Synthetic Generation of Face Videos with Plethysmograph Physiology

Zhen Wang[1]*, Yunhao Ba[1]*, Pradyumna Chari[1], Oyku Deniz Bozkurt[1],
Gianna Brown[2], Parth Patwa[1], Niranjan Vaddi[3], Laleh Jalilian[4], and Achuta Kadambi[1,3]
[1]Department of Electrical and Computer Engineering, UCLA
[2]Department of Bioengineering, UCLA
[3]Department of Computer Science, UCLA
[4]Department of Anesthesiology and Perioperative Medicine, UCLA
{zhenwang, yhba, pradyumnac}@ucla.edu, achuta@ucla.edu

Abstract

Accelerated by telemedicine, advances in Remote Photoplethysmography (rPPG) are beginning to offer a viable path toward non-contact physiological measurement. Unfortunately, the datasets for rPPG are limited as they require videos of the human face paired with ground-truth, synchronized heart rate data from a medical-grade health monitor. Also troubling is that the datasets are not inclusive of diverse populations, i.e., current real rPPG facial video datasets are imbalanced in terms of races or skin tones, leading to accuracy disparities on different demographic groups. This paper proposes a scalable biophysical learning based method to generate physio-realistic synthetic rPPG videos given any reference image and target rPPG signal and shows that it could further improve the state-of-the-art physiological measurement and reduce the bias among different groups. We also collect the largest rPPG dataset of its kind (UCLA-rPPG) with a diverse presence of subject skin tones, in the hope that this could serve as a benchmark dataset for different skin tones in this area and ensure that advances of the technique can benefit all people for healthcare equity. The dataset is available at https://visual.ee.ucla.edu/rppg_avatars.htm/.

1. Introduction

Photoplethysmography (PPG) is an optical technique that measures vital signs such as Blood Volume Pulse (BVP) by detecting the light reflected or transmitted through the skin. Remote Photoplethysmography (rPPG) based on camera videos has several advantages over the conventional PPG methods. It is non-contact thus allowing for a wide range of applications in e.g. neonatal monitoring [15, 41]. It causes no skin irratation and prevents the risk of developing into infection for those whose skins are fragile and sensitive to the adhesive sensing electrodes. As cameras are ubiquitous in electronic device nowadays (such as smartphones, laptops), rPPG can be applied for telemedicine with patients at home and no equipment setup is needed [1]. Camera-based rPPG techniques have also been used in other applications such as driver monitoring [30] and face anti-spoofing [19].

Dataset	# Subjects	# Videos	Demo. diversity	Orig. Videos Free Avail.
AFRL [10]	25	300	✗	✓
MMSE-HR [45]	40	102	✗	✗
UBFC-rPPG [6]	42	42	✗	✓
UBFC-Phys [25]	56	168	✗	✓
VIPL-HR [26]	107	3130	✗	✓
Dasari *et al.* [8]	140	140	✗	✗
Our synthetic method	480	480	High	✓

Table 1. **Comparison of rPPG real datasets and our proposed synthetic dataset.** Real datasets are limited by the number of subjects and videos and demographic diversity, while synthetic datasets have easy control of these attributes.

Traditional rPPG methods either use Blind Source Separation (BSS) [17, 36, 37] or models based on skin reflectance [9, 16, 43] to separate out the pulse signal from the color changes on the face. These methods usually require pre-processing such as face tracking, registration and skin segmentation. More recently, deep learning and convolutional neural networks (CNN) have been more popular due to its expressiveness and flexibility [7,20,21,27,28,44]. CNNs learn the mapping between the pulse signal and the color variations with end-to-end supervised training on the labeled dataset, thus achieving state-of-the-art performance on the vital sign detection. However, the performance of data-driven rPPG networks hinges on the quality of the dataset [31].

There are some efforts (as shown in Tab. 1) on collect-

*Equal contribution.

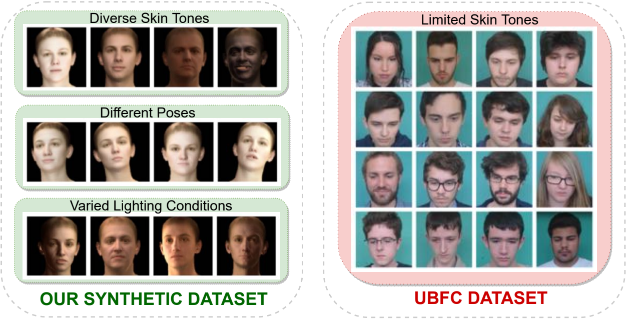

Figure 1. **Our proposed scalable model can generate synthetic rPPG videos with diverse attributes such as poses, skin tones and lighting conditions.** In contrast, existing real datasets (e.g. UBFC) only contain limited races.

ing a large rPPG dataset for better physiological measurement. Nonetheless, there exists several practical constraints towards collecting real patient data for medical purposes. These include: (1) demographic biases (such as race biases) in society that translate to data. As pointed out in [5], a diverse rPPG dataset may not be accessible for some countries/regions due to geographical distribution of skin colors as reflected in their skin tone world map for indigenous people. (2) necessity of intrusive/semi-intrusive traditional methods for collection of data, (3) patient privacy concerns, and (4) requirement of medical-grade sensors to generate the data. Hence, there is a pressing need for the concept of 'digital patients': physiologically accurate graphical renders that may assist development of algorithms and techniques for improvement of diagnostics and healthcare. We provide such a neural rendering instantiation in the rPPG field.

For decades, computer graphics has been a driving force for the visuals we see in movies and games. Imagine if we could harness computer graphics techniques to create not just photorealistic humans, but *physio-realistic* humans. We combine modalities of image and waveform to learn to generate a realistic video that can reflect underlying BVP variations as specified by the input waveform. We achieve this by an interpretable manipulation of UV albedo map obtained from the 3D Morphable Face Model (3DMM) [11]. Our model can generate rPPG videos with large variation of various attributes such as facial appearance and expression, head motions and environmental lighting as shown in Fig. 1.

