oxygen saturation, and non-invasive continuous blood pressure- of which three waveforms are collected (ECG, PPG and respiration). We use the open source tool VSCapture [KH13] to collect data onto a computer using the MX800's local area network communication protocol. The MX800's estimated numeric values for the vital signs are sampled every 1 second, while the waveforms are sampled at variable frequencies. The ECG signal is sampled between 400-600 Hz, the PPG signal between 100-150 Hz and the respiration between 40-60 Hz. Continuous non-invasive blood pressure estimates occur when the blood pressure cuff is activated, which is approximately once every 30 seconds. A total of 60 subjects participated in the study. Due to data collection errors or corrupted video, 6 subjects are excluded from the experiment. Therefore, the final VITAL dataset consists of 432 videos (∼864 minutes) of 54 subjects and their vital signs.

2.8 The r-PPG pipeline

There are four components to a typical r-PPG pipeline: (a) detection, which identifies facial regions of interest in the video frame, (b) combination, which condenses the information from regions of interest into a RGB time series signal, (c) signal inference, which uses the time series signal to estimate the pulse volume waveform, and (d) HR estimation, which estimates the HR from the pulse volume signal. This is visually described in Figure 2.2.

The video is first passed through a neural network-based face detector [ZZL16], in order to identify the face region in the frame. Using feature point detectors [KS14], the eye and mouth regions are identified and explicitly removed from the videos (since these regions do not contribute to the pulsatile signal). This is the detection step. The next steps, namely combination, inference and HR step, are carried out for smaller video-windows of 10 seconds length with an overlap of 5 seconds.

For each video frame, the skin pixels are combined to get one RGB sample for that time instance (the methods for this combination vary across papers and is the crux of this work's novelty). Across all frames, after this combination, we obtain a time series RGB signal. This is the combination step.

These RGB signals are then put through an existing signal inference technique. In this paper, we use the CHROM algorithm [HJ13] due to its versatility, as well as its easy access from openly available code [MB19]. The output obtained from this step results in a pulsatile waveform estimate for each window. This is the inference step.

The obtained pulsatile waveform is then processed to arrive at the final HR. This is the heart rate step. We first filter the waveform using a Butterworth bandpass filter with pass band frequencies of [0.7, 3.5] Hz. The power spectral density (PSD) is then computed. Temporal frequency artifacts were empirically observed in the original video as a result of aggressive compression, likely due to the unchanging green background. These erroneous peaks were appropriately removed. Next, the five highest peaks in the PSD are chosen. The peak with the highest combined fundamental and second harmonic power is chosen as the one corresponding to the HR. The final HR for the video is estimated as the average of the HR estimates for each 10 second window.

2.9 Statistical analysis

To quantitatively assess the performance of the proposed bias mitigation algorithm, the following statistical metrics are used: (i) Mean Absolute Error (MAE), (ii) Standard deviation of the error (SE) and the correlation coefficient (r) between the estimated r-PPG average HR and the ground truth PPG average HR for the entire video. We also employ Bland-Altman (B&A) plots (62) to compare differences in the benchmark and proposed method's

HR estimates and MX800 PPG HR measurements. These plots are labelled with the corresponding mean difference (m) that shows the systematic bias, and the limits of agreement (LoA) within which 95% of the differences are expected to lie, estimated as $LoA = m \pm 1.96\sigma$, assuming a normal distribution.

2.10 Deployment Cost Projections for Telemedicine

In order to calculate the estimated average deployment cost for the cheapest existing method (finger pulse oximeters), we use the following methodology:

1. We identify the estimated user base numbers for telemedicine in the US using the numbers from [Kat20] and extend these up to 2027 using the compound annual growth rate (CAGR) of 15.8% as suggested in [pol20].

2. We make the conservative assumption that all members of a given family would be active users of telemedicine services. Therefore, an estimate of the number of families using telemedicine services is given by,

   No.of families = (Number of Telemedicine users)/(Avg. Family size in the US)

   We use the average family size of 3.15 from [cps20].

3. Assuming that one pulse oximeter costs $20 (as observed from a survey of available units in the market), and assuming conservatively that one pulse oximeter has to be deployed per family, the cost of deployment is given by, Cost of deployment $=$ No. of families $\cdot$ cost per pulse oximeter unit.

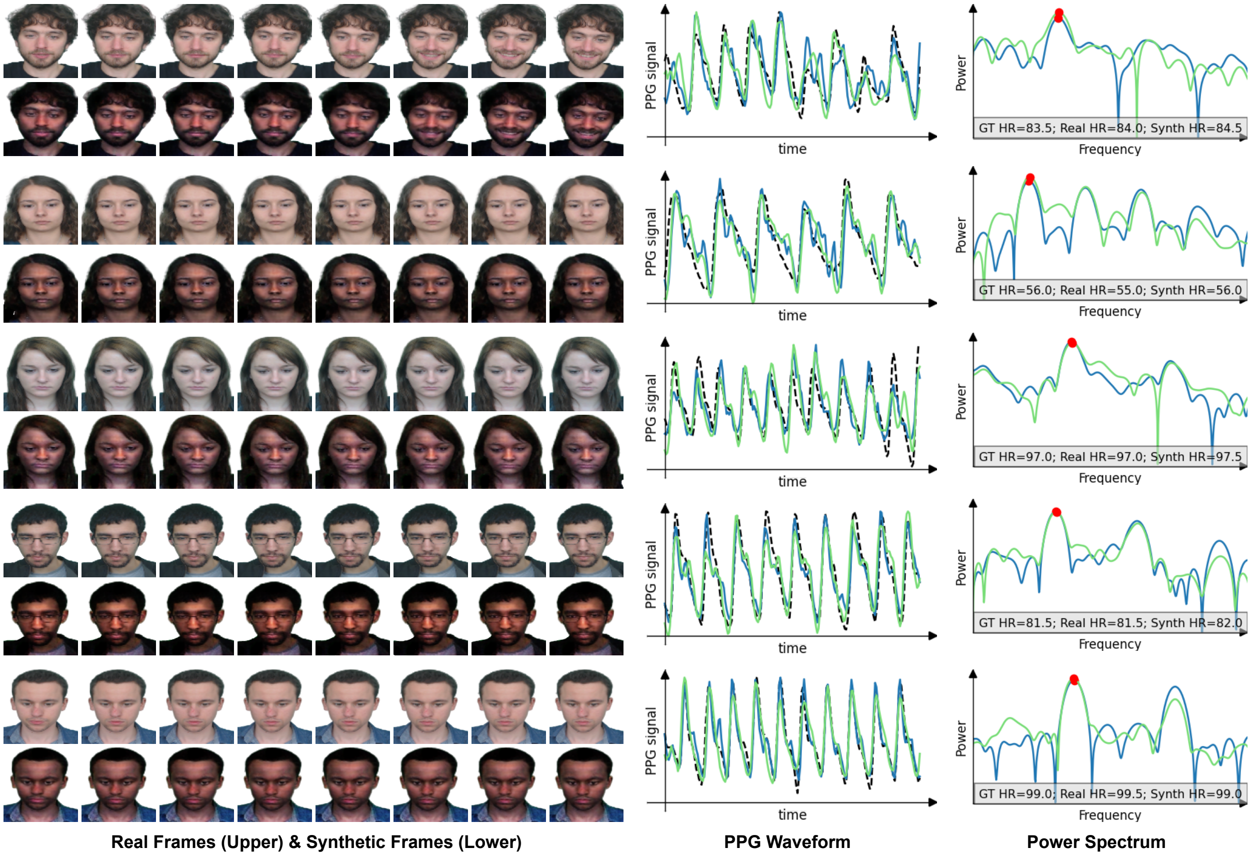


Figure 2.5: **Illustration of real frames and the corresponding synthetic frames in the UBFC-RPPG dataset.** Our proposed framework has successfully incorporated pulsatile signals when translating the skin color. The estimated pulse waves from PRN exhibit high correlation to the ground-truth waves, and the heart rates are preserved in the frequency domain.

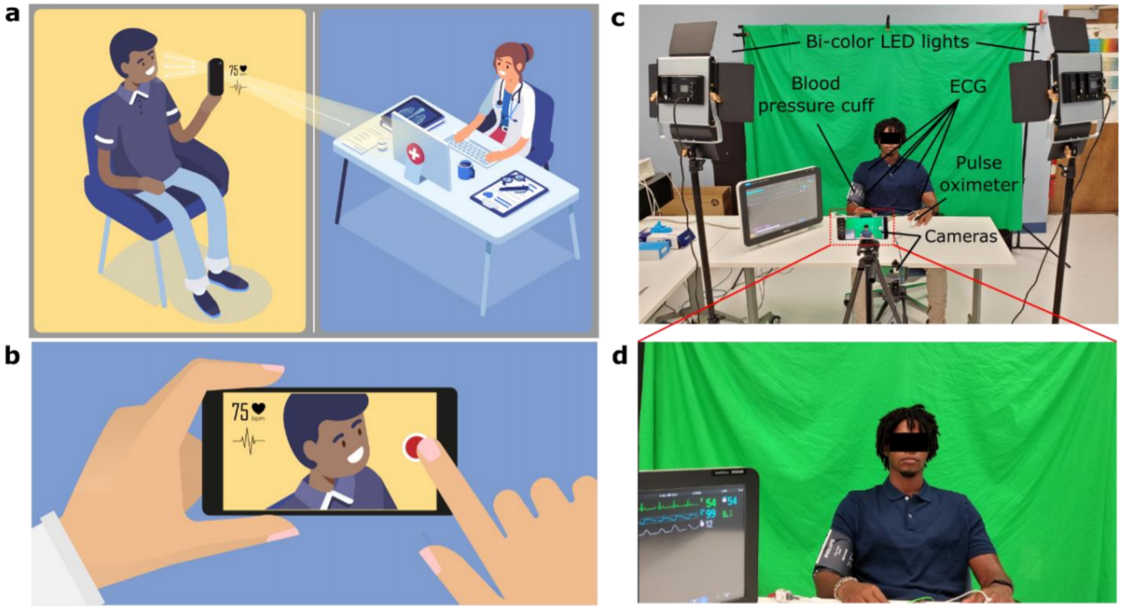


Figure 2.6: **Constructing a diverse remote vital sign monitoring dataset with a focus on telemedicine applications.** **(A)** Cartoon schematic depicting the telemedicine application for the proposed camera-based heart rate estimation. **(B)** Telemedicine video797 conferencing applications can be integrated with a software toolkit to display patient BVP and HR. **(C)** Experimental setup employed during the construction of the VITAL dataset. Two bi-color LEDs are used for controlled illumination of the subject, and laboratory tube LEDs are used for ambient illumination. The Philips IntelliVue MX800 patient monitor is utilized for ground truth vital sign monitoring. Two smartphone cameras at differing viewing angles capture video of the subject. **(D)** Example frame from video captured by the smartphone camera. The subject wears a blood pressure cuff, 5-ECG leads, and a finger pulse oximeter, which is connected to the MX800 unit. Written consent was obtained from the subject for using their image in the publication.

CHAPTER 3

Results and Discussion

3.1 The VITAL dataset

The VITAL dataset is used to validate the performance of camera-based vital sign detectors. The focus of this dataset is to represent diversity in factors that are relevant to telemedicine setups, including: (i) smartphone deployment, (ii) camera view angle, (iii) recording condition (lighting variation and talking), and (iv) patient demographic diversity. We address each of these aspects individually:

1. Smartphone deployment: The ubiquity of smartphones globally has led to the development of patient portals, many of which can be accessed via smartphone applications that can be downloaded by patients [MYS12, Ven14, BWT11]. Such applications have been used for hosting telemedicine appointments. A deployable remote HR estimation solution with a focus on telemedicine must be able to work efficiently on smartphone cameras by considering factors including video compression [YPL19, NMV21, NM19], and algorithmic complexity. Moreover, the solution must achieve success independent of camera type. Hence, the VITAL dataset uses different smartphone cameras for each view angle. The use of more than one smartphone imager inspires the development of algorithms that can scale to a variety of device-agnostic telemedicine conditions.

2. Camera view angle: In a telemedicine setting, there can also be a variety of camera angles that the algorithm must work on. In order to facilitate this verification, the

VITAL dataset consists of two camera view angles for all the videos of each subject (as seen in Figure 2.6).

3. Recording condition: Another essential factor involves testing algorithms across a range of recording conditions, to promote the development of algorithms that can operate in the "wild". The dataset consists of four recording conditions: (1) controlled lighting at 5600K ("cool" lighting) with the subject remaining stationary, (2) controlled lighting at 3200K ("warm" lighting) with the subject remaining stationary, (3) ambient room lighting- distributed white lighting- with the subject remaining stationary, and (4) ambient room lighting with the subject speaking. Additionally, a green screen backdrop is kept to potentially enable digital modification of background scenery.

