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.
Pre-training with synthetic data overcomes different costs and concerns (ethical, privacy, or copyright related) associated with using real data. Secondly, using the results of multiple baselines we presented observed relationships between performance of synthetic pre-training and a dataset's representation bias.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
The dataset was created by Yo-whan Kim, Samarth Mishra, SouYoung Jin, Rameswar Panda, Hilde Kuehne, Leonid Karlinsky, Venkatesh Saligrama, Kate Saenko, Aude Oliva, and Rogerio Feris at MIT-IBM Watson AI Lab.

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.
N/A

Any other comments?
None.

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 instances are synthetic videos collected from three different pre-existing datasets/generators: ElderSim, PHAV, and SURREACT.

How many instances are there in total (of each type, if appropriate)?
There are 150 classes in total, each containing up to 1,000 examples.

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).
The dataset is a sample of instances from a larger set. 55 actions from ElderSim, 100 actions from SURREACT, and 35 actions from PHAV were collected. We manually screened and combined overlapping classes, and randomly selected 1,000 samples for each class. For classes with samples from multiple assets, an equal number of videos were sampled from each asset to maintain an adequate ratio.

What data does each instance consist of? “Raw” data (e.g., unprocessed text or images)or features? In either case, please provide a description.
Each instance is a raw synthetic video of a human agent performing an action.

Is there a label or target associated with each instance? If so, please provide a description.
The label is one of 150 human action classes.

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.
N/A

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.
N/A

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.
Training and validation sets were generated randomly using the available examples. The motivation of SynAPT is for pre-training, so we aim to include as many characteristics of videos (e.g. lighting, humanoid type, camera angle, etc.) as possible into the training set.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
Some humanoid actions may seem “incorrect” or “noisy” to human eyes as they are computer generated.

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 composed of three different assets as described above, and the user must download the assets from the original sources after signing their requirements and agreements. After the assets are downloaded, the dataset can be accessed completely offline without APIs. These assets will remain constant as long as the providers do not take down the download links for any unforeseen reasons. We have an archive of the original assets and the synthetic dataset for SynAPT in our server. Although there aren’t any licenses or fees associated with the assets, users must sign usage agreement forms.

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)?
No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
Although some human actions might come off as offensive (e.g. kicking, punching), please note that these are videos of computer graphics generated humanoids. One of the main purposes of SynAPT is to overcome the ethical problems of real video data.

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.
No.

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.

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.

Any other comments?
None.

Collection 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 was generated by computer graphics simulators.

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?

ElderSim, SURREACT, and PHAV computer graphics simulators were used to produce the data.

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

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)?
As mentioned, the data was sampled from publicly available assets.

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.
N/A

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.
N/A, the synthetic videos for SynAPT were curated from public assets, and this process did not involve any human subject.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
As described above, the data was obtained from public sources..

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.
N/A.

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.
N/A.

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).
N/A.

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.
N/A.

Any other comments?

None.

Processing/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.
No preprocessing and cleaning were necessary. However, the three assets had identical action classes with different names (e.g. 'punching' and 'punch'). These overlapping classes were merged manually.

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.
Raw data can still be downloaded from the public assets it was sourced from.

Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.
N/A.

Any other comments?
None.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.
Looking at the biases of the pre-training datasets, we see Kinetics has a much higher representation bias than Synthetic.

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.
N/A.

What (other) tasks could the dataset be used for?
SynAPT can also be used to measure a model's performance on synthetic videos for any form of representation learning.

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?

One of the main purposes of SynAPT is to overcome the ethics, privacy, copyright concerns, and other concerns of existing real video datasets. As of now, we do not see these issues impacting future uses.

Are there tasks for which the dataset should not be used? If so, please provide a description.
Please adhere to the original usage agreements of the three assets.

Any other comments?
None.

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.
Yes, documentation on how to recreate the synthetic dataset for SynAPT is available.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
The documentation and instructions will be available in a GitHub repository.

When will the dataset be distributed?
The dataset is first released in 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.
Please agree to the ToU of the three original assets. No further IP license or ToU is required.

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.
Users must sign the ToU for the three assets.

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.
No.

Any other comments?

None.

Maintenance

Who will be supporting/hosting/maintaining the dataset?
The original asset curators will host the assets. Yo-whan Kim et al. will maintain the SynAPT recreation GitHub.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
Yo-whan Kim can be contacted via yowhan@mit.edu.

Is there an erratum? If so, please provide a link or other access point.
No.

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)?
This will be posted on the GitHub.

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.
N/A.

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.
Older versions can be viewed on the GitHub.

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.
Others may do so and should contact the original authors regarding any fixes or extensions.

Any other comments?
None.