1.1. Contributions

We summarize our contributions as follows:

- We propose a scalable physics-based learning model that can render realistic rPPG videos with high fidelity with respect to underlying blood volume variations.
- The synthetically generated videos can be directly utilized to improve the performance of the state-of-the-art deep rPPG methods. Notably, the corresponding rendering model can also be deployed to generate data for underrepresented groups, which provides an effective method to further mitigate the demographic bias in rPPG frameworks.
- To facilitate the rPPG research, we release a real rPPG dataset called UCLA-rPPG that contains diverse skin tones. This dataset can be used to benchmark performance across different demographic groups in this area.

2. Related Work

rPPG methods: rPPG techniques aim to recover the blood volume change in the skin that is synchronous with the heart rate from the subtle color variations captured by a camera. Signal decomposition methods include [17] that utilizes Principal Component Analysis (PCA) on the raw traces and chooses the decomposed signal with the largest variance as the pulse signals and Independent Component Analysis (ICA) [23, 36] that demixes the raw signals and determines the separated signals with largest periodicity as the pulse. PCA and ICA are purely statistical approaches that do not use any prior information unique to rPPG problems. A chrominance-based method (CHROM) [9] is proposed to extract the blood volume pulse by assuming a standardized skin-color to white-balance the image and then linearly combine the chrominance signals. Plane Orthogonal to Skin-tone (POS) [43] projects the temporally normalized raw traces onto a plane that is orthogonal to the light intensity change, thus canceling out the effect of that. CNNs have achieved state-of-the-art results on vital sign detection due to their flexibility [5, 7, 20, 21, 27, 28, 44]. The representation for rPPG estimation can be efficiently learned in an end-to-end manner with the annotated datasets instead of handcrafted features for traditional methods. We use two representative work PhysNet [44] and PRN [5] in our experiments to demonstrate the performance of the rPPG models on both real and synthetic datasets.

Real rPPG datasets: There are many efforts on collecting real datasets for more accurate physiological sensing [6, 8, 10, 25, 26, 45]. However, these datasets are usually very limited in the number of subject participants and also biased towards certain demographic group. Some work includes subject with darker skin types, but the number is still very limited [45]. Making machine learning methods equitable is of increasing interest in medical domain [14, 46]. There is a lack of a benchmark dataset to measure the performance of various rPPG methods on diverse skin tones, especially dark skin tones in rPPG area. Dasari *et al.* [8] proposed a dataset that only contains dark skin tones. How-

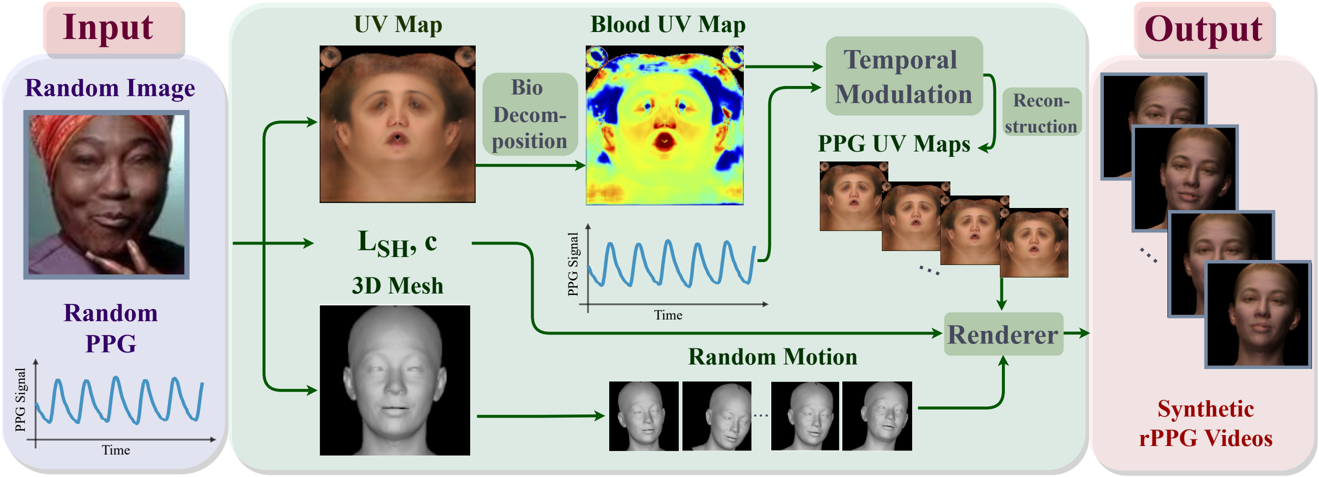

Figure 2. **Pipeline of our cross-modal synthetic generation model that can generate rPPG face videos given any face image and target rPPG signal as input.** The input image is encoded into UV albedo map, 3D mesh, illumination model L_{SH} and camera model c. We then decompose the UV albedo map into blood map, vary the UV blood map according to the target rPPG signal and generate the modified PPG UV maps. The modified PPG UV map that contains the target pulse signal variation is combined with L_{SH}, c to render the final frames with randomized motion.

ever, the actual videos are not shared but the color space values of skin region of interest. The current best-performing deep learning algorithms require sizeable input data. The rPPG model trained on such a biased dataset may easily disadvantage certain underrepresented groups in the dataset. The lack of such a benchmark dataset to systematically and rigorously evaluate various methods on diverse skin tones makes it hard to ensure that the rPPG methods deployed into the society would not cause biases against certain groups that are underrepresented. Our real dataset represents a first step towards filling this gap.

