2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-1179-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/EMBC46164.2021.9631031

Synthetic Data for Multi-Parameter Camera-Based Physiological Sensing

Daniel McDuff[1], Xin Liu[2], Javier Hernandez[1], Erroll Wood[3] and Tadas Baltrusaitis[3]

Abstract— **Synthetic data is a powerful tool in training data hungry deep learning algorithms. However, to date, camera-based physiological sensing has not taken full advantage of these techniques. In this work, we leverage a high-fidelity synthetics pipeline for generating videos of faces with faithful blood flow and breathing patterns. We present systematic experiments showing how physiologically-grounded synthetic data can be used in training camera-based multi-parameter cardiopulmonary sensing. We provide empirical evidence that heart and breathing rate measurement accuracy increases with the number of synthetic avatars in the training set. Furthermore, training with avatars with darker skin types leads to better overall performance than training with avatars with lighter skin types. Finally, we discuss the opportunities that synthetics present in the domain of camera-based physiological sensing and limitations that need to be overcome.**

I. INTRODUCTION

The application of computer vision in non-contact physiological measurement is an area of growing interest [1]. The opportunity to create ubiquitous health sensors out of webcams and smartphones is attractive as it would lower the barrier to measurement and allow for comfortable, convenient longitudinal sensing. Cameras offer the opportunity for capturing rich contextual information, through a myriad of different computer vision models (e.g., scene understanding, action and gesture recognition). Context is often tricky to derive from contact-based sensor data (e.g., wearables) and is necessary for appropriately interpreting physiological measurements.

Significant improvements in camera-based measurement accuracy have been achieved using supervised neural models [2], [3]. Training computer vision algorithms to recover physiological signals from video in this way requires synchronized video recordings and gold-standard measurements. Such data is time consuming and expensive to collect. Furthermore, these recordings contain personally identifiable biometric data. In many domains, such as object recognition [4], body pose [5] and gaze estimation [6], face and gesture recognition [7] and scene understanding [8], synthetic data has become a key tool for training models. Synthetic data pipelines enable systematic data generation with precisely synchronized labels and parameterized control of the dataset properties. A recent paper presented the first example of synthetics applied to the problem of remote imaging photoplethysmography (iPPG) measurement [9]. However, otherwise there remains much to study in this domain.

Once a synthetics pipeline has been created, it can be used to generate data in a highly scalable fashion, but how much does additional data impact the performance of camera-based physiological sensing systems? Will algorithms benefit from learning from hundreds or thousands of synthetic avatars? An attractive property of synthetics pipelines is that the data distribution can be controlled. For example, making it easier to sample uniformly from appearance characteristics such as skin type and gender which are known to differently impact the performance of computer vision models [10]. But does controlling the distribution of synthetic data actually lead to an improvement in how the resulting algorithms perform and help create less biased models? And what other insights can we gain from using synthetic data?

In this paper, we present a set of systematic experiments using synthetic data in the training of non-contact vision-based physiological measurement algorithms. First, we examine how the number of synthetically generated avatars in a training set impacts the performance of pulse rate and breathing rate measurement on two benchmark video datasets. Next, we investigate how the skin type distribution of these avatars impacts performance. To achieve this we leverage a state-of-the-art synthetics pipeline for creating video sequences of avatars with physically-grounded simulations of blood flow and breathing. We synthesize 1,000 high-fidelity videos of avatars in this way, to our knowledge the largest such synthetics dataset that exists.

To summarize the contributions of this work are to: 1) present the first multi-parameter camera-based physiological measurement results using training data created via a synthetics pipeline, 2) show that generalization of both pulse and breathing measurements from video improves with increased numbers of synthetic avatars in the training set, and 3) reveal that training on faces with darker skin types appears to improve model generalization across most skin type categories, compared to training on faces with lighter skin types.

II. BACKGROUND

A. Vision-Based Physiological Measurement

Video-based physiological measurement is an established and growing interdisciplinary field of research. Over the past two decades [11], [12], [13], [14], [15], [16], [17], [2], [3], [18] increasingly sophisticated methods have driven significant reductions in error. The field has benefited by grounding

[1] Daniel McDuff and Javier Hernandez are with Microsoft Research, Redmond, WA, USA {damcduff, javierh}@microsoft.com.

