
Alberta Wells Dataset: Pinpointing Oil and Gas Wells from Satellite Imagery

Anonymous Author(s)

Affiliation

Address

email

1 A Dataset and Code

2 A.1 Dataset

3 The dataset is currently hosted in Dropbox for anonymity reasons and can be accessed here:

4 [AWD Dataset](#)

- 5 • Example visualization of dataset samples, including the spectral bands of each image and
6 the corresponding labels: [Visualizations](#)
- 7 • Compressed training set: [Train.tar.gz](#)
- 8 • Compressed validation set: [Validation.tar.gz](#)
- 9 • Compressed test set: [Test.tar.gz](#)
- 10 • Dataset license: [License.txt](#)

11 A.2 Croissant Metadata

12 The Croissant metadata can be accessed from here: [Croissant metadata record \(AWD\)](#)

13 Our dataset is comprised of Hierarchical Data Format (HDF5) files with a multi-level hierarchy. The
14 Croissant metadata format does not currently support describing the structure within each HDF5 file,
15 as noted in a [GitHub issue](#).

16 Therefore, we provide Croissant dataset metadata that includes only dataset-level information and
17 resources, excluding RecordSets that require data from HDF5 files. We will update the metadata once
18 Croissant supports the HDF5 format.

19 Additionally, we provide another documentation framework, Datasheets for Datasets [1], which can
20 be accessed from here: [Datasheets for Datasets \(AWD\)](#).

21 We also describe the dataset structure and the structure of data in Hierarchical Data Format (HDF5)
22 files in detail in sections [C.1](#), [C.2](#) and [C.3](#).

23 A.3 Code Repository

24 The code repository with benchmark experiments and visualizations of samples can be accessed here:

25 [awd_benchmark](#)

Table 1: Dataset statistics represented across the various splits of the dataset.

| Dataset Split | No of Samples | No of Wells in Split | Original HDF5 File Size (in Gb) | Compressed .tar.gz File Size (in Gb) |
|---------------|---------------|----------------------|---------------------------------|--------------------------------------|
| Train | 167436 | 194231 | 322 | 100 |
| Validation | 9463 | 8243 | 19 | 5.7 |
| Test | 11789 | 10973 | 24 | 7.1 |
| Total | 188688 | 213447 | 365 | 112.8 |

26 B Hosting, licensing, and maintenance plan

27 B.1 Hosting & Maintenance

28 Once the dataset is made public, we plan to host it on Google Cloud storage.

29 B.2 Data Licensing

30 The AWD Dataset is released under a Creative Commons Attribution-NonCommercial 4.0 Interna-
31 tional (CC BY-NC 4.0) License (<https://creativecommons.org/licenses/by-nc/4.0/>).

32 The satellite imagery for this project was acquired through Planet Labs’ [2] Education & Research
33 license, which allows the use of the data in publications and the creation of derivative products
34 related to those publications. However, the raw imagery cannot be shared publicly. To adhere to
35 these guidelines, we provide the data in HDF5 format, with the satellite imagery pre-processed to
36 produce a derived product represented as a numpy array from Raster Vector. This process removes
37 all geographic metadata.

38 This data is for academic use only and should not be used commercially. Proper credit to the current
39 authors, Planet Labs [2], and the Alberta Energy Regulator [3] is required when using this data.

40 C Dataset Information

41 The purpose of this dataset is to assist in training deep learning systems to identify oil and gas wells,
42 including abandoned, suspended, and active ones. This will enable the detection of wells in a specific
43 area, allowing comparison with government records. If discrepancies are found, experts can conduct
44 further investigations, which can possibly lead to the discovery of an abandoned or suspended device
45 that might not be present in government records.

46 C.1 Dataset Structure

47 We provide training, validation, and testing sets, split using our proposed algorithm (as described in
48 Section 3.2 of the main paper) to create a well-distributed dataset.

49 The proposed method aims to create smaller regions of well concentration by clustering the centroids
50 of patches. These regions are designed to be (a) mutually non-intersecting, (b) part of a larger
51 geographic region by clustering the centroids of the initial clusters, and (c) containing a similar
52 distribution of non-well patches within the same region. This approach ensures that the training,
53 validation, and test sets include representations from all geographic regions, providing a diverse and
54 comprehensive evaluation. Thus, the dataset represents various geographical regions and offers a
55 diverse benchmark for evaluation and testing.