4. Patient demographic diversity: The VITAL dataset consists of 54 subjects spread across skin tone, age, gender, race, and ethnic backgrounds. Subject characteristics (gender, age, height, weight, body mass index, race, and ethnicity) are summarized in Table 3.1 using mean (SD), median (IQR), or frequency (%), unless otherwise noted. For the purpose of this study, we split the subjects into three skin tone categories based on the Fitzpatrick (FP) skin type scale [Fit88]: light, consisting of skin tones in the FP 1 and 2 scales, medium, consisting of skin tones in the FP 3 and 4 scales, and dark, consisting of skin tones in the FP 5 and 6 scales. This aggregation allows for more relevant trends, since any two consecutive FP scale categories are reasonably close.

3.2 Benchmark methods and techniques

To benchmark the performance of the proposed method, we compare the proposed method against previous remote HR estimation algorithms. All methods and techniques used are outlined in detail in the Materials and Methods section. We choose the CHROM [HJ13] signal extraction method due to its versatility and open availability of code [MB19]. We

Total number of participants in study: 54

Physical Demographics	Mean	Median
Age (years)	34 (std 10)	34 (IQR 26-41)
Height (cm)	173 (std 9)	175 (IQR 164-180)
Weight (kg)	72 (std 16)	72 (IQR 56-81)
Body Mass Index (kg m^{-2})	24 (std 5)	23 (IQR 21-26)

Sex	# of participants
Male	33 (61%)
Female	21 (39%)

Race	# of participants
White	27 (50%)
Asian	16 (29%)
Black or African American	8 (15%)
Native Hawaiian or other Pacific Islander	0 (0%)
American Indian or Alaska Native	2 (4%)
Unknown	1 (2%)

Ethnicity	# of participants
Hispanic/Latino	7 (13%)
non-Hispanic/-Latino	47 (87%)

Skin Type	# of participants
Light	19 (35%)
Medium	24 (45%)
Dark	11 (20%)

Table 3.1: Demographic characteristics of volunteers in the VITAL dataset.

compare with the two most common categories of algorithmic processing steps, which we refer to as facial aggregation (c.f. [PMP10a, HJ13, WSH16, WBS17, LRK11, HL14]) and SNR weighting (c.f. [PFL18, LBN20, KVS15, BMB19]). We believe that these two processing steps regimes encapsulate the major processing philosophies used in existing r-PPG methods.

To ensure a fair comparison with the benchmark methods, we implement identical testing conditions across techniques. For each method, the input video is passed through the same face detection algorithm (convolutional neural network-based detector [ZZL16]), following which the eyes and mouth are cropped out using facial feature points [KS14]. Some methods also use skin segmentation algorithms [WSH15, TLL18, VCJ19], but we empirically found this to perform slightly worse on the VITAL dataset. We also use a consistent HR selection technique for each method.

3.3 Algorithmic results summary

Table 3.1 insert describes the distribution of subjects across various demographic metrics. Overall, remote HR estimation performance was compared across 54 subjects, across 4 scene conditions and 2 camera angles, resulting in a total of 432 videos with an average length of 2 minutes. HR estimation is carried out for windows of duration 10 seconds, with an overlap of 5 seconds. The overall HR for the subject is then estimated by averaging these window-estimated HR. Table 3.2 contains a performance summary across all statistical metrics employed- namely the Mean Absolute Error (MAE), Standard deviation of the error (SE) and the correlation coefficient (r) (details in the Methods section). In addition, Table 3.3 contains information about improvement in the Mean Absolute Error (MAE) metric for the SNR weighting and proposed methods, over the facial aggregation method.

The experiments highlight that the proposed method: (i) shows an overall performance increase on the skin tone diverse VITAL dataset, (ii) shows debiased performance gain across skin tones, which is shown to not be the case with existing methods, (iii) is robust

Preprocessing	Statistic	Skin type			Recording condition				Camera viewpoint		Overall
		Light	Medium	Dark	3200K	5600K	Room lighting	Talking in room lighting	Front	Lower	
Facial aggregation	MAE (bpm)	3.94	4.14	6.20	3.91	4.24	3.99	5.82	5.24	3.74	4.49
	SE (bpm)	5.60	5.75	7.31	5.83	5.79	**5.48**	7.40	6.75	5.54	6.18
	r	0.78	0.81	0.44	0.74	0.77	**0.77**	0.60	0.68	0.80	0.74
Previous method (SNR weighting)	MAE (bpm)	3.86	4.45	7.24	4.42	4.60	4.36	5.87	5.38	4.24	4.81
	SE (bpm)	**5.07**	6.23	8.00	6.31	6.36	5.93	7.31	6.91	5.17	6.52
	r	**0.84**	0.76	0.30	0.69	0.69	0.71	0.61	0.66	0.74	0.70
Proposed method (Novel weighting)	MAE (bpm)	**3.74**	**3.83**	**5.65**	**3.57**	**3.87**	**3.99**	**5.25**	**4.89**	**3.44**	**4.17**
	SE (bpm)	5.13	**5.34**	**6.79**	**5.29**	**5.47**	5.61	**6.51**	**6.30**	**5.17**	**5.76**
	r	0.83	**0.85**	**0.52**	**0.80**	**0.80**	0.75	**0.72**	**0.75**	**0.83**	**0.79**

Table 3.2: **Performance of proposed method as compared to benchmark methods.** The table shows the performance comparison of the proposed method and the chosen benchmark methods. The metrics shown are Mean Absolute Error (MAE), Standard Deviation of Error (SE) and correlation coefficient (r). Both MAE and SE are given in beats per minute. The best results across methods have been bolded for each skin type, recording condition, and camera viewpoint.

to recording conditions such as lighting and talking, and (iv) is robust to camera placement with respect to the subject. Secondary observations include the nature of bias in existing methods, the accuracy under best performing conditions, and the nature of performance differentials across scene conditions and camera angles.

3.4 Algorithmic performance

Figure 3.1 shows the qualitative performance of the proposed method in comparison to the ground truth PPG and benchmark methods. The estimated pulse volume signal for the

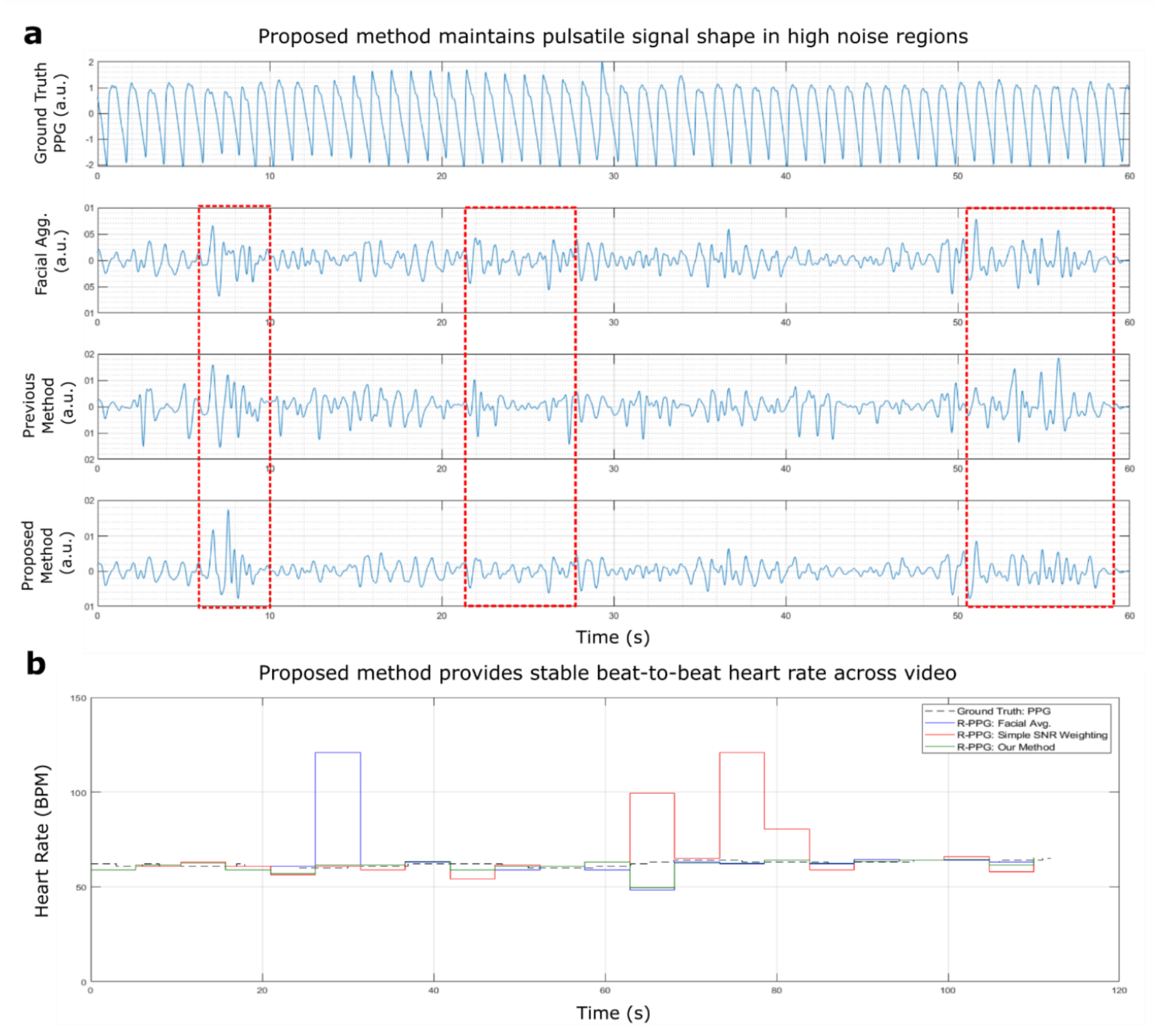


Figure 3.1: **The proposed method qualitatively recovers the pulsatile signal in a more stable manner compared to prior methods.** **(A)** Example pulsatile waveforms, including the ground truth PPG, facial aggregation r-PPG, previous method's (SNR weighting) r-PPG, and the proposed method's (novel weighting) r-PPG waveform (labelled from top to bottom). The dashed red windows show noisy regions where the r-PPG signal deteriorates. The proposed method maintains pulsatile signal shape, with pulsatile peaks seen more clearly and distinctly. **(B)** Beat-to-beat heart rate numerics over time are captured by the proposed method in a more stable manner, consistently staying within 5 bpm of the ground truth PPG.

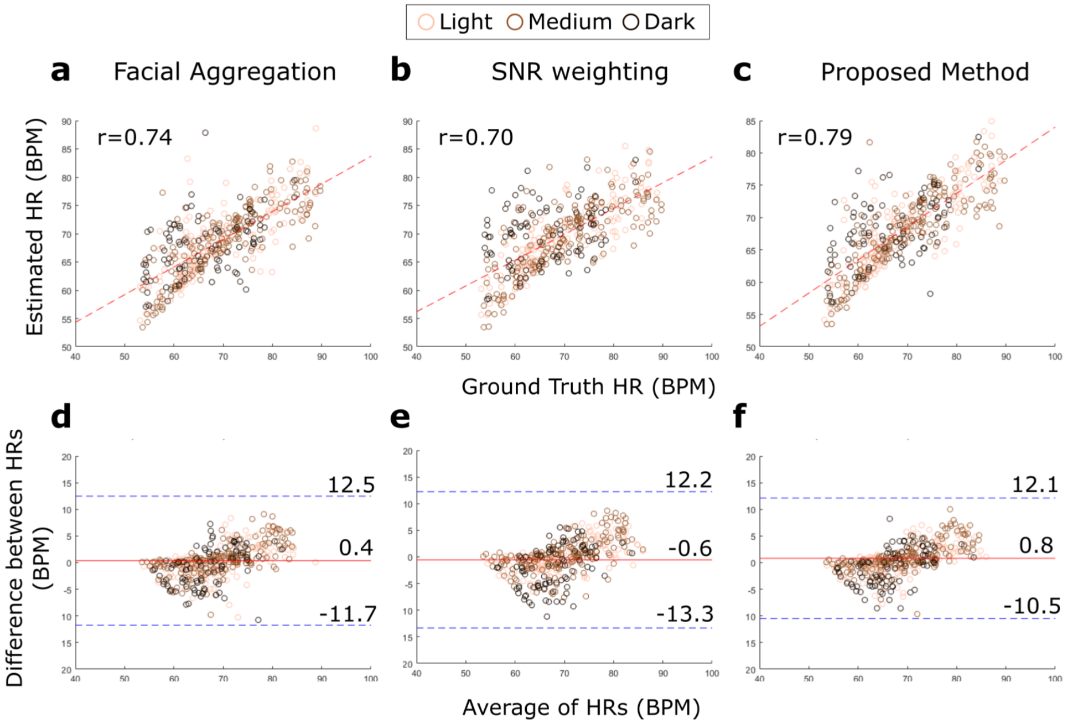


Figure 3.2: Scatter and Bland Altman plots for benchmark and proposed heart rate recovery methods. The label shows a marker for each skin type. **(A-C)** Scatter plots for different methods. The proposed method shows strong correlation with respect to ground truth heart rates from the Philips IntelliVue MX800, denoted by the Pearson Correlation Coefficient r. **(D-F)** Bland-Altman plots for different methods. The bias (m) is shown by the middle solid red line, and the limits of agreement ($LoA = m \pm 1.96\sigma$) by the upper and lower dotted blue lines.