[2] Xin Liu is with the University of Washington, Seattle, WA, USA xliu0@cs.washington.edu.

[3] Erroll Wood and Tadas Baltrusaitis are with Microsoft, Cambridge, UK {erwood, tabaltru}@microsoft.com.

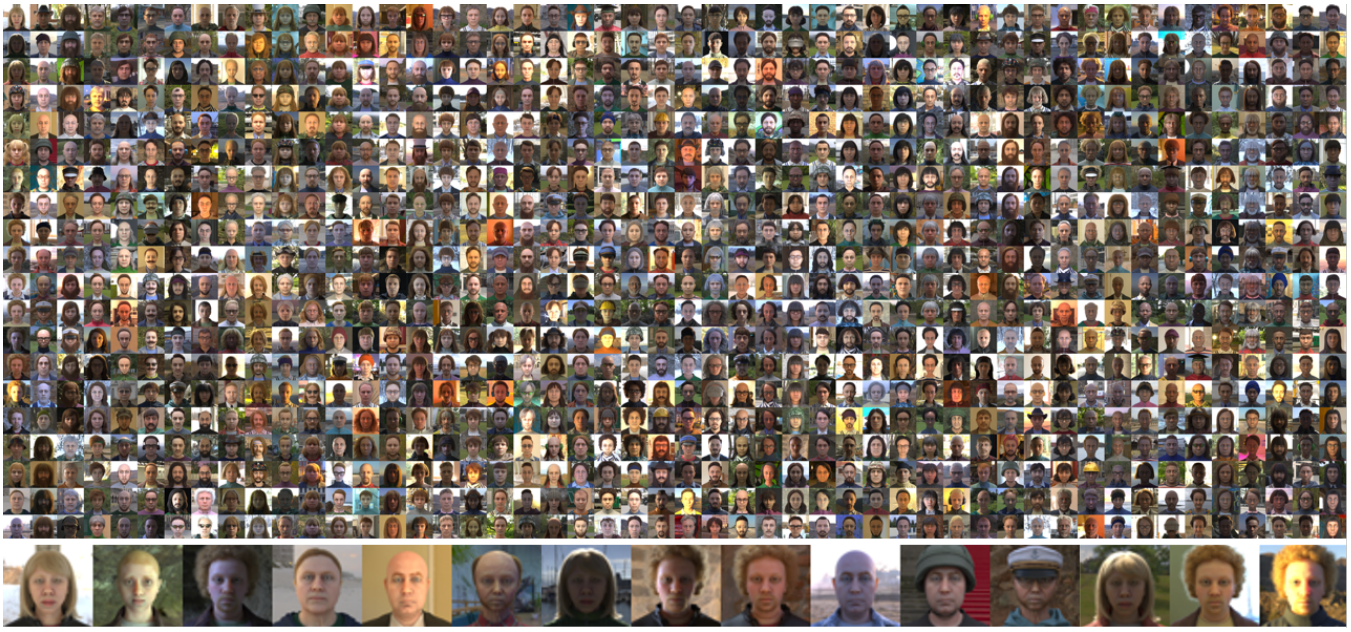

Fig. 1. Screenshots of the 1,000 avatars we created for our analyses. Larger examples are shown below to highlight the visual realism of the avatars and the diversity in appearance, lighting and surroundings.

these models via a principled approach to modeling the optics of the skin [19]. However, the best performing algorithms are by and large supervised neural networks [2], [3], [20]. These algorithms are "data hungry" and can be brittle if only trained on videos that do not reflect the diversity of real-world conditions (appearance, lighting, motion, etc.). Indeed, human appearance and physiology do contain large individual variability making generalization challenging in this domain. Models trained on one dataset and tested on another lead to substantially poorer performance than a model tested on data withheld from the same set as the training data. Specific examples of this include over-fitting to the video codec and/or rate factor of the videos in the training set [21] and to the skin type of the subjects [22].