Synthetic generation of rPPG videos: The real rPPG dataset construction is a laborious process and generally takes a large amount of time for collection and administrative work for Institutional Review Board (IRB) approval. Therefore, it is tempting to have a scalable method that can generate large-scale synthetic rPPG datasets for data augmentation. Realizing the difficulty of this, there are a few groups working on generating synthetic rPPG facial videos to augment real data [5, 24, 32, 40]. Mcduff *et al.* [24] propose to render rPPG face videos using facial avatars and simulate the blood volume change with Blender. However, as discussed in the limitation of their method, the rendering of a frame is extremely slow (20 seconds per frame), thus preventing synthetic generation of large-scale videos. The initial overhead for creating the pipeline is also expensive and labor-intensive. A skin tone augmentation method is proposed in [5] where they use a generative neural network to transfer light skin tones to dark skin tones while retaining the pulsatile signals so that the performance on dark skin tones can be improved with the augmented dataset more balanced. Like the other augmentation method on rPPG signals [40], they are both limited as they can only be utilized on current datasets and have to be retrained with new datasets. In contrast, our synthetic generation method can generate diverse appearance with any in-the-wild image and target rPPG signal as input and the generation is merely a forward pass of the neural network.

3. Methods

In this section, we propose a scalable method that can generate synthetic dataset with any given reference image and target rPPG signal in Sec. 3.1. The generated videos can be used to train the state-of-the-art rPPG networks, which we introduce in Sec. 3.2.

3.1. Synthesizing Biorealistic Face Videos

We first describe the 3DMM model used to obtain the facial albedo maps and then demonstrate how to further obtain facial blood maps from the extracted albedo by analyzing light transport in the skin. Details about how to generate synthetic facial videos with the decomposed blood maps and the source of the input facial images and PPG waveforms are also provided in this section. Please see Fig. 2 for an illustration of the entire synthetic generation pipeline.

Non-linear 3DMM: To generate faces with different poses, illuminations and desirable rPPG signal variations, we have to infer the 3D shape and albedo parameters of the face. We use DECA [11] to predict subject-specific albedo, shape, pose, and lighting parameters from an image. In details, it uses a statistical 3D head model FLAME [18] to output a mesh M with $n = 5023$ vertices. The camera model $\mathbf{c}$ is learned to map the mesh M to image space. Since there is no appearance model in FLAME, the linear albedo subspace of Basel Face Model (BFM) [34] is used and the UV layout of BFM is converted to be compatible with FLAME. It outputs a UV albedo map A with a learnable coefficient $\boldsymbol{\alpha}$. By expressing illumination model as the Spherical Harmonics (SH) [39], the shaded face image can be represented as the following equation:

$$B\left(\boldsymbol{\alpha}, \mathbf{l}, N_{uv}\right)_{i,j} = A(\boldsymbol{\alpha})_{i,j} \odot \sum_{k=1}^{9} \mathbf{l}_k H_k\left(N_{i,j}\right), \tag{1}$$

where H_k is the SH basis, $\mathbf{l}_k$ are the corresponding coefficients and $\odot$ denotes the Hadamard product. $N_{i,j}$ is the normal map expressed in the UV form. The final texture image is obtained by rendering the image using the mesh M, shaded image B, and the camera model $\mathbf{c}$ through a rendering function $\mathcal{R}(\cdot)$:

$$I_r = \mathcal{R}(M, B, \mathbf{c}). \tag{2}$$

As rPPG is essentially the change of blood volume in the face, our idea is to first obtain the spatial concentration of blood f_{blood} of the UV albedo A and then temporally modulate the UV blood albedo map in a way that is consistent with the rPPG signals. We will next show how this biophysically interpretable manipulation is achieved.

Light transport in the skin: In order to obtain blood map f_{blood} on the face, we first study light transport in the skin to build the connection between face albedo and f_{blood}. Following a spectral image formation model, the original UV face albedo A_c with $c \in \{R, G, B\}$ is reconstructed by integrating the product of the camera spectral sensitivities S_c, the spectral reflectance R, and the spectral power distribution of the illuminant E over wavelength λ [2]:

$$A_c = \int_{\lambda} E(\lambda) R(f_{\text{mel}}, f_{\text{blood}}, \lambda) S_c(\lambda) d\lambda. \tag{3}$$

An optical skin reflectance model [4] with hemoglobin f_{blood} and melanin map f_{mel} as parameters is utilized to define the wavelength-dependent skin reflectance $R(f_{\text{mel}}, f_{\text{blood}}, \lambda)$. Specifically, we assume a two-layer skin model that characterizes the transmission through the epidermis $T_{\text{epidermis}}$ and reflection from the dermis R_{dermis}:

$$R\left(f_{\text{mel}}, f_{\text{blood}}, \lambda\right) = T_{\text{epidermis}}\left(f_{\text{mel}}, \lambda\right)^2 R_{\text{dermis}}\left(f_{\text{blood}}, \lambda\right). \tag{4}$$

The transmittance in epidermis is modeled by Lambert-Beer law [38] as light not absorbed by the melanin in this layer is propagated to the dermis [3]:

$$T_{\text{epidermis}}(f_{\text{mel}}, \lambda) = e^{-\mu_{a.\text{epidermis}}(f_{\text{mel}}, \lambda)}, \tag{5}$$

where $\mu_{a.\text{epidermis}}(f_{\text{mel}}, \lambda)$ is the absorption coefficient of the epidermis. More specifically,

$$\mu_{a.\text{epidermis}}(f_{\text{mel}}, \lambda) = f_{\text{mel}} \mu_{a.\text{mel}}(\lambda) + (1 - f_{\text{mel}}) \mu_{\text{skinbaseline}}(\lambda), \tag{6}$$

where $\mu_{a.\text{mel}}$ is the absorption coefficient of melanin and $\mu_{\text{skinbaseline}}$ is baseline skin absorption coefficient.