Preprocessing	Skin type			Recording condition				Camera viewpoint		Overall
	Light	Medium	Dark	3200K	5600K	Room lighting	Talking in room lighting	Front	Lower	
Previous method (SNR weighting)	0.08	-0.31	-1.04	-0.51	-0.36	-0.36	-0.05	-0.14	-0.50	-0.32
Proposed method (Novel Weighting)	0.20	0.31	0.55	0.35	0.37	0.00	0.58	0.35	0.30	0.32

Table 3.3: **Performance improvement with respect to facial aggregation benchmark of the previous (SNR weighting) method and the proposed method.** The metric shown is the Mean Absolute Error (MAE) improvement, in beats per minute.

proposed method is found to visually contain peaks at the same frequency as the ground truth PPG signal. In some instances, the dicrotic notch is also present, although less prominent. Particularly noisy regions of the video are highlighted by the dashed red lines in Figure 3.1A. In these time windows, the proposed method is found to visually recover peaks more distinctly with less high frequency artifacts in comparison to the benchmark r-PPG methods. Additionally, Figure 3.1B shows the beat-to-beat time evolution of the HR estimate, across the 10 second windows. Both the estimates from the ground truth signal and the output of the proposed method follow similar trends, consistently staying within 5 beats per minute (bpm) of each other. However, because of the high frequency noise artifacts in existing methods, the estimated HR suffers from large errors in localized regions, worsening the overall HR estimate across the 2-minute video. Such qualitative improvements also translate quantitatively, where the proposed method shows a sub-6 beats per minute MAE for all skin tones, with an overall average MAE of 4.17 beats per minute.

Figure 3.2 shows the corresponding scatter and B&A plots for the proposed method, facial aggregation method and SNR weighting methods across all collected videos. The proposed method shows a higher correlation (r = 0.79) in comparison to the benchmark facial aggregation (r = 0.74) and SNR weighting (r = 0.70) methods. The B&A plots show a less than 1 bpm bias across all methods. The proposed method shows the best limits of agreement with almost all videos falling within 10 bpm of the ground truth HR.

3.4.1 Skin tone performance

For all three methods, performance degrades from light to dark skin. The facial aggregation approach obtains a MAE of 3.94, 4.14, and 6.20 bpm for light, medium and dark skin tone subjects, resulting in an overall average performance of 4.49 bpm. When comparing the facial aggregation results to the SNR weighting approach, a MAE improvement of +0.08 bpm is obtained for light skin tones, and a successive MAE degradation of -0.31 bpm and -1.04 bpm is obtained for medium and dark skin tones respectively. Hence, on a skin tone-diverse dataset such as VITAL, this leads to a comparative decrease in overall performance of -0.32 bpm. In contrast, the proposed method shows significant improvement across all skin tones when compared to the facial aggregation method, with a MAE improvement of +0.20 bpm, +0.31 bpm and +0.55 bpm obtained for light, medium and dark skin tones respectively. Consequently, the overall performance of the proposed method on the VITAL dataset improves by +0.32 bpm.

Figure 3.3A-C highlights the high correlation between the proposed method's r-PPG HR estimates and ground truth PPG HR for light (r = 0.83) and medium skin tones (r = 0.85), and moderate correlation for dark skin tones (r = 0.52). The B&A plots in Figure 3.3D-F show a less than 2 bpm bias across all skin tones, and that all the proposed method's r-PPG HR estimates are mostly within 10 bpm of the ground truth. These correlation metrics are an improvement to the benchmark methods of facial aggregation and SNR weighting. Figure 3.4 and Figure 3.5 show the corresponding scatter and B&A plots for the facial aggregation

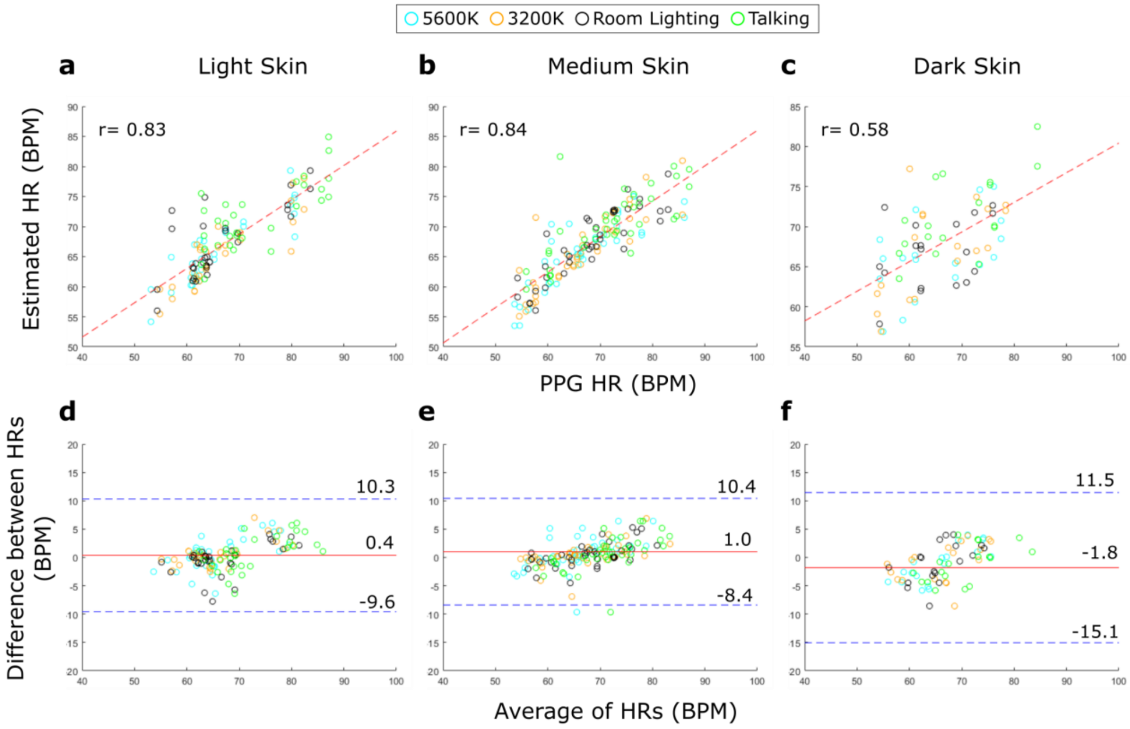


Figure 3.3: **Scatter and Bland Altman plots for proposed method, varied across skin tone categories.** The label shows a marker for each video recording condition. **(A-C)** Scatter plots for different skin types highlighting the correlation between estimated and ground truth heart rate, denoted by the Pearson Correlation Coefficient r. **(D-F)** Bland-Altman plots for different skin types. The bias (m) is shown by the middle solid red line, and the limits of agreement ($LoA = m \pm 1.96\sigma$) by the upper and lower dotted blue lines.

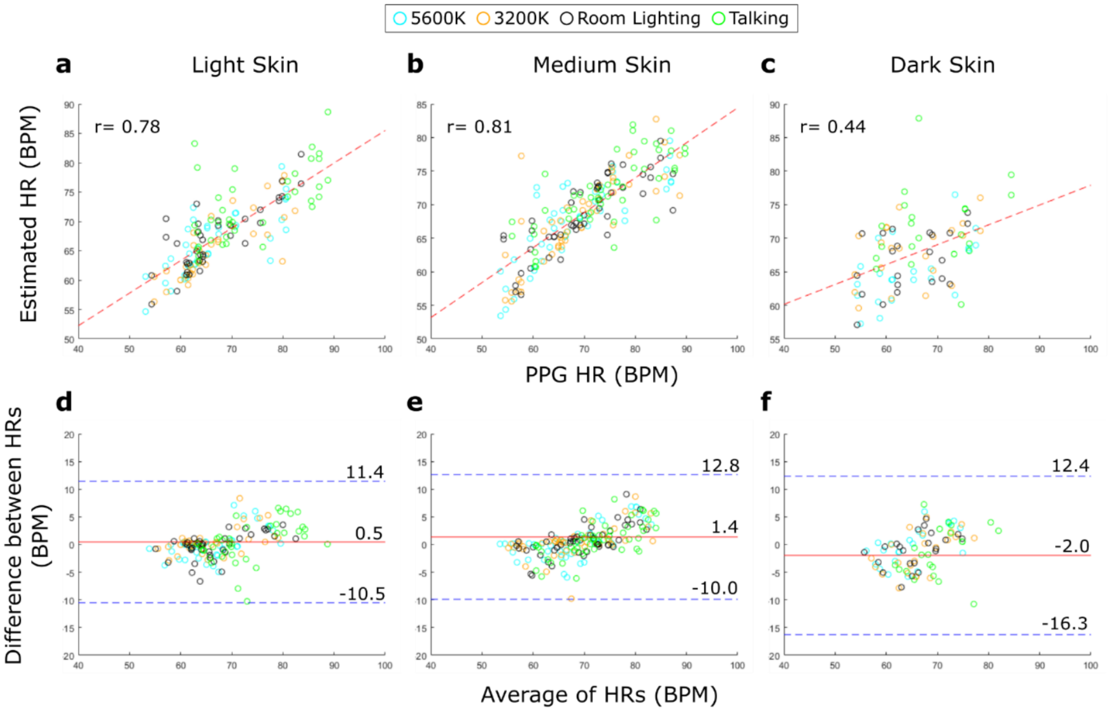


Figure 3.4: **Scatter and Bland Altman plots for facial aggregation method, varied across skin tone categories.** The label shows a marker for each video recording condition. **(A-C)** Scatter plots for different skin types highlighting the correlation between estimated and ground truth heart rate, denoted by the Pearson Correlation Coefficient r. **(D-F)** Bland-Altman plots for different skin types. The bias (m) is shown by the middle solid red line, and the limits of agreement ($LoA = m \pm 1.96\sigma$) by the upper and lower dotted blue lines.

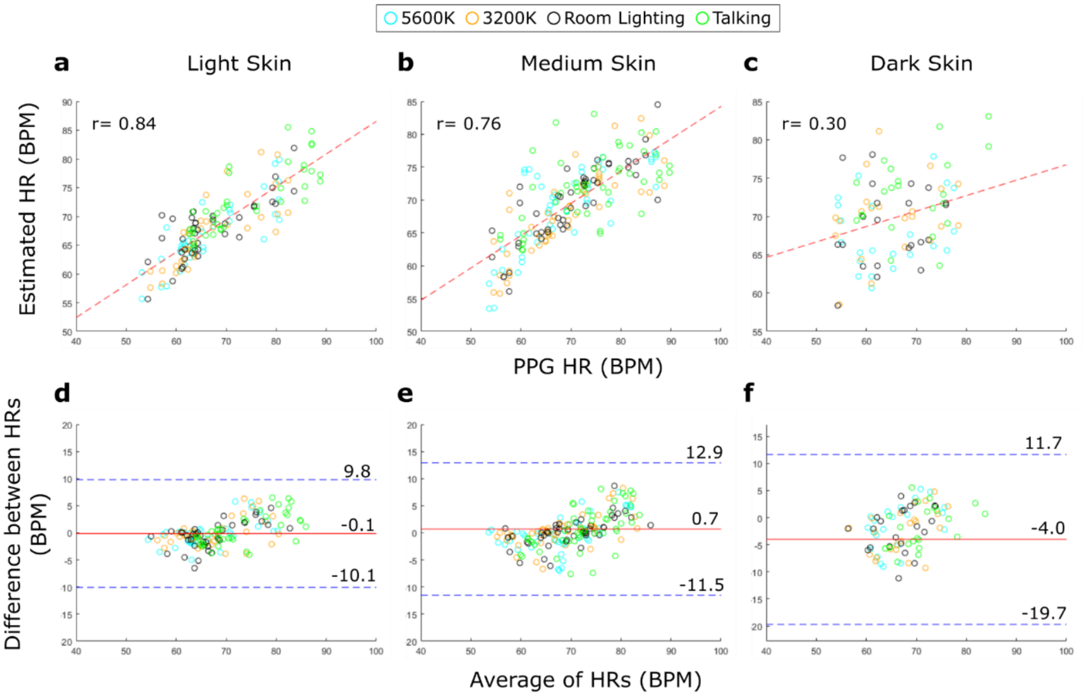


Figure 3.5: **Scatter and Bland Altman plots for SNR weighting method, varied across skin tone categories.** The label shows a marker for each video recording condition. **(A-C)** Scatter plots for different skin types highlighting the correlation between estimated and ground truth heart rate, denoted by the Pearson Correlation Coefficient r. **(D-F)** Bland-Altman plots for different skin types. The bias (m) is shown by the middle solid red line, and the limits of agreement ($LoA = m \pm 1.96\sigma$) by the upper and lower dotted blue lines.

