Datasheet for Datasets (Alberta Wells Dataset: Pinpointing Oil and Gas Wells from Satellite Imagery)

Anonymous Author(s)
Affiliation
Address
`email`

A Documentation frameworks: Datasheet for Datasets

A.1 Motivation

1. **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 Alberta Wells Dataset (AWD) was created to identify oil and gas wells—whether abandoned, suspended, or active—using medium-resolution multi-spectral satellite imagery. While the issue of detecting oil and gas wells has been addressed by several authors, existing datasets are typically small (500-5,000 samples) and limited to specific regions, often including only active wells. This limitation reduces their effectiveness in identifying abandoned or suspended wells. The AWD aims to fill this gap in the literature by offering a comprehensive dataset with over 188,000 samples (including over 94,000 samples containing wells) from PlanetLabs satellite imagery, encompassing more than 213,000 individual wells.
2. **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**
   The raw data is sourced from the Alberta Energy Regulator (AER), specifically from the monthly AER ST37 publication. This dataset includes comprehensive details about all reported wells in Alberta, such as their geographic location, mode of operation, license status, and the type of product extracted, among other attributes. The data is provided in shapefile format along with accompanying metadata. However, the dataset cannot be used directly because the license status or mode of operation often does not reflect the well's actual status. Therefore, the authors include domain experts from <Anonymous>, who specialize in field measurements of methane and air pollutant emissions from oil, gas, and urban systems, as well as in the geospatial and statistical data analysis of emissions and energy infrastructure, to ensure the quality of the dataset.
3. **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 <Anonymous>.

A.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 interactions between them; nodes and edges)? Please provide a description.
  We provide a dataset file stored in Hierarchical Data Format 5 (HDF5, i.e., a .h5 file), which contains multispectral 4-band RGBN satellite images in raster format and data labels with

both identified by unique instance names. These satellite images, acquired from Planet Labs, have a resolution of 3 meters per pixel and include corresponding metadata. The metadata contains information about the number and types of wells present in a patch. For data labels, we offer binary segmentation maps, multi-class segmentation maps (each class representing a well in an active, abandoned, or suspended state), and COCO format object detection labels. The images were taken from the province of Alberta, Canada, with each satellite imagery patch representing a square with a side length of 1050 meters (1.05 km), covering an area of 1.025 square kilometers. The entire dataset spans over 193,000 square kilometers.

- **How many instances are there in total (of each type, if appropriate)?**
The proposed dataset comprises 188,688 instances, of which 94,344 contain one or more wells, totaling 213,447 well points. Each instance includes corresponding multispectral satellite imagery, segmentation maps (both binary and multi-class, with classes indicating active, suspended, or abandoned states), and bounding box annotations with the state of operations as the object class ID in COCO format. We standardized the diameter of a well site to 90 meters (typically ranging from 70 to 120 meters) for creating annotations, resulting in a diameter of 30 pixels in the labels. More details about the distribution of wells in each split are provided in the supplementary materials as well as the main paper.

- **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 AWD Dataset is based on the AER ST37 monthly status data of wells in the Alberta region of Canada. It includes wells that are in active, suspended, or abandoned states of operation. To ensure the dataset's quality, the authors with appropriate domain expertise conducted extensive quality control, filtering, and duplicate removal. This process was necessary because the full dataset included cases of well sites being restored and reclaimed, as well as various duplicates, noise, and data on other types of wells involving different natural resources. Therefore, the AWD Dataset, which includes multi-spectral satellite imagery, segmentation, and detection labels, is constructed from a refined subset of the original AER ST37 data, specifically targeting oil and gas wells that can be precisely identified.

- **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 Image instance in our dataset, formatted in HDF5, contains satellite imagery represented as a numpy array from Raster Vector. We preprocessed this imagery by reprojecting it to the EPSG 32611 coordinate reference system and removed all geographic metadata, such as image bounds and coordinates, from the shared data. However, we do provide attributes like Sample Name, wells present, no of wells, Abandoned well present, Active well present, and Suspended well present. We utilized Planet Labs' 4-band (RGBN) satellite imagery product (ortho_analytic_4b_sr), which incorporates the latest PSB.SD instrument with a 47-megapixel sensor. Each satellite imagery patch acquired represents a square with a side length of 1050 meters (1.05 km), covering an area of 1.025 square kilometers. The entire dataset spans over 193,000 square kilometers.

