Style Transfer with Bio-realistic Appearance Manipulation for Skin-tone Inclusive rPPG

Yunhao Ba*, Zhen Wang*, Kerim Doruk Karinca, Oyku Deniz Bozkurt, and Achuta Kadambi

Abstract—Data-driven remote vital sign estimation provides an efficient alternative to on-site clinical monitoring, however, its performance can be biased due to the imbalanced training sets. In this work, we take remote photoplethysmography (rPPG) as an example to examine the performance bias from skin tone variations in non-contact heart rate estimation. In rPPG, recent deep learning models have significantly improved the accuracy of the physiological measurement, however, the existing datasets MMSE-HR, AFRL, and UBFC-RPPG only contain roughly 10%, 0%, and 5% of dark-skinned subjects respectively. The imbalanced training sets result in a poor generalization capability of these models and lead to unwanted bias toward different demographic groups. In Western academia, it is regrettably difficult in a university setting to collect data on these dark-skinned subjects. Here we show a first attempt to overcome the lack of dark-skinned subjects by synthetic augmentation. A joint optimization framework is utilized to translate real videos from light-skinned subjects to dark skin tones while retaining their pulsatile signals. In the experiment, our method exhibits around 38% reduction in mean absolute error for the dark-skinned group and 49% improvement on bias mitigation, as compared with the previous work trained with just real samples. Project website: https://visual.ee.ucla.edu/rppg_augmentation.htm/.

Index Terms—Remote Photoplethysmography, Bio-realistic Skin-tone Translation, Fairness in Healthcare, Bias Mitigation

◆

1 INTRODUCTION

DURING the pandemic, telehealth consults have increased more than 50-fold for certain groups (e.g., those with chronic diseases) [1] due to the concerns that the congregation of people may increase the risk of contraction. Although contact sensors (e.g., electrocardiograms, oximeters) provide a gold-standard measurement of human body functions, these contact devices are not widely available, which makes a non-contact way of detecting vital signs crucial for telehealth settings [2], [3], [4]. Non-contact health sensing can also benefit applications in clinical settings, such as neonatal intensive care unit (ICU) sensing [5], as the contact sensors may cause infection for these vulnerable groups. For non-contact health sensing systems to be deployed at scale in society, it is important to ensure their performance consistency across a broad range of ethnic groups [6]. In this paper, we use remote photoplethysmography (rPPG) as an example to explore how to push Pareto frontier by promoting both accuracy and fairness in heart rate estimation with synthetic augmentation as shown in Figure 1. We select camera-based rPPG [7], [8] since it provides a solution to the above scenarios given that web cameras are more ubiquitously available, contactless, and low-cost. In the meantime, the existing rPPG datasets are usually overwhelmed by subjects of light skin tones, which makes it problematic to deploy rPPG for various demographic groups.

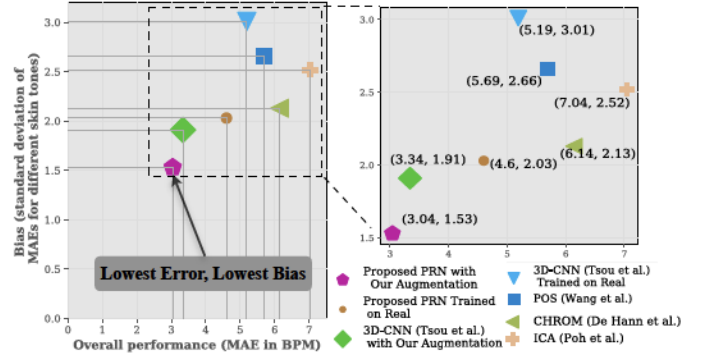Camera-based rPPG uses subtle skin color variations on

Fig. 1. **Our proposed augmentation method pushes the Pareto frontier toward both axes: accuracy and equity for rPPG.** We use the mean absolute error (MAE) of the heart rate (HR) estimation for all skin tones as the overall performance metric and the standard deviation of MAEs across different skin-tone groups as the bias metric. Our proposed augmentation method has the lowest estimation error with minimized bias as compared with the existing solutions. HR MAE is measured in the unit of beats per minute (BPM) in the plot.

the face to obtain physiological signals. When the light hits the face, the amount of light reflected or absorbed is determined by the physiological processes, and the color change corresponding to the Blood Volume Pulse (BVP) is synchronized with the heart rate (HR), which provides the feasibility to extract HR from facial videos. While data-driven neural networks have exhibited remarkable estimation accuracy for non-contact camera-based sensing [9], [10], [11], [12], 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); (2) the requirement of medical-grade sensors and necessity of intrusive/semi-

- Y. Ba, Z. Wang, O. D. Bozkurt, and A. Kadambi are with the Department of Electrical and Computer Engineering, University of California, Los Angeles.
- K. D. Karinca and A. Kadambi are with the Department of Computer Science, University of California, Los Angeles.
- Emails in authors' order: {yhba, zhenwang}@ucla.edu, dorukkar-inca@cs.ucla.edu, denizbozkurt@ucla.edu, achuta@ee.ucla.edu.
- *Equal contribution.

intrusive traditional methods for data collection; and (3) patient privacy concerns (e.g., OBF dataset [13] is not publicly available due to the licence issue).

Recent studies have shown that computer vision algorithms have been disadvantaging the underrepresented groups in some applications, such as face recognition [15]. Non-contact rPPG estimation is not an exception given the unbalanced and relatively small datasets in the field [16]. There are very rare subjects with dark skin tones in the existing benchmark datasets. More specifically, MMSE-HR [17], AFRL [18], and UBFC-RPPG [19] 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 [16]. 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.

Realizing the difficulty of recruiting patients to collect large-scale rPPG datasets in the university setting, synthetic augmentation of facial videos has become an active research topic recently. McDuff *et al.* [20] 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 [21]. Tsou *et al.* [22] augment source rPPG videos with other specified pulse 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 photorealistic 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.* [14] introduce a race translation model across various racial domains with a CycleGAN-based architecture [23]. However, their work is not designed to incorporate pulsatile signals. As illustrated in Figure 2, this vanilla skin tone translation network [14] 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 method with reduced bias across different skin color groups.

1.1 Contributions

To summarize, the contributions of our work include:

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

2 RELATED WORK

2.1 Imaging Photoplethysmography

Imaging PPG methods aim to recover the pulsatile signal from the subtle color changes in the face videos. Algorithms of detecting non-contact PPG signal can roughly be divided into three categories: Signal decomposition [8], [24], [25], [26], [27], model-based methods [28], [29], [30], [31], and deep learning methods [9], [10], [11], [12]. Signal decomposition techniques based on Blind Source Separation (BSS) techniques decompose/demix the face videos into different sources utilizing PCA [24] or ICA [8]. However, these methods do not exploit skin reflectance properties that are specific to rPPG problems.