B. Synthetic Data in Computer Vision

There is a long history of the use of synthetic data in training and evaluating computer vision systems [23], [24], [25], [26], [27], [28], [29], [30], [7], [5], [8], [4], [6]. Synthetics have been employed extensively in models for face and body analysis specifically [31], [7], [28], [29], [32]. But the same is not true for camera-based physiological sensing. One attractive feature of synthetic data generation using parameteric models is the ability to control the distribution of samples. These data can then be used to help address problematic biases that exist in models. For example, Kortylewaski et al. [31], [7] show that the damage of real-world dataset biases on facial recognition systems can be partially addressed by pre-training on synthetic data. In this work we built on this prior work and examine specifically what synthetic data can offer camera-based physiological measurement. In particular, we focused on two questions: How much does increasing the number of subjects with different facial appearances improve cross-dataset generalization performance? And how is performance on subjects with different skin types impacted by the distribution of skin types in the synthetic training set. To avoid conflating the impact of real videos with synthetic data we train *only* on synthetics. Previous work shows that we can expect some additional benefit by combining synthetic and real data [9] but that is not our focus here.

III. SYNTHETIC DATASET

Generating our synthetic dataset involved creating facial avatars that had simulations of facial blood flow and breathing motions. This section provides more details about the synthetics pipeline. We synthesized a large corpora of avatars with unique combinations of facial appearance, head motion and expression, and environment (including ambient lighting configuration). These dimensions were selected based on some of the most significant generalization challenges that existing non-contact physiological models face (e.g., generalizing to different appearance, motion and illumination conditions).

A. Base Avatars

Facial Appearance. Our first goal was to assess how increasing the number of avatars impacted generalization performance. Therefore, we created 1,000 unique appearances which required randomly picking a skin material with a particular albedo texture. For approximately half of the appearances, we included some form of facial hair (beard and/or moustache). In addition, we modified the skin color properties of the different faces. Fig. 1 shows some examples.

Facial Motion. We introduced rigid head motions by rotating the head about the vertical axis at angular velocities of

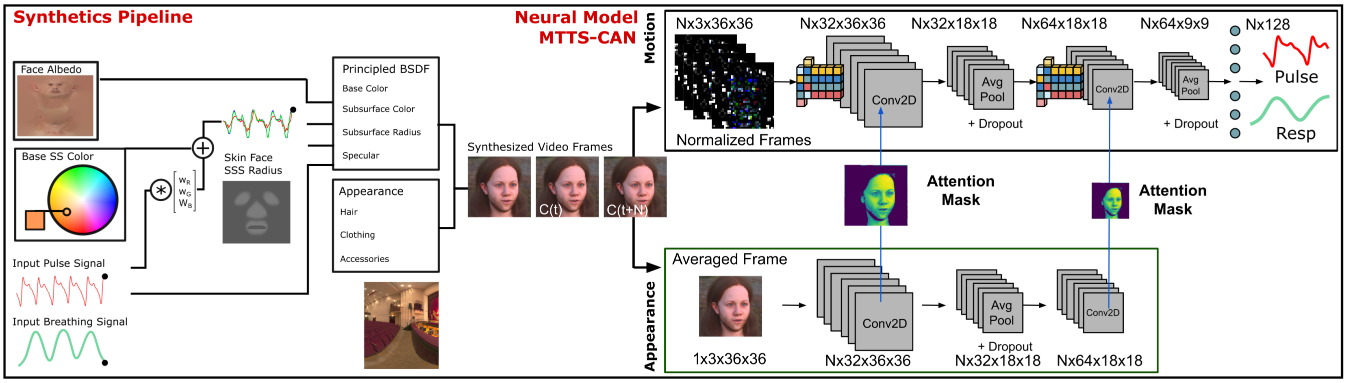

Fig. 2. A combined illustration of our multi-parameter data synthesis pipeline and multi-task temporal shift convolutional attention network for camera-based physiological measurement. In this work the models were trained entirely on synthetic data and tested on real videos.

0, 10, 20, and 30 degrees/second. In addition, we synthesized videos with smiling, blinking, and mouth opening by using a collection of artist-created blend shapes.

