Using High-Fidelity Avatars to Advance Camera-based Cardiac Pulse Measurement

Daniel McDuff, *Member, IEEE*, Javier Hernandez, *Member, IEEE*, Xin Liu, Erroll Wood and Tadas Baltrusaitis

***Abstract*— Non-contact physiological measurement has the potential to provide low-cost, non-invasive health monitoring. However, machine vision approaches are often limited by the availability and diversity of annotated video datasets resulting in poor generalization to complex real-life conditions. To address these challenges, this work proposes the use of synthetic avatars that display facial blood flow changes and allow for systematic generation of samples under a wide variety of conditions. Our results show that training on both simulated and real video data can lead to performance gains under challenging conditions. We show strong performance on three large benchmark datasets and improved robustness to skin type and motion. These results highlight the promise of synthetic data for training camera-based pulse measurement; however, further research and validation is needed to establish whether synthetic data alone could be sufficient for training models.**

***Index Terms*— Simulation, Synthetics, Camera, Non-Contact, Photoplethysmography**

I. Introduction

Photoplethysmography (PPG) is a non-invasive method for measuring peripheral hemodynamics and vital signals such as Blood Volume Pulse (BVP) via light reflected from, or transmitted through, the skin. While traditional PPG sensors are used in contact with the skin, digital imagers can be used offering some unique benefits [5], [39], [51], [58], [64]. First, for subjects with delicate skin (e.g., infants in a Neonatal Intensive Care Unit (NICU), burn patients, or the elderly) contact sensors can damage their skin, cause discomfort, and/or increase their likelihood of infection. Second, cameras are ubiquitous (available on many tablets, personal computers, and cellphones) offering unobtrusive and pervasive health monitoring [59]. Third, unlike traditional contact measurement devices (e.g., a smartwatch) remote cameras allow for spatial mapping of the pulse signal that can be used to approximate pulse wave velocity and capture spatial patterns in the peripheral hemodynamics [24], [25], [44].

While there are many benefits of non-contact PPG measurement (a.k.a., imaging photoplethysmography (iPPG) [64]), this approach is especially vulnerable to different environmental factors posing relevant research challenges. For instance, recent research has focused on making iPPG measurements more robust under dynamic lighting and motion [53], [61]. Historically, iPPG methods often relied on unsupervised methods (e.g., ICA or PCA) [33], [39] or hand-crafted demixing algorithms [12], [61]. Recently, supervised neural models have been proposed providing state-of-the-art performance in the context of heart rate measurement [11], [27], [28], [65]. These performance gains are often a direct result of the model scaling well with the volume of training data; however, as with many tasks the volume and diversity of the available data soon become the limiting factor.

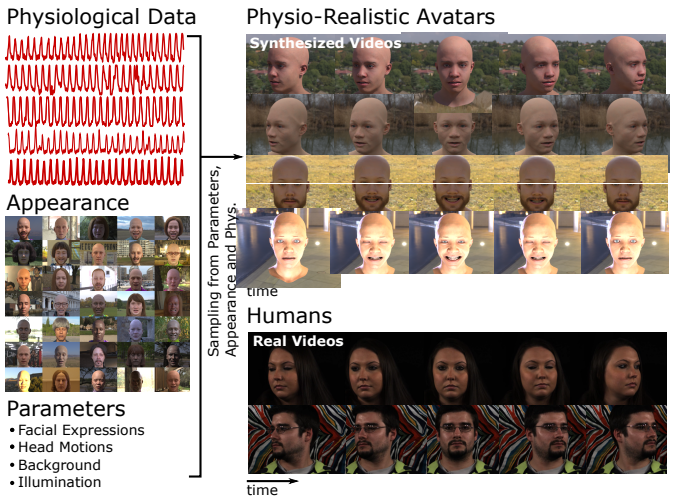

Fig. 1. We propose the use of synthetic avatars to improve non-contact physiological measurement via imaging photoplethysmography. Our approach leverages a physically-based model of the subsurface absorption and scattering of light in the skin to display facial blood flow under different conditions: varied motions, backgrounds and appearances. We observe that training a network on a combination of real and synthetic videos leads to the best overall performance on real videos.

Collecting high-quality physiological data presents numerous challenges. First, recruiting and instrumenting participants is often expensive and requires advanced technical expertise which severely limits its potential volume. Second, training datasets that have already been collected may not contain the types of motion, illumination changes, or appearances that feature in the application context. Thus, a model trained on these data may be brittle and not generalize well. Third, the data can reveal the identity of the subjects and/or sensitive health information. For imaging methods this is exacerbated by the fact that most datasets of video recordings include

D. McDuff, J. Hernandez are with Microsoft Research Redmond, WA 98004 USA (e-mail: damcduff@microsoft.com, javierh@microsoft.com).
X. Liu is with with the University of Washington, Seattle, WA 98105 USA (e-mail: xliu0@cs.washington.edu).
E. Wood and T. Baltrusaitis are with Microsoft Research Cambridge, UK (e-mail: erwood@microsoft.com, tabaltru@microsoft.com).

the subjects face in some or all of the frames [18], [46], [67]. If we could use synthetic data to train iPPG systems it would, to an extent, side-step all three of these challenges and make for an attractive prospect. Once a graphics pipeline is in place, generation of synthetic data is much more scalable than recording videos as computation is relatively inexpensive and can be procured at will using cloud computing. In addition, rare events or typically underrepresented populations can be simulated in videos, assuming we have some knowledge of the statistical properties of the events or a set of examples. Furthermore, synthetic datasets would not need to contain faces or physiological signals with the likeness of any specific individual. Finally, parameterized simulations allow us to systematically vary certain variables of interest (e.g., velocity of motion or intensity of the illumination within a video) which is both useful to train more robust methods as well as evaluating performance under different conditions [54].

We propose to use high-fidelity computer simulations to augment training data that can be used to improve non-contact iPPG measurement (see Fig. 1). This involves answering several research questions: Can we simulate sufficiently high-fidelity data for training iPPG algorithms? Do model parameters learned on synthetic data generalize to real videos? Can using synthetic data help improve generalizability of the learned model? We hypothesize that this is indeed the case and that data synthesis will play a more important role when creating future non-contact physiological measurement methods. The main contributions of this paper are to: 1) propose an approach for synthesizing avatars with realistic facial blood flow as synthetic data for training non-contact physiological measurement models, 2) evaluate a set of models trained on combinations of real and synthetic data on benchmark datasets, 3) show empirical results that synthetic data can help improve overall performance and offer improvements in cases where data is underrepresented in real-world datasets (e.g., task specific motions, or people with darker skin types).

II. Related Work

Non-Contact Physiological Measurement. The BVP can be measured via the light reflected from, or transmitted through, the skin [4]. Imaging-PPG is a set of techniques for measuring this signal using non-contact imagers (e.g., a webcam) and ambient light. Foundational work showed that this signal could be measured via a CCD camera sensor [6], [64]. This work showed that both infra-red and RGB cameras were capable of capturing this subtle information from a video. Following this, further studies were able to replicate these results using different imaging device and measuring the signal from several regions on the body including the face [51], [58], [68], [69]. More recent research has focused on making these algorithms more robust to motion (e.g., rigid head motions and speech) and dynamic illumination [39], [53], [60]. Imaging PPG has enabled the non-contact measurement of several important vital signs and physiological signals including: pulse rate [39], respiration [39], [52], pulse rate variability [39], [49] and pulse transit time [44]. Preliminary work has also shown that it is possible to measure blood oxygen saturation using imagers; however, this still requires calibration as the device and ambient illumination are important parameters in the calculations [52]. Morphological changes in the BVP signal have been shown to be indicators of high blood pressure [1], [17] and could be helpful in assessing the impact of certain chronic conditions, such as hypertension. Several datasets have been collected and shared with the research community [7], [34], [46], [67]. These datasets contain hundreds of videos with ground-truth physiological recordings (either PPG, ECG or both). However, despite the size and availability of these data there remain limitations. The diversity in skin types, systematic variations of noise signals (e.g., motion or lighting changes), and the presence of physiological abnormalities (e.g., arrhythmias) are not very high.