Model-based methods, such as CHROM [29], apply color space transforms to linearly combine the chrominance signals to obtain the final PPG signals. The Pulse Blood Vector [28] method uses characteristic blood volume changes to weight different color channels. This method can be further improved by first projecting the temporally-normalized skin tone onto the plane which is orthogonal to the intensity variation term and then linearly combine the projected signals [30]. These methods use all the face skin pixels for the rPPG measurement, which may achieve sub-optimal results as each pixel may have very different contribution to the pulse signals.

More recently, data-driven method has gained more attention [10], [22], [32], [33], [34], [35]. More specifically, DeepPhys [32] proposes a Convolutional Attention Network (CAN) which uses appearance information to guide motion estimation to recover physiological signals. PhysNet [10]
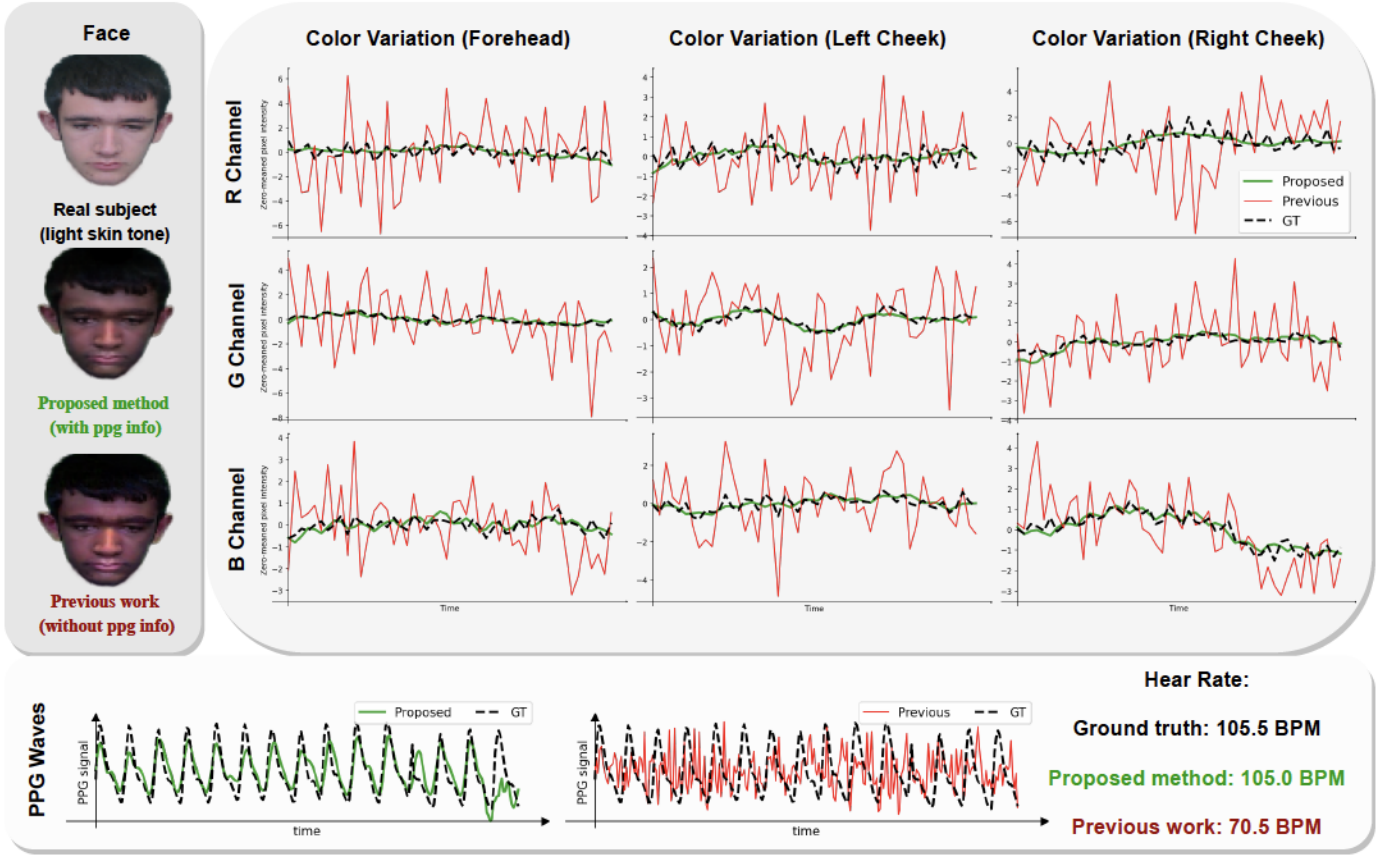
Fig. 2. **The proposed method successfully incorporates pulsatile signals into the generated videos, while the existing work [14] only focuses on the visual appearance.** For different facial regions, frames generated by the proposed method exhibit similar pixel intensity variations as compared with 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.

captures the temporal correlation of the pulse signals in the rPPG face videos using a 3D spatial-temporal Convolutional Neural Network (3D-CNN) or a Recurrent Neural Network (RNN). While these methods exhibit remarkable performance improvement as compared with model-based solutions, their generalization capability is highly affected by the diversity of the training samples.

2.2 Synthetic Augmentation in Healthcare

Medical images have been widely used in clinics and played a critical role in various clinical applications. Due to the significant cost of collecting high-quality medical images, most datasets are very limited in size, and this has impeded the scientific progress. Traditional data augmentation schemes, such as horizontal/vertical flipping, rotation, translation, are used and have become a standard procedure for training deep neural networks in computer vision applications [36]. However, the diversity of the dataset can not be improved significantly by such schemes. Medical image synthesis can be of great benefit to address this problem [37], such as synthetic skin lesion images [38] and synthetic Magnetic Resonance (MR) images for brain tumors [39].

In the rPPG field, McDuff *et al.* [20] use synthetic avatars with blood volume changes to generate rPPG face videos under various settings. The infrastructure for their pipeline is expensive and labor-intensive, which makes it difficult to

scale up their generation process. Tsou *et al.* [22] propose to augment the source rPPG videos with a specified rPPG signal present in another video and show improvement on the heart rate estimation task with the augmented dataset. Their model cannot augment the original dataset with different face appearance, such as skin tones. In contrast, we use a generator to synthesize bio-realistic videos with dark skin tones to reflect the underlying subtle PPG signal variations in a scalable way and show that it is beneficial to improving the measurement of heart rate for remote clinical use.