Environment. We randomly picked a static background and illumination on the face [33] from a large existing collection [34]. Fig. 1 shows some examples.

B. Physiological Changes

To simulate the facial blood flow, we leveraged photoplethysmographic signals from PhysioNet [35]. In particular, we used the BIDMC PPG and Respiration Dataset [36] which includes 53 8-minute contact PPG and respiration recordings from the original MIMIC-II dataset [37]. We then added each of the signals to the avatars by modifying two main properties of a physically-based shading material[1].

Subsurface Skin Color: We simulated skin tone changes by globally varying subsurface color across all skin pixels on the albedo map which is a texture map transferred from a high-quality 3D face scan.

Subsurface Scattering: We manipulated the subsurface radius for the channels to capture the changes in subsurface scattering as the blood volume varies. In particular, we used an artist-created spatially-weighted scattering radius texture (see Fig. 2) which captures variations in the thickness of the skin across the face.

Breathing Motion: We controlled both the torso and head motions to simulate motions of the body due to breathing. Specifically, the pitch of the chest was rotated subtly using the breathing input signal. The rotation of the head was dampened slightly to create greater variance in the appearance changes and so that the breathing motions were most dominant in the chest and shoulders.

IV. Testing Datasets

We performed all our testing on videos of real people with gold-standard measurements for reference.

AFRL [38]: Videos were recorded at 658x492 pixel resolution and 120 frames per second (fps). Twenty-five participants (17 males) were recruited to participate in the study. Gold-standard PPG measurements were captured from a fingertip sensor and breathing measurements were captured from a chest strap. These were recorded using a research-grade biopotential acquisition unit. Each participant was recorded six times for 5-minutes each. The angular velocity of the head motion was increased in each task and this process was repeated twice in front of two background screens.

MMSE-HR [39]: 102 videos of 40 participants were recorded at 25 fps capturing 1040x1392 resolution images during spontaneous emotion elicitation experiments. The ground truth contact signal was measured via a Biopac2 MP150 system[2] 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. Gold-standard breathing measurements are not included in the MMSE-HR dataset.

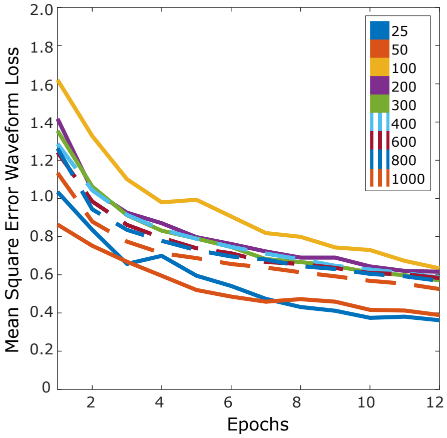

Fig. 3. Training loss over 12 epochs for models trained on [25, 50, 100, 200, 300, 400. 600. 800, 1000] avatars.

V. Model Training and Testing

Our goal here was not to propose a new inference model, therefore we used an existing state-of-the-art neural architec-

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

[2] https://www.biopac.com/

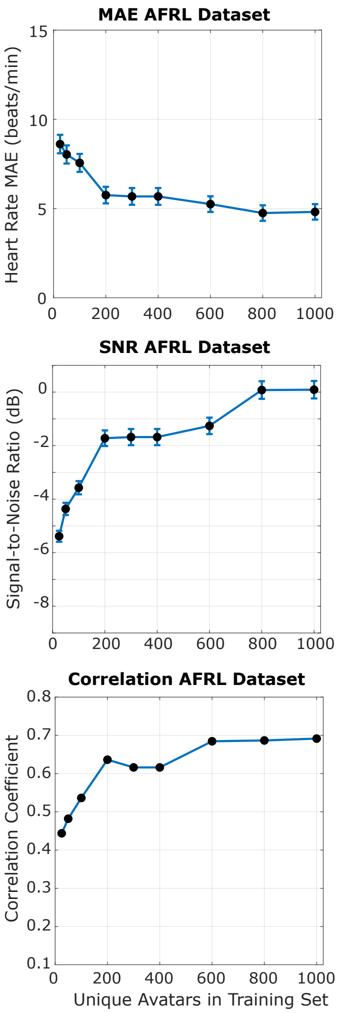

Fig. 4. Pulse rate error (beats/min) for models trained with [25, 50, 100, 200, 300, 400. 600. 800, 1000] randomly sampled avatars. Standard error bars shown for MAE and SNR plots.

