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.

Goatlife is created to promote the development of robust computer vision models and algorithms in goat farms. This dataset is designed for goat recognition and animal behavioral recognition tasks, which will primarily include detection and identification. The contributions of our dataset to the computer vision community are to expand the image content to goat production scene and encourage the development of generalized and robust model development for goat-specific computer vision applications. The current submission will be a subset of the larger *Goatlife* dataset, named *Goatlife-estrus* or "*GOES*", which is specifically designed for tasks related to goat estrus recognition in individually housed goats.

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 AIFARMS Tuskegee University Autonomous Microsite

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 project was funded by George Washington Carver Agricultural Experiment Station and USDA/NIFA Evans Allen Program (Grant No. ALX-FVC-18) and USDA/AFRI (Grant No. 2020-67021-32799/Project Accession No.1024178

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.

These instances in the dataset includes images and videos that contains tail position of Kiko goats during non-estrus cycle.

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

4,428 total instances or photos extracted from video segments.

Tail Up: 3,603

Tail Down: 353

Undetermined: 472

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 *GOES* dataset were extracted from recorded surveillance videos in the goat barn located at the caprine research unit at Tuskegee University. All images in *GOES* dataset were selected from video files with the sampling rate of 15 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 *GOES* 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 *GOES* dataset, only goat tail positions were annotated and labeled in images. Kiko in each image were annotated based on three categories, tail up, tail down, and undetermined.

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 or frame was individually labelled by a different file name. The filename is labeled by the goat coat color (light, brown, and black), followed by an underscore and a number (ie: light0001; brown0012, black0012)

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.

No data split in this dataset. The end users are free to split the dataset in a different way.

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

Errors: The annotations of each class were generated manually, so there may be human errors in creating precise bounding boxes.

Redundancies: No two images are identical; however, the images extracted from the same video may be similar The sequential relationships are indicated in 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.

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.

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 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.

N/A

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 extracted images from raw video recordings were manually annotated and completed by the works of undergraduate, graduate students and postdoc within the Bolden-Tiller lab.

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) by GoPro Hero 8 camera at the caprine research unit at Tuskegee University. The captured recordings wre stored in the SD card of the camera and later transferred to the hard drive of a laptop computer.

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

The Bolden-Tiller lab research team manually reviewed raw videos of each animal in the study. Video clips were chosen based on animals that exhibited the most tail wagging. Additionally, animals with varying coat colors were selected to enhance dataset diversity. Images were then extracted from the raw video clips at a sampling rate of 15 frames 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)?

Undergraduate Students, Graduate Students, and Postdoc in the Bolden-Tiller lab research team.

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 *GOES* raw video and video segments were collected in February and March of 2024 in 4K resolution. The images from the dataset were extracted and annotated in September 2024.

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 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)?

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.

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.

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).

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.

Any other comments?

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.

The extracted images from the dataset were labeled based on three classes of tail position: tail up, tail down, undetermined. It was also resized to 640 x 640 pixels

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.

Roboflow was used to label the dataset: The project link is shown below:

https://app.roboflow.com/join/eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJ3b3Jrc3BhY2VJZCI6IjQ2MUoyakp3OW1oc1RLRDZGVnN4WGdmdzNIODIiLCJyb2xlIjoib3duZXIiLCJpbnZpdGVyIjoiam11bmdpbkB0dXNrZWdlZS5lZHUiLCJpYXQiOjE3MzgyNjM3NTV9.juaYdI8YUBoLApZQ9r9h7PPFxDPJn1aml4Rl_-aBVjI

Any other comments?

No

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.

N/A

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

This dataset was designed for detection tasks. The dataset can be used for precision livestock application related to behavior recognition, animal health monitoring, and/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?

The *GOES* dataset is specifically representative of goat production, particularly in the context of estrus detection. The data and estrus status of the does reflect production conditions exclusively for individually housed Kiko goats at the Tuskegee University Caprine Research Unit. Users are encouraged to be aware of this limitation and to contact the relevant parties if they intend 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 aimed to be available at https://data.aifarms.org. Upon accessing the web site, the end user 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 2025

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.

N/A

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.

N/A

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.

N/A

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)?

Any questions or concerns related to the dataset will need to contact Olga Bolden-Tiller (oboldentiller@tuskegee.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 only if needed. The updates regarding the code for training dataset will be continually released by team members of the Bolden-Tiller lab.

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 a correction is made to the dataset, the old version will be replaced; However, if a new subset of the dataset is created, it will be released under a 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 *GOES* dataset. Any users who generate annotations will be liable for hosting and distributing their annotations.

Any other comments?

No