2.3 Neural Style Transfer for Medicine

Neural style translation has been applied to various medical applications, such as digital histopathological, since the images of the same tissue recorded from different labs and hospitals usually exhibit a large variation in terms of their colors [40], [41], [42], [43]. Color translation frameworks based on neural networks [44], [45], [46] have been proposed to learn not only the certain color distribution but also the corresponding histopathological patterns. The performance of tissue segmentation and classification is improved with the color-augmented histopathological datasets. Inspired by these successful applications, our paper provides a first attempt to bridge the gap between neural style transfer and rPPG for bio-realistic skin tone augmentation.
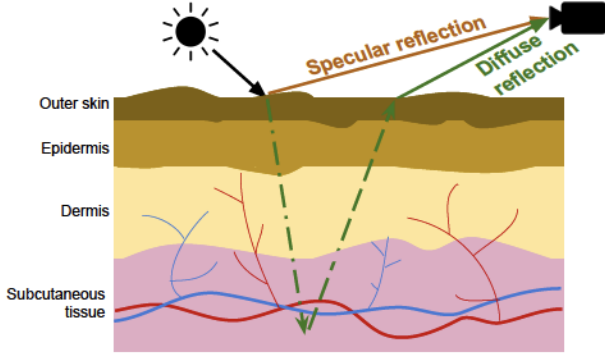
Fig. 3. **Illustration of the dichromatic skin model.** The specular component is due to the reflection from the skin surface, and the diffuse component is related to the absorption and scattering properties of the skin tissues. Our bio-realistic skin tone translation model aims to conduct skin tone translation while preserving the relative variations between BVP and the skin appearance.

3 METHOD

Our bio-realistic skin translation framework is designed to adhere to the light transport analysis of human skins. In Section 3.1, we briefly review the existing skin reflection theory that models pulsatile blood variations. In Section 3.2, we detail our pipeline to translate videos of real subjects with light skin tones to synthetic dark skin tones. The implementation details are provided in Section 3.3.

3.1 Optical Model for Pulsatile Blood Variations

Under the assumption of a light source with a constant spectral composition and varying intensity, RGB channels $\mathbf{C}_k(t)$ at the kth skin pixel measured by a remote color camera can be described by the dichromatic reflection model as a time-varying function [30] as illustrated in Figure 3:

$$\mathbf{C}_k(t) = I(t) \cdot \left(\mathbf{v}_s(t) + \mathbf{v}_d(t)\right) + \mathbf{v}_n(t), \quad (1)$$

where $I(t)$ is the luminance intensity level, $\mathbf{v}_s(t)$ and $\mathbf{v}_d(t)$ are the time-varying specular and diffuse reflections respectively, and $\mathbf{v}_n(t)$ is quantization noise. Specular component $\mathbf{v}_s(t)$ in Equation (1) is a result of the mirror-like reflection from the skin surface, which is usually considered to be BVP independent. We can write $\mathbf{v}_s(t)$ as the following equation [30]:

$$\mathbf{v}_s(t) = \mathbf{u}_s \cdot \left(s_0 + s(t)\right), \quad (2)$$

where $\mathbf{u}_s$ is the unit color vector of incident light, s_0 is the stationary part of the specular reflection, and $s(t)$ is varying part of the specular reflection induced by motion. Diffuse reflection $\mathbf{v}_d(t)$ in Equation (1) is related to the absorption and scattering properties of the skin tissues, and its varying component is identified as a key indicator to the blood volume changes [30]:

$$\mathbf{v}_d(t) = \mathbf{u}_d \cdot d_0 + \mathbf{u}_p \cdot p(t), \quad (3)$$

where $\mathbf{u}_d$ is the unit color vector of the skin, d_0 is the stationary reflection strength, $\mathbf{u}_p$ is the relative pulsatile strengths in RGB channels, and $p(t)$ is the pulse signal.

Substituting Equation (2) and Equation (3) into Equation (1), we can write $\mathbf{C}_k(t)$ as follows:

$$\mathbf{C}_k(t) = I(t) \cdot \left(\mathbf{u}_s \cdot (s_0 + s(t)) + \mathbf{u}_d \cdot d_0 + \mathbf{u}_p \cdot p(t)\right) + \mathbf{v}_n(t). \quad (4)$$

The stationary parts of the specular and diffuse components can be combined into a single skin stationary term:

$$\mathbf{u}_c \cdot c_0 = \mathbf{u}_s \cdot s_0 + \mathbf{u}_d \cdot d_0, \quad (5)$$

where $\mathbf{u}_c$ is the unit color vector of the skin reflection, and c_0 denotes the reflection strength. This further simplifies Equation (4) as:

$$\mathbf{C}_k(t) = I_0 \cdot (1 + i(t)) \cdot (\mathbf{u}_c \cdot c_0 + \mathbf{u}_s \cdot s(t) + \mathbf{u}_p \cdot p(t)) + \mathbf{v}_n(t), \quad (6)$$

where $I(t)$ is expressed as the sum of a stationary part I_0 and a time-varying motion-induced part $I_0 \cdot i(t)$. Video-based PPG measurement algorithms aim to estimate the pulse signal $p(t)$ from the pixel intensity $\mathbf{C}_k(t)$ by separating the physiological and non-physiological variations, while the primary focus of this paper is to establish an inverse mapping between $p(t)$ and $\mathbf{C}_k(t)$ for dark-skin realistic human faces in a data-driven manner.

3.2 Bio-realistic Skin Tone Translation

In order to translate real subjects with light skin tones to synthetic subjects with dark skin tones, we utilize two interconnected networks: a video generator G and an rPPG estimator E, as illustrated in Figure 4. We next describe the proposed 3D convolutional video generator, the rPPG estimation network, and our joint optimization scheme.

3.2.1 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). Please refer to the supplementary material for a detailed description of the network architecture.

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 [14] pretrained on VGGFace2 [47] to generate the pseudo target frames $\mathbf{I}_{dark}$. More specifically, the generator *Caucasian-to-African* in [14] is utilized to translate videos of light-skinned subjects in the existing rPPG dataset to dark skin tones.