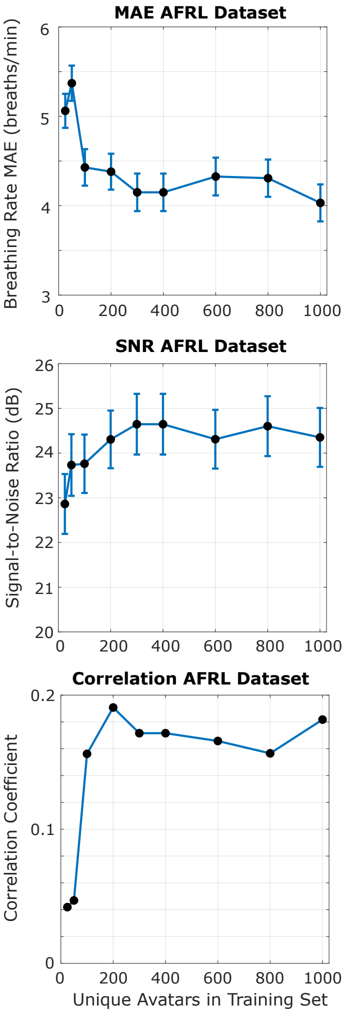

Fig. 5. Breathing results for models trained with [25, 50, 100, 200, 300, 400. 600. 800, 1000] randomly sampled avatars. Standard error bars shown for MAE and SNR plots.

ture for evaluating how synthetic data can improve camera-based physiological sensing. This model, multi-task temporal shift convolutional attention network (MTTS-CAN) [3], captures rich spatial and temporal relationships in the data and is therefore a good candidate for investigating the impact of the synthetic data. MTTS-CAN has a two-branch network, illustrated in Fig. 2, comprising of a motion branch and an appearance branch. The motion branch efficiently models spatial-temporal features by shifting the frames along the temporal axis. The appearance branch provides an attention mechanism to help guide the motion representation to focus on spatial regions of interests (e.g., skin) containing physiological signals instead of others (e.g., hair). The loss function during training was the average mean squared error of pulse and breathing waveform predictions compared to the ground truth waveforms. We trained using the first-derivative of the waveforms as prior work has shown that this is effective for predicting both PPG and breathing signals from video [2].

Using our synthetics pipeline, we generated videos of 1,000 avatars with blood flow (PPG) and breathing signals.

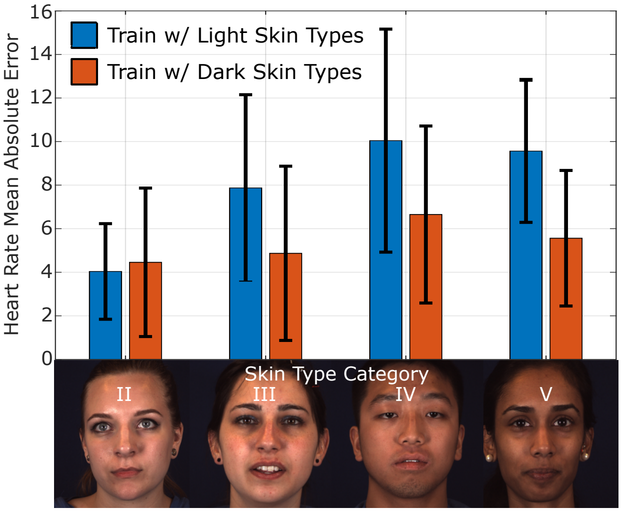

Fig. 6. Pulse rate error (beats/min) for models trained with light skin type (blue) and dark skin type (red) avatars. Results are shown by Fitzpatrick skin type (II, III, IV, V) for subjects in the MMSE-HR dataset.