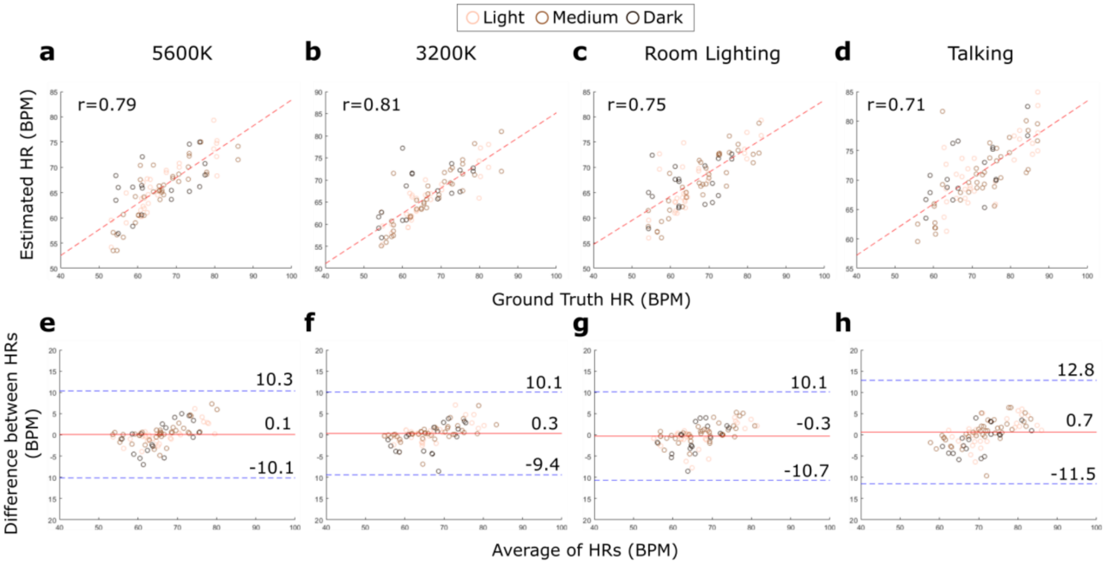


Figure 3.6: **Scatter and Bland Altman plots for proposed method, varied across scene condition categories.** The label shows a marker for each skin type. **(A-D)** Scatter plots for different recording conditions highlighting the correlation between estimated and ground truth heart rate, denoted by the Pearson Correlation Coefficient r. **(E-H)**. Bland-Altman plots for different recording conditions. The bias (m) is shown by the middle solid red line, and the limits of agreement ($LoA = m \pm 1.96\sigma$) by the upper and lower dotted blue lines.

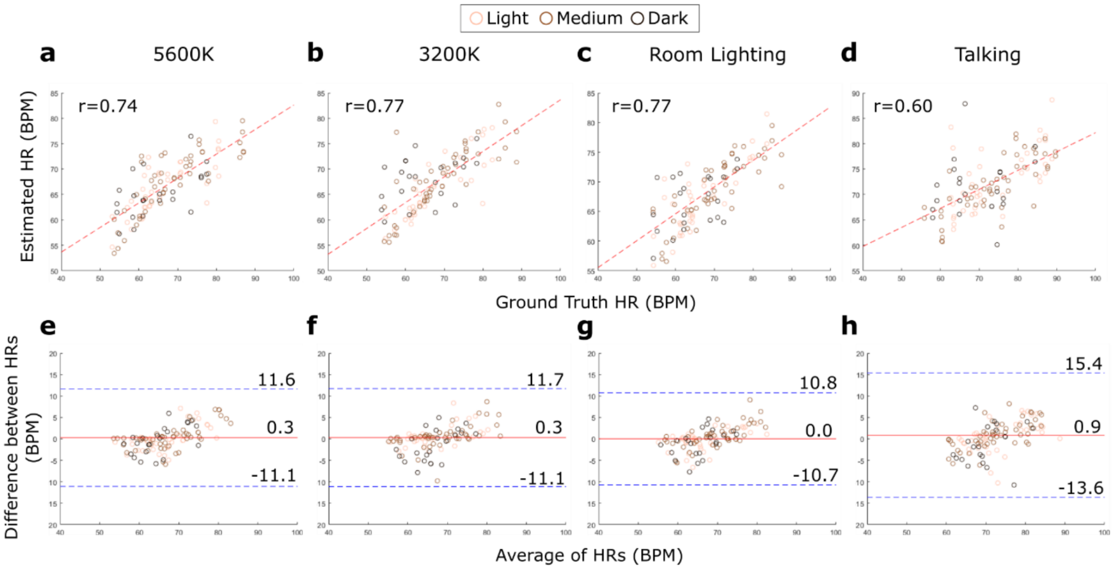


Figure 3.7: **Scatter and Bland Altman plots for facial aggregation method, varied across scene condition categories.** The label shows a marker for each skin type. **(A-D)** Scatter plots for different recording conditions highlighting the correlation between estimated and ground truth heart rate, denoted by the Pearson Correlation Coefficient r. **(E-H)** Bland-Altman plots for different recording conditions. The bias (m) is shown by the middle solid red line, and the limits of agreement ($LoA = m \pm 1.96\sigma$) by the upper and lower dotted blue lines.

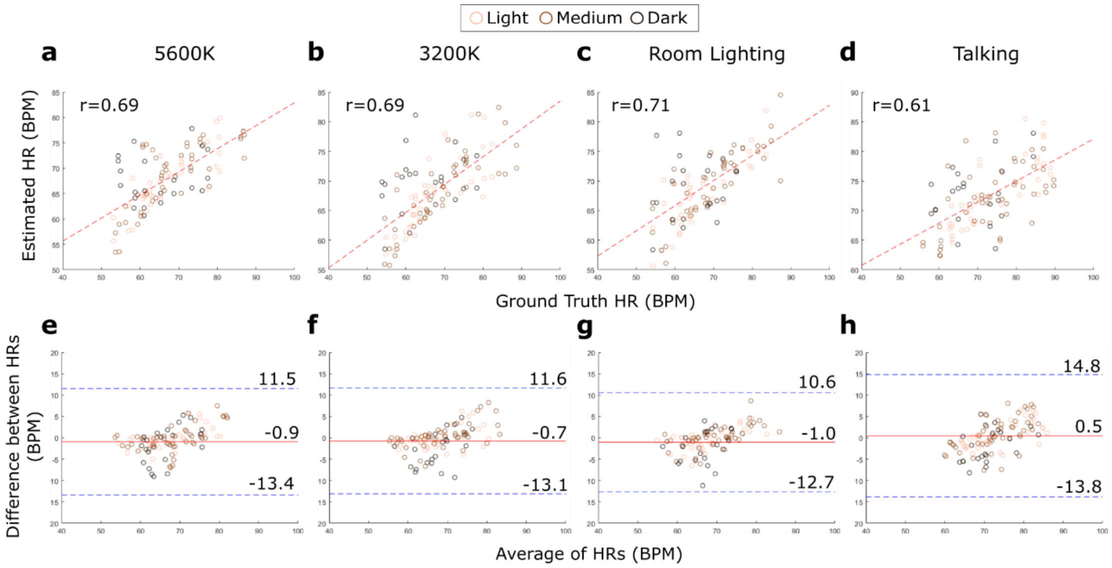


Figure 3.8: **Scatter and Bland Altman plots for SNR weighting method, varied across scene condition categories.** The label shows a marker for each skin type. **(A-D)** Scatter plots for different recording conditions highlighting the correlation between estimated and ground truth heart rate, denoted by the Pearson Correlation Coefficient r. **(E-H)** Bland-Altman plots for different recording conditions. The bias (m) is shown by the middle solid red line, and the limits of agreement ($LoA = m \pm 1.96\sigma$) by the upper and lower dotted blue lines.

and SNR weighting methods respectively.

3.4.2 Recording condition performance

Each of the three methods performs similarly across the three lighting conditions. The facial aggregation method shows an average MAE of 4.05 bpm across the lighting conditions, while the SNR weighting method shows an average performance of 4.46 bpm. In contrast to this, the proposed method shows an average performance of 3.81 bpm across the three lighting conditions, representing an improvement of +0.24 bpm MAE. The performance on the 'talking' activity is worse as compared to that on other scene conditions for all three methods. Similar to other trends, the SNR weighting method shows a performance reduction of -0.05 bpm over the facial aggregation benchmark. However, the proposed method shows a large improvement of +0.57 bpm when compared to the facial aggregation benchmark. Figure 3.6A-D highlights the high correlation between the proposed method's r-PPG HR estimates and ground truth PPG HR across the various recording conditions. The dark skin tone markers across all recording conditions make up the majority of outlying data. The B&A plots in Figure 3.6E-G show a bias of less than 1 bpm across the three lighting conditions, and Figure 3.6H shows a bias of less than 2 bpm during subject talking. These figures also show that the proposed method's r-PPG heart estimates are mostly within 10 bpm of the ground truth across all recording conditions. These correlation metrics are an improvement to the benchmark methods of facial aggregation and SNR weighting. Figure 3.7 and Figure 3.8 show the corresponding scatter and B&A plots for the facial aggregation and SNR weighting methods respectively.

3.4.3 Camera viewpoint performance

For all three methods, the bottom camera viewpoint performs the best. The facial aggregation method shows a MAE of 5.24 bpm for the front setting, and 3.74 bpm for the bottom

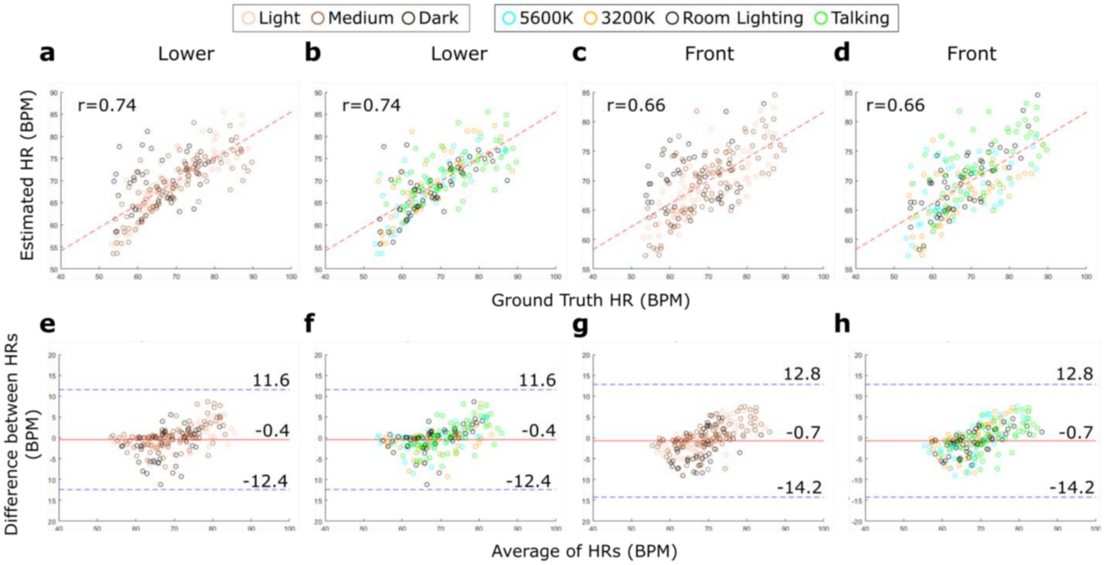


Figure 3.9: **Scatter and Bland Altman plots for proposed method's dependence on camera angle, varied across skin tone categories and recording conditions. (A-B)** Scatter plots for the lower camera angle, varying across skin tone categories and recording conditions, respectively. **(C-D)** Scatter plots for the front camera angle, varying across skin tone categories and recording conditions, respectively. **(E-F)** Bland Altman plots for the lower camera angle, varying across skin tone categories and recording conditions, respectively. **(G-H)** Bland Altman plots for the front camera angle, varying across skin tone categories and recording conditions, respectively. For all Bland Altman plots, the bias (m) is shown by the middle solid red line, and the limits of agreement ($LoA = m \pm 1.96\sigma$) by the upper and lower dotted blue lines.