Supervised Camera-based Cardiac Pulse Measurement. Convolutional networks are the most common form of supervised learning used for camera physiological measurement [11], [31], [47], [48], [65]. Chen and McDuff [11] and Špetlík et al. both proposed two part networks, the former using parallel branches and a gate attention and the latter using sequential "extractor" and "predictor" networks. Given the characteristic morphology and periodicity of many physiological signals sequence learning (e.g., via an LSTM or RNN) can help remove noise from predicted waveforms [27], [30], [35], [65]. Temporal information can also be captured with 3D convolution operations [9], [28], [65].

Generative adversarial training has also proven useful, Pulse GAN [47] is one such example, in which the authors used a chromiance signal as an intermediate representation during the training process. In another example, the Dual-GAN [31] method involves segmentation of multiple facial regions of interest using a set of facial landmarks. These regions of interest are then spatially averaged and transformed into both RGB and YUV colorspaces. Using these data spatio-temporal maps (STMaps) are constructed which form the input to a convolutional network. This method produces strong results thanks to careful segmentation and the ability to leverage multiple color space representations.

Given the high individual variability in both visual appearance and physiological signals, personalization or customization of models becomes attractive. Personalization techniques have been proposed for camera physiological measurement. Meta-RPPG [27] was the first such approach which focuses on using transductive inference based meta-learning. Liu et al. [29] proposed a meta-learning framework built on top of the CAN architecture we use in this work.

Training-based on Simulation. One of the most notable properties of neural models is how they scale efficiently with the number of training examples. A large amount of engineering and research efforts have been invested in scaling learning infrastructures so that models with vast numbers (millions or billions) of parameters can be trained with time efficiency. However, it is becoming increasingly difficult to collect sufficient volumes of labeled data to exploit this scale, especially for video-based applications.

Using parameterized graphics simulations to augment existing datasets have been extensively explored in different computer vision domains [21], [45], [54]–[57] such as training

pose recognition [45], scene segmentation for self-driving cars [40], improving object recognition [37], detecting pedestrians under different conditions [54], and for performance evaluation of learned models [21]. AirSim is a graphics-based simulation environment [42] that has been successfully used in the context of training autonomous drone navigation [8]. In the context of physiological sensing; however, synthetic data has been mostly used for evaluation purposes of different algorithms considering other modalities (e.g., [10], [16], [36]). To the best of our knowledge, our work is the first example of using high-fidelity physiological simulations to train iPPG methods.

Synthetics have proven particularly valuable for face and body analyses. In training, synthetics have been used successfully to create models for landmark localization and face parsing [62], body pose estimation [45] and eye tracking [63]. Although not completely representative of real observations, synthetics are also valuable in testing (e.g., for face detection or eye tracking [50]).

Our work is made possible thanks to the ability to render high-fidelity frames/videos with an optical basis for manipulating blood volume in the skin. Creating realistic blood flow simulations is achieved by modelling the appearance of multiple translucent skin layers [3], [15], [23]. These dynamic appearance models usually capture the subsurface scattering that occurs when light interacts with the outer layers of the skin, and are motivated by in-vivo measurements of melanin and hemoglobin concentrations [22].

III. Optical Basis for Synthesized Data

Camera-based vital sign measurement using photoplethysmography involves capturing subtle color changes in skin pixels. Our graphics simulation is inspired by Shafer's dichromatic reflection model (DRM) [61]. We start by assuming there is a light source that has a constant spectral composition but varying intensity, the RGB values of the k-th skin pixel in an image sequence can then be defined by a time-varying function:

$$\boldsymbol{C}_k(t) = I(t) \cdot (\boldsymbol{v}_s(t) + \boldsymbol{v}_d(t)) + \boldsymbol{v}_n(t) \tag{1}$$

$$\boldsymbol{C}_k(t) = I(t) \cdot (\boldsymbol{v}_s(t) + \boldsymbol{v}_{abs}(t) + \boldsymbol{v}_{sub}(t)) + \boldsymbol{v}_n(t) \tag{2}$$

where $\boldsymbol{C}_k(t)$ denotes a vector of the RGB color channel values; $I(t)$ is the luminance intensity level, which changes with the light source as well as the distance between the light source, skin tissue and camera; $I(t)$ is modulated by two components in the DRM: specular (glossy) reflection $\boldsymbol{v}_s(t)$, mirror-like light reflection from the skin surface, and diffuse reflection $\boldsymbol{v}_d(t)$. The diffuse reflection in turn has two parts: the absorption $\boldsymbol{v}_{abs}(t)$ and sub-surface scattering of light in skin-tissues $\boldsymbol{v}_{sub}(t)$; $\boldsymbol{v}_n(t)$ denotes the quantization noise of the camera sensor. $I(t)$, $\boldsymbol{v}_s(t)$ and $\boldsymbol{v}_d(t)$ can all be decomposed into a stationary and a time-dependent part through a linear transformation [61]:

$$\boldsymbol{v}_d(t) = \boldsymbol{u}_d \cdot d_0 + (\boldsymbol{u}_{abs} + \boldsymbol{u}_{sub}) \cdot p(t) \tag{3}$$

where $\boldsymbol{u}_d$ denotes the unit color vector of the skin-tissue; d_0 denotes the stationary reflection strength; $\boldsymbol{v}_{abs}(t)$ and $\boldsymbol{v}_{sub}(t)$ denote the relative pulsatile strengths caused by both changes in hemoglobin and melanin absorption and changes in subsurface scattering respectively, as the blood volume changes; $p(t)$ denotes the BVP.

$$\boldsymbol{v}_s(t) = \boldsymbol{u}_s \cdot (s_0 + \Phi(m(t), p(t))) \tag{4}$$

where $\boldsymbol{u}_s$ is the unit color vector of the light source spectrum; s_0 and $\Phi(m(t), p(t))$ denote the stationary and varying parts of specular reflections; $m(t)$ denotes all the non-physiological variations such as flickering of the light source, head rotation, facial expressions and actions (e.g., blinking, smiling).

$$I(t) = I_0 \cdot (1 + \Psi(m(t), p(t))) \tag{5}$$

where I_0 is the stationary part of the luminance intensity, and $I_0 \cdot \Psi(m(t), p(t))$ is the intensity variation observed by the camera.