56 Each dataset split is saved in an HDF5 format file, structured as described in the following sections,
57 and then compressed into a .tar.gz file for faster transfer. Details on the number of samples in each
58 set and the size of the dataset, both original and compressed, are presented in Table 1.

59 C.2 Dataset File Directory Structure

60 The following directory structure is used for each dataset file being stored in a Hierarchical Data
61 Format 5 (HDF5) file:

```
62 <Train/Test/Val>Set.h5
63     |---image
64         |---<sample_name>
65             |---Satellite Image (Multispectral Rasterio Image [.tiff] Data)
66             |---Meta Data of <sample_name>
67     |---label
68         |---binary_seg_maps
69             |---<sample_name>
70                 |---Binary Segmentation Map (Rasterio Image [.jpg] Data)
71         |---multi_class_seg_maps
72             |---<sample_name>
73                 |---Multiclass Segmentation Map (Rasterio Image [.jpg] Data)
74         |---bounding_box_annotations
75             |---<sample_name>
76                 |---Bounding Box JSON Data (COCO Format)
77     |---author:Anonymous Author(s)
78     |---description: Alberta Wells Dataset:
79                     Pinpointing Oil and Gas Wells from Satellite Imagery
```

80 C.3 Structure of Dataset Directory

81 To enhance the efficiency of the data loader, we split the larger .h5 dataset into smaller .h5 files, each
82 corresponding to a unique sample (image patch). By splitting the dataset in such a manner, we are
83 able to improve the speed per iteration of the dataloader by over 100%. This results in the following
84 data structure:

```
85 <Sample_Id>.h5
86     |---image
87         |---Satellite Image (Multispectral Rasterio Image [.tiff] Data)
88         |---Meta Data
89     |---label
90         |---binary_seg_maps
91             |---Binary Segmentation Map (Rasterio Image [.jpg] Data)
92         |---multi_class_seg_maps
93             |---Multiclass Segmentation Map (Rasterio Image [.jpg] Data)
94         |---bounding_box_annotations
95             |---Bounding Box JSON Data (COCO Format)
96     |---author:Anonymous Author(s)
97     |---description: Alberta Wells Dataset:
98                     Pinpointing Oil and Gas Wells from Satellite Imagery
```

99 C.4 Dataset Size & Distribution of Samples

100 The proposed dataset comprises over 94,000 patches of satellite imagery containing wells, with a total
101 of 188,000 patches sourced from Planet Labs [2]. This dataset covers more than 213,000 individual
102 wells. Details about the distribution of the number of patches, wells present, and dataset split sizes
103 are provided in Table 1, with the distribution of the number of wells per sample being described in
104 Table 2. We also include an equal number of images that contain no wells in each dataset split.

Table 2: The distribution of individual wells in positive samples from the dataset. We also include an equal number of images that contain no wells in each dataset split.

| No of Wells in a Sample | Frequency of Well Instances in a Sample | | |
|----------------------------|---|------------------|------------|
| | Training Split | Validation Split | Test Split |
| 1 | 44299 | 3393 | 4128 |
| 2 - 3 | 25378 | 979 | 1242 |
| 4 - 5 | 7899 | 190 | 328 |
| 6 - 10 | 4927 | 123 | 227 |
| 11 - 15 | 751 | 23 | 38 |
| 16 - 25 | 333 | 11 | 19 |
| 26 - 35 | 67 | 10 | 2 |
| 36 - 55 | 45 | 3 | 0 |
| 56 - 75 | 18 | 0 | 0 |
| 76 - 125 | 1 | 0 | 0 |
| Total | 83718 | 4732 | 5984 |