- **Is there a label or target associated with each instance?** If so, please provide a description.
There are three types of labeled data for each image: binary segmentation maps (in Rasterio Image [.jpg] format) indicating the presence or absence of oil and gas wells, multiclass segmentation maps (also in Rasterio Image [.jpg] format) potentially identifying various classes of objects, and bounding box annotations (in COCO format) specifying the location and size of objects, such as wells, within the image. These components together form a comprehensive dataset suitable for training and evaluating machine learning models for

tasks like object detection and segmentation in satellite imagery, particularly focused on pinpointing oil and gas wells in Alberta

- **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.
  The satellite imagery used in this project was obtained under Planet Labs' [1] Education & Research license, which prohibits sharing raw satellite imagery. We re-projected the raw data to EPSG:32611 using the nearest resampling method and removed all geographic metadata, such as image bounds and coordinates, from the shared data imagery to create a derived product that complies with the license terms.
- **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.
  N/A
- **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 we propose comprises more than 94,000 patches of satellite imagery containing wells, totaling 188,000 patches sourced from Planet Labs. This dataset covers over 213,000 individual wells. To ensure a balanced dataset, we divided it into training, validation, and testing sets using our algorithm outlined in Section 3.2 of the main paper. Our proposed method for splitting the data aims to create smaller, non-overlapping regions of concentrated wells by clustering patch centroids. These regions are intended to (a) not intersect, (b) be part of a larger geographic area by clustering initial cluster centroids, and (c) contain a similar distribution of non-well patches. This approach ensures that the training, validation, and test sets cover all geographic regions, providing a diverse and thorough evaluation. The dataset splits represent various geographical areas, making it a comprehensive benchmark for evaluation and testing. Each dataset split is stored in an HDF5 format file.
- **Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.
  One limitation of our study is our reliance on well locations provided by the Alberta Energy Regulator, which may not encompass all sites, leading to potential omissions in the ground-truth data. This could result in a lower reported validation and test accuracy, with some correctly predicted well locations being mistakenly categorized as false.
- **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 does not rely on the persistence of external resources.
- **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.

A.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 the data was indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
The AER publishes AER ST37, a monthly list of wells in Alberta, including location, operation mode, license status, and product type. However, the data needs rigorous quality control as license status, or operation mode may not accurately reflect the actual well status. The authors, with extensive domain expertise, removed duplicate well entries in the metadata and shapefile, keeping the most recent update. We then merge and filter the datasets, categorizing wells as active, abandoned, or suspended based on expert criteria. Duplicate coordinates are resolved by keeping the instance with the latest drill date. We verify all wells are within Alberta's boundaries. After thorough quality control by domain experts, we calculate the geographical bounds covered by wells and divide the region into non-overlapping square patches. These patches include varying numbers of wells, with an equal number of patches with and without wells.

- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensors, manual human curation, software programs, software APIs)?** How were these mechanisms or procedures validated?
  We acquired multispectral satellite imagery data from Planet Labs, which comprises four bands (RGBN) with a 3-meter-per-pixel resolution obtained through their proprietary API. This data was processed using quality-controlled and cleaned well data to generate segmentation and object detection annotations. The annotations were created using custom Python code, leveraging libraries like Shapely, GeoPandas, and Rasterio, and were validated through visualization using folium and matplotlib.

- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**
  No.

- **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 dataset was a collaborative effort involving the Alberta Energy Regulator, Planet Labs, and the authors. Without the contributions from individuals in these three organizations, this dataset would not have been possible. Proper credit must be given to the authors, Planet Labs, and the Alberta Energy Regulator when using this data.

- **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.
  We acquired the data from the Alberta Energy Regulator, specifically from its monthly well bulletin AER ST37 [2], dated March 2024. Leveraging domain expertise, we filtered the data to reflect the condition of wells as of September 30, 2023. This decision was made because imagery acquired from Alberta during the winter months tends to have high cloud cover. Therefore, we filtered the data to ensure we could collect the best data for each patch based on satellite data acquired between the summer months of June and September in the region.

- **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

A.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 missing values)?** If so, please provide a description.

In the Dataset section of our submission, we provide a detailed description of the quality control, cleaning, and labeling processes applied to the data obtained from the Alberta Energy Regulator, which forms the basis of our dataset. The satellite imagery utilized in this project was acquired under the Education & Research license from Planet Labs. We reprojected the raw data to EPSG:32611 using the nearest resampling method. Additionally, we removed all geographic metadata, such as image bounds and coordinates, from the shared data imagery to ensure compliance.

- **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 raw satellite imagery data has been saved for internal use; however, it cannot be shared in its current form. Before sharing, the data must undergo preprocessing to remove metadata, as stipulated by the agreement mentioned earlier.