The generator is first supervised by the L1 distance between the pseudo target frames $\mathbf{I}_{dark}$ and the generated frames $\hat{\mathbf{I}}_{dark} = G(\mathbf{I}_{light})$ to learn the visual appearance of the synthetic dark-skinned subjects. At this stage, the output frames $\hat{\mathbf{I}}_{dark}$ do not contain pulsatile signal, since the target frames $\mathbf{I}_{dark}$ from [14] 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.
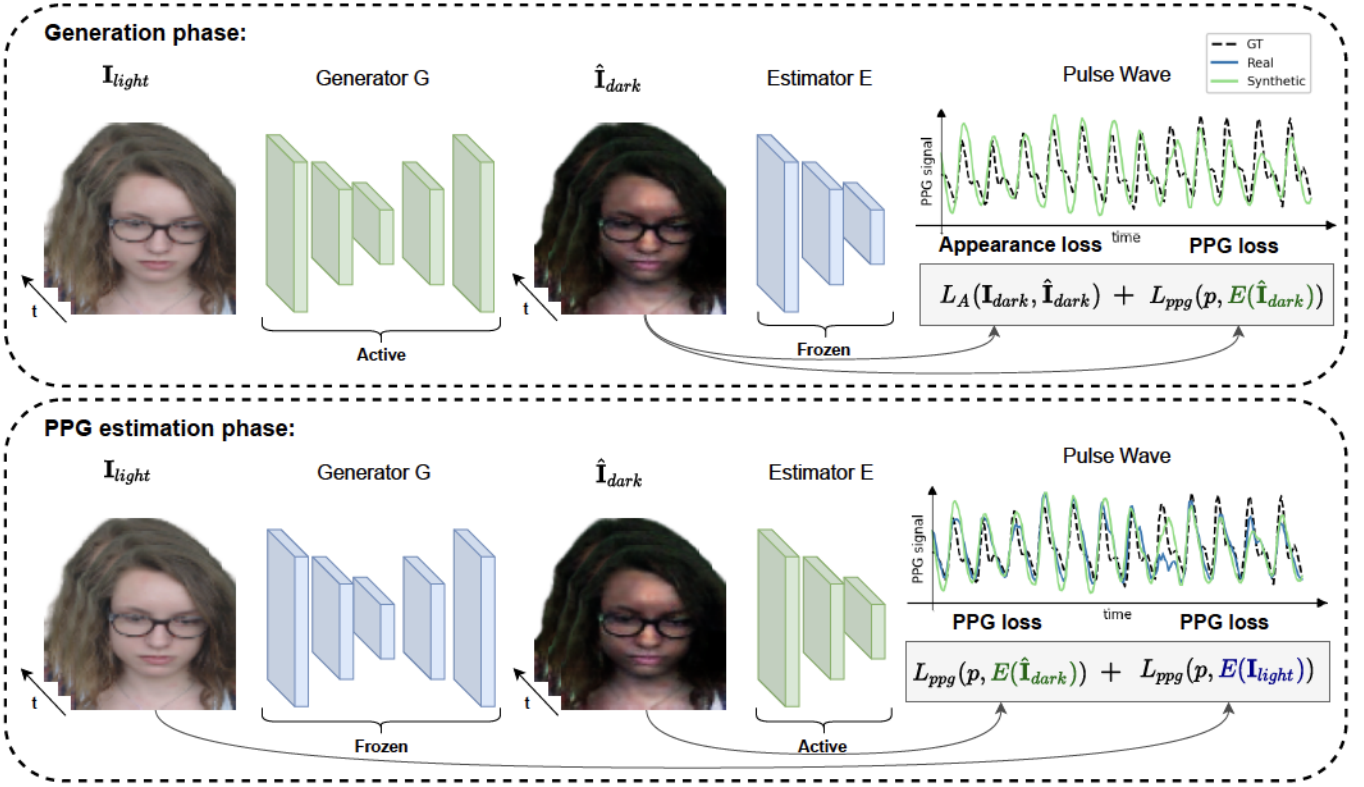
Fig. 4. **Illustration of the proposed joint optimization framework.** Our framework is capable of translating light-skinned facial videos to dark skin tones while maintaining the original pulsatile signals. With a two-phase weight updating scheme, the rPPG estimation network can benefit from the synthetic dark-skinned videos and gradually learn to conduct inference on dark-skinned subjects without accessing real facial videos with dark skin tones.

3.2.2 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 down-sampling purpose. A detailed description of each block can be found in the supplementary material.

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

$$L_{ppg}(p, \hat{p}) = 1 - \frac{T \sum_i p_i \hat{p}_i - \sum_i p_i \sum_i \hat{p}_i}{\sqrt{\left(T \sum_i p_i^2 - (\sum_i p_i)^2\right)\left(T \sum_i (\hat{p}_i)^2 - (\sum_i \hat{p}_i)^2\right)}}, \quad (7)$$

where the summation $\sum_i$ is over the frame length T. This negative Pearson correlation loss has shown to be more effective as compared with the point-wise mean squared error (MSE) loss in the previous work [10]. 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 next part, we detail how to further incorporate the synthetic subjects into the training process.

3.2.3 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 design a joint optimization mechanism to incorporate pulsatile signals into the synthetic videos and improve the generalizability of the rPPG estimator simultaneously.

We use a two-phase weight updating scheme to train the video generator and the rPPG estimator simultaneously. These two phases are alternated within each mini-batch as illustrated in Figure 4. 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:

$$L_G(\mathbf{I}_{light}, p) = L_{ppg}(p, E(\hat{\mathbf{I}}_{dark})) + \lambda * L_A(\mathbf{I}_{dark}, \hat{\mathbf{I}}_{dark}), \quad (8)$$

$$L_A(\mathbf{I}_{dark}, \hat{\mathbf{I}}_{dark}) = \frac{1}{\sum_i z_i} \sum_i z_i |\mathbf{I}_{dark_i} - \hat{\mathbf{I}}_{dark_i}|, \quad (9)$$

$$z_i = \begin{cases} 0 & \text{if } |\mathbf{I}_{dark_i} - \hat{\mathbf{I}}_{dark_i}| < \epsilon, \\ 1 & \text{otherwise} \end{cases}, \quad (10)$$

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 [14] 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:

$$L_E(\mathbf{I}_{light}, \hat{\mathbf{I}}_{dark}), p) = L_{ppg}(p, E(\hat{\mathbf{I}}_{dark})) + L_{ppg}(p, E(\mathbf{I}_{light})).$$
(11)

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

3.3 Implementation Details

The facial bounding box for each video is estimated by applying a face detector based on Multitask Cascaded Convolutional Neural Networks (MTCNN) [48] to its first frame, and a square region with 160% width and height of the detected bounding box is cropped and resized to 80×80 using linear interpolation. The learning rate for the generator and the rPPG network are 0.0001 and 0.0003 respectively. The learning rates are modified base on a cosine annealing schedule during training [49]. The networks are initialized with Kaiming initialization [50] with a batch size of two and ReLU activation. We use Adam [51] solver with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The network architectures are implemented with batch normalization [52] in PyTorch [53], and the experiments are conducted on a single NVIDIA Tesla V100 GPU.