The reflectance in dermis can be modeled using the Kubelka-Munk theory [13], and the proportion of light remitted from a layer is given by [3]:

$$R_{\text{dermis}}\left(f_{\text{blood}}, \lambda\right) = \frac{\left(1-\beta^2\right)\left(e^{K d_{\text{pd}}} - e^{-K d_{\text{pd}}}\right)}{\left(1+\beta^2\right) e^{K d_{\text{pd}}} - (1-\beta)^2 e^{-K d_{\text{pd}}}}, \tag{7}$$

where d_{pd} is the thickness of the dermis, and K and β are related to the absorption of the medium contained within the dermis (i.e. blood). For simplicity of notation, we drop the dependence of K and β on f_{blood} and λ in Eq. (7).

Biophysical decomposition and variation of UV albedo map: With the light transport theory of the skin, we follow a physics-based learning framework (BioFaceNet [2]) to obtain f_{blood} from albedo A. The wavelengths are discretized into 33 parts from 400nm to 720nm with 10nm equal spacing. We utilize an autoencoder architecture and use a fully-convolutional network as encoder to predict the hemoglobin and melanin maps and fully-connected networks to encode the parameters for lighting E and camera spectral sensitivities S_c. The model-based decoder is then to reconstruct the albedo with all the learned parameters according to Eq. (3).

Different from the previous work [2], we obtain biophysical parameters directly from the UV albedo maps instead of the facial images. This arrangement allows us to model the underlying blood volume changes more precisely regardless of the environmental illumination variations. Our model is trained to minimize the following loss function:

$$\mathcal{L} = w_1 \mathcal{L}_{\text{appearance}} + w_2 \mathcal{L}_{\text{CameraPrior}}, \tag{8}$$

where the appearance loss $\mathcal{L}_{\text{appearance}}$ is the $L2$ distance between the reconstructed UV map A_{linRecon} and the original one in the linear RGB space A_{linRGB}. We convert A to linear space by inverting the Gamma transformation with $\gamma = 2.2$. To make the problem more constrained, we also introduce the additional camera prior loss: $\mathcal{L}_{\text{CameraPrior}} = \|\mathbf{b}\|_2^2$, where $\mathbf{b}$ is the prior for the camera spectral sensitivities. w_1 and w_2 are the weights for the reconstructed loss and camera prior loss, respectively.

To reflect the change of the target rPPG signal on the face, we temporally vary the UV blood map f_{blood} linearly with the target rPPG signal in the test phase. Given the blood map of a reference UV map (e.g. the UV blood map of first frame), we generate the UV blood map of the consequent frames as the multiplication of the UV blood map of the reference frame and a ratio scalar that is calculated as the ratio of p_t (rPPG signal at time t) and p_{ref} (rPPG signal at the reference time). Then the modified UV blood map of each frame that contains the desired rPPG signal is reconstructed using the BioFaceNet decoder to get UV map. The final image is rendered using the UV map combined with illumination and camera model according to Eq. (2).

For the purpose of simulating real-world scenarios where the subject might move in the collection process, we randomize the poses in the generation of the sequence of the frames by adding a small random value to the pose and expression parameter of the previous frame.

Face image dataset: To generate synthetic rPPG videos with diverse face appearances, we use the public in-the-wild face datasets BUPT-Balancedface [42]. It is categorized according to ethnicity (i.e. Caucasian, Indian, Asian and African). We use these images as the reference images for generating the synthetic videos as shown in Fig. 2.

PPG recordings: To synthesize videos of a given input PPG signal, we use PPG waveforms recordings from BIDMC PPG and Respiration Dataset [35]. It contains 53 8-minute contact PPG recordings with sampling frequency 125Hz. We sample it correspondingly with the video frame rate (30Hz) and the first sequences of time length L are used where L is the duration of the generated video.

3.2. Physiological Measurement Networks

We use two state-of-the-art deep rPPG networks PhysNet [44] and PRN [5] to benchmark the performance on both real and synthetic datasets. PhysNet and PRN both utilize 3D convolutional neural networks (3D-CNN) architecture to learn spatio-temporal representation of the rPPG videos and predict the rPPG signal in the facial videos. PRN differs in that it uses residual connection for convolutional layers. They take consecutive frames of length T as the input, and its output is the corresponding BVP value for each input frame. The Negative Pearson loss is used to measure the difference between the ground-truth PPG signal p and the estimated rPPG signal $\hat{p}$:

$$L_{ppg}(p,\hat{p}) = 1 - \frac{T\sum_i p_i\hat{p}_i - \sum_i p_i \sum_i \hat{p}_i}{\sqrt{\left(T\sum_i p_i^2 - (\sum_i p_i)^2\right)\left(T\sum_i \hat{p}_i^2 - (\sum_i \hat{p}_i)^2\right)}}, \quad (9)$$

where all the summation is over the length of frames T.

Implementation details: For the training of BioFaceNet, we use 3000 face albedo images with 750 images in each race. We use 80% images for training and 20% for validation. The weight w_1 and w_2 for the loss is $1e^{-3}$ and $1e^{-4}$ respectively. The learning rate is set as $1e^{-4}$ and the number of epochs is 200. For the generation of synthetic videos, we set the length of generated frames L as 2100.

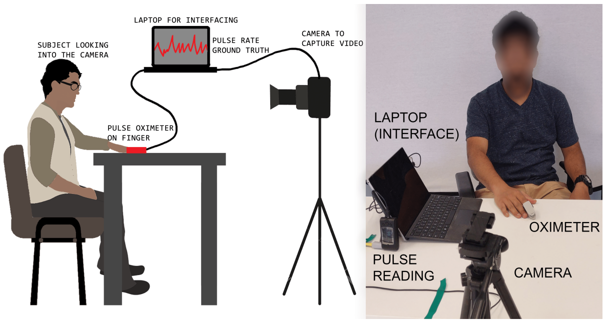

Figure 3. **Experimental setup of data collection.** The subject wears an oximeter on their finger and sits looking directly into the camera. The camera and the oximeter are connected to a laptop to get synchronous video and ground-truth pulse reading. Face blurred to preserve anonymity.

