Datasheets for Datasets

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.

  PigLife dataset is created to promote the development of robust computer vision models and algorithms in pig farms. This dataset is designed for pig recognition tasks and animal behavioral recognition tasks, which include but not limited to detection, segmentation, tracking, identification. The contributions of our dataset to the computer vision community are to expand the image content to pig production scene and encourage the development of generalized and robust model development for pig-specific computer vision applications.

- 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 AIFARMS of University of Illinois.

- 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 is supported by Agriculture and Food Research Initiative (AFRI) grant no. 2020-67021-32799/project accession no.1024178 from the USDA National Institute of Food and Agriculture.

- 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.

  This instances in the dataset includes images and videos that contains pigs across the most production cycle in different systems: breeding, gestation, farrow, wean, nursery, growth, and finish. Only pig instances were annotated in this dataset; all the images are distinct but may describes the same scene.

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

  There are 21,869 annotated pig instances and 2,316 images in PigLife sub-1 dataset.

- 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).

  All videos in *PigLife sub-1* dataset were extracted from continuous recorded surveillance videos in pig farms. All images in *PigLife sub-1* dataset were selected from video files with the sampling rate of 1 image (frame) per second.

- 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 in *PigLife sub-1* dataset is an image. The images are not processed and compressed. The raw video files are also included in this dataset.

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

  Yes. In *PigLife sub-1* dataset, only pigs were annotated and labeled in images. Pigs in each image were annotated with masks. There are no other categories within the mask. The average image has ~10 masks.

- 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
- 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 relationship between each image was encoded into file name. filename is encrypted by the combination of four-digit codes and separators (sub-category appendant [s]; sequence or number [-]) to demonstrate the content of a video or an image. Filename contains 5 parts: growth stage (SID, 1000 ~ 1100), image description (IID, 1101 ~ 2000), housing condition (EID, 2001 ~ 5000), animal description (AID, 5001 ~ 6000), frame number (FN, 0 ~ INF).

- 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 release the recommended training and test splits and respective annotations that go with these splits. The users are free to split the dataset in a different way. No

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

  Errors: The masks are generated manually, so there may be human errors in the masks. Redundancies: no two images are the same. but images from the same video are more similar to each other. The sequential relationships are explicit by the filename.

- 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. To the best of our ability, we tried to mask all appearances of an individual in a video. The dataset should not contain any identifiable individual or any other confidential or sensitive information. We

underwent an internal privacy review to evaluate the potential risks with respect to the privacy of people in the photos. All video clips were masked to remove the background with human and irrelevant information. 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 relate to people? If not, you may skip the remaining questions in this section. 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. NA
- 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. NA
- Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial 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. NA
- Any other comments? 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.

  The masks associated with each image were manually annotated by our annotation group. The masks of each pig instance were confirmed by the second annotator. Both annotators are animal science experts affiliated with the University of Illinois Urbana-Champaign.

- 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?

  Raw videos were collected through 4K resolution (3840 x 2160) and 15 FPS using surveillance cameras (IP8M-T2599E, Amcrest) at the research farms in University of Illinois, Urbana and Champaign.

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

  Raw videos were manually watched by an animal expert and roughly marked by duration of each behavior for each animal. The video clips were selected to balance the behavioral labels and named following the rules to describe the general scene of the video. Images were extracted from masked video clips with the sampling rate of 1 image (frame) per second.

- 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)?

  Student volunteers.

- 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 *PigLife sub-1* dataset was derived from the videos taken from 2021 to 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.

  We underwent an internal privacy review to evaluate the potential risks with respect to the privacy of people in the photos. All video clips were masked to remove the background with human and irrelevant information.

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

  No

- Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? NA
- 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. NA
- 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. NA
- 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). NA
- 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. NA
- 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 remainder of the questions in this section.

- All video clips were masked to remove the background with human and irrelevant information on the side of view.
- 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
- Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

  We used the pig annotation was processed manually through VGG Image Annotation (https://www.robots.ox.ac.uk/~vgg/software/via/). Subsequently, we also used the self-developed annotation tool (Animal Video Analysis Tool, https://aifarms.github.io/AVAT/)
- Any other comments? No

Uses

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

  The dataset was used to test the benchmark detection and segmentation models that were described in this 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.

  No. But all users of the dataset must cite this paper. It is trackable via citation.
- What (other) tasks could the dataset be used for?

  This dataset was designed for detection task and segmentation tasks. We encourage research community to discovery more application related to the shape of pigs, such as behavior recognition or body measures.
- 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?

  PigLife sub-1 dataset is more representative to pig production scene than most of the publicly existing datasets at this time. The data and status of pigs only represent several particular production scenes with respect to the pig housing setup at the University of Illinois, Urbana and Champaign. We encourage users to be mindful of the limitation of the dataset and also, if they are interested, to contact us if they would like to expand the dataset.
- Are there tasks for which the dataset should not be used? If so, please provide a description.

  No
- 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 will be available for the research community.

- 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 available at https://data.aifarms.org through the following link. Upon accessing the web site, you will be presented with a license that stipulates the allowed uses of the dataset. After accepting the license, you will be able to download the dataset.

- When will the dataset be distributed?

  The dataset will be released in 2023.

- 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.

  Yes, https://data.aifarms.org/download/piglife

- 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.

  Yes, the data is proprietary to University of Illinois.

  Confidential Information (https://data.aifarms.org/download/piglife).

  1. You acknowledge that the Data is proprietary to ILLINOIS. You agree to protect the Data from disclosure or unauthorized use and to treat the Data with at least the same level of care as You use to protect Your own proprietary Data and/or confidential information, but in no event no less than a reasonable standard of care.

  2. If You become aware of any unauthorized licensing, copying, or use of the Data, You shall promptly notify ILLINOIS in writing at otm@illinois.edu.

  3. You agree to use the Data only in the manner and for the specific uses authorized in this 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.

  Export Controls (https://data.aifarms.org/download/piglife): The Data delivered under this License may be subject to U.S. export control laws and may be subject to export or import regulations in other countries. You agree to comply strictly with all such applicable laws and regulations and acknowledge that You have the responsibility, at Your own expense, to obtain such licenses to export, re-export, or import as may be required.

- Any other comments? No

Maintenance

- Who will be supporting/hosting/maintaining the dataset?

  The dataset will be hosted at https://data.aifarms.org and maintained by AIFARMS at the University of Illinois Urbana-Champaign.
- How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

  If you find any errors or want to contribute to this sub-dataset, please contact Angela Green-Miller (angelag@illinois.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 users (e.g., mailing list, GitHub)?
- Yes. The research group will expand the dataset, add extra labels and correct the errors, if and when they arise. The updates will be continually released on Github by Jiangong Li (jli153@cau.edu.cn).

- 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.
  This dataset does not relate to any human.
- 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.
  No. If the correction was made on the dataset, old version will be replaced. If the new subset of dataset was created, then will be released with the new name.
- 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.
  We encourage all research group to gather further data and annotations for *PigLife* dataset. Any users who generate annotations will be liable for hosting and distributing their annotations.
- Any other comments? No