4 EXPERIMENTS

To demonstrate the effectiveness of the proposed method, we conduct a comprehensive evaluation on several commonly used rPPG datasets. We describe the datasets for our experiment in Section 4.1, the comparison methods in Section 4.2, and the evaluation metrics in Section 4.3. Some illustration of the generated the synthetic videos is provided in Section 4.4. The performance of different comparison models and the proposed solutions are listed in Section 4.5 and Section 4.6. The bias mitigation analysis is shown in Section 4.7.

4.1 Datasets

4.1.1 UBFC-RPPG [19]:

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

4.1.2 VITAL dataset [54]:

Facial videos are recorded at 1920x1080 pixel resolution and 30 frames per second for 60 subjects at room lighting in the highly compressed MP4 format. Each video is roughly 2 minutes long. A Philips IntelliVue MX800 patient monitor is utilized for ground-truth vital sign monitoring. The subject wears a blood pressure cuff, 5-ECG leads, and a finger pulse oximeter, which is connected to the MX800 unit. Diverse skin tones and varied demographic groups are represented in the dataset. We use 58 subjects in the VITAL dataset (subject 26 and subject 40 are left out due to data errors in the collecting process). For the skin types quantified by Fitzpatrick scales [55], there are 5, 16, 14, 11, 5, 7 subjects respectively from I (lightest) to VI (darkest).

4.2 Comparison Methods

We compare our model with three conventional methods: POS [30], CHROM [29] and ICA [8]. These rPPG baseline methods are implemented based on the publicly available MATLAB toolbox [56], and we follow the procedures in the toolbox to obtain facial pixels of interest, i.e., converting facial frames from RGB to YC_RC_B and identifying skin pixels based on a predefined threshold. We also compare with a data-driven state-of-the-art rPPG algorithm 3D-CNN [22]. It is implemented based on the architecture description as detailed in the original publication.

4.3 Evaluation Metrics

After obtaining the estimated pulse waves from each model, we apply a Butterworth filter to the output signals with cutoff frequencies of 0.7 and 2.5 Hz for heart rate estimation. The filtered waves are divided with sliding windows of 30-second length and 1-second stride, and a heart rate is estimated based on the position of the peak frequency for each window. For each subject, four error metrics are calculated and averaged over all windows. The four metrics include mean absolute error (MAE), root mean square error (RMSE), Pearson's correlation coefficient (PCC) between the estimated hear rate and the ground-truth hear rate, and signal-to-noise ratio (SNR) of the estimated PPG waves. The ground-truth HR for UBFC-RPPG is obtained by applying the same procedures as described above to the ground-truth pulse waves, and the ground-truth HR for the VITAL dataset is obtained from the MX800 patient monitor through ECG signals. Details of these metrics are provided as follows:

$$\text{MAE} = \frac{\sum_{i=1}^{N} \left| \text{HR}_i - \hat{\text{HR}}_i \right|}{N}, \tag{12}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} \left(\text{HR}_i - \hat{\text{HR}}_i \right)^2}{N}}, \tag{13}$$

$$\text{PCC} = \frac{T \sum_i p_i \hat{p}_i - \sum_i p_i \sum_i \hat{p}_i}{\sqrt{\left(T \sum_i p_i^2 - (\sum_i p_i)^2 \right) \left(T \sum_i (\hat{p}_i)^2 - (\sum_i \hat{p}_i)^2 \right)}}, \tag{14}$$

$$\text{SNR} = 10 \log_{10} \left(\frac{\sum_{f=0.75}^{2.5} \left(U_t(f) \hat{S}(f) \right)^2}{\sum_{f=0.75}^{2.5} \left((1 - U_t(f)) \hat{S}(f) \right)^2} \right), \tag{15}$$

where HR is the ground-truth heart rate, $\hat{\text{HR}}$ is the estimated heart rate, 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 estimated 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.

4.4 Generating Synthetic Dark-skinned Subjects

We demonstrate the superiority of our proposed method with empirical results on UBFC-RPPG [19] and VITAL [54] for HR estimation using the above four metrics. 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. Some generated frames in the UBFC-RPPG validation set are illustrated in Figure 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.

4.5 Performance on UBFC-RPPG

Performance metrics of different models in the UBFC-RPPG validation set are listed in Table 1. We list the HR estimation accuracy of PRN trained with the proposed joint optimization pipeline (referred as PRN augmented), real samples (referred as PRN w/ Real), and synthetic samples (referred as PRN w/ Synth). The synthetic samples are generated by our generator G through translating the real samples in the UBFC-RPPG training set when the joint optimization converges. As a comparison, we also include the performance of a state-of-the-art deep learning model 3D-CNN [22] that is trained with both real and synthetic samples (referred as

TABLE 1
Performance of HR estimation on UBFC-RPPG. Boldface font represents the preferred results.

Method	MAE↓	RMSE↓	PCC↑	SNR↑
PRN augmented	**0.68**	**1.31**	0.86	5.76
PRN w/ Real	0.75	1.64	0.83	**7.91**
PRN w/ Synth	4.32	6.56	0.54	-1.93
3D-CNN [22] w/ Real&Synth	0.89	1.66	**0.88**	7.74
3D-CNN [22] w/ Real	1.09	1.91	0.84	7.80
3D-CNN [22] w/ Synth	0.95	1.80	0.82	3.48
POS [30]	3.69	5.31	0.75	3.07
CHROM [29]	1.84	3.40	0.77	4.84
ICA [8]	8.28	9.82	0.55	1.45

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 [30], CHROM [29] and ICA [8]) are also provided in the table.

Notably, the proposed PRN architecture has already outperformed other rPPG estimation methods even without synthetic skin color augmentation. More specifically, the proposed PRN has around 31% improvement on MAE and around 14% improvement on RMSE over the state-of-the-art 3D-CNN using real training samples. With the synthetic augmentation, the performance of PRN can be further improved. PRN trained with augmentation achieves 9% improvement on MAE (from 0.75 BPM to 0.68 BPM) as compared with PRN trained with just real samples. This suggests that even for UBFC-RPPG dataset which is overwhelmed by subjects with light skin tones, increasing the diversity of training samples is still able to enhance the performance. This finding is consistent with the recent research [57] 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 the existing data-driven rPPG estimation algorithm without conducting the joint optimization process again to adapt to another new network architecture.