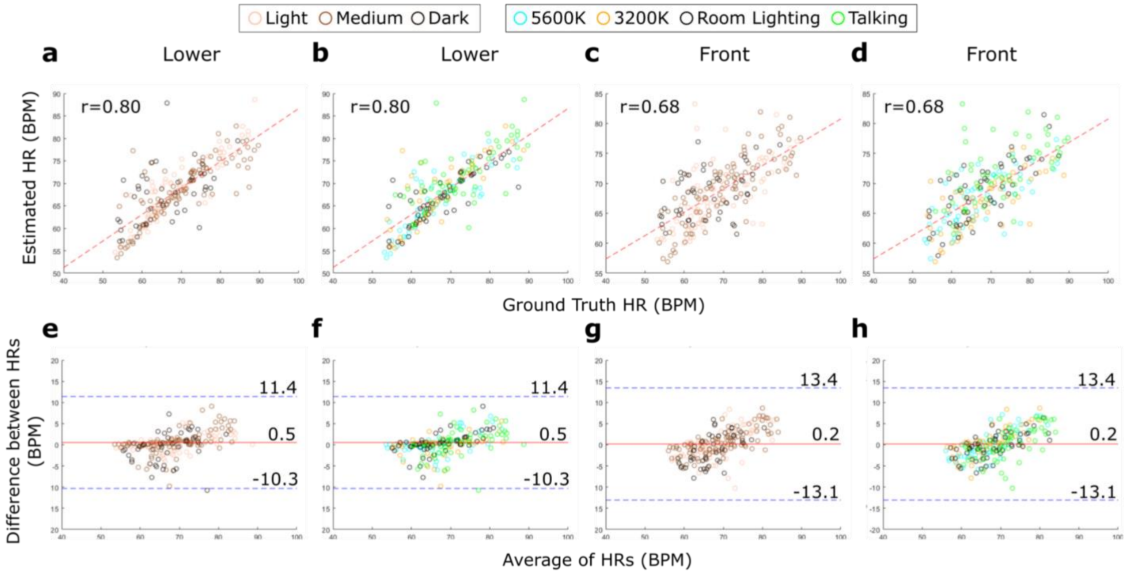


Figure 3.10: **Scatter and Bland Altman plots for the facial aggregation method's dependence on camera angle, varied across skin tone categories and recording conditions.** **(A-B)** Scatter plots for the lower camera angle, varying across skin tone categories and recording conditions, respectively. **(C-D)** Scatter plots for the front camera angle, varying across skin tone categories and recording conditions, respectively. **(E-F)** Bland Altman plots for the lower camera angle, varying across skin tone categories and recording conditions, respectively. **(G-H)** Bland Altman plots for the front camera angle, varying across skin tone categories and recording conditions, respectively. For all Bland Altman plots, the bias (m) is shown by the middle solid red line, and the limits of agreement ($LoA = m \pm 1.96\sigma$) by the upper and lower dotted blue lines.

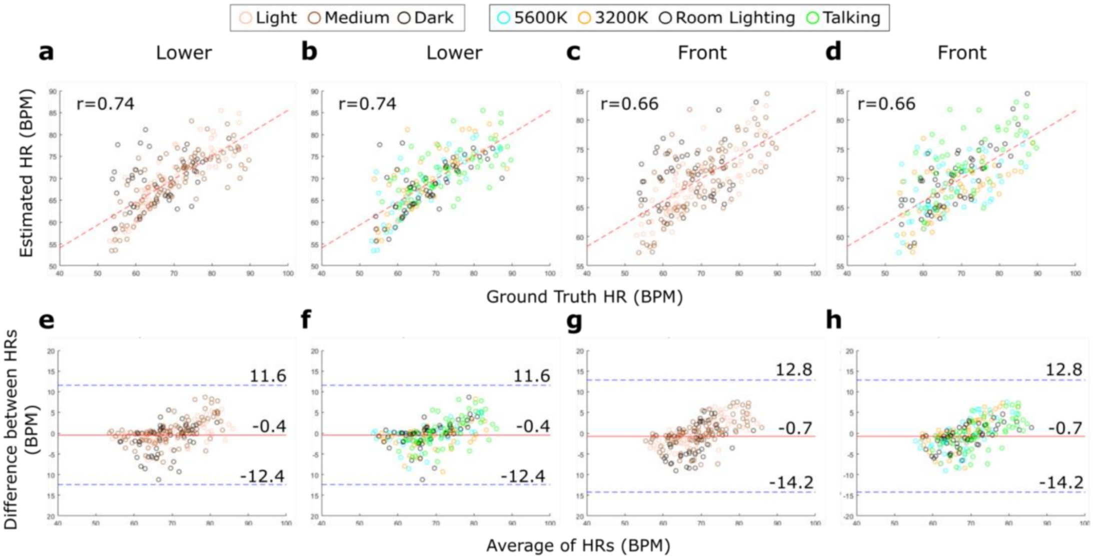


Figure 3.11: **Scatter and Bland Altman plots for the SNR weighting method's dependence on camera angle, varied across skin tone categories and recording conditions.** **(A-B)** Scatter plots for the lower camera angle, varying across skin tone categories and recording conditions, respectively. **(C-D)** Scatter plots for the front camera angle, varying across skin tone categories and recording conditions, respectively. **(E-F)** Bland Altman plots for the lower camera angle, varying across skin tone categories and recording conditions, respectively. (G-H) Bland Altman plots for the front camera angle, varying across skin tone categories and recording conditions, respectively. For all Bland Altman plots, the bias (m) is shown by the middle solid red line, and the limits of agreement ($LoA = m \pm 1.96\sigma$) by the upper and lower dotted blue lines.

setting. The SNR weighting method on the other hand shows MAE of 5.38 bpm and 4.24 bpm, while the proposed method shows a MAE of 4.89 bpm and 3.44 bpm. Figures 3.9, 3.10, and 3.11 show the corresponding scatter and B&A plots for the proposed method, facial aggregation method and SNR weighting method respectively. The correlation between estimated and ground truth HR seen by the proposed method for the front and bottom viewpoints (0.75 and 0.83) is a clear improvement over the same for the facial aggregation (0.68 and 0.80) and the SNR weighting (0.66 and 0.74) methods.

3.4.4 Best Performance

The best performing camera viewpoint and recording condition on the VITAL dataset is using the bottom camera angle with lighting at 5600K, where the label "best performing" is chosen with respect to both overall performance and skin tone bias. Figure 3.12 highlights that the proposed method achieves a MAE performance of below 3 bpm across all skin tone categories. Specifically, a MAE of 1.97, 2.86 and 3.01 bpm, and correlation of 0.93, 0.91, and 0.87, is achieved for the light, medium and dark skin tones respectively. This is a significant improvement over the two existing methods with regards to both overall performance and skin tone bias. The facial averaging method shows an MAE of 2.40, 3.47 and 4.09 bpm, and correlation of 0.89, 0.84, and 0.75, while the SNR weighting method shows an MAE of 1.48, 3.30 and 5.66 bpm, and correlation of 0.98, 0.85, and 0.58, for the same respective skin tone categories.

3.5 Synthetic generation performance

3.5.1 Performance on UBFC-RPPG

Performance metrics of different models in the UBFC-RPPG validation set are listed in Table 3.4. We exhibit the HR estimation accuracy of PRN trained with the proposed joint

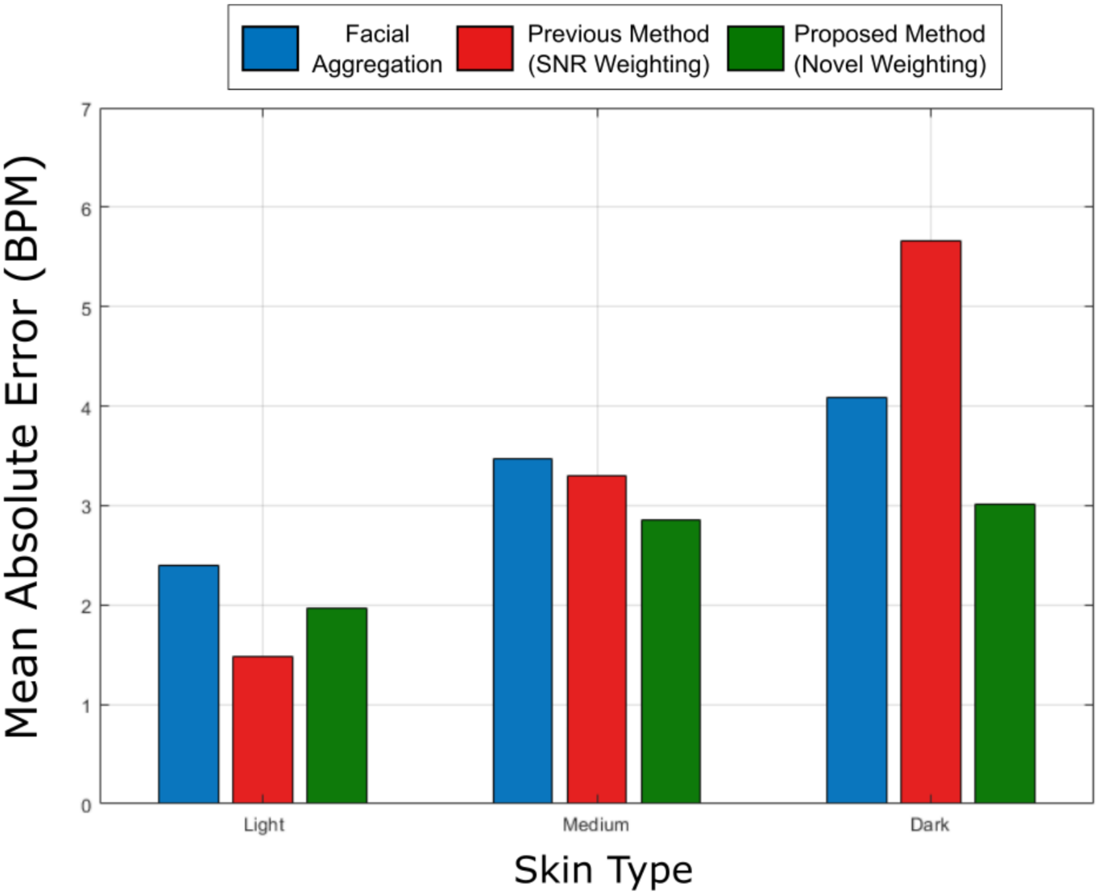


Figure 3.12: **Bar plot highlighting algorithmic comparison for the best-performing scene configuration.** It is seen that the proposed method shows increasing performance gains over both the facial aggregation and the SNR weighting methods. Specifically, for the best-performing scene configuration using the bottom camera angle viewpoint with 5600K lighting the proposed method is the only method able to have a close to 3 or below 3 bpm MAE performance in the best case, thereby establishing its capability towards medically relevant HR measurements.

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 [TLH20] w/ Real&Synth	0.89	1.66	0.88	7.74
3D-CNN [TLH20] w/ Real	1.09	1.91	0.84	7.80
3D-CNN [TLH20] w/ Synth	0.95	1.80	0.82	3.48
POS [WBS17]	3.69	5.31	0.75	3.07
CHROM [HJ13]	1.84	3.40	0.77	4.84
ICA [PMP11]	8.28	9.82	0.55	1.45

Table 3.4: **Performance of HR estimation on UBFC-RPPG.** Boldface font represents the preferred results.

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 [TLH20] 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 [WBS17], CHROM [HJ13] and ICA [PMP11]) are also provided in the table.

Notably, the proposed PRN architecture has already outperformed other rPPG estimation methods even without synthetic skin color augmentation in all four evaluation metrics. 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 of light skin tones, increasing the diversity of training samples is still able to enhance the performance. This finding is consistent with the recent research [LNP20] 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. We train 3D-CNN with both real and corresponding synthetic samples from G. As compared with the 3D-CNN model trained with just real samples, 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 other data-driven rPPG estimation algorithm without conducting the joint optimization process again to adapt to another new network architecture.

3.5.2 Cross-dataset performance on VITAL

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. This type of cross-dataset verification can provide more visibility on the generalization capability of the models. Similarly, we report MAE and RMSE of various models trained with real and synthetics samples as shown in Figure 3.13. Since VITAL dataset contains testing subjects of diverse skin tones and 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 the performance across different demographic groups. Table 3.5 shows the SNR and PCC metrics.