The bounding boxes of the videos are generated using a pretrained Haar cascade face detection model. For each video, one bounding box is detected and increased 60% in each direction before the frames are cropped. To be consistent with the original papers, each frame is resized to 128×128 pixels using bilinear interpolation for PhysNet and 80×80 for PRN. The length of training clips T is 128 for PhysNet and 256 for PRN. The Adam optimizer is used and the learning rate is set as $1e^{-4}$. All the code is implemented in PyTorch [33] and trained on Nvidia V100 GPU.

4. Experiments

In this section, we introduce the datasets we use for the experiments and evaluation protocol in Sec. 4.1. We report and analyze the experimental results for our real dataset in Sec. 4.2 and UBFC-rPPG dataset in Sec. 4.3.

4.1. Datasets and Evaluation Protocol

Our real dataset UCLA-rPPG: In order to benchmark the performance of current rPPG estimation methods, we collect a real dataset of 104 subjects. The setting is faulty for two of them so we dropped their samples. Finally, the dataset consists of 102 subjects of various skin tone, age, gender, ethnicity and race. The Fitzpatrick (FP) skin type scale [12] of the subjects varies from 1-6. For each subject, we record 5 videos of about 1 minute each (1790 frames at 30fps). After removing erroneous videos we have total 503 videos. All the videos in our dataset are uncompressed and synchronized with the ground truth heart rate.

Fig. 3 illustrates the data collection process of our real dataset UCLA-rPPG. The left part of the figure is a cartoon

Method	F1-2		F3-4		F5-6		Overall	
	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓
PhysNet [44] w/ Real&Synth	**0.54**	0.84	0.38	0.70	**1.55**	**2.17**	**0.71**	**1.10**
PhysNet [44] w/ Real	0.81	1.21	0.43	0.77	2.61	3.34	1.06	1.51
PhysNet [44] w/ Synth	1.06	1.52	1.16	1.66	4.96	6.20	2.06	2.73
PRN [5] w/ Real&Synth	**0.54**	**0.79**	**0.36**	**0.65**	3.41	4.09	1.15	1.53
PRN [5] w/ Real	0.65	1.02	0.40	0.71	4.35	5.26	1.43	1.90
PRN [5] w/ Synth	1.47	2.00	0.63	1.07	8.89	9.88	2.87	3.47
POS [43]	3.40	4.34	3.03	3.98	8.07	10.23	4.27	5.49
CHROM [9]	4.06	5.11	3.99	5.25	7.45	9.74	4.79	6.22
ICA [36]	3.75	4.73	3.26	4.19	7.51	9.34	4.35	5.50

	F1-2		F3-4		F5-6		Overall	
	PCC ↑	SNR ↑	PCC ↑	SNR ↑	PCC ↑	SNR ↑	PCC ↑	SNR ↑
PhysNet [44] w/ Real&Synth	**0.84**	**14.40**	**0.80**	**17.11**	**0.60**	**9.19**	**0.76**	**14.45**
PhysNet [44] w/ Real	0.81	13.13	0.77	15.83	0.59	6.54	0.74	12.84
PhysNet [44] w/ Synth	0.74	7.19	0.64	6.11	0.23	-3.33	0.57	4.10
PRN [5] w/ Real&Synth	0.81	12.24	0.79	14.61	0.57	4.84	0.74	11.59
PRN [5] w/ Real	0.77	10.73	0.77	13.22	0.48	2.38	0.70	9.91
PRN [5] w/ Synth	0.69	5.14	0.67	5.27	0.21	-5.81	0.56	2.53
POS [43]	0.50	-0.30	0.42	-0.09	0.27	-5.38	0.41	-1.34
CHROM [9]	0.41	-1.81	0.31	-1.60	0.26	-5.31	0.33	-2.49
ICA [36]	0.45	-0.60	0.38	-0.19	0.27	-5.24	0.37	-1.44

Table 2. **Heart rate estimation results on our real dataset UCLA-rPPG show that both PhysNet and PRN trained with real and synthetic datasets performs consistently better than the models trained with only real data.** The improved performance shows the benefit of the synthetic video dataset we generate.

illustration of the data collection process. The right part of the figure is a photo depicting the actual data collection process. The human subjects wear an oximeter on finger and looks into the camera. Both the camera and the oximeter are connected to a laptop to get synchronous data.

UBFC-rPPG [6]: UBFC-rPPG database contains 42 front facing videos of 42 subjects and corresponding ground truth PPG data recorded from a pulse oximeter. The videos are recorded at 30 frames per second with a resolution of 640 × 480. Each video is roughly one minute long.

Metrics: To evaluate how the heart rate estimates compare with gold-standard heart rates obtained from gold-standard pulse waves, we use the following four metrics Mean absolute error (MAE), Root Mean Squared Error (RMSE), Pearson's Correlation Coefficient (PCC) and Signal-to-Noise Ratio (SNR). Pearson's Correlation Coefficient (PCC) and Signal-to-Noise Ratio (SNR) is defined as in [29].

For traditional baseline methods POS, CHROM and ICA we compare, we use iPhys toolbox [22] to get the estimated rPPG waveforms. The output rPPG signals are normalized by subtracting the mean and dividing by the standard deviation. We filter all the model outputs using a 6th-order Butterworth filter with cut-off frequencies 0.7 and 2.5 Hz. The filtered signals are divided into 30-second windows with 1-second stride and the above four evaluation metrics are calculated on these windows and averaged.

4.2. Performance on UCLA-rPPG

For the study of this work, we split the subjects into three skin tone groups based on the Fitzpatrick skin type [12]. They are light skin tones, consisting of skin tones in the FP 1 and 2 scales, medium skin tones, consisting of skin tones in the FP 3 and 4 scales, and dark skin tones, consisting of skin tones in the FP 5 and 6 scales. This aggregation helps compare experimental results on skin tones more objectively. Since our ultimate goal is to improve the performance on our dataset, we first train on all the synthetic data and then finetune on the real data for the models trained with both real and synthetic data. For training and testing deep rPPG networks PhysNet and PRN on real dataset, we randomly split all the subjects into training, validation and test set with 50%, 10% and 40% and all the test results are averaged on three random splits. The validation set is used to select the best epoch for testing the model.