4.6 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. VITAL dataset contains different subjects and is captured in an entirely different environment as compared to the UBFC-RPPG dataset. This type of cross-dataset verification can provide more visibility on the generalization capability of the models.
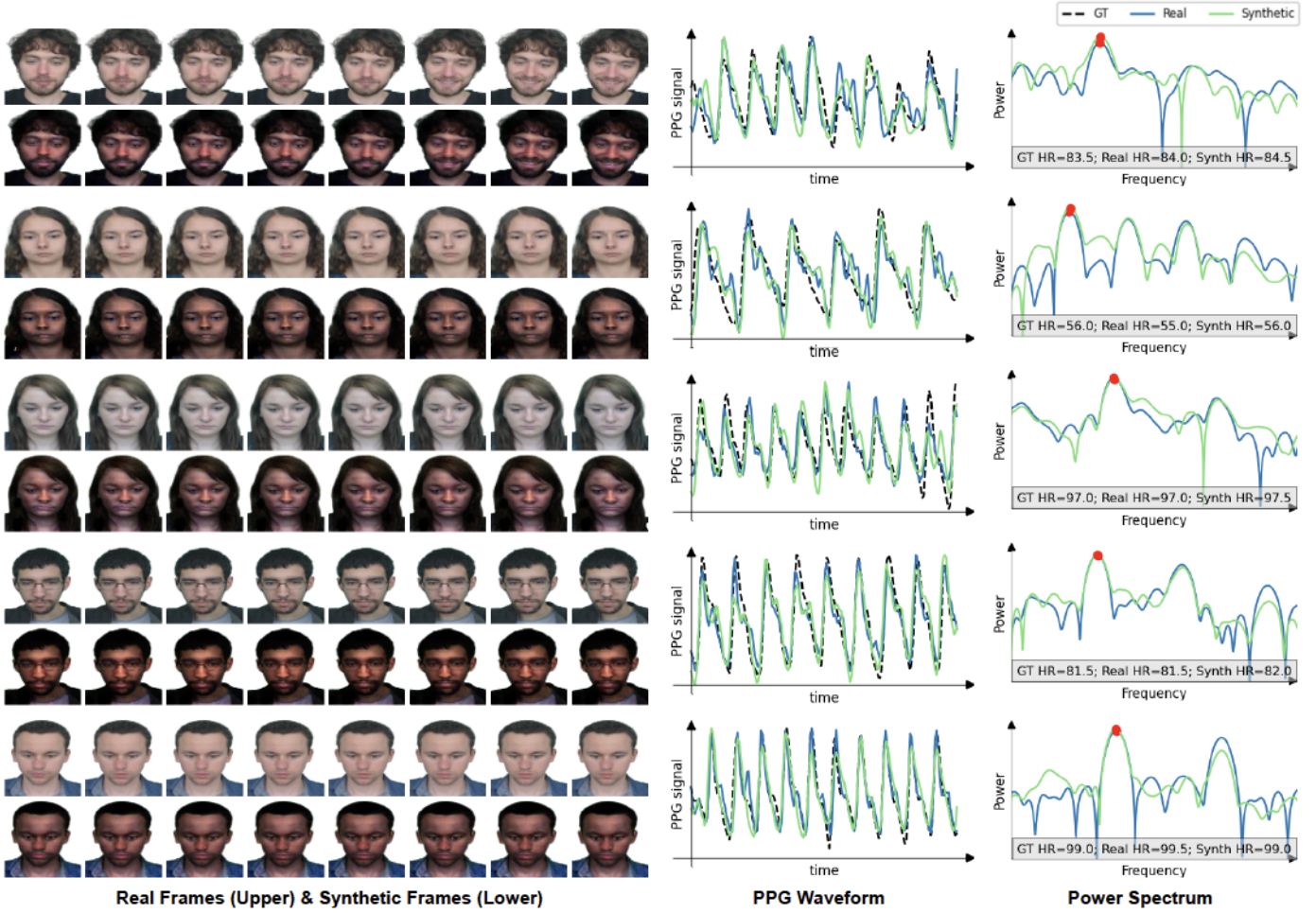
Fig. 5. **Illustration of real frames and the corresponding synthetic frames in the UBFC-RPPG dataset.** Our proposed framework has successfully incorporated pulsatile signals when translating the skin color. The estimated pulse waves from PRN exhibit high correlation to the ground-truth waves, and the heart rates are preserved in the frequency domain.

Similarly, we report MAE, RMSE, PCC, and SNR of various models trained with real and synthetic samples in Table 2. Since VITAL dataset contains testing subjects of diverse skin tones with the associated Fitzpatrick scale labels (F1-6), we group the subjects into three categories, i.e., F1-2 (light skin color), F3-4 (medium skin color), and F5-6 (dark skin color), to measure the performance across different demographic groups.

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

synthetic data shows lower estimation errors as compared with real data, and the errors are reversed for the light skin color group. This validates the fact that data-driven rPPG estimation models are heavily impacted by the skin color distribution of training samples, and it is critical to create a diverse and balanced training set for generalizability and real-world deployment of rPPG algorithms.

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

POS [30], CHROM [29] and ICA [8] show relatively large HR estimation errors as compared with the data-driven models, where their MAEs on the light skin color group is usually larger than 4 BPM. Their MAEs are even higher for other groups. Unlike the end-to-end rPPG estima-

Method	F1-2		F3-4		F5-6		Overall	
	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓
PRN augmented	2.37	3.13	**2.95**	**3.82**	4.39	5.98	3.04	4.01
PRN w/ Real	3.38	4.46	4.67	5.83	6.61	8.48	4.60	5.88
PRN w/ Synth	4.27	6.01	4.52	6.18	5.64	8.33	4.66	6.57
3D-CNN [22] w/ Real&Synth	**2.32**	**3.11**	3.18	4.09	5.45	7.07	3.34	4.35
3D-CNN [22] w/ Real	3.31	4.64	5.86	6.78	7.07	8.89	5.19	6.44
3D-CNN [22] w/ Synth	3.88	5.23	4.68	6.07	7.81	9.88	5.04	6.56
POS [30]	4.97	6.28	5.36	6.86	7.25	9.74	5.69	7.25
CHROM [29]	6.51	8.92	5.01	6.38	7.83	14.56	6.14	8.99
ICA [8]	7.65	9.66	7.14	8.40	5.75	7.31	7.04	8.63
	F1-2		F3-4		F5-6		Overall	
	PCC↑	SNR↑	PCC↑	SNR↑	PCC↑	SNR↑	PCC↑	SNR↑
PRN augmented	0.40	3.45	0.63	5.73	0.30	-3.38	**0.48**	3.02
PRN (w/ Real)	0.36	0.32	0.50	0.03	0.08	-7.00	0.36	-1.32
PRN (w/ Synth)	0.29	-0.64	0.42	-0.44	0.11	-6.35	0.31	-1.74
3D-CNN [22] (w/ Real&Synth)	**0.42**	**3.96**	**0.65**	5.21	0.17	-4.84	0.47	2.68
3D-CNN [22] (w/ Real)	0.30	-0.61	0.48	-1.26	0.11	-8.26	0.34	-2.47
3D-CNN [22] (w/ Synth)	0.07	-2.04	0.38	-1.34	0.10	-6.38	0.21	-2.64
POS [30]	0.26	-2.22	0.42	-1.04	0.27	-5.59	0.33	-2.41
CHROM [29]	0.15	-2.14	0.46	-1.11	-0.10	-5.53	0.23	-2.40
ICA [8]	0.24	-2.06	0.32	-1.73	0.06	-5.04	0.23	-2.53