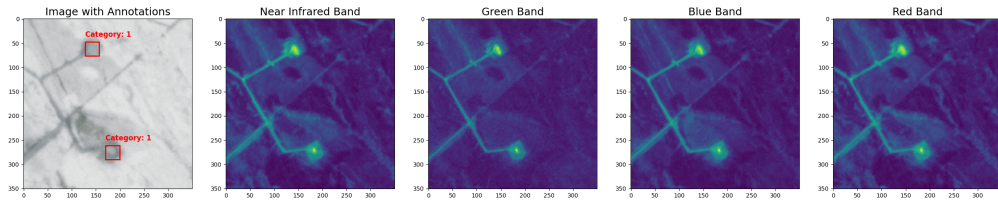
105 C.5 PlanetScope Satellite Imagery

106 For our experiments, we selected a 4-band (RGBN) satellite imagery product (ortho_analytic_4b_sr)
 107 from Planet Labs [2] as illustrated in Figure 1. This product uses Planet’s PSB.SD instrument, which
 108 features a telescope with a larger 47-megapixel sensor and is designed to be interoperable with
 109 Sentinel-2 imagery in several bands. The frequency of each band of image is described in Table 3.
 110 The instrument provides a frame size of 32.5 km x 19.6 km, an image capture capacity of 200 million
 111 km²/day, and an imagery bit depth of 12-bit, with a ground sample distance (nadir) ranging from 3.7
 112 m to 4.2 m.

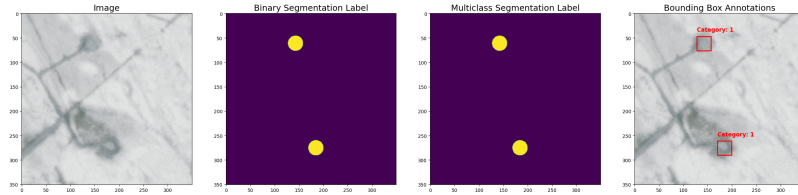
113 The satellite images are corrected for atmospheric conditions and spectral response consistency.
 114 These multispectral products are tailored for monitoring in agriculture and forestry, offering precise
 115 geolocation and cartographic projection. They are ideal for tasks such as land cover classification,
 116 with radiometric corrections ensuring accurate data transformation.

117 C.6 Label Data Description

118 For our experiments, we create single-channel segmentation maps, which are binary maps used
 119 to locate instances of wells. We also generate multi-class segmentation maps, where each class
 120 denotes a well in an active, abandoned, or suspended state. Furthermore, we provide COCO format



(a) A sample patch with Bbox annotations and the corresponding imagery in its different spectral bands.



(b) A sample patch with its segmentation labels (binary and multi-class) and bounding box annotations.

Figure 1: A Sample Patch from the Evaluation Set with 2 active wells.

Table 3: The Frequency of Each Spectral Band of a Planetscope PS.SD acquired Image

| Band of Image | Frequency (in nm) of Spectral Band |
|------------------------|------------------------------------|
| Band 1 = Blue | 465 - 515 |
| Band 2 = Green | 547 - 585 |
| Band 3 = Red | 650 - 680 |
| Band 4 = Near-infrared | 845 - 885 |

Table 4: Sample of Meta-Data Associated with each Instance in the Dataset

| Meta-Data Attribute Name | Value |
|--------------------------|-----------|
| Sample_Name | eval_6934 |
| wells_present | True |
| no_of_wells | 10 |
| Abandoned_well_present | True |
| Active_well_present | True |
| Suspended_well_present | True |

object detection labels for wells. In both segmentation and detection labels, we represent various states with class IDs as 'Active': 1, 'Suspended': 2, 'Abandoned': 3. To maintain consistency, we standardize the diameter of a well site to 90 meters (typically ranging from 70 to 120 meters) when annotating, resulting in a 30-pixel diameter in the labels. Figure 1(b) illustrates image patches with their corresponding labels, Figure 1(a) illustrates various spectral bands present in an image alongside the original image with bounding box annotations for reference and an example of a bounding box label in COCO format is shown below.

Sample of Bounding Box Annotation:

```
[
  {
    'id': 0,
    'image_id': 'eval_7028',
    'category_id': 1,
    'bbox': [46, 145, 29, 29],
    'iscrowd': 0
  },
  {
    'id': 1,
    'image_id': 'eval_7028',
    'category_id': 2,
    'bbox': [45, 127, 29, 29],
    'iscrowd': 0
  }
]
```

C.7 Meta Data Description

Each dataset sample is accompanied by metadata, including the sample name (sample ID in string format), the presence of a well in the sample, the number of wells in the sample, and whether a well of a specific category is present in the sample. Table 4 provides an illustration of metadata associated with a sample.