- **Is the software that was used to preprocess/clean/label the data available?** If so, please provide a link or other access point.
  We plan to share the relevant code used for dataset quality control, patch creation, dataset splitting, data acquisition, and label and HDF5 file creation with the public release of the dataset in the future.

- **Any other comments?**
  N/A

A.5 Uses

- **Has the dataset been used for any tasks already?** If so, please provide a description.
  Currently, there are no public demonstrations of the AWD Dataset in use. In this work, we showcase its application for Binary Segmentation and Binary Object Detection of Well Sites to train algorithms for accurately locating well sites. These algorithms can be scaled across larger regions of interest to compare against existing databases, identifying potentially undocumented wells. Flagging wells not present in databases is crucial, as these could be abandoned wells that are significant emitters of greenhouse gases, making them candidates for plugging.

- **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?**
  Additionally, we provide multi-class labels indicating the operational state of the wells for both cases. These labels can be utilized in future projects for locating wells and classifying their operational status, which will aid in identifying well sites that are not present in government records.

- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**
  This dataset focuses on Alberta, Canada, known for its diverse oil reserves and varied landscapes, providing a representative sample comparable to regions in the Appalachian and Mountain West areas of the United States and some former Soviet states with oil wells and unidentified site issues. A limitation of our study is the reliance on well locations from the Alberta Energy Regulator, which may miss some sites, leading to potential false negatives in the ground-truth data. However, this should have minimal impact on algorithm training, as these labels are a minor part of the dataset, and deep learning algorithms can handle moderate label noise well (see e.g., [3]). The main effect may be underreported test accuracy, with some correctly predicted well locations wrongly counted as false. We plan to investigate this further in future work. Additionally, the use of multi-spectral optical data in the AWD dataset may limit the models' applicability in regions with frequent cloud cover.

- **Are there tasks for which the dataset should not be used?** If so, please provide a description.
  This dataset is intended for non-commercial use only and should not be utilized in any application that could negatively impact biodiversity.

- **Any other comments?**
  N/A

A.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 will be made public (open-source) in the future.

- **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** Does the dataset have a digital object identifier (DOI)?
  The data is currently accessible through a Dropbox folder, which will eventually be migrated to Google Cloud. The link to access the data will be provided on our project's GitHub repository.

- **When will the dataset be distributed?**
  The dataset can be downloaded from Dropbox, with the link specified in the main paper and mentioned in the README of the shared codebase for benchmark experiments. Once the submission is made public, the dataset will be hosted on Google Cloud, and the link will be provided in the public GitHub repository.

- **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 AWD Dataset is released under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License (https://creativecommons.org/licenses/by-nc/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 satellite imagery used in this project was acquired under the Education & Research license of Planet Labs [1]. This license allows for the use of the data in publications and the creation of derivative products, which can be shared in association with publications. However, raw imagery cannot be shared publicly. To comply with these guidelines, we share the data in HDF5 format, with satellite imagery represented as a numpy array from Raster Vector. We have removed all geographic metadata, such as image bounds and coordinates, from the shared data. The data is intended for academic use only and should not be used for commercial purposes. Proper credit must be given to the current authors, Planet Labs, and the Alberta Energy Regulator when using this data.

- **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?**
  N/A

A.7 Maintenance

- **Who is supporting/hosting/maintaining the dataset?**
  We are currently hosting the dataset on Dropbox to ensure anonymity. Once it is made public, we plan to host it on Google Cloud storage.
- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**
  You can reach the authors through the email addresses provided in the paper once it is made public. Additionally, you can raise any issues on the GitHub repository, which will be made public in the future.
- **Is there an erratum?** If so, please provide a link or other access point.
  Not to the best of our knowledge.
- **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)?
  As our dataset is based on data from a fixed timeframe and consists of satellite imagery collected during a specific period, we do not currently have plans to update it in the near future. However, if there are any changes to these plans, updates to the dataset will be posted on the corresponding GitHub repository once it is made public.
- **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.
  If there are newer versions of the dataset, they will maintain the same format. We will ensure that the code associated with the project on GitHub supports these updates, and we will update the READMEs to reflect any changes to the dataset.
- **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 users? If so, please provide a description.
  We plan to share the relevant code in the future. However, to ensure the ability to compare against our results, we encourage those who wish to build on the dataset to publish their work separately rather than adding to our data repository.
- **Any other comments?**
  N/A

References

[1] Planet Labs PBC. Planet application program interface: In space for life on earth. https://api.planet.com.

[2] ST37 — aer.ca. https://www.aer.ca/providing-information/data-and-reports/statistical-reports/st37. [Accessed 06-06-2024].

[3] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017.