The interaction between physiological and non-physiological motions, $\Phi(\cdot)$ and $\Psi(\cdot)$, are usually complex non-linear functions. The stationary components from the specular and diffuse reflections can be combined into a single component representing the stationary skin reflection:

$$\boldsymbol{u}_c \cdot c_0 = \boldsymbol{u}_s \cdot s_0 + \boldsymbol{u}_d \cdot d_0 \tag{6}$$

where $\boldsymbol{u}_c$ denotes the unit color vector of the skin reflection and c_0 denotes the reflection strength. Substituting (3), (4), (5) and (6) into (1), produces:

$$\boldsymbol{C}_k(t) = I_0 \cdot (1 + \Psi(m(t), p(t))) \cdot (\boldsymbol{u}_c \cdot c_0 + \boldsymbol{u}_s \cdot \Phi(m(t), p(t)) + (\boldsymbol{u}_{abs} + \boldsymbol{u}_{sub}) \cdot p(t)) + \boldsymbol{v}_n(t) \tag{7}$$

As the time-varying components are orders of magnitude smaller than the stationary components in (7), we can approximate $\boldsymbol{C}_k(t)$ as:

$$\boldsymbol{C}_k(t) \approx \boldsymbol{u}_c \cdot I_0 \cdot c_0 + \boldsymbol{u}_c \cdot I_0 \cdot c_0 \cdot \Psi(m(t), p(t)) + \boldsymbol{u}_s \cdot I_0 \cdot \Phi(m(t), p(t)) + (\boldsymbol{u}_{abs} + \boldsymbol{u}_{sub}) \cdot I_0 \cdot p(t) + \boldsymbol{v}_n(t) \tag{8}$$

For synthesizing data for physiological measurement methods, we want to create skin with RGB changes that vary with $p(t)$. Using a principled bidirectional scattering distribution function (BSDF) shader, we are able to capture both of the components of $\boldsymbol{u}_p$, $\boldsymbol{u}_{abs}$ and $\boldsymbol{u}_{sub}$, using the subsurface color and subsurface radius parameters. The specular reflections are controlled by the specular parameter. Thus, for a given pulse signal, $p(t)$, we can synthesize the skin's appearance over time. Furthermore, we can synthesize these changes in a wide variety of other variations, which for the purposes of vital sign measurement will represent noise sources.

For any of the video-based physiological measurement methods, the task is to extract $p(t)$ from $\boldsymbol{C}_k(t)$. The motivation for using a machine learning model to capture the relationship between $\boldsymbol{C}_k(t)$ and $p(t)$ in (8) is that a neural model can capture a more complex relationships than hand-crafted demixing or source separation algorithms (e.g., ICA, PCA)

that have ignored $p(t)$ inside $\Phi(\cdot)$ and $\Psi(\cdot)$, and assumed a linear relationship between $\boldsymbol{C}_k(t)$ and $p(t)$.

IV. Avatar Synthesis

We use high-fidelity facial avatars and a physiologically-based animation model for simulating videos of faces with a realistic blood flow (pulse) signal. These videos are then used to train a neural model for recovering the BVP from video sequences. The resulting model is tested on real video benchmark datasets. This process is shown in Fig. 1.

A. Physiological Recordings

To synthesize the appearance of the avatars, we use photoplethysmographic waveforms recordings from PhysioNet [20]. Specifically, we use the BIDMC PPG and Respiration Dataset [38] which include 53 8-minute contact PPG recordings sampled at 125Hz from different individuals. These recordings were taken from the larger MIMIC-II dataset [41]. We sample PPG recordings from different subjects for each of the 50 avatars that we synthesize. As we only synthesize/render short sequences (nine 10-second sequences described below) for each avatar we only use the first 90 seconds (9×10 seconds) of each recording.

B. Synthesizing Videos with Pulse Signals

A key part of our work is a realistic model of facial blood flow. We simulate blood flow by adjusting properties of the physically-based shading material we use for the face[1]. The albedo component of the material is a texture map transferred from a high-quality 3D face scan. The facial hair has been removed from these textures by an artist so that the skin properties can be easily manipulated (3D hair can be added later in the process). Specular effects are controlled with an artist-created roughness map, to make some parts of the face (e.g. the lips) shinier than others. An example of our material setup can be seen in Fig. 2.

Subsurface Skin Color: As blood flows through the skin, the composition of the skin changes and causes variations in subsurface color. We manipulate skin tone changes using the subsurface color parameters. The weights for this are derived from the absorption spectrum of hemoglobin and typical frequency bands from an exemplar digital camera[2] (Red: 550-700 nm, Green: 400-650 nm, Blue: 350-550 nm). In this work we globally vary these across all skin pixels on the albedo map (but not non-skin pixels). Specifically, we used relative subsurface scattering coefficients of 0.36, 0.41 and 0.23 for the red, green and blue channels respectively. We varied these by up to 0.1 for each of the synthesized videos within a normal distribution about each of the values.

Subsurface Scattering: We manipulate the subsurface radius for the channels to capture the changes in subsurface scattering as the blood volume varies. The subsurface scattering is spatially weighted using an artist-created subsurface scattering radius texture (see Fig. 2) which captures variations in the thickness of the skin across the face. We vary the BSDF subsurface radii for the RGB channels using the same weighting prior as above. Empirically we find these parameters work for synthesizing data for training camera-based vital sign measurement. We found that varying the subsurface scattering alone, without changes in subsurface color, were too subtle and could not recreate the effects the BVP on reflected light observed in real videos.

C. Systematic Variations

To obtain machine learning systems that are robust to certain forms of variation encountered in the real world, we introduced the following types of variation into our dataset:

Facial Appearance. We synthesized faces with 50 different appearances (examples can be seen in Fig. 3). The facial identities are made up from a combination of component parts, including: geometry from a generative parametric face model constructed from face scans, texture from albedo and displacement textures constructed directly from face scans, eye color from a selection of eye colors, hair from a library of hair grooms. Each of these components are sampled independently to arrive at a random facial identity. We used 3D face scans purchased from publicly available sources[3]. These scans have been retopologized (aligned) and cleaned by us to make them suitable for parametric model training and use as a source for textures. Specifically, for each face, we set up the skin material with an albedo texture picked at random from our collection of 159 textures. The textures were created by using the raw scans available to us as the source, and the retopologized version of those scans as the target for albedo and displacement map projection. In addition, we performed some cleaning of the source scans to remove undesired artifacts such as hairnets, head hair and facial hair. In order to model wrinkle-scale geometry, we also apply a matching high-resolution displacement map that was transferred from the scan data. Characterizing differences in facial appearance is challenging. However, skin tone is particularly important in imaging PPG measurement. The Fitzpatrick scale [19] is a dermatological tool that captures the melanin content of the skin and is used to help describe the impact of UV radiation. As such it only describes the skin type and does not capture any other differences about appearance such as facial structure which vary around the world, but currently it is the most widely used approach for systematically analyzing differences in methods to skin tone. We are currently investigating alternative coding schemes but at the present time feel that the Fitzpatrick scale is the most appropriate to use. The approximate Fitzpartick skin type distribution for the 50 faces was: Type I - 9, II - 15, III - 12, IV - 4, V - 5, VI - 5. While this distribution is still not uniform, it represents a much more balanced distribution than in existing imaging PPG datasets. Just under half (21) of the avatars were synthesized with some form of facial hair (beard and/or moustache) to further increase the variety in appearance.

Head Motion. Since motion is one of the greatest sources of noise in imaging PPG measurement, we simulate a set of

