Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
SCAMPS is a dataset of high fidelity synthetic human simulations that are designed for the purposes of training and testing camera physiological measurement methods.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
The data were created by the Human Synthetics, HUE and Biomedical Computing Teams at Microsoft.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.
All work associated with the SCAMPS dataset was funded by Microsoft.

Any other comments?
No.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
The dataset consists of RGB frames and segmentation maps with synchronized ground-truth signals: interbeat and breath intervals, PPG, ECG and breathing waveforms precisely aligned with each video frame. Facial actions, blinking and head pose labels are also provided.

How many instances are there in total (of each type, if appropriate)?
The dataset contained 2,800 videos, each of 600 frames.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
We could not generate all possible combinations of the physiological parameters, appearance properties and environments as this would have been an extremely large number of videos and computationally extremely expensive. However, we created a broad range of physiological parameters and appearance characteristics, which is one of the advantages of creating data from a simulation. While these data are not entirely representative of the real-world they do contain significantly more diversity than existing video PPG datasets.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
The dataset consists of raw images/videos and physiological waveforms.

Is there a label or target associated with each instance? If so, please provide a description.
The labels are the physiological waveforms and statistics (e.g., heart rate and breathing rate).

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
The dataset is complete (i.e., there are no missing labels or partial instances).

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.
Not applicable.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
We have provided a training, validation and testing split for the dataset. This split is avatar/person independent and all sets contain a diverse set of examples (different head motions, facial movements, etc.) Other train, validation and test splits could be created to test generalization to distribution shifts (i.e.,

train on videos with no motion, test on video with motion) too.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
There are no sources of noise that were added to the dataset. Simulations have the advantage of providing "perfect" labels in the sense that the generated data is defined by the labels. However, one might consider the approximations baked into a simulation as introducing noise compared to observations in the real-world. So it is important to be aware of the sim2real gap between a model trained on synthetic data.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
The dataset is self contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' nonpublic communications)? If so, please provide a description.
The dataset does not contain data that might be considered confidential.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
The dataset does not contain data that might be offensive, insulting, threatening or cause anxiety.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
We do not provide age or gender labels with these videos. While the facial scans do cover a range of ages and gender identities the synthesized faces can be quite different from the scans and therefore no longer reflect the age or gender.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
No, the data is entirely synthetic.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
No, the data is entirely synthetic.

Collection/Generation Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
The data is synthetic (i.e., created by a computer graphics engine). Therefore, it might not be considered derived from other data (i.e., the facial scans). However, this distinction is probably not particularly appropriate or helpful in this instance.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?
The data were created using a 3D synthetics engine. The engine builds on 3D assets.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The dataset is sampled from the simulation engine and is not an exhaustive set of the possible parameter combinations. We sampled parameters uniformly and randomly. The generation of high-fidelity synthetics is computationally expensive, so generating examples from all possible parameter combinations was not possible.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?
The dataset creation was carried out by Microsoft employees.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.
The final data were generated from January - March 2022.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
The dataset creation was reviewed by the Microsoft Office for Responsible AI. Our dataset, transparency note and paper contain information about the dataset requested during this process.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
Assets from Triplegangers (https://www.triplegangers.com/) and 3DScanStore (https://www.3dscanstore.com/) were used in the synthetic pipeline. However, while the synthetic data we release was constructed using these assets the rendered data do not bear a resemblance to any of the subjects in the scans and the physiological data is entirely synthetic.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
Not applicable.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
Not applicable.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).
Not applicable.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
Not applicable.

Any other comments?
No.

Preprocessing/Cleaning/Labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.
The dataset released was not preprocessed in anyway after generation from the simulation engine.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
Yes, the "raw" data was saved and is released, without further preprocessing.

Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.
Not applicable.

Any other comments?
No.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.
The dataset has been used for the example task of video heart rate measurement (rPPG) in our paper.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
This is a new dataset and there are currently no papers or systems that use the dataset. However, in future we will link to papers and systems from our project page.

What (other) tasks could the dataset be used for?
The dataset could be used for training video methods for cardiac and pulmonary measurement and facial action unit detection.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?
This dataset was created for research and experimentation on camera measurement of physiological signals. While the dataset is useful for testing models, is not designed as a test set for evaluating the clinical efficacy of a model, just because a model performs well on synthetic data does not mean it will generalize to videos of real people. Therefore, no medical device trained or tested exclusively on these data would be suitable for use in the real-world.

Are there tasks for which the dataset should not be used? If so, please provide a description.
The SCAMPS dataset was not designed for computer vision tasks such as face recognition, gender recognition, facial attribute recognition, or emotion recognition. We do not believe this dataset would be suitable for these applications without further validation.

Any other comments?
No.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
The dataset is publicly available, as described below. The only restrictions are on commercial use.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
The dataset will be distributed on our GitHub page: https://github.com/danmcduff/scampsdataset. The download is available as a tar file.

When will the dataset be distributed?
The dataset is currently available as of June 2022.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
The dataset will be released under a custom research license. The license does not restrict who can access the data, but does require that the dataset only be used in non-commercial research activities (which can include non-commercial research undertaken by or funded via a commercial entity). The data may not be used in any commercial offering, including as part of a product or service.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions,

and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
We are not aware of any export controls or regulatory restrictions that apply to this dataset or any individual instance.

Any other comments?
No.

Maintenance

Who will be supporting/hosting/maintaining the dataset?
Daniel McDuff (danjmcduff@gmail.com)

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
By email: danjmcduff@gmail.com or GitHub by filing a bug/issue report.

Is there an erratum? If so, please provide a link or other access point.
We have a erratum document in our GitHub repo: https://github.com/danmcduff/scampsdataset/blob/main/erratum.txt This will be used for documenting errata and updates to the dataset.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?
The dataset and documentation will be updated to address any errors. Updates will be made promptly and as and when needed.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
Not applicable.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.
If new versions of the dataset are created, we will maintain access to the original dataset on our project page.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.
We are very happy to link to, or host, additional preprocessed versions of these data. Many models require data to be preprocessed in a particular way. For reproducibility it is very helpful to have the code and examples of the preprocessed data.

Any other comments?
No.