Fig. 1 shows a frame from each of the avatar videos. The videos were six seconds in length and were created at a frame-rate of 30Hz, providing 180 frames per avatar and a total of 180,000 frames. This is still a relatively small number of frames compared to many video datasets, but arguably has much greater variability in terms of appearance. Synthesizing face videos is still relatively computationally expensive and time consuming. Creating these high fidelity avatar videos took approximately one month using three computers each with an Nvidia M40 GPU.

For each model in our experiment, we trained the network for a fixed number (12) epochs. Fig. 3 shows that although we vary the number of avatars in our experiments, the training loss consistently converges by this point. To evaluate the trained models we tested on the AFRL and MMSE datasets. For the AFRL data we divide each video into 5 60-second non-overlapping windows. For the MMSE dataset we use the full video for each prediction. We use three commonly employed evaluation metrics for pulse and breathing: mean absolute error between the predicted rate and the ground truth, signal-to-noise ratio (SNR), and correlation between predicted rate and the ground truth.

VI. SYNTHETICS IN TRAINING

A. Number of Subjects in the Training Set

First, we tested models in which we trained networks with data from [25, 50, 100, 200, 300, 400, 600, 800, 1000] avatars. We randomly sample from the pool of avatars to create the subsets for training. Fig. 4 shows the mean absolute pulse rate error, SNR and correlation for the AFRL and MMSE datasets (error bars show standard error). Fig. 5 shows the mean absolute breathing rate error, SNR and correlation for the AFRL dataset (the MMSE dataset does not include breathing ground truth data). In both cases (pulse rate and breathing rate estimation) the testing mean absolute error dropped significantly for networks trained with 600 avatars compared to 25 avatars. For pulse rate measurement we saw diminishing returns with more than 600 avatars and with breathing rate measurement we saw diminishing returns with more than 200 avatars. These numbers probably reflect the limits of the variance in samples that can be created with out current simulation. After generating 600 unique avatars any additional avatars we generated were likely to have some similarities with one or more of the avatars in the set, thus these additional training samples offered little additional new information. We plan to develop our synthetics pipeline further to address this bottleneck.

B. Sim-to-Real Gap

From the results in Figs. 4 and 5 it was apparent that even when training with 1000 avatars the heart rate and breathing rate measurement performance was still not comparable with the state-of-the-art [3], [18], [9]. We hypothesize that this is due to the "sim-to-real" gap. Training only on synthetic data has limitations because there remains a domain gap between the appearance of synthetic avatars and real videos. Combining our synthetic data with real videos will likely lead to improvements [9] as would other methods for addressing "sim-to-real" generalization. Our goal here was to primarily to examine the impact of synthetic dataset properties and therefore, we leave these steps for future work.

C. Diversity of Subjects in the Training Set

Our synthetic data allowed us to examine empirically how the distribution of appearance of the avatars impacts generalization performance. Perhaps the most obvious appearance characteristic when measuring the blood volume pulse optically is skin type. Previous work has found systematic biases in the performance of non-contact photoplethymography measurement algorithms with skin type [40], [22]. Specifically, performance is over signficantly poorer for darker skin types. The albedo textures were sorted by skin tone. We then trained two models one with "lighter" skin tone avatars and another with "darker" skin tone avatars. Fig. 6 shows the pulse rate mean absolute error (beats/min) for the two models on the MMSE-HR dataset by Fitzpatrick skin type [41]. We expected to observe that the model trained on light skin type avatars would perform best on light skin type participants (groups II and III) and the model trained on darker skin type avatars would perform best on dark skin type participants (groups IV and V). However, we observed that the model trained on dark skin type avatars performed as well as, or better than the other model on all categories. This suggests that training on data with darker skin types leads to a more robust model, perhaps because the task is harder - forcing the model to learn better representations or a more robust attention mechanism. This is an interesting observation that warrants further investigation.