[1]https://www.blender.org/

[2]https://www.bnl.gov/atf/docs/scout-g$_u sers_m anual.pdf$

[3]https://www.3dscanstore.com/

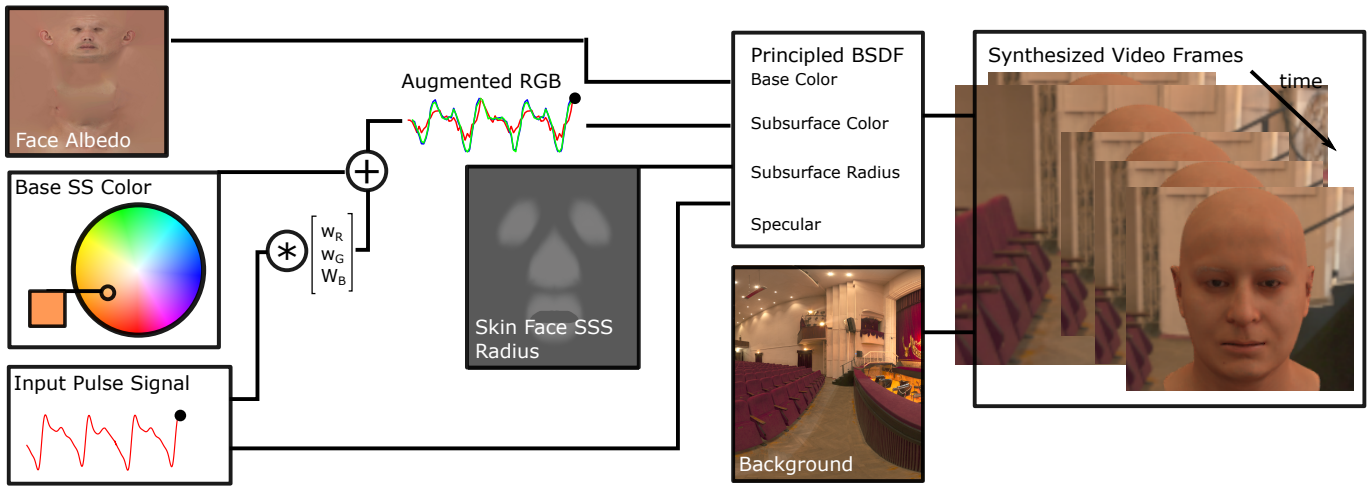

Fig. 2. Our approach to synthesizing videos of faces with dynamic blood flow signals. We start with a face albedo, base subsurface color and input pulse signal. The skin properties are varied temporally based on hemoglobin properties. The subsurface skin color captures changes in absorption, $\boldsymbol{v}_{abs}(t)$, with variations in hemoglobin. The subsurface scattering, $\boldsymbol{v}_{sub}(t)$, captures how light scattering changes with the volume of blood.

Fig. 3. Examples of the appearances of the avatars we synthesized for our dataset.

rigid head motions to augment training examples that capture these conditions. In particular, we smoothly rotate the head about the vertical axis at angular velocities of 0, 10, 20, and 30 degrees/second similar to prior work [18]. Six of the nine videos synthesized for each avatar features motion, two at each angular velocity.

Facial Expression. Similar to head motions, facial expressions movements are also a frequent source of noise in PPG measurement. We synthesized videos with smiling, blinking, and mouth opening (similar to speaking), which are some of the most common facial expressions exhibited in everyday life. We apply smiles and blinks to the face using our collection of artist-created blend shapes, and we open the mouth by rotating the jaw bone with linear blend skinning. Four of the nine videos synthesized had smiling, mouth opening, and blinking motions.

Environment. We render faces in different image-based environments to create a realistic variety in both background appearance and illumination on the face [14]. For each sequence, we pick one high dynamic range spherical environment map from our collection [66] (see Fig. 2 for examples). The lighting in each environment is dependent on the HDRI used to synthesize each video. Thus each synthetic video has a different lighting spectral composition, direction and intensity. In this work we synthesized static background scenes only, but future work may benefit from considering backgrounds with motion, or even facial occlusions that more closely resemble challenging real-life conditions.

D. The Final Synthetic Dataset

We rendered nine video sequences for each of our 50 different facial identities, resulting in 450 video sequences in total. Each sequence was 10 seconds long, with a frame-rate of 30Hz. The nine clips feature rotational head motions, facial expressions, and different backgrounds as described above. Each frame took approximately 20 seconds to render with Blender Cycles[4] on an Nvidia GTX 1080Ti GPU. These videos were used to train a convolutional attention network described in Section V-B. The trained network was then tested on three benchmark video datasets of non-synthetic (a.k.a real) videos described in Section V-A.

V. Experiments

A. Benchmark Datasets

AFRL [18]: Videos were recorded at 658x492 pixel resolution and 120 frames per second (fps) using a Basler Scout scA640-120gc GigE-standard, color camera. These videos

[4]https://docs.blender.org/manual/en/latest/render/cycles/index.html

were compressed with a constant rate factor (CRF) of 12. Twenty-five participants (17 males) were recruited to participate in the study. Fingertip PPG was recorded as ground truth signals using a research-grade biopotential acquisition unit. Each participant was recorded six times for 5-minutes each with increasing head motion in each experiment and this process was repeated twice in front of two background screens.

MMSE-HR [67]: 102 videos of 40 participants were recorded at 25 fps capturing 1040x1392 resolution images in an uncompressed format using a 3D dynamic imaging system[5] during spontaneous emotion elicitation experiments. The gold standard contact signal was measured via a Biopac2 MP150 system[6] which provided pulse rate at 1000 fps and was updated after each heartbeat. These videos feature smaller but more spontaneous motions than those in the AFRL dataset.

UBFC-RPPG [7]: 42 videos of 42 participants were recorded at 640x480 resolution and 30 fps in uncompressed 8-bit RGB format using a Logitech C920 HD pro. A fingertip oximeter was used to obtain the gold standard PPG.

B. Physiological Measurement Network

To evaluate the impact of synthetic data on the quality of recovered pulse signals from video, we used an existing end-to-end learning model, Convolutional Attention Network (CAN) [11], which uses motion and appearance representations learned jointly through an attention mechanism. An illustration of a CAN network architecture is shown in Figure 4. The approach consists of a two-branch convolutional neural network to represent motion and appearance.

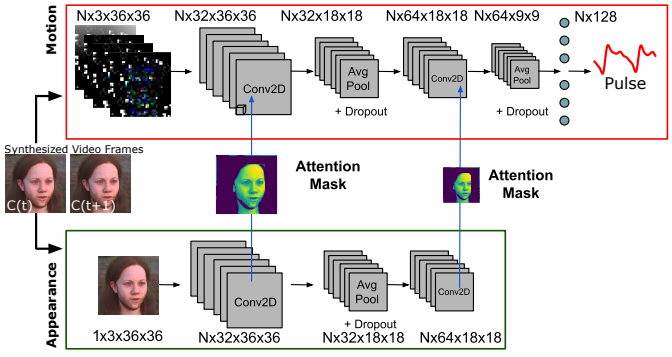

Fig. 4. An illustration of the convolutional attention network (CAN) architecture that we used in our experiments. To make the figure clearer in this image each convolutional layer shown reflects a pair layers one following the other and the fully connected layer reflects a pair of fully connected layers.

The motion representation branch allows the network to differentiate between intensity variations caused by noise, e.g., from motion from subtle characteristic intensity variations induced by blood flow. The input to the motion representation branch is calculated as the difference of two consecutive video frames. To reduce the noise from changes in ambient illumination and the distance of the face to the illumination source, the frame difference is first normalized based on the skin reflection model [61]. The normalization is applied to each video sequence by subtracting the pixel mean and dividing by the standard deviation. We perform normalization on real and synthetic frames.

The appearance representation captures the regions in the image that contribute strong iPPG signals. Via the attention mechanism, the appearance representation guides the motion representation and helps differentiate the PPG signal from the other sources of noise. The input frames are similarly normalized by subtracting the mean and dividing by the standard deviation. Again the same procedure is used for the real and synthetic frames.

C. Performance Metrics

To provide a comprehensive characterization of the performance of the models trained on simulated data, we used different metrics to evaluate two main components.

Heart Rate: We compute heart rate estimates for non-overlapping 30-second windows and calculate the mean absolute error (MAE) and root mean squared error (RMSE) between these estimates and the gold-standard heart rate calculated from the contact sensor measurements in each dataset. As follows:

$$MAE = \frac{1}{T}\sum_{1}^{T}|HR_i - HR_i'| \tag{9}$$

$$RMSE = \sqrt{\frac{1}{T}\sum_{1}^{T}(HR_i - HR_i')^2} \tag{10}$$

Where HR is the gold-standard heart rate and HR' is the estimated heart rate from the video. The gold-standard HR frequency was determined from the manually corrected ECG peaks in the AFRL dataset and the HR estimates provided with the dataset for the MMSE-HR dataset.

We also compute the Pearson correlation between the estimated heart rates and the gold-standard heart rates from the contact sensor measurements.

