Hardware Design and Accurate Simulation for Benchmarking of 3D Reconstruction Algorithms: Dataset Documentation

Sebastian Koch*
TU Berlin
s.koch@tu-berlin.de

Yurii Piadyk*
New York University
ypiadyk@nyu.edu

Markus Worchel
TU Berlin
m.worchel@campus.tu-berlin.de

Marc Alexa
TU Berlin
marc.alexa@tu-berlin.de

Claudio Silva
New York University
csilva@nyu.edu

Denis Zorin
New York University
dzorin@cs.nyu.edu

Daniele Panozzo
New York University
panozzo@nyu.edu

1 Dataset Documentation

In the following, we describe our dataset in the common datasheets format [7].

1.1 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. The dataset was generated to train and evaluate common post-processing algorithms in 3D scanning, such as noise removal, shape completion or surface reconstruction. To the best of our knowledge, there is no such dataset available which contains actual scans together with matching simulated scans and high precision ground truth information.

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 in a joint effort by Sebastian Koch, Yurii Piadyk, Markus Worchel, Marc Alexa, Claudio Silva, Denis Zorin and Daniele Panozzo. The authors are researchers affiliated with Technische Universität Berlin or New York University at the time of the release.

Who funded the creation of the dataset? This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise. This work was partially supported by the NSF CAREER award 1652515, the NSF Grants IIS-1320635, DMS-1436591, DMS-1821334, OAC-1835712, OIA-1937043, CHS-1908767, CHS-1901091, a gift from Adobe Research, a gift from nTopology, and a gift from Advanced Micro Devices, Inc.

1.2 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

*Both authors contributed equally to this work.

Submitted to the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks. Do not distribute.

interactions be- tween them; nodes and edges)? Please provide a description. The instances represent 3D scans of different objects. Each scan is represented by the input data that was recorded to capture the geometry of an object as well as the reconstructed ground truth geometry from the data.

How many instances are there in total (of each type, if appropriate)? There are a total of 7 color textured objects scanned and 4 calibration objects in different configuration (e.g matte vs glossy material) from the physical scanner and 10000 scans from the simulated scanner.

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 does not contain all possible instances. The geometric objects that we selected are a subset from the much larger ABC dataset [8] and probably not representative. However, since we supply a data generator users are free to extend the selection with arbitrarily many objects.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description. There are two different instance types of data, the data coming from the physical scanning process and the data coming from the simulated scanning process.

The physical scan instances comprise the following data (with inconsequential components like a photo of the object missing for some scans):

- Unprocessed HDR images acquired for different rotating stage positions and collections of patterns, such as gray codes and uniform color (exr-format)
- Unprocessed HDR images acquired in the initial or final position of the rotating stage for additional collections of patterns, such as micro-phase shifting or unstructured light (exr-format)
- A background image with scanned object removed (exr-format)
- A photo of the setup with object installed (jpg-format)
- A brief description of the scanning conditions and settings (txt-format)
- A scanning script used to acquire the HDR images (script-format)

The ground truth (CAD) models of the scanned objects and processed calibration data are provided separately with the software repository. We do also provide the raw data acquired during the calibration process.

The simulated scan instances comprise the following data:

- Scan images of 3D objects illuminated with ambient light and white projector light (png-format)
- Depth maps from reconstruction and ground-truth (npy-format)
- Reconstructed point cloud (ply-format)
- The original object geometry/mesh (obj-format)

Is there a label or target associated with each instance? If so, please provide a description. No.

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. No, everything is included.

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. Different instances show the same object in different orientations. The relation can be retrieved from the data.

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 itself contains all the scans. For the benchmarks, the dataset is split into training and test set and stored in lists which reference the original scans from the dataset.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description. Since the orientation of the objects is chosen randomly, there is the chance that objects are scanned in similar orientations.

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. One part of the original object files are taken from the publicly available ABC dataset [8] and included in our dataset together with the scans. The other part of the object comprises textured objects which were retrieved from Sketchfab [6] and also included into our dataset.

Our dataset is therefore self-contained, however the original licenses of the object files have to be taken into consideration. For the models from ABC, the license is listed on the dataset website. For the models from Sketchfab, the following licenses and copyrights apply

Dodo Creator: Horniman Museum, License: CC BY-NC-ND 4.0 [2], https://skfb.ly/DSGo

Vessel Creator: Minneapolis Institute of Art, License: CC0 1.0 [3], https://skfb.ly/6RBty

House Creator: Deshan, License: CC BY 4.0 [1], https://skfb.ly/6SZrX

Radio Creator: gorzi, License: CC BY 4.0 [1], https://skfb.ly/6VKNJ

Sculpture Creator: Keith Morgan, License: CC BY 4.0 [1], https://skfb.ly/6vXUN

Chair Creator: Interiors3D, License: CC BY 4.0 [1], https://skfb.ly/6VAOQ

Vase Creator: Minneapolis Institute of Art, License: CC0 1.0 [3], https://skfb.ly/6VyUB

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 source geometry data was published before.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why. No, the objects represent mechanical/technical parts and common objects.

Does the dataset relate to people? If not, you may skip the remaining questions in this section. No.

1.3 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. The associated data was either generated by the physical scanner or the matching simulated scanner and the subsequent processing pipeline.

The input for both approaches were digital 3D objects or their physical realisations (either through milling or 3D printing).

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated? The scans from the physical and simulated scanner were collected as described in the technical description of our system. For validation, we ran multiple experiments as described in the paper to make sure both physical and simulated scans match on a pixel-wise basis and subsequently at all levels of the processing pipeline.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? We sampled the objects that were used for the dataset generation according to a few criteria depending on their purpose. For the calibration objects, manufacturability and the possibility to robustly retrieve the pose from the machined object were the key criteria. For the textured objects, the key criteria were manufacturability and permissive licenses. The mechanical objects from ABC were chosen according to their bounding box dimensions (objects with cubic bounding boxes were preferred).