We report results on the three groups and overall performance using evaluation metrics of MAE, RMSE, PCC and SNR in Tab. 2. In general, models trained with both

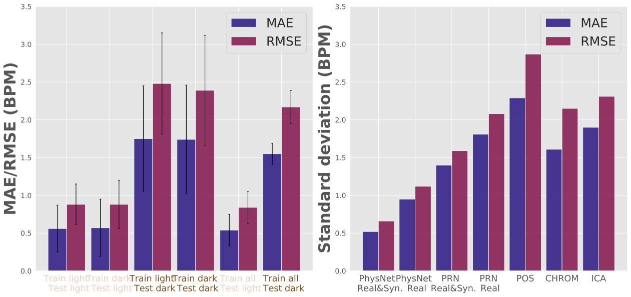

Figure 4. **Left: Ablation study.** The model pre-trained with all synthetic dataset outperforms these pre-trained on either light or dark skin tones alone. **Right: Bias mitigation.** The standard deviation of MAE and RMSE of the deep rPPG models trained with real and synthetic dataset are smaller than real data alone and the traditional models.

real and synthetic data perform consistently better than using real data alone on all the skin tones for all evaluation metrics. PhysNet trained with both real and synthetic data achieved the best overall MAE result 0.71 BPM, with 33% reduction in error compared with PhysNet trained with only real data (1.06 BPM). Notably, the performance improvement is most significant on dark skin stones F5-6 group with 41% and 35% reduction in MAE and RMSE respectively for PhysNet. The same phenomenon is also observed for PRN, where the improvement is most noticeable for darker skin tones. We attribute this to the introduction of synthetic videos we generate in Sec. 3.1. The other two metrics PCC and SNR also validate the superiority of the model trained with both real and synthetic datasets. The results for traditional methods POS, CHROM and ICA are far worse than the deep learning methods, as these methods usually takes the average of all the pixels and ignore the inhomogeneous spatial contribution of the pixels to pulsatile signals.

Bias mitigation: To evaluate the bias of various rPPG methods on subjects with diverse skin tones, we use the standard deviation of the MAE and RMSE results on three skin tone groups. From the right of Fig. 4, we can see the standard deviation of PhysNet with both real and synthetic dataset is the smallest and the MAE disparity among all the three groups are reduced by 45% (from 0.95 BPM to 0.52 BPM) compared with the model trained with only real dataset. Similarly, the standard deviations of both metrics MAE and RMSE for PRN are also reduced for the model trained with both real and synthetic datasets.

Ablation study: We first pre-train the PhysNet with either light skin tones (subjects with race Caucasian in the synthetic dataset) or dark skin tones (subjects with race African), then finetune the model on real dataset and test the model on real subjects with either light skin tones or dark skin tones. From the left of Fig. 4, we can see the model with the pre-trained rPPG network on diverse races are consistently better than these on a single race. The improvement is more obvious on dark skin tones test set. This demonstrates the benefits of a diverse synthetic dataset.

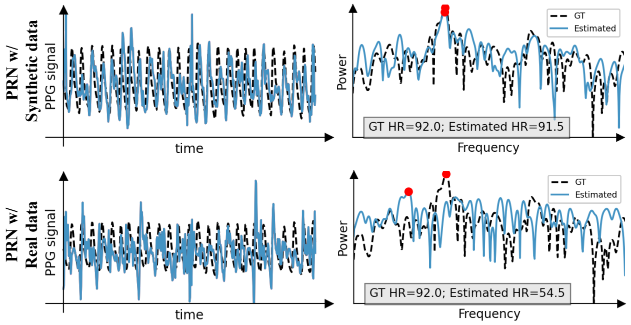

Figure 5. **The example shows that PRN [5] trained with synthetic data (above) generalizes better than PRN trained with real data (bottom) on UBFC-rPPG dataset.** The waves are more aligned with the ground-truth PPG wave (dashed black line) and the power spectrum plot is also more consistent with the ground-truth for the PRN trained with synthetic data.

Method	MAE ↓	RMSE ↓	PCC ↑	SNR ↑
PhysNet [44] w/ Real&Synth	0.90	1.80	**0.84**	6.28
PhysNet [44] w/ Real	1.42	2.74	0.78	5.64
PhysNet [44] w/ Synth	**0.84**	**1.76**	0.83	**6.70**
PRN [5] w/ Real&Synth	1.15	2.38	0.82	5.36
PRN [5] w/ Real	2.36	4.21	0.66	-1.24
PRN [5] w/ Synth	1.09	1.99	0.83	3.00
POS [43]	3.69	5.31	0.75	3.07
CHROM [9]	1.84	3.40	0.77	4.84
ICA [36]	8.28	9.82	0.55	1.45

Table 3. **Performance of HR estimation on UBFC-rPPG shows the superiority of the synthetic datasets.** Boldface font represents the preferred results.