Method	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**	0.22	-4.24	0.46	**2.84**
PRN (w/ Real)	0.36	0.32	0.50	0.03	0.04	-6.79	0.35	-1.28
PRN (w/ Synth)	0.29	-0.45	0.42	-0.44	0.11	-6.34	0.31	-1.66
3D-CNN [TLH20] (w/ Real&Synth)	**0.42**	**3.96**	**0.65**	5.21	**0.25**	-4.77	**0.48**	2.69
3D-CNN [TLH20] (w/ Real)	0.30	-0.61	0.48	-1.26	0.19	-8.10	0.35	-2.44
3D-CNN [TLH20] (w/ Synth)	0.07	-2.04	0.38	-1.36	0.18	-5.82	0.23	-2.53
POS [WBS17]	0.16	-1.31	0.36	-0.78	0.09	-4.50	0.23	-1.74
CHROM [HJ13]	0.19	-0.69	0.36	-0.54	-0.09	**-4.22**	0.21	-1.36
ICA [PMP11]	0.18	-1.24	0.25	-1.98	0.03	-4.25	0.18	-2.18

Table 3.5: **Performance of HR estimation on VITAL**. Boldface font denotes the best results.

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 1.95 BPM reduction on MAE and 2.13 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 medium and 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 training set skin color distribution, 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 [WBS17], CHROM [HJ13] and ICA [PMP11] 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 6 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 whole 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 [MHW20]. Synthetic augmentation techniques thus become particularly effective for cross-domain learning and can improve the generalization capability of HR estimation for real-world applications.

3.5.3 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 standard deviation of MAE and RMSE in each Fitzpatrick scale as the measurement. This measurement has also been used in some prior work [MHW20, YAA20]. The standard deviation for each method in the VITAL dataset is illustrated in Figure 3.13, together with a sample portrait for each skin scale from F1 to F6. CHROM exhibits the largest variation (MAE: 3.14 BPM, RMSE: 3.24 BPM) across different Fitzpatrick scales, while jointly optimized PRN shows a minimal bias (MAE: 1.64 BPM, RMSE: 2.19 BPM) as compared with all the conventional methods. In contrast to PRN trained with just real samples (MAE: 2.14 BPM), the augmented training offers a 23% 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 augmenting the synthetic samples in the training set. We attribute this improvement to the more diverse and balanced datasets augmented with our generator.

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**	**4.97**	**6.83**	**3.16**	**4.19**
PRN w/ Real	3.38	4.46	4.67	5.83	6.92	8.96	4.67	5.98
PRN w/ Synth	4.36	6.19	4.52	6.18	5.61	8.15	4.69	6.59
3D-CNN [TLH20] w/ Real&Synth	**2.32**	**3.11**	3.18	4.09	5.89	7.83	3.43	4.51
3D-CNN [TLH20] w/ Real	3.31	4.64	5.86	6.78	7.19	9.02	5.21	6.47
3D-CNN [TLH20] w/ Synth	3.88	5.23	4.00	5.71	6.34	8.35	4.44	6.08
POS [WBS17]	6.20	7.56	7.80	9.20	6.90	9.01	7.03	8.57
CHROM [HJ13]	6.02	7.39	7.01	8.33	6.93	8.22	6.64	7.97
ICA [PMP11]	7.72	8.64	9.57	10.82	6.74	8.07	8.31	9.46

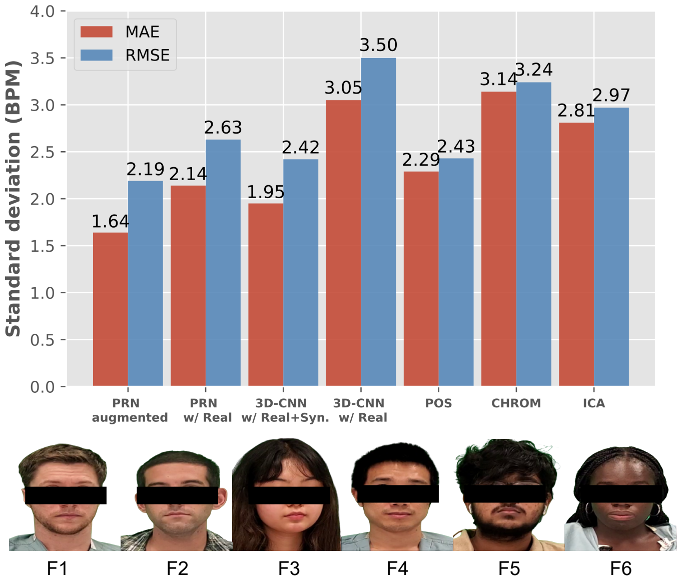


Figure 3.13: **Top: The proposed method shows an improved HR estimation accuracy on the VITAL dataset.** Boldface font denotes the preferred results. **Bottom: Synthetic dark-skinned videos can help to reduce bias in HR estimation.** The augmented PRN and the 3D-CNN [TLH20] trained on both real and synthetic videos show a reduced standard deviation on MAE and RMSE across Fitzpatrick scales F1-6 in the VITAL dataset.

CHAPTER 4

Conclusion

In summary, we propose a first attempt to translate facial frames from light-skinned subjects to dark skin tones while preserving the subtle color variations corresponding to the pulsatile signals. The proposed jointly optimized rPPG estimator can outperform the existing state-of-the-art methods with reduced estimation bias across different demographic groups. Our generated synthetic videos maintain both photo-realistic and bi-realistic features and can be directly used to improve the performance of some existing deep learning rPPG estimation model. The cross-dataset evaluation also verifies the effectiveness of this type synthetic augmentation to improve the generalization capability.

The current benchmark datasets MMSE-HR [ZGW16], AFRL [EBM14], and UBFC-RPPG [BMB19] have roughly 10%, 0%, and 5% of dark-skinned subjects respectively. It is generally very hard to collect data on these subjects due to the drastically different skin color distribution worldwide. This motivates the skin color translation from light skin tones to dark skin tones in this work. Our current pipeline is only a first attempt that focuses on the skin color translation, and all the remaining factors (e.g., pulse signals, body motion, and other facial attributes) are directly copied from the original videos. To maximize the benefit of synthetic augmentation, it is also critical to extend the generation framework to incorporate arbitrary facial attributes and pulse waves. We hope the method presented in this paper could inspire following work on synthetic generation for a more diverse dataset. 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 requirement of various subsequent rPPG estimation models without frame size interpolation.

Video synthesis, such as deepfakes, has raised public concerns in the community [ML21]. 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 the existing medicinal applications.

With respect to algorithmic development, this work addresses the aforementioned biases in skin-tone, illumination conditions, and subject motions using physics-rooted knowledge and camera noise analysis. From our theory, we derive 3 key conclusions: (i) imaging noise creates skin tone bias (and lighting bias), (ii) imaging noise and specular reflections degrade the r-PPG signal, and (iii) denoising is to be done before signal inference. Therefore, we primarily focus our attention to signal processing strategies as opposed to signal extraction modifications. The first attempted work to reduce r-PPG skin tone bias was done by Kumar et al. (DistancePPG) [KVS15], in which a weighted average of BVP signals from various facial regions-of-interest (ROI). However, to the best of our knowledge, no work yet has continued development of r-PPG algorithms that tackle the important issue of performance bias on darker skin tones. The proposed r-PPG algorithm draws from existing r-PPG denoising methods that use a similar weighted ROI philosophy as in DistancePPG (c.f. [PFL18, LBN20, BMB19]). Specifically, it modifies the strategy by weighting in RGB space rather than blood volume signal space, and by introducing a skin diffuse component weighting. This enables the proposed algorithm to mitigate performance losses for subjects with darker skin tones, subjects in varying illumination conditions, and subjects who may be moving their face such as when they are talking.

The proposed algorithm achieves the best overall average MAE performance across the VITAL dataset of 4.17 bpm, as opposed to 4.49 bpm by the facial aggregation method [PMP10a, HJ13, WSH16, WBS17, LRK11, HL14] and 4.81 bpm for the SNR weighting method [PFL18, LBN20, KVS15, BMB19]. This achievement can be attributed to the performance gains seen across all skin tones in comparison to the facial aggregation method. The SNR weighting method shows performance gain only for the light skin tone subjects (+0.08 bpm) and a performance drop for the medium and dark skin tones (-0.31 and -1.04 bpm respectively), thereby actually increasing the skin tone performance bias. Consequently, the method's overall performance suffers on a more diversely represented dataset such as VITAL. This illustrates the importance for the need of a truly diverse dataset when developing r-PPG technology.

Nevertheless, as with previous methods, the performance of the proposed algorithm still exhibits a skin-tone bias. However, we highlight that the proposed algorithm achieves the largest MAE improvements over the facial aggregation method of +0.55 bpm for the traditionally worse performing dark skin tone in comparison with the light (+0.20 bpm) and medium (+0.31 bpm) skin tones. This outcome attests to the fairness of the algorithm. The proposed algorithm is the only method able to perform with an overall MAE less than 6 bpm across all skin tones. For the best performing setting (bottom camera viewpoint with 5600K lighting), the proposed method obtains a less than 3 bpm MAE across all skin tones. This establishes the viability and performance accuracy of the proposed method for medically relevant HR estimation. These inferences are further enforced by the largest increase in the correlation coefficient and largest decrease in the SE for dark skin tones by the proposed method, as opposed to the SNR weighting method which sees performance reduction for medium and dark skin tones. Hence, in addition to the overall improvement in performance across all skin tones, the proposed method successfully steps towards reducing the performance bias that exists between skin tones.

Large improvements in performance of the proposed method are also observed for the talking activity over the facial aggregation benchmark, as compared to the SNR weighting method which shows an overall performance drop. This technology may one day allow for real-time continuous contact-less HR monitoring during a telemedicine visit, which would provide greater information to outpatient clinicians. This advance may also be relevant for in-hospital continuous contactless monitoring in ICU settings or hospital floor care.

Improvements in performance are also observed across camera viewpoints. The proposed method shows considerable improvements for the front and bottom angles. A typical telemedicine visit, through a cell phone platform, may involve the patient holding the camera at varying angles with respect to the face. The shown robustness and performance improvement of the proposed method therefore makes it increasingly amenable to such tasks. Interestingly, for all methods tested (existing and novel), the bottom angle shows improved performance as compared to the front angle. This could be because interfering factors such as hair, spectacles and so on occupy a smaller portion of the usable frame in the bottom angle, as well as differing face scales in the two angles.

In relation to the clinical significance of this work, remote vital sign monitoring has risen in prominence over recent years, with an acceleration in clinical development due to the COVID-19 pandemic. In response to the pandemic, health systems across the country implemented a large-scale restriction of non-urgent in-person appointments [JJM20], transitioned many outpatient services to telemedicine visits [CSH20], and developed remote monitoring care pathways [APH20] in order to facilitate social distancing yet maintain continuity of care. To remotely monitor COVID-19 patients, many health systems shipped home vital sign equipment to patients in order to obtain quantitative physiological data that could facilitate high quality remote management via telemedicine. At a population level, however, supplying and shipping vital sign monitoring devices to patients is expensive and not scalable, making such a solution nonviable. Figure 4.1 shows the projected cost of deploying finger pulse oximeters for telemedicine application, the most viable and inexpensive existing solution to assess patient HR and oxygen saturation. For the scales at which telemedicine is projected to

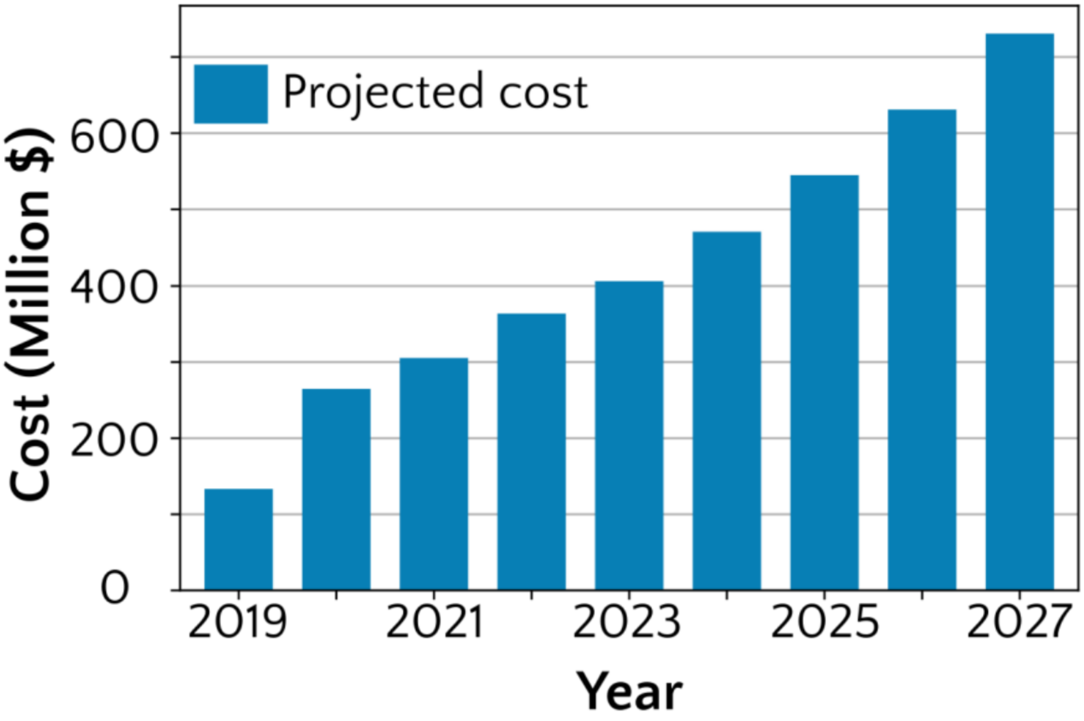


Figure 4.1: **Projected cost of deploying finger pulse oximeters for telemedicine application.** HR sensing solutions for telemedicine and remote patient monitoring have relied on the adoption of wearable sensors. Currently, the most viable and inexpensive existing wearable solution to assess patient HR and oxygen saturation are finger pulse oximeters. For the scales at which telemedicine is projected to grow, such a solution would involve a deployment cost in excess of $700 million in the US alone. In contrast, the proposed camera-based method offers a purely algorithmic solution that can be integrated into existing healthcare system telemedicine video-conferencing applications.

grow, even this solution would involve a deployment cost in excess of $700 million in the US alone (see Section 2.10 for calculation details) [pol20]. Given the high penetration of mobile phone technology globally, there is great interest in transforming smartphones into low-cost portable HR, respiratory rate, and pulse oximeter monitors, thereby increasing accessibility to vital monitoring equipment and alleviating healthcare inequity. Using in-built camera modules and computer vision algorithms to obtain quantitative vital sign data remotely offers a purely algorithmic solution with potentially zero marginal cost.