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)? Nobody.

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 data was selected in 2020, but the timeframe of collection and association doesn't matter in our case.

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.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section. No.

1.4 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 miss- ing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section. Yes, the 3D objects were run through preprocessing steps. For CAD objects, the preprocessing consisted of surface mesh generation with high resolution. The textured objects were already retrieved as meshes, and were just converted to suitable formats for 3D printing and the rendering simulation.

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. The original data is also supplied from our website where applicable, or from the website where it was originally published.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point. Yes, the software for preprocessing is available from our dataset website.

1.5 Uses

Has the dataset been used for any tasks already? If so, please provide a description. Yes, the dataset was used to train and benchmark postprocessing algorithms for 3D scanning. In the main paper and the technical supplementary material we show the use of our data for the tasks of scan denoising, shape completion and surface reconstruction.

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? Besides the 3D scan processing tasks, we have identified the following possible tasks for our dataset and data generation software:

- Pattern development and evaluation for structured light scanning.
- Development and testing of 3D reconstruction algorithms.
- Generation of semantically annotated range scans.
- Generation of range scans with explicitly annotated sharp features.

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, we don't see any.

Are there tasks for which the dataset should not be used? If so, please provide a description. No, probably not.

1.6 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 and the software is publicly 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 dataset, our data generation code and benchmarks are distributed at https://geometryprocessing.github.io/scanner-sim. The data itself is hosted on the long term storage NYU Faculty Digitial Archive under the following address: https://archive.nyu.edu/handle/2451/62251.

When will the dataset be distributed? Most data is already published. The remaining chunks of data will be published before Neurips 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. Our datasets are available under the CC BY 4.0 license [1] except for the 3D models of 7 colored 3D printed objects which are licensed by their respective creators under various Creative Commons licenses as listed above.

Our code is published partly under the MIT license [5] and partly under the GPL 3.0 license [4]. The parts are clearly separated in our code repository and marked with the respective license. The rendering engine we used and modified is GPL-licensed, therefore the simulation framework is licensed under GPL and the calibration and physical scanner software is licensed under MIT. However, the images that are created with the simulation framework are not affected by this, and we choose to license the ones in our dataset with the CC BY 4.0 [1] license.

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. No, there are no restrictions.

1.7 Maintenance

Who is supporting/hosting/maintaining the dataset? The website and software is hosted on github as described above. The large data chunks are hosted on the long term storage archive (with a guarantee of at least 10 years) of New York University. The main authors are supporting and maintaining the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? The curators of the dataset can be contacted via the contact channels on the dataset website.

Is there an erratum? If so, please provide a link or other access point. If it turns out to become necessary, we will publish errata on the dataset website.

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)? In case of updates, users will be notified over the GitHub notification channels and by checking the dataset website.

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. No.

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, in case of updates, the old versions will still be hosted and supported.

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. Yes, contributions and extensions are welcome via the supported GitHub channels.

References

[1] CC BY 4.0. https://creativecommons.org/licenses/by/4.0/. Accessed: 2021-01-25.

[2] CC BY-NC-ND 4.0. https://creativecommons.org/licenses/by-nc-nd/4.0/. Accessed: 2021-01-25.

[3] CC0 1.0. https://creativecommons.org/publicdomain/zero/1.0/. Accessed: 2021-01-25.

[4] GPL. https://www.gnu.org/licenses/gpl-3.0.de.html. Accessed: 2021-01-25.

[5] MIT. https://opensource.org/licenses/MIT. Accessed: 2021-01-25.

[6] Sketchfab. https://sketchfab.com/. Accessed: 2021-01-25.

[7] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *CoRR*, abs/1803.09010, 2018.

[8] Sebastian Koch, Albert Matveev, Zhongshi Jiang, Francis Williams, Alexey Artemov, Evgeny Burnaev, Marc Alexa, Denis Zorin, and Daniele Panozzo. Abc: A big cad model dataset for geometric deep learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

The Machine Learning Reproducibility Checklist (v2.0, Apr.7 2020)

For all **models** and **algorithms** presented, check if you include:

- ❑ A clear description of the mathematical setting, algorithm, and/or model.
- ❑ A clear explanation of any assumptions.
- ❑ An analysis of the complexity (time, space, sample size) of any algorithm.

For any **theoretical claim**, check if you include:

- ❑ A clear statement of the claim.
- ❑ A complete proof of the claim.

For all **datasets** used, check if you include:

- ■ The relevant statistics, such as number of examples.
- ■ The details of train / validation / test splits.
- ■ An explanation of any data that were excluded, and all pre-processing step.
- ■ A link to a downloadable version of the dataset or simulation environment.
- ■ For new data collected, a complete description of the data collection process, such as instructions to annotators and methods for quality control.

For all shared **code** related to this work, check if you include:

- ■ Specification of dependencies.
- ■ Training code.
- ■ Evaluation code.
- ■ (Pre-)trained model(s).
- ■ README file includes table of results accompanied by precise command to run to produce those results.

For all reported **experimental results**, check if you include:

- ■ The range of hyper-parameters considered, method to select the best hyper-parameter configuration, and specification of all hyper-parameters used to generate results.
- ■ The exact number of training and evaluation runs.
- ■ A clear definition of the specific measure or statistics used to report results.
- ❑ A description of results with central tendency (e.g. mean) & variation (e.g. error bars).
- ■ The average runtime for each result, or estimated energy cost.
- ❑ A description of the computing infrastructure used.

Reproduced from: *www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist-v2.0.pdf*