4.3. Performance on UBFC-rPPG

We use the model with best performance on our real dataset to test them on UBFC-rPPG dataset [6] along with the traditional methods. Since this is a cross-dataset evaluation for the model trained on UCLA-rPPG, we test the deep learning models on all the subjects in UBFC-rPPG. All the results with four evaluation metrics are reported in Tab. 3. While the synthetic dataset performs worse than the models trained in our real dataset, the performance gain is more obvious in UBFC dataset. The MAE of PhysNet trained on synthetic dataset achieved the lowest MAE and RMSE (0.84 BPM and 1.76 BPM respectively). The explanation for this observation is that when the distribution of the dataset is similar to the distribution of the test data as in the intra-dataset setting in our real dataset, the benefits of synthetic datasets are not straightforward. The models trained on real

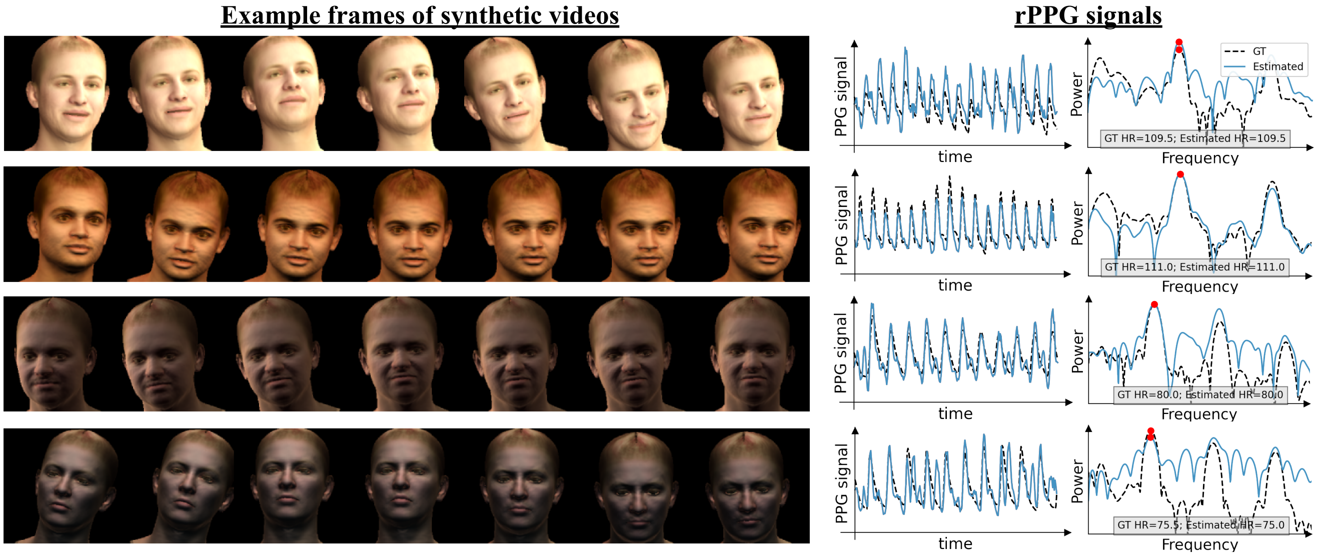

Figure 6. **Illustration of example frames of our generated synthetic videos.** Our proposed framework has successfully incorporated PPG signals into the reference image. The estimated pulse waves from PRN for generated synthetic videos are highly correlated to the ground-truth waves, and the heart rates are preserved as shown in the power spectrum plot.

dataset perform worse on generalizing to another dataset due to different environmental setting such as lighting. We also give a qualitative study in Fig. 5 that shows that the rPPG wave extracted using our synthetic dataset resemble more closely to the ground-truth than that using real dataset. As a result, it gives more accurate heart rate estimation.

4.4. Visualization

As shown in Fig. 6, our model can successfully produce synthetic avatar videos that reflect the associated underlying blood volume changes. Estimated pulse waves from the synthetic videos are closely aligned with the ground truth. The power spectrum of the PPG waves with a clear peak near the gold-standard HR value also validates the effectiveness of the incorporation of pulsatile signals.

5. Discussion

Limitations: Though our synthetic dataset could be used to achieve state-of-the-art results (on UBFC-rPPG datasets, it alone can generalize even better than the model trained on real dataset) for heart rate estimation, the facial appearance is not photo-realistic, which may still degrade the performance due to sim2real gap. We are not focused on modeling the background in the generated videos in this work. However, it is found in [29] that the background can be utilized for better pulsatile signals extraction. Also we vary the UV blood map linearly according to the target rPPG signals in the synthetic generation method. While this yields reasonable empirical results, we believe biophysical model based manipulation of the UV blood map could further improve the performance of the synthetic generation.

Ethics Statement: This paper's novelty is to generate synthetic face videos that are physiologically consistent with heartbeat, and we hope it can be a tool to address some social issues, such as biases around race and gender in medicine. It should also be noted that even though the research here was solely used to improve remote health technologies, it might be used to fool rPPG-based deepfake detectors. We strongly advise against using this technology for such applications.

Conclusion: We propose a method to generate large-scale synthetic rPPG videos with high-fidelity to the underlying rPPG signals. The synthetic generation pipeline enables the scalable generation of rPPG facial videos with any given image and rPPG signal. We validate the effectiveness of the synthetic videos on UCLA-rPPG dataset we collect that contains diverse skin tones and UBFC-rPPG dataset. The experimental results show that the synthetic dataset can improve the performance on both datasets and help reduce the bias among different demographic groups.

Acknowledgements: We thank the members of the Visual Machines Group (VMG) at UCLA for feedback and support. Authors on this research were supported by the Army Research Lab funding program W911NF-20-2-0158. A.K. is supported by an NSF CAREER award IIS-2046737.