BVP Signal-to-Noise Ratio (SNR):

The BVP signal-to-noise (SNR) is calculated according to the method proposed by De Haan et al. [13]. This captures the signal quality of the recovered pulse estimate. Again, the gold-standard HR frequency was determined from the manually corrected ECG peaks in the AFRL dataset and the HR estimates provided with the dataset for the MMSE-HR dataset.

$$SNR = 10\log_{10}\left(\frac{\sum_{f=30}^{240}((U_t(f)\hat{S}(f))^2}{\sum_{f=30}^{240}(1 - U_t(f))\hat{S}(f))^2)}\right) \tag{11}$$

where $\hat{S}$ is the power spectrum of the BVP signal (S), f is the frequency (in BPM) and $U_t(f)$ is a binary template that is one for the heart rate region from HR-6 BPM to HR+6BPM and its first harmonic region from 2*HR-12BPM to 2*HR+12BPM, and 0 elsewhere. The HR and BVP SNR (measured in dB) is calculated for non-overlapping 30 second time windows.

[5]https://di4d.com/
[6]https://www.biopac.com/

D. Training and Testing

In all our experiments we use a person independent training regime and create training, validation and test partitions.

For experiments on the AFRL dataset, we perform a five-fold evaluation in which the 25 participants in the AFRL dataset [18] were randomly divided into five folds, with 15 participants in the training set, five in the validation set, and five in the test set. The learning models were then trained to evaluate how our models can be generalized to new participants. The validation set was used to select the epoch for which the model would be used for testing. During training and model selection the mean squared error (MSE) between the predicted and gold-standard pulse waveforms was used as the loss/performance metric.

The evaluation metrics for AFRL performance shown in the results tables are all averaged over the five folds. Prior work has shown that participant-independent training is a more challenging task than participant-dependent training [11] and it is a more realistic scenario for real-world applications. For experiments on the MMSE-HR and UBFC datasets, we use the model that performed best on the AFRL dataset and test it without fine-tuning (i.e., dataset independent evaluation). We compare our proposed approach to three other methods [12], [39], [61] for recovering the BVP. These methods are unsupervised and therefore results are reported across all participants without the need for cross-validation on either dataset.

For the convolutional neural network architecture motion representation model, we used nine layers with 128 hidden units, average pooling and tanh as the activation functions. The last layer of the motion model had linear activation units and the MSE loss. For the appearance model, we used the same architecture as the motion model but without the last three layers, consistent with [11]. Finally, a 6th-order Butterworth filter was applied to all model outputs (cut-off frequencies of 0.7 and 2.5 Hz) before computing the frequency spectra and heart rate. The baseline methods were implemented using the public MATLAB toolbox [32].

VI. Results

Training with Synthetic Data. Our first experiments are to validate the effect of using synthetic data to train the vital signs measurement algorithm. Table I shows results of models trained on non-synthetic (real) data, synthetic data, and a combination of real and synthetic data. Results are shown for the AFRL dataset for which we perform the five-fold participant-independent cross-validation. For the MMSE results we report performance of the model trained on the AFRL data and thus this is both participant independent (because no people feature in both dataset) and can be viewed as an example of cross-dataset transfer learning. In all cases the results are participant independent. The models trained on real and synthetic data outperform the models trained only on real data for both datasets. This is true for the BVP SNR reflecting that the underlying pulse signal is cleaner and for HR MAE and RMSE reflecting that the HR estimates are more accurate. On the AFRL dataset the results are not very different because training with real data from the same dataset already performs very well (and the HR correlation is marginally higher). This is because synthetic data does not provide a great benefit if the distribution of the test data is very similar to that of the training data. In the AFRL dataset the participants all have similar skin types (tones) and the lighting is very constant across all videos. Thus, if examples of all tasks from this dataset are included in the training set, even if they feature different participants, the margin for improvement is small.

However, when we perform a cross-dataset using a combination of synthetic and real data set provides a much more considerable improvement. The MAE in HR estimates on the MMSE dataset is 2.26 (compared to 3.74 for the next best approach), a 40% reduction in error. In this case, the distribution of the testing data is quite different from the data in the training set and, consequently, the benefit of using synthetics becomes apparent. The synthetic data help improve the generalization of the model when there is a larger domain gap between the training and testing data. To provide a qualitative example, Fig. 5 shows an example of the recovered pulse waveforms and corresponding power spectra for two videos in the MMSE-HR dataset. Notice how the pulse spectra are much cleaner and more closely resemble the gold-standard when using the model trained on real and synthetic data. Interestingly, the performance on the UBFC dataset is strongest when training with only synthetic data, we hypothesize this is because the domain gap between the real training data and the UBFC test data is larger. Future work will investigate how to characterize the difference between dataset distributions in this domain.

Cross-Task Performance. Body motions are one of the most common and problematic sources of noise in non-contact vital signs measurement. In the previous analyses on the AFRL dataset we included examples of every task in the training, validation and test sets. However, when we train and validate only on videos with static subjects and then test on videos with head motions the improvements gained from using synthetic data are much more dramatic (shown in Table II). The avatar data includes heads with motions, the result highlights that synthetic data can help bridge the gap between heads with motion. For example, if no real video data with gold-standard measurements were available with motions similar to those in the test scenario we can synthesize data to bridge the gap.

Comparison with Benchmarks. Next let us compare performance on both datasets against the other baseline methods. Table I shows the results of the CAN alongside ICA [39], CHROM [13] and POS [61]. On both datasets the neural network trained on real and synthetic data outperforms all the other methods. The POS method performs well on the AFRL dataset with similar results in HR estimation, but a lower BVP SNR. On the MMSE dataset the CAN outperforms the other methods by a considerable margin (MAE = 2.26 BPM vs. 3.74 BPM from the next best method). Unsupervised methods have previously been used more frequently than supervised algorithms for imaging-based measurement of vital signs because of concerns about the generalizability of "trained" models. However, our results suggest that with sufficient diversity in the training set that supervised methods could have an advantage.

TABLE I

BENCHMARK PERFORMANCE OF PULSE MEASUREMENT ON THE AFRL [18], MMSE-HR [67] AND UBFC [7] DATASETS. NOTE THAT THE AFRL DATASET CONTAINS VIDEOS WITH SPECIFIC AND LARGE ROTATIONAL HEAD MOTIONS, WHILE OUR SYNTHETIC DATA DID CONTAIN ROTATIONAL HEAD MOTIONS THE LARGE DIFFERENCE IN PERFORMANCE SUGGESTS THAT IT MAY NOT HAVE FAITHFULLY REFLECTED REALISTIC MOTION.

	Within Dataset (Subj. Ind.) AFRL (All Tasks) [18]				Across Dataset (Train on AFRL) MMSE-HR [67]				Across Dataset (Train on AFRL) UBFC [7]			
Method	MAE	RMSE	ρ	SNR (dB)	MAE	RMSE	ρ	SNR (dB)	MAE	RMSE	ρ	SNR (dB)
CAN (w/ Real+Synth)	2.42	4.37	0.88	6.57	2.26	3.70	0.97	4.85	5.15	9.51	0.77	0.79
CAN (w/ Synth)	9.23	13.4	0.36	-7.17	4.98	11.8	0.70	-1.98	5.01	8.90	0.80	-2.58
CAN (w/ Real) [11]	2.43	4.39	0.87	6.21	4.43	9.98	0.80	-0.66	5.83	10.9	0.68	-0.63
POS [61]	2.48	5.07	0.89	2.32	3.90	9.61	0.78	2.33	8.24	19.9	0.57	-1.19
CHROM [12]	6.42	12.4	0.60	-4.83	3.74	8.11	0.82	1.90	7.46	15.5	0.72	-1.10
ICA [39]	4.36	7.84	0.77	3.64	5.44	12.00	0.66	3.03	14.3	26.1	0.28	-2.67

MAE = Mean Absolute Error in HR estimation, RMSE = Root Mean Squared Error in HR estimation, SNR = BVP Signal-to-Noise Ratio, ρ = Pearson Correlation in HR estimation.