tion networks, these conventional methods usually require preprocessing steps which may diminish the subtle color changes on the face and degrade the performance. Besides, these models need to average the pixel intensities over the skin region, and this might be a sub-optimal solution since skin pixels at different facial regions can contribute differently to the pulse signals.

The cross-dataset experiment indicates that the improvement of our proposed framework is more substantial as compared with intra-dataset evaluation where all the samples are obtained within the same environment. This suggests that synthetic videos can provide more significant benefit by diversifying the training samples when there exist some data distribution shifts between real training and testing videos. This finding is also consistent with the observation for ray-tracing based augmentation method [20]. Synthetic augmentation techniques thus become particularly effective for cross-domain learning and can improve the generalization capability of HR estimation for real-world applications.

Fig. 6. **Synthetic dark-skinned videos can help to reduce bias in HR estimation.** The augmented PRN and the 3D-CNN [22] 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.

4.7 Bias Mitigation

It is critical for an algorithm to have consistent performance across different demographic groups in real-world medical deployment. To quantify the performance gap for each group, we use the standard deviation of MAE and RMSE for each Fitzpatrick scale as the measurement. This measurement has also been used in some prior work [14], [20]. The standard deviation for each method in the VITAL dataset is illustrated in Figure 6, together with a sample portrait for each skin scale from F1 to F6. Conventional POS method exhibits large variation (MAE: 2.66 BPM, RMSE: 3.19 BPM) across different Fitzpatrick scales, while the jointly optimized PRN shows the lowest bias (MAE: 1.53 BPM, RMSE:
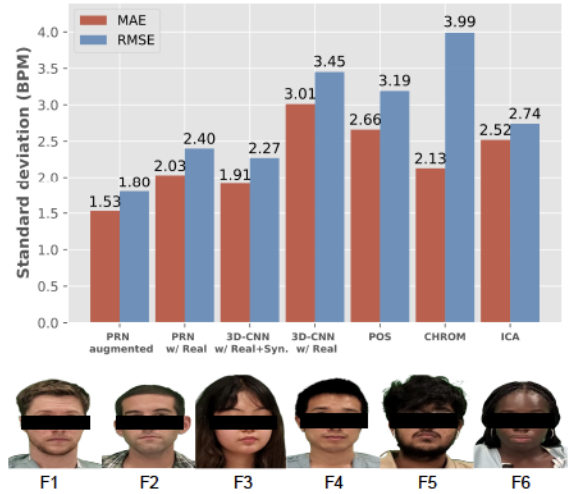
1.80 BPM) as compared with all the conventional methods. In contrast to PRN trained with just real samples (MAE: 2.03 BPM), the augmented training offers a 25% improvement of bias mitigation among different groups while simultaneously improving the overall performance of all the groups. This suggests our joint training framework can provide a more desired trade-off between performance and bias. For 3D-CNN, the standard deviations for MAE and RMSE are also reduced by adding the synthetic samples into the training set. We attribute this improvement to the more diverse and balanced dataset augmented by our generator.

5 Discussion and Limitations

The paper has made an attempt to tackle bias in rPPG. The lack of dark-skinned subjects in existing rPPG datasets (MMSE-HR, AFRL, and UBFC-RPPG have roughly 10%, 0%, and 5% dark-skinned subjects) has produced unwanted bias against some underrepresented groups, and there exist several practical constraints towards collecting a large-scale balanced dataset for rPPG. To address this issue, an attempt is proposed to translate facial frames from light-skinned subjects to dark skin tones while preserving the subtle color variations corresponding to the pulsatile signals. The jointly optimized rPPG estimator can outperform the existing state-of-the-art methods with reduced estimation bias across different demographic groups. More specifically, PRN trained with augmentation has around 38% reduction in MAE for the dark-skinned group along with 49% improvement on bias mitigation in the VITAL dataset, as compared with 3D-CNN [22] trained with just real samples. Our generated synthetic videos maintain both photo-realistic and bio-realistic features and can be directly used to improve the performance of the existing deep learning rPPG estimation model.

Video synthesis, such as deepfakes, has raised public concerns in the community [58]. 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 medical applications.

We now discuss a few limitations of this work. Our current pipeline is an initial 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 requirements of various subsequent rPPG estimation models without frame size interpolation. The primary goal of this work is to overcome the shortness of real dark-skinned subjects by synthetic generation. Therefore, the current framework is designed based on Caucasian-to-African translation. Future work may extend this to other appropriate racial group(s) to further diversify the training data. Our framework replies on a generator designed based on 3D convolutions, where its output is not directly supervised by videos from real dark-skinned subjects. While the improved heart rate estimation results support the effectiveness of the proposed solution, inductively generalizing claims in this paper of reducing bias need to be validated in much larger-scale clinical trials than what are possible in an academic paper introducing a new method.

In this paper, we used existing metrics to evaluate rPPG quality, such as standard waveform measures of MAE and RMSE. These metrics were carefully chosen so they are regressable against previous rPPG papers. It could be that the metrics could themselves be biased (e.g., if the rPPG waveform has a unique shape amongst demographics and/or if the synthetic data has an unusual shape). Ultimately, we felt more comfortable using the same error metrics used in previous works, to aid in comparisons. Identifying biases in a metric and/or proposing solutions requires thought and experiment, particularly when the context involves fairness. An option for future work is to evaluate if there is possibly a better metric for the rPPG problem.

6 Conclusion

To conclude, we perform appearance transfer while retaining the subtle transient characteristics of realistic blood flow. During training, we demonstrate that heart rate estimation can be improved in *both* performance and equity. Other than heart rate estimation, we hope that future work can apply physiologically-sound appearance transfer to other vital signs, such as blood pressure, blood oxygen saturation, and respiration rate.