References

[1] Edem Allado, Mathias Poussel, Anthony Moussu, Véronique Saunier, Yohann Bernard, Eliane Albuisson, and Bruno Chenuel. Innovative measurement of routine physiological variables (heart rate, respiratory rate and oxygen saturation) using a remote photoplethysmography imaging system: A prospective comparative trial protocol. *BMJ open*, 11(8):e047896, 2021. 1

[2] Sarah Alotaibi and William Smith. Biofacenet: Deep biophysical face image interpretation. In *British Machine Vision Conference (BMVC)*, 2019. 4

[3] Sarah Alotaibi and William AP Smith. A biophysical 3d morphable model of face appearance. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 824–832, 2017. 4

[4] Sarah Alotaibi and William AP Smith. Decomposing multispectral face images into diffuse and specular shading and biophysical parameters. In *IEEE International Conference on Image Processing (ICIP)*, pages 3138–3142. IEEE, 2019. 4

[5] Yunhao Ba, Zhen Wang, Kerim Doruk Karinca, Oyku Deniz Bozkurt, and Achuta Kadambi. Overcoming difficulty in obtaining dark-skinned subjects for remote-ppg by synthetic augmentation. *arXiv preprint arXiv:2106.06007*, 2021. 2, 3, 5, 6, 7

[6] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124:82–90, 2019. 1, 2, 6, 7

[7] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 349–365, 2018. 1, 2

[8] Ananyananda Dasari, Sakthi Kumar Arul Prakash, László A Jeni, and Conrad S Tucker. Evaluation of biases in remote photoplethysmography methods. *NPJ digital medicine*, 4(1):1–13, 2021. 1, 2

[9] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013. 1, 2, 6, 7

[10] 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 *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1462–1469. IEEE, 2014. 1, 2

[11] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 2, 3

[12] Thomas B Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology*, 124(6):869–871, 1988. 5, 6

[13] Takanori Igarashi, Ko Nishino, and Shree K Nayar. *The appearance of human skin: A survey*. Now Publishers Inc, 2007. 4

[14] Achuta Kadambi. Achieving fairness in medical devices. *Science*, 372(6537):30–31, 2021. 2

[15] Fatema-Tuz-Zohra Khanam, Asanka G Perera, Ali Al-Naji, Kim Gibson, Javaan Chahl, et al. Non-contact automatic vital signs monitoring of infants in a neonatal intensive care unit based on neural networks. *Journal of Imaging*, 7(8):122, 2021. 1

[16] Mayank Kumar, Ashok Veeraraghavan, and Ashutosh Sabharwal. Distanceppg: Robust non-contact vital signs monitoring using a camera. *Biomedical optics express*, 6(5):1565–1588, 2015. 1

[17] Magdalena Lewandowska, Jacek Rumiński, Tomasz Kocejko, and Jędrzej Nowak. Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity. In *Federated conference on computer science and information systems (FedCSIS)*, pages 405–410. IEEE, 2011. 1, 2

[18] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 4

[19] Siqi Liu, Pong C Yuen, Shengping Zhang, and Guoying Zhao. 3d mask face anti-spoofing with remote photoplethysmography. In *Proceedings of the European Conference on Computer Vision*, pages 85–100. Springer, 2016. 1

[20] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *arXiv preprint arXiv:2006.03790*, 2020. 1, 2

[21] Hao Lu, Hu Han, and S Kevin 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*, pages 12404–12413, 2021. 1, 2

[22] Daniel McDuff and Ethan Blackford. iphys: An open non-contact imaging-based physiological measurement toolbox. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6521–6524. IEEE, 2019. 6

[23] Daniel McDuff, Sarah Gontarek, and Rosalind W Picard. Improvements in remote cardiopulmonary measurement using a five band digital camera. *IEEE Transactions on Biomedical Engineering*, 61(10):2593–2601, 2014. 2

[24] 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. 3

[25] Rita Meziatisabour, Yannick Benezeth, Pierre De Oliveira, Julien Chappe, and Fan Yang. Ubfc-phys: A multimodal database for psychophysiological studies of social stress. *IEEE Transactions on Affective Computing*, 2021. 1, 2

[26] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video. In *Asian Conference on Computer Vision*, pages 562–576. Springer, 2018. 1, 2

[27] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Transactions on Image Processing*, 29:2409–2423, 2019. 1, 2

[28] Xuesong Niu, Zitong Yu, Hu Han, Xiaobai Li, Shiguang Shan, and Guoying Zhao. Video-based remote physiological measurement via cross-verified feature disentangling. In *Proceedings of the European Conference on Computer Vision*, pages 295–310. Springer, 2020. 1, 2

[29] Ewa Nowara, Daniel McDuff, and Ashok Veeraraghavan. The benefit of distraction: Denoising remote vitals measurements using inverse attention. *arXiv preprint arXiv:2010.07770*, 2020. 6, 8

[30] Ewa Magdalena Nowara, Tim K. Marks, Hassan Mansour, and Ashok Veeraraghavan. Sparseppg: Towards driver monitoring using camera-based vital signs estimation in near-infrared. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1353–135309, 2018. 1

[31] 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*, pages 284–285, 2020. 1

[32] Ewa M. Nowara, Daniel McDuff, and Ashok Veeraraghavan. Combining magnification and measurement for non-contact cardiac monitoring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3810–3819, 2021. 3

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 5

[34] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *IEEE international conference on advanced video and signal based surveillance*, pages 296–301, 2009. 4

[35] Marco AF Pimentel, Alistair EW Johnson, Peter H Charlton, Drew Birrenkott, Peter J Watkinson, Lionel Tarassenko, and David A Clifton. Toward a robust estimation of respiratory rate from pulse oximeters. *IEEE Transactions on Biomedical Engineering*, 64(8):1914–1923, 2016. 5

[36] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, 58(1):7–11, 2010. 1, 2, 6, 7

[37] 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. 1

[38] Stephen J Preece and Ela Claridge. Spectral filter optimization for the recovery of parameters which describe human skin. *IEEE transactions on pattern analysis and machine intelligence*, 26(7):913–922, 2004. 4

[39] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 497–500, 2001. 4

[40] 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. 3

[41] 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, 2019. 1

[42] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 692–702, 2019. 5

[43] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard de Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2016. 1, 2, 6, 7

[44] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In *British Machine Vision Conference (BMVC)*, 2019. 1, 2, 5, 6, 7

[45] 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*, pages 3438–3446, 2016. 1, 2

[46] James Zou and Londa Schiebinger. Ensuring that biomedical ai benefits diverse populations. *EBioMedicine*, page 103358, 2021. 2