VII. DISCUSSION

Our results highlight that synthetic data can be used to train multi-parameter camera-based physiological measurement algorithms. Our experiments have shown that increasing the diversity of appearance of avatars can positively impact generalization performance on real videos for both pulse and breathing measurement. A rich synthetics pipeline presents the possibility to exploit this further, by generating data with more varied facial expressions, body motions, and illumination conditions. In addition to leveraging these at training time, they could also be used for systematic testing. However, it is clear that there is a gap between the performance that can be obtained when training only on our current simulation data.

A. Opportunities for Synthetics

Synthetics open up a number of promising directions for camera-based physiological measurement. As shown in this work, the controlled generation of different facial attributes and potential variations (e.g., motion, environment) allowed us to systematically study the effect of different factors and create more generalizable models. In addition, this methodology allowed us to gain more understanding about the impact of each of the variations in a standardized setting which could be used for prioritizing research questions such as "what is the source of variation that most heavily impacts the algorithm?." The generation of synthetic data enabled us to quickly increase the volume of data. This is likely be even more advantageous for training even more data hungry models such as transformers. Finally, the proposed methodology offers the potential to improve other relevant physiological domains. For instance, this work considered data from healthy individuals to generate the physiological signals but future work may consider leveraging datasets with cardio-respiratory abnormalities that may be difficult to replicate during data collection from real people, especially in the context of camera-based physiological sensing.

B. Limitations of Synthetics

Despite the many benefits, it is important to note that the proposed approach has some limitations too. One of the biggest challenges is that the creation of hyper-realistic avatars like the ones considered in this work requires a large overhead both in terms of time and expense. Once the synthetics pipeline is created, generating the avatars themselves is computationally intensive. However, we expect these costs to be reduced in the future. Another limitation is that using synthetic data can only get us so far in terms of performance. Even though we leveraged state-of-the-art avatar generation, there is still some domain gap between real people and synthesized ones. Future efforts may leverage advances in generative networks like StyleGAN [42] to further increase their realism and bridge the "sim-to-real" gap.

VIII. Conclusion

We have shown, via empirical evidence that increasing the number of unique avatars in a synthetic dataset can lead to reductions in physiological parameter estimation. We hope that this evidence inspires more research using synthetic data for training camera-based physiological sensing algorithms and that in turn we are able to realize more of the potential for this technology.

References

[1] D. McDuff, J. R. Estepp, A. M. Piasecki, and E. B. Blackford, "A survey of remote optical photoplethysmographic imaging methods," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2015, pp. 6398–6404.

[2] W. Chen and D. McDuff, "Deepphys: Video-based physiological measurement using convolutional attention networks," *arXiv preprint arXiv:1805.07888*, 2018.

[3] X. Liu, J. Fromm, S. Patel, and D. McDuff, "Multi-task temporal shift attention networks for on-device contactless vitals measurement," *Neural Information Processing Systems (NeurIPS)*, 2020.

[4] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2017, pp. 23–30.

[5] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *CVPR 2011*. Ieee, 2011, pp. 1297–1304.

[6] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2107–2116.

[7] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, and T. Vetter, "Analyzing and reducing the damage of dataset bias to face recognition with synthetic data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[8] A. Handa, V. Patraucean, V. Badrinarayanan, S. Stent, and R. Cipolla, "Understanding real world indoor scenes with synthetic data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4077–4085.

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

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

[11] T. Wu, V. Blazek, and H. J. Schmitt, "Photoplethysmography imaging: a new noninvasive and noncontact method for mapping of the dermal perfusion changes," in *EOS/SPIE European Biomedical Optics Week*. International Society for Optics and Photonics, 2000, pp. 62–70.

[12] V. Blazek, T. Wu, and D. Hoelscher, "Near-infrared ccd imaging: Possibilities for noninvasive and contactless 2d mapping of dermal venous hemodynamics," in *Optical Diagnostics of Biological Fluids V*, vol. 3923. International Society for Optics and Photonics, 2000, pp. 2–9.

[13] C. Takano and Y. Ohta, "Heart rate measurement based on a time-lapse image," *Medical engineering & physics*, vol. 29, no. 8, pp. 853–857, 2007.

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

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

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

[17] W. Wang, S. Stuijk, and G. de Haan, "Exploiting spatial redundancy of image sensor for motion robust rppg." *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 2, pp. 415–425, 2015.