Outside of a pandemic situation, knowledge of vital signs is also important information for clinicians who are managing medical conditions that require such data for health management, and remotely obtaining vital signs may allow care teams to perform remote surveillance and home monitoring of patients with greater confidence. Notably, several minority and lower socioeconomic status patient populations may benefit from more remote care, especially as it has been established that the COVID-19 pandemic has disproportionately affected such communities, both nationally and in states the most affected by the pandemic [AOA20, ASR20]. In New York City and Michigan, African American and Latino residents have the highest age-adjusted rates of hospitalized and non-hospitalized COVID-19, and age-adjusted death rates for African Americans are more than twice those for white and Asian residents [HBT20, GMS20]. African American communities have also been found to have higher prevalence of cardiovascular and related complications, when compared with traditionally light skin toned people [Men18]. These patient populations may therefore stand to benefit the most from skin tone robust contactless vital sign (specifically heart rate) sensing technologies that facilitate high-quality remote care pathways.

Finally, we believe contactless vital sign sensing technology would be useful at the start of in person clinic or hospital encounters or for continuous patient monitoring in a hospital floor or ICU setting. Cameras, as opposed to hospital staff, may one day obtain key vital signs without contact, thereby reducing exposure of patients to staff, enabling improved infection control, and freeing up hospital staff to attend to other important patient care needs.

With regards to limitations and future work, while our algorithm has been tested on an adult population, additional work is needed to enable clinical adoption. Further research investigating HR estimation using our proposed method is still needed in pediatric and geriatric populations and patient populations with known cardiopulmonary disease. Future work must also focus on improving computer vision methods to detect extremes of HR and discern heart arrhythmia. Additionally, the proposed method does not obviate skin tone bias but rather is the first work that can be demonstrated to mitigate skin tone bias in the VITAL dataset. Therefore, research must be undertaken to further reduce bias and assure fairness by building upon our work, as well as to continue improving overall performance on subjects and videos in real life scenarios.

From an algorithmic perspective, we believe that one of the most important factors towards large scale deployment of such methods for clinical use is the inherent fairness of the algorithm. As healthcare increasingly accelerates towards a digitally connected and virtual future, early consideration must be given to developing equitable health technology that does not exacerbate healthcare disparities or create new disparities. Ultimately, we hope this work motivates the community towards exciting and essential research avenues looking into inherent system biases associated with r-PPG. By reducing biases, we move a step closer towards deploying high quality, medically inclusive non-contact vital sensing techniques that can aid clinicians in delivering remote patient care, during times of peace and pandemic alike.

REFERENCES

[ANT16] B. Aubakir, B. Nurimbetov, I. Tursynbek, and H. A. Varol. "Vital sign monitoring utilizing Eulerian video magnification and thermography." In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3527–3530, August 2016. ISSN: 1558-4615.

[AOA20] Vida Abedi, Oluwaseyi Olulana, Venkatesh Avula, Durgesh Chaudhary, Ayesha Khan, Shima Shahjouei, Jiang Li, and Ramin Zand. "Racial, Economic, and Health Inequality and COVID-19 Infection in the United States." *Journal of Racial and Ethnic Health Disparities*, September 2020.

[APH20] Tucker Annis, Susan Pleasants, Gretchen Hultman, Elizabeth Lindemann, Joshua A. Thompson, Stephanie Billecke, Sameer Badlani, and Genevieve B. Melton. "Rapid implementation of a COVID-19 remote patient monitoring program." *Journal of the American Medical Informatics Association*, **27**(8):1326–1330, August 2020. Publisher: Oxford Academic.

[AS17] S. Alotaibi and W. A. P. Smith. "A Biophysical 3D Morphable Model of Face Appearance." In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 824–832, October 2017. ISSN: 2473-9944.

[ASR20] Kristen M. J. Azar, Zijun Shen, Robert J. Romanelli, Stephen H. Lockhart, Kelly Smits, Sarah Robinson, Stephanie Brown, and Alice R. Pressman. "Disparities In Outcomes Among COVID-19 Patients In A Large Health Care System In California." *Health Affairs*, **39**(7):1253–1262, May 2020. Publisher: Health Affairs.

[BDG13] G. Balakrishnan, F. Durand, and J. Guttag. "Detecting Pulse from Head Motions in Video." In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3430–3437, June 2013. ISSN: 1063-6919.

[BG18] Joy Buolamwini and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." In *Conference on fairness, accountability and transparency*, pp. 77–91, 2018.

[BMB19] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. "Unsupervised skin tissue segmentation for remote photoplethysmography." *Pattern Recognition Letters*, **124**:82–90, June 2019.

[BWK21] Yunhao Ba, Zhen Wang, Kerim Doruk Karinca, Oyku Deniz Bozkurt, and Achuta Kadambi. "Overcoming difficuly in obtaining dark-skinned subjects for remote-PPG by synthetic augmentation." *arXiv*, 2021.

[BWT11] Maged N. Kamel Boulos, Steve Wheeler, Carlos Tavares, and Ray Jones. "How smartphones are changing the face of mobile and participatory healthcare: an overview, with example from eCAALYX." *BioMedical Engineering OnLine*, **10**(1):24, April 2011.

[Cha04] George Chaplin. "Geographic distribution of environmental factors influencing human skin coloration." *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, **125**(3):292–302, 2004.

[CKK20a] Pradyumna Chari, Krish Kabra, Doruk Karinca, Soumyarup Lahiri, Diplav Srivastava, Kimaya Kulkarni, Tianyuan Chen, Maxime Cannesson, Laleh Jalilian, and Achuta Kadambi. "Diverse R-PPG: Camera-Based Heart Rate Estimation for Diverse Subject Skin-Tones and Scenes." *arXiv preprint arXiv:2010.12769*, 2020.

[CKK20b] Pradyumna Chari, Krish Kabra, Doruk Karinca, Soumyarup Lahiri, Diplav Srivastava, Kimaya Kulkarni, Tianyuan Chen, Maxime Cannesson, Laleh Jalilian, and Achuta Kadambi. "Diverse R-PPG: Camera-Based Heart Rate Estimation for Diverse Subject Skin-Tones and Scenes." *arXiv preprint arXiv:2010.12769*, 2020.

[CM18] Weixuan Chen and Daniel McDuff. "DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks." In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, pp. 356–373, Cham, 2018. Springer International Publishing.

[cps20] "Current Population Survey (CPS)." The United States Census Bureau, Nov 2020.

[CSH20] Samantha L. Connolly, Kelly L. Stolzmann, Leonie Heyworth, Kendra R. Weaver, Mark S. Bauer, and Christopher J. Miller. "Rapid Increase in Telemental Health Within the Department of Veterans Affairs During the COVID-19 Pandemic." *Telemedicine and e-Health*, September 2020. Publisher: Mary Ann Liebert, Inc., publishers.

[CSX18] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. "Vggface2: A dataset for recognising faces across pose and age." In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 67–74. IEEE, 2018.

[DCC19] Catherine Dinh-Le, Rachel Chuang, Sara Chokshi, and Devin Mann. "Wearable Health Technology and Electronic Health Record Integration: Scoping Review and Future Directions." *JMIR mHealth and uHealth*, **7**(9):e12861, September

2019. Company: JMIR mHealth and uHealth Distributor: JMIR mHealth and uHealth Institution: JMIR mHealth and uHealth Label: JMIR mHealth and uHealth Publisher: JMIR Publications Inc., Toronto, Canada.

[EBM14] Justin R Estepp, Ethan B Blackford, and Christopher M Meier. "Recovering pulse rate during motion artifact with a multi-imager array for non-contact imaging photoplethysmography." In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1462–1469. IEEE, 2014.

[Fit88] Thomas B. Fitzpatrick. "The Validity and Practicality of Sun-Reactive Skin Types I Through VI." *Archives of Dermatology*, **124**(6):869–871, June 1988. Publisher: American Medical Association.

[GMS20] Tian Gu, Jasmine A. Mack, Maxwell Salvatore, Swaraaj Prabhu Sankar, Thomas S. Valley, Karandeep Singh, Brahmajee K. Nallamothu, Sachin Kheterpal, Lynda Lisabeth, Lars G. Fritsche, and Bhramar Mukherjee. "Characteristics Associated With Racial/Ethnic Disparities in COVID-19 Outcomes in an Academic Health Care System." *JAMA network open*, **3**(10):e2025197, 2020.

[HBT20] David R. Holtgrave, Meredith A. Barranco, James M. Tesoriero, Debra S. Blog, and Eli S. Rosenberg. "Assessing racial and ethnic disparities using a COVID-19 outcomes continuum for New York State." *Annals of Epidemiology*, **48**:9–14, August 2020.

[HDF10] S. W. Hasinoff, F. Durand, and W. T. Freeman. "Noise-optimal capture for high dynamic range photography." In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 553–560, June 2010. ISSN: 1063-6919.

[HJ13] G. de Haan and V. Jeanne. "Robust Pulse Rate From Chrominance-Based rPPG." *IEEE Transactions on Biomedical Engineering*, **60**(10):2878–2886, October 2013. Conference Name: IEEE Transactions on Biomedical Engineering.

[HL14] G. de Haan and A. van Leest. "Improved motion robustness of remote-PPG by using the blood volume pulse signature." *Physiological Measurement*, **35**(9):1913–1926, 2014.

[JJM20] Ferguson Jm, Jacobs J, Yefimova M, Greene L, Heyworth L, and Zulman Dm. "Virtual Care Expansion in the Veterans Health Administration During the COVID-19 Pandemic: Clinical Services and Patient Characteristics Associated with Utilization." *Journal of the American Medical Informatics Association : JAMIA*, October 2020.

[Kat20] Rimma Kats. "US telemedicine users will surpass 40 million this year." *eMarketer*, November 2020.

[KH13] John George Karippacheril and Tam Yuk Ho. "Data acquisition from S/5 GE Datex anesthesia monitor using VSCapture: An open source.NET/Mono tool." *Journal of Anaesthesiology, Clinical Pharmacology*, **29**(3):423–424, 2013.

[KNP13] S. Kumar, W. Nilsen, M. Pavel, and M. Srivastava. "Mobile Health: Revolutionizing Healthcare Through Transdisciplinary Research." *Computer*, **46**(1):28–35, January 2013. Conference Name: Computer.

[KS14] V. Kazemi and J. Sullivan. "One millisecond face alignment with an ensemble of regression trees." In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1867–1874, June 2014. ISSN: 1063-6919.

[KVS15] Mayank Kumar, Ashok Veeraraghavan, and Ashutosh Sabharwal. "DistancePPG: Robust non-contact vital signs monitoring using a camera." *Biomedical Optics Express*, **6**(5):1565–1588, May 2015. Publisher: Optical Society of America.

[LAS18] Xiaobai Li, Iman Alikhani, Jingang Shi, Tapio Seppanen, Juhani Junttila, Kirsi Majamaa-Voltti, Mikko Tulppo, and Guoying Zhao. "The OBF database: A large face video database for remote physiological signal measurement and atrial fibrillation detection." In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 242–249. IEEE, 2018.

[LBN20] Peixi Li, Yannick Benezeth, Keisuke Nakamura, Randy Gomez, and Fan Yang. "Model-based Region of Interest Segmentation for Remote Photoplethysmography." In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, volume 4, pp. 383–388, Prague, Czech Republic, October 2020.