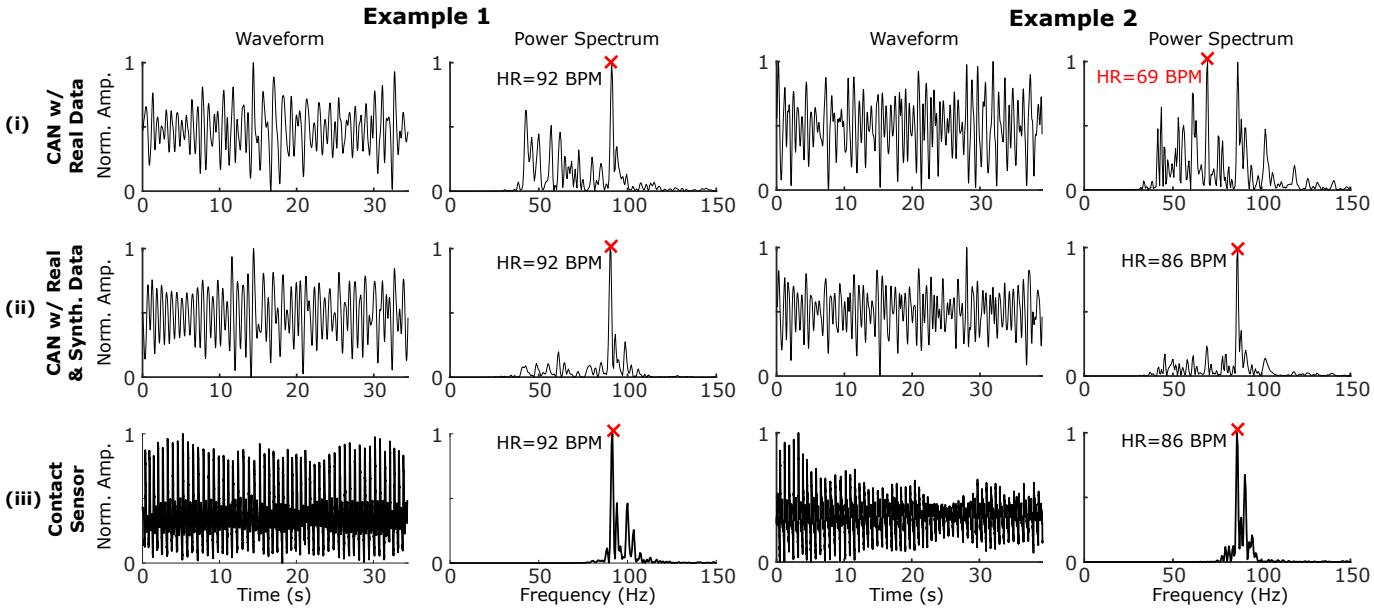

Fig. 5. Examples of the blood volume pulse recovered using a camera and the trained neural network (thin black lines), with only (i) real and (ii) synthetic and real data, in comparison with (iii) a finger contact sensor (thick black lines). PPG waveforms normalized from 0 to 1 and pulse power spectra normalized from 0 to 1 are shown. Notice how the pulse spectra are much cleaner and more closely resemble the gold-standard when using the model trained on real and synthetic data, compared to real data alone. These are qualitative examples of why we see an improvement in BVP SNR as shown in Tables I.

Robustness to Skin Tone. Skin type influences the signal-to-noise ratio (SNR) of the recovered BVP in many camera-based vital sign measurement algorithms [2], [43], [61]. A larger melanin concentration in people with darker skin absorbs more light, making the intensity of light returning to the camera lower and thus the iPPG signal weaker. To exacerbate this problem subjects with darker skin types are often underrepresented in computer vision datasets, including those used for camera-based physiological measurement. Synthetic data can be used to identify biases in CV systems and help address them.

Figure 6 and Table III shows the performance when testing the model on subjects with different skin types based on the Fitzpatrick skin type scale [19] from the MMSE-HR dataset [67]. The synthetic data provides a substantial improvement in HR MAE, especially for the lightest (II) and darkest (VI) skin types. These are the skin types that are typically underrepresented in real video datasets used for non-contact vital sign measurement algorithms, including the AFRL dataset. Not only are the overall heart rate estimation errors lower for all skin types, the standard deviation in average heart rate MAE across skin types is approximately halved when training with real and synthetic data compared to when training with only real data. To summarize, our results show that by using synthetic data we can also improve performance on subjects whose appearance type (in this case skin type) was underrepresented in the "real" portion of the training data set. We should note, however, that when training only on synthetic data the performance on the real videos featuring subjects of skin types V and VI is poor. We have a couple of hypotheses about the cause of this result. 1) Our synthetic dataset, while arguably containing much greater variance than the real training set (AFRL dataset), is relatively quite a bit smaller. Therefore, there is a chance that overfitting occurs and that this overfitting is most severe for the darker skin type videos. 2) The MMSE-HR test set, while one of the more diverse, contains only a few subjects with skin types V and VI therefore, there is quite a large variance in performance. Thus, sometimes when the model reaches a local minima this

can lead to large errors on those subjects. We are encouraged that combining synthetic and real data appears to help rectify these errors somewhat.

TABLE II

TASK-INDEPENDENT PERFORMANCE: PULSE MEASUREMENT ON VIDEOS WITHOUT HEAD MOTION (TASKS 1 & 2) AND WITH HEAD MOTION (TASKS 3, 4, 5 & 6) FROM THE AFRL [18] DATASET WHEN TRAINING ON VIDEOS WITHOUT MOTION (TASKS 1 & 2).

	AFRL Motion Tasks (3-6)			
Method	MAE	RMSE	SNR	ρ
CAN (w/ Real (Static) + Synthetic)	**6.52**	**9.82**	**-0.30**	**0.63**
CAN (w/ Real (Static))	8.21	11.8	-1.68	0.50
CAN (w/ Synthetic)	14.1	18.9	-10.2	0.16

MAE = Mean Absolute Error in HR estimation, RMSE = Root Mean Squared Error in HR estimation, SNR = BVP Signal-to-Noise Ratio, ρ = Pearson Correlation in HR estimation, WMAE = Waveform MAE.

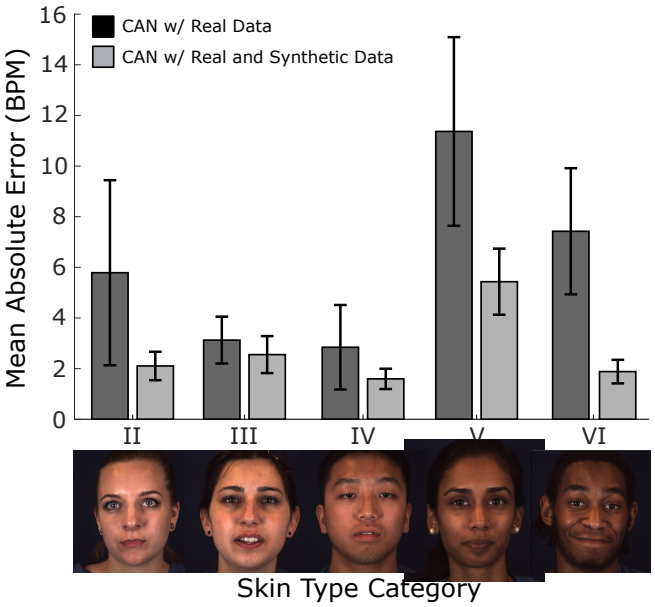

Fig. 6. Heart rate mean absolute error (BPM) by skin tone on the MMSE-HR dataset [67]. Training with synthetic data reduces the errors for the lightest (II) and darkest (VI) skin types the most, those that are often underrepresented in real video training datasets. No. of participants: II=8, III=11, IV=18, V=2, VI=2.