Ethics Statement: We envision positive benefits of bio-realistic avatars, as a way to expand training datasets for medical instruments, like remote vital sign monitors. We condemn the use of this technique to fool DeepFake catchers.

References

[1] J. P. Weiner, S. Bandeian, E. Hatef, D. Lans, A. Liu, and K. 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*, vol. 4, no. 3, pp. e212 618–e212 618, 2021.

[2] B. R. Bloem, E. R. Dorsey, and M. S. Okun, "The coronavirus disease 2019 crisis as catalyst for telemedicine for chronic neurological disorders," *JAMA neurology*, vol. 77, no. 8, pp. 927–928, 2020.

[3] X. Song, X. Liu, and C. Wang, "The role of telemedicine during the covid-19 epidemic in china—experience from shandong province," pp. 1–4, 2020.

[4] A. J. Bokolo, "Exploring the adoption of telemedicine and virtual software for care of outpatients during and after covid-19 pandemic," *Irish Journal of Medical Science (1971-)*, vol. 190, no. 1, pp. 1–10, 2021.

[5] M. Villarroel, S. Chaichulee, J. Jorge, S. Davis, G. Green, C. Arteta, A. Zisserman, K. McCormick, P. Watkinson, and L. Tarassenko, "Non-contact physiological monitoring of preterm infants in the neonatal intensive care unit," *NPJ digital medicine*, vol. 2, no. 1, pp. 1–18, 2019.

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

[7] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethys-mographic imaging using ambient light." *Optics express*, vol. 16, no. 26, pp. 21 434–21 445, 2008.

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

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

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

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

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

[13] X. Li, I. Alikhani, J. Shi, T. Seppanen, J. Junttila, K. Majamaa-Voltti, M. Tulppo, and G. 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)*. IEEE, 2018, pp. 242–249.

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

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

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

[17] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang *et al.*, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3438–3446.

[18] J. R. Estepp, E. B. Blackford, and C. 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)*. IEEE, 2014, pp. 1462–1469.

[19] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois, "Unsupervised skin tissue segmentation for remote photoplethys-mography," *Pattern Recognition Letters*, vol. 124, pp. 82–90, 2019.

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

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

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

[23] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

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

[25] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind

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

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

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

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

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

[31] M. Kumar, A. Veeraraghavan, and A. Sabharwal, "Distanceppg: Robust non-contact vital signs monitoring using a camera," *Biomedical optics express*, vol. 6, no. 5, pp. 1565–1588, 2015.

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

[33] R. Song, H. Chen, J. Cheng, C. Li, Y. Liu, and X. Chen, "Pulsegan: Learning to generate realistic pulse waveforms in remote photoplethysmography," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1373–1384, 2021.

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

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

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[37] D. Nie, R. Trullo, J. Lian, L. Wang, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, "Medical image synthesis with deep convolutional adversarial networks," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 12, pp. 2720–2730, 2018.

[38] Z. Qin, Z. Liu, P. Zhu, and Y. Xue, "A gan-based image synthesis method for skin lesion classification," *Computer Methods and Programs in Biomedicine*, p. 105568, 2020.

[39] M. Foroozandeh and A. Eklund, "Synthesizing brain tumor images and annotations by combining progressive growing gan and spade," *arXiv preprint arXiv:2009.05946*, 2020.

[40] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas, "A method for normalizing histology slides for quantitative analysis," in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE, 2009, pp. 1107–1110.

[41] B. E. Bejnordi, G. Litjens, N. Timofeeva, I. Otte-Höller, A. Homeyer, N. Karssemeijer, and J. A. van der Laak, "Stain specific standardization of whole-slide histopathological images," *IEEE transactions on medical imaging*, vol. 35, no. 2, pp. 404–415, 2015.

[42] X. Li and K. N. Plataniotis, "A complete color normalization approach to histopathology images using color cues computed from saturation-weighted statistics," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 7, pp. 1862–1873, 2015.

[43] A. Badano, C. Revie, A. Casertano, W.-C. Cheng, P. Green, T. Kimpe, E. Krupinski, C. Sisson, S. Skrøvseth, D. Treanor *et al.*, "Consistency and standardization of color in medical imaging: a consensus report," *Journal of digital imaging*, vol. 28, no. 1, pp. 41–52, 2015.

[44] Y. Xiao, E. Decencière, S. Velasco-Forero, H. Burdin, T. Bornschlögl, F. Bernerd, E. Warrick, and T. Baldeweck, "A new color augmentation method for deep learning segmentation of histological images," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 886–890.

[45] H. Cho, S. Lim, G. Choi, and H. Min, "Neural stain-style transfer

source separation." *Optics express*, vol. 18, no. 10, pp. 10 762–10 774, 2010.

learning using gan for histopathological images," *arXiv preprint arXiv:1710.08543*, 2017.

[46] H. Liang, K. N. Plataniotis, and X. Li, "Stain style transfer of histopathology images via structure-preserved generative learning," *arXiv preprint arXiv:2007.12578*, 2020.

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

[48] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

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

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

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

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

[53] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035.

[54] 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.

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

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

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

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

Zhen Wang received the B.E. degree from the University of Electronic Science and Technology of China and the B.E. degree with first class honour from the University of Glasgow. He received his M.S. degree from Columbia University. He is currently pursuing his Ph.D. degree at University of California, Los Angeles. His research interests include computer vision and computational imaging.

Kerim Doruk Karinca currently works at Intrinsic AI, building computer vision systems for robots. He graduated with an M.S. in Computer Science at University of California, Los Angeles, focusing on deep learning for healthcare. Before that, as an undergraduate researcher, he helped develop portable medical AI devices.

Oyku Deniz Bozkurt is an M.S. student in Electrical and Computer Engineering at University of California, Los Angeles. She received her B.S. degree from the department of Electrical Engineering at Georgia Institute of Technology.

Achuta Kadambi received the Ph.D. degree from MIT (2018) and joined UCLA, where he is Assistant Professor of Electrical and Computer Engineering, and of, Computer Science. He has received early career recognitions from NSF (CAREER), DARPA (YFA), ARO (YIP), Forbes (30 under 30), and has cofounded startups in computational imaging and computer vision. Achuta recently co-authored a textbook Computational Imaging, available through MIT Press.

Yunhao Ba received his B.E. (First Class Honours) and M.S. degrees from the Hong Kong Polytechnic University and University of California, Los Angeles. He is currently perusing his Ph.D. degree at University of California, Los Angeles. His research interests include computational imaging, physics-based learning, and imaging through adverse conditions.