Datasheet for The Drunkard's Dataset

David Recasens
University of Zaragoza

Martin R. Oswald
ETH Zurich, University of Amsterdam

Marc Pollefeys
ETH Zurich, Microsoft

Javier Civera
University of Zaragoza

Abstract

This datasheet [1] seeks to state the reasons for creating the Drunkard's Dataset and to show transparency about the funding interests.

A 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 created to push the limits of estimating the camera motion in deformable scenes as it is a challenging research problem relatively under- explored in the literature, and for which a lack of clear benchmarks slows down its progress.
- **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 the authors of this paper in the context of an international collaboration. The entity to which each author belongs is indicated in the header of this document.
- **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 projects EndoMapper GA 863146 (EU-H2020), PGC2018-096367-B-I00 (Spanish Government) and DGA-T45 17R/FSE (Aragón Government).
- **Any other comments?**
  None.

B 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 video sequence simulations in synthetic realistic indoor 3D buildings.
- **How many instances are there in total (of each type, if appropriate)?**
  There are a total of 19 different buildings, and for each one four trajectories under different deforming conditions were recorded. In total, they contain over 400K 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).

37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks.

It contains all possible instances.

- **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 of the 400K frames contains color and depth images, optical flow and normal maps and the camera trajectory ground truth with the following format:

  - **Color**: RGB uint8 .png images.

  - **Depth**: uint16 .png grayscale images whose pixel values must be multiplied by $(2^{16} - 1)30$ to obtain metric scale in meters.

  - **Optical flow**: .npy image numpy arrays that are .npz compressed. They have two channels: horizontal and vertical pixel translation to go from current frame to the next one.

  - **Normals**: .npy image numpy arrays that are .npz compressed. There are three channels: x, y and z to represent the normal vector to the surface where the pixel falls.

  - **Camera poses**: .txt file containing at each line a different SE(3) world-to-camera transformation for every frame. Format: timestamp, translation (tx, ty, tz), quaternions (qx, qy, qz, qw).

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

  For each color image there is a depth image, optical flow and normal maps and the camera pose ground truth.

- **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.

  Everything is included. No data is missing

- **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.

  Each frame data is named with the timestamp, so two adjacent frames have consecutive numbering.

- **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.

  For the Drunkard's Odometry we are using scenes 0, 4 and 5 for testing and the other 16 for training. In this way, $\sim 10\%$ and $\sim 90\%$ of the data is for testing and training, respectively. However, different testing scenes can be used to benchmark algorithms that are focused in other tasks. A dissimilar task, such as relocalization, could be more interested in testing scenes 4, 9 and 14 as there the camera traverses the building three times, but in each loop visiting the rooms in different order.

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

  Very few frames of some specific scenes were corrupted during rendering. This is indicated in the dataset by a text file called *wrong_frames.txt*. This issue is being addressed currently and it does not affect the experimentation, as the test scenes are completely free of corrupted frames and during training only consecutive frames are needed, so those few corrupted ones can be skipped.

- **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 entirely 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.

- **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.

- **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.

  N/A.

- **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.

  N/A.

- **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.

  N/A.

- **Any other comments?**

  None.

C 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.

  Each frame of the Drunkard's Dataset comes from the animation of the 3D meshes of the Habitat-Matterport 3D dataset [2] and their rendering at specific time instants in specific camera frames following pre-defined trajectories.

- **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?

  Each frame was rendered in Blender [3] from applying dynamic deformations to the 3D meshes and manually setting a camera trajectory inside them, using a personal computer with a RTX Nvidia 2080 Ti.

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

  It is not a sample from a larger set.

- **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 generation process was fully done by the authors.

- **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 dataset was created approximately from January to June 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.

  N/A.

- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**
  N/A.
- **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.

D 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.
  No pre-preprocessing was applied.
- **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.
  N/A.
- **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.

E Uses

- **Has the dataset been used for any tasks already?** If so, please provide a description.
  Yes, we implemented a novel method called the Drunkard's Odometry that we release together with the Drunkard's Dataset for data validation and as a specific use case.
- **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 yet, as this paper is introducing the dataset. However, as soon as there are some, we will list in the project website[1] the papers that use our dataset.

[1] https://davidrecasens.github.io/TheDrunkard'sOdometry/

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

  Apart from camera odometry/tracking, this dataset can be used to benchmark algorithms related to SLAM, localization, loop closure, Structure-from-Motion, 3D reconstruction, motion planning or collision avoidance for example.

- **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?

  The Drunkard's Dataset and the imported 3D meshes from the Habitat-Matterport 3D dataset are MIT licensed. Check the license files in the project page for further details.

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

  No.

- **Any other comments?**

  No.

F 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 publicly available in the project website.

- **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** Does the dataset have a digital object identifier (DOI)?

  In the project website there is a link to a Google Drive folder containing the data. Due to its huge size, it is difficult to do it in any other manner, and so it does not have a DOI. The Google Drive account is associated to our institution, and specifically to the EndoMapper project[2] that supports this work.

- **When will the dataset be distributed?**

  The dataset is already available at the time of the submission.

- **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 Drunkard's Dataset is MIT licensed. Check the license in the project page for further details.

- **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.

- **Any other comments?**

  None.

[2] https://sites.google.com/unizar.es/endomapper/home

G Maintenance

- **Who will be supporting/hosting/maintaining the dataset?**
  The dataset will be maintained by David Recasens. The data is safely hosted at Google Drive in an institutional account (Universidad de Zaragoza).
- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**
  Information to contact David Recasens can be found in his personal website https://davidrecasens.github.io/.
- **Is there an erratum?** If so, please provide a link or other access point.
  No, but any updated will be announced in the project 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 dataset consumers (e.g., mailing list, GitHub)?
  Any spotted error will be addressed as soon as possible and reported in the project website.
- **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.
  Any update regarding this aspect will be announced in the project website. The data might be extended if we find relevant additional aspects or sequences to include, but we do not foresee replacing the data we already have, so we will not have these "old versions" problems.
- **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 can contribute. Indeed, we release the code we used to create the data. The procedure will involve contacting the authors, preferably David Recasens, in order to have a unified official version.
- **Any other comments?**
  None.

H Author Responsibility Statement

The authors of this paper hereby declare that they bear all responsibility in case of any violation of rights, including but not limited to copyright infringement or other legal issues, arising from the content presented in this paper or associated work. The authors acknowledge that they have obtained all necessary permissions and rights for any third-party materials used in the paper and have complied with all relevant ethical guidelines and regulations.

References

[1] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021. 1

[2] S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. Turner, E. Undersander, W. Galuba, A. Westbury, A. X. Chang, *et al.*, "Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai," *arXiv preprint arXiv:2109.08238*, 2021. 3

[3] B. O. Community, *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 3