VII. DISCUSSION

Collecting datasets for training non-contact vital signal measurement algorithms has several challenges. We have presented an approach for synthesizing avatars that helps alleviate the need for real videos. Our results show that training with synthetic data can successfully improve the performance of non-contact vital sign measurement. Specifically, including synthesized and real video data in the training set can lead to an improvement of the pulse SNR ratio as well as lowering heart rate measurements errors compared to training with just real video data alone. In particular, the recovered BVP signal quality was much improved across both datasets (see Table I).

TABLE III

MEAN ABSOLUTE ERROR IN HEART RATE ESTIMATION BY PARTICIPANT FITZPATRICK SKIN TONE CATEGORY. THE STANDARD DEVIATION IN AVERAGE ERRORS ACROSS SKIN TONES IS SHOWN IN THE FINAL COLUMN.

	Fitz. Skin Tone					
Method	II	III	IV	V	VI	σ
CAN (w/ Real + Synthetic)	2.10	2.55	1.59	5.43	1.88	1.56
CAN (w/ Real)	5.79	3.12	2.84	11.4	7.42	3.50
CAN (w/ Synthetic)	2.60	2.83	3.47	13.9	31.8	12.6

When training and testing on the same datasets (in a participant-independent manner), the improvements were modest. This suggests that when the testing data has a similar distribution (similar motions, lighting, skin types) there is not much benefit to be gained from synthetic data. Synthetic data is particularly effective at reducing errors in cross-domain learning, improving cross-task, cross-dataset and cross-appearance generalization. Synthesizing data allows us to create many combinations of facial appearances (skin tones, hair styles, facial hair), expressions, speech, head motions (rotational and translational), ambient lighting conditions and backgrounds. Finally, we show that synthetic data can substantially improve (and reduce the variance in) performance of non-contact vital sign measurement for skin tones underrepresented in training data.

VIII. LIMITATIONS AND FUTURE WORK

While synthetics are flexible and scalable once you have created a pipeline, the initial overhead for this infrastructure is expensive and labor-intensive to create. Our synthetics pipeline involved a multi-year effort to create and unfortunately we cannot publicly release the dataset at this time. Furthermore, while we demonstrate that our synthetics pipeline can offer a tangible benefit, we did not push the limits of the improvements that synthetic data can provide. It is possible that greater improvements could have been obtained if we had synthesized more face videos. However, the videos were synthesized on a frame-by-frame basis taking approximately an hour to synthesize a single 10s video.

While our results are promising and justify further exploration of synthetic data as a tool for training video-based cardiac measurement models, they do not conclusively prove that synthetic data alone is sufficient for achieving state-of-the-art results. More research and validation will be required to ascertain the full potential of these tools and to design methods that could overcome the gap between simulated and real videos.

While our approach could be used for creating more motion robust iPPG algorithms for non-contact measurement in fitness centers or telehealth systems. Many of the applications of non-contact vital signal measurement do not necessarily involve analysis of adult faces. Modeling infants for training models to be deployed in a NICU would be a great extension of this work.

There are several ways it may be possible to improve upon the results we have presented in this paper. In this work, we used the synthetically generated images as they were rendered,

we did not rigorously analyzed the amount of sensor noise that needs to be added to the synthetic images to mimic the quantization levels of different cameras. Noise such as this could be added as a form of data augmentation. Similarly, there are other forms of data augmentation that might help with creating a more generalizable model including: applying gamma or white balance correction to the synthetic frames and mimicking motion blur or changes in camera focus.

The density of the microvascular bed of tissue is not uniform across the skin and whether capillaries are open or closed can also vary across time [26]. Our synthetics pipeline cannot currently model these spatial and temporal changes in the microvascular tissue, the light absorption is changed uniformly for the whole skin texture. So while this does mean that a model trained on the synthetic data learns good skin segmentation, which is certainly beneficial for obtaining the PPG signal, it could be improved further.

These promising results for training a remote PPG measurement model using synthetic data leads us to believe that it might be possible to use a similar approach for training models for measuring blood oxygen saturation. However, modeling oxygen saturation (volume of oxygenated versus deoxygenated blood) is a more complicated task than measuring overall blood volume and our current implementation does not support this.

IX. Conclusion

This work proposes the use of synthetic avatars to synthesize novel samples of facial blood volume changes that can improve the robustness of non-contact physiological sensing methods. We are looking forward to a future when similar methodology can be used to not only improve the generalization performance under challenging real-life scenarios but also minimize potential performance differences across underrepresented groups or people.

References

[1] Paul S Addison. Slope transit time (stt): A pulse transit time proxy requiring only a single signal fiducial point. *IEEE Transactions on Biomedical Engineering*, 63(11):2441–2444, 2016.

[2] Paul S Addison, Dominique Jacquel, David MH Foo, and Ulf R Borg. Video-based heart rate monitoring across a range of skin pigmentations during an acute hypoxic challenge. *Journal of clinical monitoring and computing*, 32(5):871–880, 2018.

[3] Mohammed Hazim Alkawaz, Ahmad Hoirul Basori, and Siti Zaiton Mohd Hashim. Oxygenation absorption and light scattering driven facial animation of natural virtual human. *Multimedia Tools and Applications*, 76(7):9587–9623, 2017.

[4] John Allen. Photoplethysmography and its application in clinical physiological measurement. *Physiological measurement*, 28(3):R1, 2007.

[5] Ethan B Blackford, Justin R Estepp, Alyssa M Piasecki, Margaret A Bowers, and Samantha L Klosterman. Long-range non-contact imaging photoplethysmography: cardiac pulse wave sensing at a distance. In *Optical Diagnostics and Sensing XVI: Toward Point-of-Care Diagnostics*, volume 9715, page 971512. International Society for Optics and Photonics, 2016.

[6] Vladimir Blazek, Ting Wu, and Dominik Hoelscher. Near-infrared ccd imaging: Possibilities for noninvasive and contactless 2d mapping of dermal venous hemodynamics. In *Optical Diagnostics of Biological Fluids V*, volume 3923, pages 2–9. International Society for Optics and Photonics, 2000.

[7] 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.

[8] Elizabeth Bondi, Debadeepta Dey, Ashish Kapoor, Jim Piavis, Shital Shah, Fei Fang, Bistra Dilkina, Robert Hannaford, Arvind Iyer, Lucas Joppa, et al. Airsim-w: A simulation environment for wildlife conservation with uavs. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 1–12, 2018.

[9] Frédéric Bousefsaf, Alain Pruski, and Choubeila Maaoui. 3d convolutional neural networks for remote pulse rate measurement and mapping from facial video. *Applied Sciences*, 9(20):4364, 2019.

[10] Peter H. Charlton, Timothy Bonnici, Lionel Tarassenko, David A. Clifton, Richard Beale, and Peter J. Watkinson. An assessment of algorithms to estimate respiratory rate from the electrocardiogram and photoplethysmogram. *Physiological Measurement*, 2016.

[11] 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.

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

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

[14] Paul Debevec. Image-based lighting. In *ACM SIGGRAPH 2006 Courses*, pages 4–es. 2006.

[15] Eugene d'Eon, David Luebke, and Eric Enderton. Efficient rendering of human skin. In *Proceedings of the 18th Eurographics conference on Rendering Techniques*, pages 147–157. Eurographics Association, 2007.

