Dataset of 3D Garments with Sewing Patterns v1

Maria Korosteleva, Sung-Hee Lee
Graduate School of Culture Technology
KAIST
mariako, sunghee.lee@kaist.ac.kr

2021

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.

Dataset of 3D Garments with Sewing Patterns was created to fill in the gap in the data availability for research on Deep Learning for garments, and as a case of a more general problem of Deep Learning for structured deformable objects. To the best of our knowledge, this is the first dataset that both densely explores the design spaces of common garments and provides sewing patterns, corresponding draped garments, and their noisy versions at the same time.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

As released in 2021, the initial version of this dataset is created by Maria Korosteleva and Sung-Hee Lee, who are part of the Motion Computing Lab, Graduate School of Culture Technology, KAIST, with support of the lab members.

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.

This work was supported by National Research Foundation, Korea (NRF-2020R1A2C2011541), and Korea Creative Content Agency, Korea (R2020040211).

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.

Each instance of the dataset is a garment design sample, described as a sewing patterns, draped 3D models, one clean, one noisy imitating artifacts of 3D scanning process, and renders of the clean 3D model as draped over the body. Every instance is a variation of one of the 19 base garment designs. All instances are draped over the same 3D body model of an average woman body shape provided by SMPL Body Model.[1]

All instances of the dataset use the same base body for draping and utilize the same values of physics simulation parameters. All the 3D models are aligned with each other

[1] https://smpl.is.tue.mpg.de/

and the base body in 3D. All the 2D rendered images are aligned with each other as well.

How many instances are there in total (of each type, if appropriate)?

Base Garment Type	#Samples	#Quality Samples
Total	**23500**	**22647**
Training groups	**22450**	**21636**
T-Shirt	2300	2238
T-Shirt Sleeveless	1800	1799
Jacket	2200	2131
Hood Jacket (Hoody)	2700	2166
Skirt 2 panels	1200	1199
Skirt 4 panels	1600	1600
Skirt 8 panels	1000	973
Pants	1000	970
Waistband Pants	1500	1474
Dress Sleeveless	2550	2532
Waistband Dress	2600	2561
Jumpsuit Sleeveless	2000	1993
Test groups	**1050**	**1011**
Jacket sleeveless	150	150
Hoody sleeveless	150	138
Poolover hoody	150	143
Waistband skirt	150	133
Dress with sleeves	150	150
Jumpsuit with sleeves	150	149
Waistband Jumpsuit	150	148

Table 1: Per-type breakdown of the dataset instances counts. The table shows the total number of instances generated for each type (*#Samples*) and the number of instances after filtering failure cases (*#Quality Samples*). Failure cases include garments with simulation artifacts such as severe body- and self-intersections, garments sliding off during draping, etc. The base types are grouped into those designed for model training (*Training groups*) and those designed for evaluation (*Test groups)* with smaller overall counts of samples.

Table 1 shows the total count and per-type breakdown of the dataset instances.

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 garment designs that may be encountered in the real world. The dataset is not representative of the larger set of garment designs, as the distribution of the latter is highly diverse, dynamic and not known exactly. Instead, the motivation was to resemble the variety of designs that exist within a garment type while covering this diversity uniformly. As for the garment types, the creators build upon the garment choices of earlier research work on 3D garments while adding some variation to it. This dataset should be considered as one of the early steps of representing garment designs with sewing patterns and 3D models.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Every instance contains the following components:

- garment sewing pattern (specified in JSON file);
- rendered images of sewing pattern panels (as PNG and SVG files)
- 3D model of a garment as draped over an average women body provided by SMPL Body Model in OBJ file format;
- front and back rendered images of the said 3D model worn by the body model (as PNG files);
- a corrupted version of the 3D model with removed occluded surfaces imitating the artifacts of a 3D scanning process (as OBJ file);
- per-vertex segmentation labels for each clean and corrupted 3D model indicating the panel each vertex belongs too (as TXT file).

Is there a label or target associated with each instance? If so, please provide a description.

An instance describes a single garment designs in multiple modalities, each of which could be used as input or as target depending on the task (e.g., from pattern to 3D model, from 2D image to pattern, etc.)

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.

Yes, a small fraction of instances lacks 3D models due to the errors in simulation process. All these instances are explicitly marked as failure cases in the properties files that accompany every type, and thus can be trivially filtered out.

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.

Yes, the instances are explicitly grouped by the base type they were sampled from. The grouping is reflected in the directory structure.

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.

The dataset contains 19 garment groups, seven of which are recommended to be used for evaluation only while others could be used for training, as indicated in Table 1. The base garments of the test groups are constructed to use the same components as the training garments, but those are arranged to form new sewing pattern topologies (e.g., training set contains sleeveless jumpsuits and T-shirts with sleeves while test set contains jumpsuits with sleeves). This split is designed to evaluate generalization across sewing pattern topologies in the tasks that require it.