[18] X. Liu, Z. Jiang, J. Fromm, X. Xu, S. Patel, and D. McDuff, "Metaphys: few-shot adaptation for non-contact physiological measurement," in *Proceedings of the Conference on Health, Inference, and Learning*, 2021, pp. 154–163.

[19] W. Wang, A. Den Brinker, S. Stuijk, and G. De Haan, "Algorithmic Principles of Remote-PPG," *IEEE Transactions on Biomedical Engineering*, vol. PP, no. 99, pp. 1–12, 2016.

[20] Q. Zhan, W. Wang, and G. de Haan, "Analysis of cnn-based remote-ppg to understand limitations and sensitivities," *Biomedical optics express*, vol. 11, no. 3, pp. 1268–1283, 2020.

[21] E. M. Nowara, D. McDuff, and A. Veeraraghavan, "Systematic analysis of video-based pulse measurement from compressed videos," *Biomedical Optics Express*, vol. 12, no. 1, pp. 494–508, 2021.

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

[23] V. Veeravasarapu, R. N. Hota, C. Rothkopf, and R. Visvanathan, "Model validation for vision systems via graphics simulation," *arXiv preprint arXiv:1512.01401*, 2015.

[24] ——, "Simulations for validation of vision systems," *arXiv preprint arXiv:1512.01030*, 2015.

[25] V. Veeravasarapu, C. Rothkopf, and V. Ramesh, "Model-driven simulations for deep convolutional neural networks," *arXiv preprint arXiv:1605.09582*, 2016.

[26] D. Vazquez, A. M. Lopez, J. Marin, D. Ponsa, and D. Geronimo, "Virtual and real world adaptation for pedestrian detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 4, pp. 797–809, 2014.

[27] W. Qiu and A. Yuille, "Unrealcv: Connecting computer vision to unreal engine," in *European Conference on Computer Vision*. Springer, 2016, pp. 909–916.

[28] D. McDuff, R. Cheng, and A. Kapoor, "Identifying bias in ai using simulation," *arXiv preprint arXiv:1810.00471*, 2018.

[29] D. McDuff, S. Ma, Y. Song, and A. Kapoor, "Characterizing bias in classifiers using generative models," *Advances in Neural Information Processing Systems*, vol. 32, pp. 5403–5414, 2019.

[30] R. M. Haralick, “Performance characterization in computer vision,” in *BMVC92*. Springer, 1992, pp. 1–8.

[31] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, and T. Vetter, “Empirically analyzing the effect of dataset biases on deep face recognition systems,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2093–2102.

[32] T. Baltrusaitis, E. Wood, V. Estellers, C. Hewitt, S. Dziadzio, M. Kowalski, M. Johnson, T. J. Cashman, and J. Shotton, “A high fidelity synthetic face framework for computer vision,” *arXiv preprint arXiv:2007.08364*, 2020.

[33] P. Debevec, “Image-based lighting,” in *ACM SIGGRAPH 2006 Courses*, 2006, pp. 4–es.

[34] G. Zaal, “HDRI haven,” 2018.

[35] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, “Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals,” *circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[36] M. A. Pimentel, A. E. Johnson, P. H. Charlton, D. Birrenkott, P. J. Watkinson, L. Tarassenko, and D. A. Clifton, “Toward a robust estimation of respiratory rate from pulse oximeters,” *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 8, pp. 1914–1923, 2016.

[37] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark, “Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database,” *Critical care medicine*, vol. 39, no. 5, p. 952, 2011.

[38] J. R. Estepp, E. B. Blackford, and C. M. Meier, “Recovering pulse rate during motion artifact with a multi-imager array for non-contact imaging photoplethysmography,” in *Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1462–1469.

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

[40] P. S. Addison, D. Jacquel, D. M. Foo, and U. R. Borg, “Video-based heart rate monitoring across a range of skin pigmentations during an acute hypoxic challenge,” *Journal of clinical monitoring and computing*, vol. 32, no. 5, pp. 871–880, 2018.

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

[42] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June. IEEE Computer Society, jun 2019, pp. 4396–4405.