[16] Felipe I. Donoso, Rosa L. Figueroa, Eduardo A. Lecannelier, Esteban J. Pino, and Alejandro J. Rojas. Atrial activity selection for atrial fibrillation ECG recordings. *Computers in Biology and Medicine*, 43(10):1628–1636, oct 2013.

[17] Mohamed Elgendi, Richard Fletcher, Yongbo Liang, Newton Howard, Nigel H Lovell, Derek Abbott, Kenneth Lim, and Rabab Ward. The use of photoplethysmography for assessing hypertension. *NPJ digital medicine*, 2(1):1–11, 2019.

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

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

[20] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

[21] Robert M Haralick. Performance characterization in computer vision. In *BMVC92*, pages 1–8. Springer, 1992.

[22] Jorge Jimenez, Timothy Scully, Nuno Barbosa, Craig Donner, Xenxo Alvarez, Teresa Vieira, Paul Matts, Verónica Orvalho, Diego Gutierrez, and Tim Weyrich. A practical appearance model for dynamic facial color. *ACM Transactions on Graphics (TOG)*, 29(6):141, 2010.

[23] Jorge Jimenez, David Whelan, Veronica Sundstedt, and Diego Gutierrez. Real-time realistic skin translucency. *IEEE Computer Graphics and Applications*, 30(4):32–41, 2010.

[24] Alexei A Kamshilin, Serguei Miridonov, Victor Teplov, Riku Saarenheimo, and Ervin Nippolainen. Photoplethysmographic imaging of high spatial resolution. *Biomedical optics express*, 2(4):996–1006, 2011.

[25] Mayank Kumar, James Suliburk, Ashok Veeraraghavan, and Ashutosh Sabharwal. Pulsecam: High-resolution blood perfusion imaging using a camera and a pulse oximeter. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3904–3909. IEEE, 2016.

[26] Eugene M Landis. The capillaries of the skin: A review. *Journal of Investigative Dermatology*, 1(4):295–311, 1938.

[27] Eugene Lee, Evan Chen, and Chen-Yi Lee. Meta-rppg: Remote heart rate estimation using a transductive meta-learner. In *European Conference on Computer Vision*, pages 392–409. Springer, 2020.

[28] 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.

[29] Xin Liu, Ziheng Jiang, Josh Fromm, Xuhai Xu, Shwetak Patel, and Daniel McDuff. Metaphys: few-shot adaptation for non-contact physiological measurement. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 154–163, 2021.

[30] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 389–398, 2018.

[31] 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.

[32] Daniel McDuff and Ethan 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)*, pages 6521–6524. IEEE, 2019.

[33] 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.

[34] 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.

[35] Ewa M Nowara, Daniel McDuff, and Ashok Veeraraghavan. The benefit of distraction: Denoising camera-based physiological measurements using inverse attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4955–4964, 2021.

[36] Joonas Paalasmaa, Hannu Toivonen, and Markku Partinen. Adaptive Heartbeat Modeling for Beat-to-Beat Heart Rate Measurement in Ballistocardiograms. 2014.

[37] Xingchao Peng, Baochen Sun, Karim Ali, and Kate Saenko. Learning deep object detectors from 3d models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1278–1286, 2015.

[38] 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.

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

[40] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016.

[41] Mohammed Saeed, Mauricio Villarroel, Andrew T Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, and Roger G Mark. Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. *Critical care medicine*, 39(5):952, 2011.

[42] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics*, pages 621–635. Springer, 2018.

[43] Dangdang Shao, Francis Tsow, Chenbin Liu, Yuting Yang, and Nongjian Tao. Simultaneous monitoring of ballistocardiogram and photoplethysmogram using a camera. *IEEE Transactions on Biomedical Engineering*, 64(5):1003–1010, 2016.

[44] Dangdang Shao, Yuting Yang, Chenbin Liu, Francis Tsow, Hui Yu, and Nongjian Tao. Noncontact monitoring breathing pattern, exhalation flow rate and pulse transit time. *IEEE Transactions on Biomedical Engineering*, 61(11):2760–2767, 2014.

[45] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *CVPR 2011*, pages 1297–1304. Ieee, 2011.

[46] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2011.

[47] Rencheng Song, Huan Chen, Juan Cheng, Chang Li, Yu Liu, and Xun Chen. Pulsegan: Learning to generate realistic pulse waveforms in remote photoplethysmography. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1373–1384, 2021.

[48] Radim Špetlík, Vojtech Franc, and Jirí Matas. Visual heart rate estimation with convolutional neural network. In *Proceedings of the British Machine Vision Conference, Newcastle, UK*, pages 3–6, 2018.

[49] Yu Sun, Sijung Hu, Vicente Azorin-Peris, Roy Kalawsky, and Stephen E Greenwald. Noncontact imaging photoplethysmography to effectively access pulse rate variability. *Journal of biomedical optics*, 18(6):061205, 2012.

[50] Lech Świrski and Neil Dodgson. Rendering synthetic ground truth images for eye tracker evaluation. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 219–222, 2014.

[51] Chihiro Takano and Yuji Ohta. Heart rate measurement based on a time-lapse image. *Medical engineering & physics*, 29(8):853–857, 2007.

[52] L Tarassenko, M Villarroel, A Guazzi, J Jorge, DA Clifton, and C Pugh. Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. *Physiological measurement*, 35(5):807, 2014.

[53] Mark van Gastel, Sander Stuijk, and Gerard de Haan. Motion robust remote-ppg in infrared. *IEEE Transactions on Biomedical Engineering*, 62(5):1425–1433, 2015.

[54] David Vazquez, Antonio M Lopez, Javier Marin, Daniel Ponsa, and David Geronimo. Virtual and real world adaptation for pedestrian detection. *IEEE transactions on pattern analysis and machine intelligence*, 36(4):797–809, 2014.

[55] VSR Veeravasarapu, Rudra Narayan Hota, Constantin Rothkopf, and Ramesh Visvanathan. Model validation for vision systems via graphics simulation. *arXiv preprint arXiv:1512.01401*, 2015.

[56] VSR Veeravasarapu, Rudra Narayan Hota, Constantin Rothkopf, and Ramesh Visvanathan. Simulations for validation of vision systems. *arXiv preprint arXiv:1512.01030*, 2015.

[57] VSR Veeravasarapu, Constantin Rothkopf, and Visvanathan Ramesh. Model-driven simulations for deep convolutional neural networks. *arXiv preprint arXiv:1605.09582*, 2016.

[58] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008.

[59] 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.

[60] 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.

[61] 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, 2017.

[62] Erroll Wood, Tadas Baltrusaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: Face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3681–3691, 2021.

[63] Erroll Wood, Tadas Baltrusaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling. Rendering of eyes for eye-shape registration and gaze estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3756–3764, 2015.

[64] Ting Wu, Vladimir Blazek, and Hans Juergen Schmitt. Photoplethysmography imaging: a new noninvasive and noncontact method for mapping of the dermal perfusion changes. In *Optical Techniques and Instrumentation for the Measurement of Blood Composition, Structure, and Dynamics*, volume 4163, pages 62–70. International Society for Optics and Photonics, 2000.

[65] Zitong Yu, Wei Peng, Xiaobai Li, Xiaopeng Hong, and Guoying 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 International Conference on Computer Vision*, pages 151–160, 2019.

[66] Greg Zaal. HDRI haven, 2018.

[67] 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.

[68] Jia Zheng and Sijung Hu. The preliminary investigation of imaging photoplethysmographic system. In *Journal of Physics: Conference Series*, volume 85, page 012031. IoP Publishing, 2007.

[69] Jia Zheng, Sijung Hu, Vassilios Chouliaras, and Ron Summers. Feasibility of imaging photoplethysmography. In *2008 International Conference on BioMedical Engineering and Informatics*, volume 2, pages 72–75. IEEE, 2008.