However, the dataset users can freely design other splits as required for their tasks.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Some instances of the dataset contain sewing patterns with extreme design choices which the automatic simulation process cannot properly resolve, thus producing 3D models with cases of body- and self-intersection or fallen-off garments. A small number of models are missing due to the unexpected failures of the simulator. All these failure cases are explicitly listed in the `dataset_properties.json` files that accompany every type, and thus could be trivially filtered out. Table 1 provides the number of total and clean instances in the dataset.

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 future user? 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 non-public communications)? If so, please provide a description.

No. The garment designs and associated files are fully artificial, and the base body mesh of SMPL is not associated with any particular person.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

We expect little risk in these matters. In some of the rendered images the garments do not cover the body fully hence revealing some of the sensitive body areas. However, the base body model is shaped as if wearing tight clothing and does not depict any particular details.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No

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 data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Sewing patterns associated with every instance are sampled from the manually designed base sewing pattern templates. These base templates are provided for each type in corresponding subfolders. Other garment representations are obtained using these sampled sewing patterns.

The dataset was generated using an original software developed by the authors (available on GitHub) that integrates multiple widely used solutions into the generation process. Draped garment models were obtained using commercial Qualoth physics simulator and rendered with Arnold. Scan imitation was performed though visibility testing using the Autodesk Maya Python API ray tracing routines.

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

Dataset instances are uniformly sampled from corresponding parametric sewing pattern templates, which are manually designed by the authors. The number of samples for training templates is decided upon heuristically. The complexity of a template design space approximately correlates with a number of parameters defined in that template. Hence, to ensure proper coverage of each design space, about 200 samples per parameter were generated from each base type, with added 150 samples for validation, 150 samples for testing, and 50-200 samples for simulation failure allowance. For example,

$$1000 = 200 * 3 + 150 + 150 + 100$$

samples were drawn from pants template that has three parameters. The number of samples from templates of test group are fixed at 150 samples drawn from each template. Final counts are given in Table 1.

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 data collection process was automatic for the most part. The manual labor was involved in designing the base parametric templates, which were created by the dataset authors themselves.

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 bulk of the data was generated in March, 2021-May 2021. The data generation pipeline was in development from February 2020 to May 2021.

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.

No, there was no need for ethical review as the dataset is fully synthetic.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No

Any other comments?

The data generation software and related materials are publicly available on GitHub: https://github.com/maria-korosteleva/Garment-Pattern-Generator.

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 remainder of the questions in this section.

No, all the needed operations on the data instances were performed during the data generation process.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No

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.

No

What (other) tasks could the dataset be used for?

The dataset provides multiple representation of the same instance of garment design, hence is suitable to explore connections between any of the given modalities. For example, the dataset could be used for the tasks of neural draping with sewing patterns as inputs, estimating structure of garments from 3D models or 2D images, garment segmentation into panel pieces, as well as geometry evaluation from images.

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 future user 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 undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

No

Are there tasks for which the dataset should not be used? If so, please provide a description.

The dataset should not be used in any tasks that heavily rely on realistic distribution of garment designs.

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, the dataset is available publicly for anyone interested to 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 is distributed through Zenodo[2], which will ensure the long term data availability. Dataset DOI: https://doi.org/10.5281/zenodo.5267549.

When will the dataset be distributed?

The dataset is planned to release before September 2021.

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.

This dataset is licensed under CC BY 4.0.[3]

[2] https://zenodo.org/

[3] https://creativecommons.org/licenses/by/4.0/

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.

The dataset utilizes renders of the SMPL body model which is shared under CC BY 4.0 license.

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?

Authors bear all responsibility in case of violation of rights due to publication of the dataset.

Maintenance

Who will be supporting/hosting/maintaining the dataset?

Dataset hosting is outsourced to the dataset hosting provider. Additional support and management will be provided by the dataset authors and other members of Motion Computing Lab, KAIST.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Main contact:
Maria Korosteleva (mariako@kaist.ac.kr).
Additional contact:
Sung-Hee Lee (sunghee.lee@kaist.ac.kr).

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 users (e.g., mailing list, GitHub)?

Yes, the development of the dataset is planned to continue, and contributions from users are also welcomed. New versions the dataset will be shared and announced on the GitHub repo front page as soon as they are available.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were 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.

The dataset does not relate to people.

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 users.

Yes, we plan to support versioning of the dataset so that all the versions are available to potential users. Zenodo platform will maintain the history of version, and each version will have a unique DOI assigned.

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 other users? If so, please provide a description.

Contributions of both the sewing pattern templates and generated datasets are highly appreciated. Additional templates as well as new features will be accepted though the pool request process to the data generator repository on GitHub. New garment data samples can be added to the dataset, upon request on GitHub or though an email to dataset managers. The availability of this option might be limited by Zenodo data storage limits. All contributions will be properly acknowledged.