[LCZ14] X. Li, J. Chen, G. Zhao, and M. Pietikäinen. "Remote Heart Rate Measurement from Face Videos under Realistic Situations." In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4264–4271, June 2014. ISSN: 1063-6919.

[LNP20] Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis." *Proceedings of the National Academy of Sciences*, **117**(23):12592–12594, 2020.

[LRK11] 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)*, pp. 405–410, September 2011.

[LWT19] Ka Hou Christien Li, Francesca Anne White, Timothy Tipoe, Tong Liu, Martin CS Wong, Aaron Jesuthasan, Adrian Baranchuk, Gary Tse, and Bryan P. Yan. "The Current State of Mobile Phone Apps for Monitoring Heart Rate, Heart Rate Variability, and Atrial Fibrillation: Narrative Review." *JMIR mHealth and uHealth*, **7**(2):e11606, February 2019. Company: JMIR mHealth and uHealth Distributor: JMIR mHealth and uHealth Institution: JMIR mHealth and uHealth Label: JMIR mHealth and uHealth Publisher: JMIR Publications Inc., Toronto, Canada.

[LXY20] Heather Lukas, Changhao Xu, You Yu, and Wei Gao. "Emerging Telemedicine Tools for Remote COVID-19 Diagnosis, Monitoring, and Management." *ACS Nano*, **14**(12):16180–16193, December 2020. Publisher: American Chemical Society.

[MB19] Daniel McDuff and Ethan Blackford. "iPhys: An Open Non-Contact Imaging-Based Physiological Measurement Toolbox." *arXiv:1901.04366 [cs]*, January 2019. arXiv: 1901.04366.

[McD18] Daniel McDuff. "Deep super resolution for recovering physiological information from videos." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1367–1374, 2018.

[Men18] George A. Mensah. "Cardiovascular Diseases in African Americans: Fostering Community Partnerships to Stem the Tide." *American Journal of Kidney Diseases*, **72**(5, Supplement 1):S37–S42, November 2018.

[MHW20] Daniel McDuff, Javier Hernandez, Erroll Wood, Xin Liu, and Tadas Baltrusaitis. "Advancing Non-Contact Vital Sign Measurement using Synthetic Avatars." *arXiv preprint arXiv:2010.12949*, 2020.

[ML21] Yisroel Mirsky and Wenke Lee. "The creation and detection of deepfakes: A survey." *ACM Computing Surveys (CSUR)*, **54**(1):1–41, 2021.

[MN21] Daniel McDuff and Ewa 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.

[MSd16] A.V. Moço, S. Stuijk, and G. de Haan. "Motion robust PPG-imaging through color channel mapping." *Biomedical Optics Express*, **7**(5):1737–1754, 2016.

[MYS12] Abu Saleh Mohammad Mosa, Illhoi Yoo, and Lincoln Sheets. "A Systematic Review of Healthcare Applications for Smartphones." *BMC Medical Informatics and Decision Making*, **12**(1):67, July 2012.

[NFR18] Caterina Nahler, Bernhard Feldhofer, Matthias Ruether, Gerald Holweg, and Norbert Druml. “Exploring the Usage of Time-of-Flight Cameras for Contact and Remote Photoplethysmography.” In *2018 21st Euromicro Conference on Digital System Design (DSD)*, pp. 433–441, 2018.

[NM19] Ewa Nowara and Daniel McDuff. “Combating the Impact of Video Compression on Non-Contact Vital Sign Measurement Using Supervised Learning.” In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 1706–1712, Oct 2019.

[NMM20] E. M. Nowara, T. K. Marks, H. Mansour, and A. Veeraraghavan. “Near-Infrared Imaging Photoplethysmography During Driving.” *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–12, 2020. Conference Name: IEEE Transactions on Intelligent Transportation Systems.

[NMV20a] Ewa Nowara, Daniel McDuff, and Ashok Veeraraghavan. “The Benefit of Distraction: Denoising Remote Vitals Measurements using Inverse Attention.” *arXiv:2010.07770 [cs, eess]*, October 2020. arXiv: 2010.07770.

[NMV20b] Ewa M. Nowara, Daniel McDuff, and Ashok 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*, pp. 284–285, 2020.

[NMV21] Ewa M. Nowara, Daniel McDuff, and Ashok Veeraraghavan. “Systematic analysis of video-based pulse measurement from compressed videos.” *Biomedical Optics Express*, **12**(1):494–508, January 2021. Publisher: Optical Society of America.

[NSH20] 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*, **29**:2409–2423, 2020. Conference Name: IEEE Transactions on Image Processing.

[NWH20] Masataka Nishiga, Dao Wen Wang, Yaling Han, David B. Lewis, and Joseph C. Wu. “COVID-19 and cardiovascular disease: from basic mechanisms to clinical perspectives.” *Nature Reviews Cardiology*, **17**(9):543–558, September 2020. Number: 9 Publisher: Nature Publishing Group.

[PFL18] Lai-Man Po, Litong Feng, Yuming Li, Xuyuan Xu, Terence Chun-Ho Cheung, and Kwok-Wai Cheung. “Block-based adaptive ROI for remote photoplethysmography.” *Multimedia Tools and Applications*, **77**(6):6503–6529, March 2018.

[PMH19] Tine Proesmans, Christophe Mortelmans, Ruth Van Haelst, Frederik Verbrugge, Pieter Vandervoort, and Bert Vaes. “Mobile Phone–Based Use of the Photoplethysmography Technique to Detect Atrial Fibrillation in Primary Care: Diagnostic Accuracy Study of the FibriCheck App.” *JMIR mHealth and uHealth*,

7(3):e12284, March 2019. Company: JMIR mHealth and uHealth Distributor: JMIR mHealth and uHealth Institution: JMIR mHealth and uHealth Label: JMIR mHealth and uHealth Publisher: JMIR Publications Inc., Toronto, Canada.

[PMP10a] Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard. "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation." *Opt. Express*, **18**(10):10762–10774, May 2010.

[PMP10b] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation." *Optics express*, **18**(10):10762–10774, 2010.

[PMP11] M. Poh, D. J. McDuff, and R. W. Picard. "Advancements in Noncontact, Multiparameter Physiological Measurements Using a Webcam." *IEEE Transactions on Biomedical Engineering*, **58**(1):7–11, January 2011. Conference Name: IEEE Transactions on Biomedical Engineering.

[pol20] "U.S. Telemedicine Market Share, Size, Trends, Industry Analysis Report By Component (Hardware, Software & Services); By Application (Teleradiology, Telepsychiatry, Telestroke, Tele-ICU, Teledermatology, Teleconsultation); Mode of Delivery (Mobile Health Apps, Virtual, Telehealth Portals & Kiosks), By End User (Providers, Payers, Patients); Segment Forecast, 2020 - 2027." Technical Report PM1672, Polaris Market Research, New York, August 2020.

[PWG20] Omkar Patil, Wei Wang, Yang Gao, and Zhanpeng Jin. "MobiEye: turning your smartphones into a ubiquitous unobtrusive vital sign monitoring system." *CCF Transactions on Pervasive Computing and Interaction*, **2**(2):97–112, June 2020.

[RIS19] Attila Reiss, Ina Indlekofer, Philip Schmidt, and Kristof Van Laerhoven. "Deep ppg: Large-scale heart rate estimation with convolutional neural networks." *Sensors*, **19**(14):3079, 2019.

[Saw20] Jennifer Sawyer. "Wearable Internet of Medical Things Sensor Devices, Artificial Intelligence-driven Smart Healthcare Services, and Personalized Clinical Care in COVID-19 Telemedicine." *American Journal of Medical Research*, **7**(2):71–77, 2020. Publisher: Addleton Academic Publishers.

[SFC18] Radim Spetlík, Vojtech Franc, J. Cech, and Jiri Matas. "Visual Heart Rate Estimation with Convolutional Neural Network." In *BMVC*, 2018.

[SMT15] Steven R. Steinhubl, Evan D. Muse, and Eric J. Topol. "The emerging field of mobile health." *Science Translational Medicine*, **7**(283):283rv3–283rv3, April 2015. Publisher: American Association for the Advancement of Science Section: Review.

[SND17] Guanghao Sun, Yosuke Nakayama, Sumiyakhand Dagdanpurev, Shigeto Abe, Hidekazu Nishimura, Tetsuo Kirimoto, and Takemi Matsui. "Remote sensing of multiple vital signs using a CMOS camera-equipped infrared thermography system and its clinical application in rapidly screening patients with suspected infectious diseases." *International journal of infectious diseases: IJID: official publication of the International Society for Infectious Diseases*, **55**:113–117, February 2017.

[SZC20] Rencheng Song, Senle Zhang, Juan Cheng, Chang Li, and Xun Chen. "New insights on super-high resolution for video-based heart rate estimation with a semi-blind source separation method." *Computers in Biology and Medicine*, **116**:103535, January 2020.

[TAR16] 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 *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2396–2404, June 2016. ISSN: 1063-6919.

[TKD12] Gill R. Tsouri, Survi Kyal, Sohail A. Dianat, and Lalilt K. Mestha. "Constrained independent component analysis approach to nonobtrusive pulse rate measurements." *Journal of Biomedical Optics*, **17**(7):077011, July 2012. Publisher: International Society for Optics and Photonics.

[TLH20] Yun-Yun Tsou, Yi-An Lee, and Chiou-Ting Hsu. "Multi-Task Learning for Simultaneous Video Generation and Remote Photoplethysmography Estimation." In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[TLL18] C. Tang, J. Lu, and J. Liu. "Non-contact Heart Rate Monitoring by Combining Convolutional Neural Network Skin Detection and Remote Photoplethysmography via a Low-Cost Camera." *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1390–13906, 2018.

[VCJ19] Mauricio Villarroel, Sitthichok Chaichulee, João Jorge, Sara Davis, Gabrielle Green, Carlos Arteta, Andrew Zisserman, Kenny McCormick, Peter Watkinson, and Lionel Tarassenko. "Non-contact physiological monitoring of preterm infants in the Neonatal Intensive Care Unit." *npj Digital Medicine*, **2**(1):1–18, December 2019. Number: 1 Publisher: Nature Publishing Group.

[Ven14] C. Lee Ventola. "Mobile Devices and Apps for Health Care Professionals: Uses and Benefits." *Pharmacy and Therapeutics*, **39**(5):356–364, May 2014.

[VSN08] Wim Verkruysse, Lars O. Svaasand, and J. Stuart Nelson. "Remote plethysmographic imaging using ambient light." *Optics Express*, **16**(26):21434–21445, December 2008. Publisher: Optical Society of America.

[WBH21] Jonathan P Weiner, Stephen Bandeian, Elham Hatef, Daniel Lans, Angela Liu, and Klaus W Lemke. "In-Person and Telehealth Ambulatory Contacts and Costs in a Large US Insured Cohort Before and During the COVID-19 Pandemic." *JAMA network open*, **4**(3):e212618–e212618, 2021.

[WBS17] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan. "Algorithmic Principles of Remote PPG." *IEEE Transactions on Biomedical Engineering*, **64**(7):1479–1491, July 2017. Conference Name: IEEE Transactions on Biomedical Engineering.

[WSH15] W. Wang, S. Stuijk, and G. de Haan. "Exploiting Spatial Redundancy of Image Sensor for Motion Robust rPPG." *IEEE Transactions on Biomedical Engineering*, **62**(2):415–425, February 2015. Conference Name: IEEE Transactions on Biomedical Engineering.

[WSH16] W. Wang, S. Stuijk, and G. de Haan. "A Novel Algorithm for Remote Photoplethysmography: Spatial Subspace Rotation." *IEEE Transactions on Biomedical Engineering*, **63**(9):1974–1984, September 2016. Conference Name: IEEE Transactions on Biomedical Engineering.

[YAA20] Seyma Yucer, Samet Akçay, Noura Al-Moubayed, and Toby 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*, pp. 18–19, 2020.

[YLZ19] Zitong Yu, Xiaobai Li, and Guoying Zhao. "Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks." *arXiv e-prints*, **1905**:arXiv:1905.02419, May 2019.

[YPL19] 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 *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 151–160, October 2019. ISSN: 2380-7504.

[YWA10] Qingxiong Yang, Shengnan Wang, and Narendra Ahuja. "Real-Time Specular Highlight Removal Using Bilateral Filtering." In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision – ECCV 2010*, Lecture Notes in Computer Science, pp. 87–100, Berlin, Heidelberg, 2010. Springer.

[ZGW16] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. "Multimodal spontaneous emotion corpus for human behavior analysis." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3438–3446, 2016.

[ZPI17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. "Unpaired image-to-image translation using cycle-consistent adversarial networks." In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

[ZZL16] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks." *IEEE Signal Processing Letters*, **23**(10):1499–1503, October 2016. Conference Name: IEEE Signal Processing Letters.