D Dataset Samples Illustration

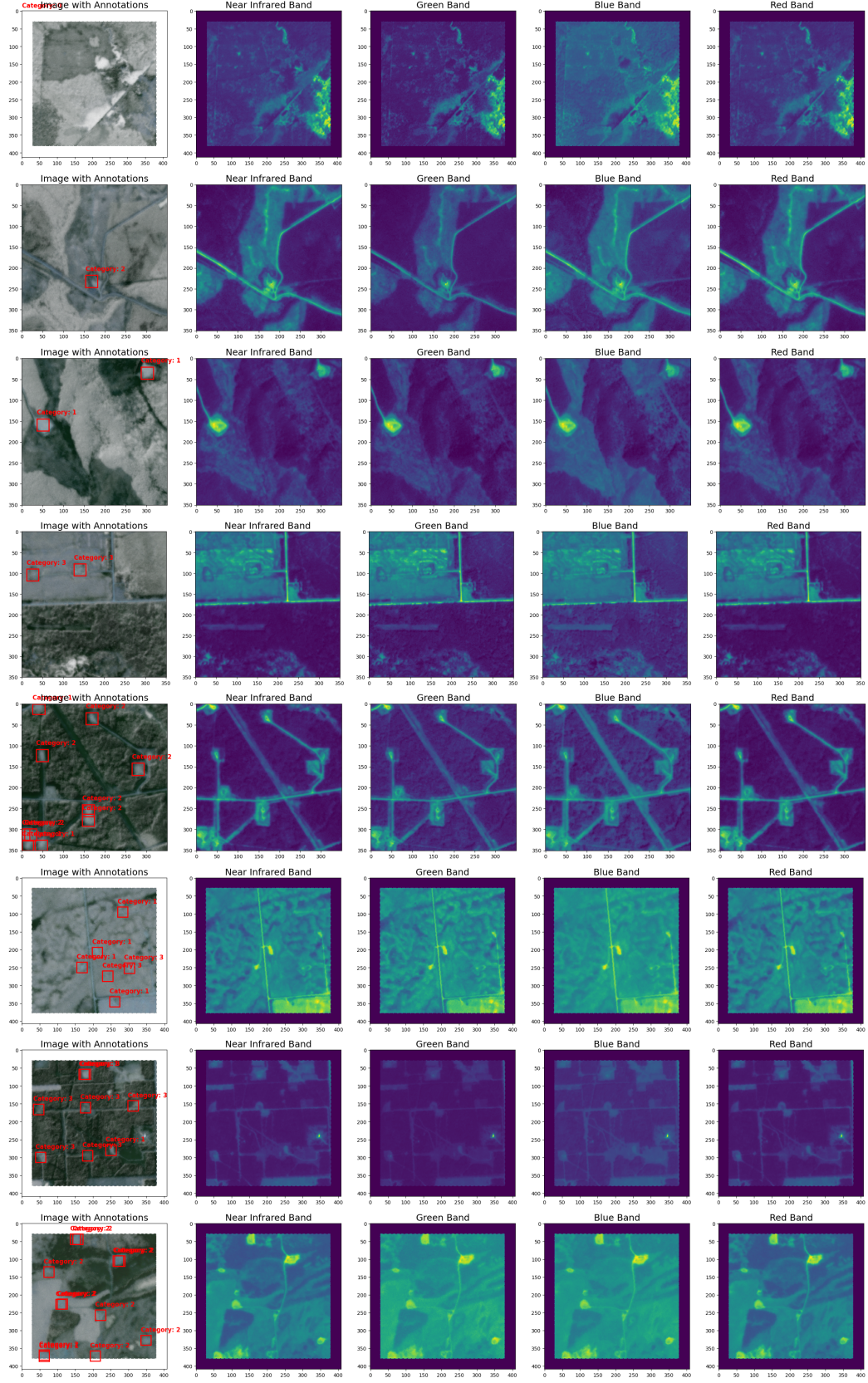


Figure 2: Qualitative results from the dataset illustrate the diverse distribution of wells in dataset samples, including Bbox annotations and corresponding imagery in different spectral bands.

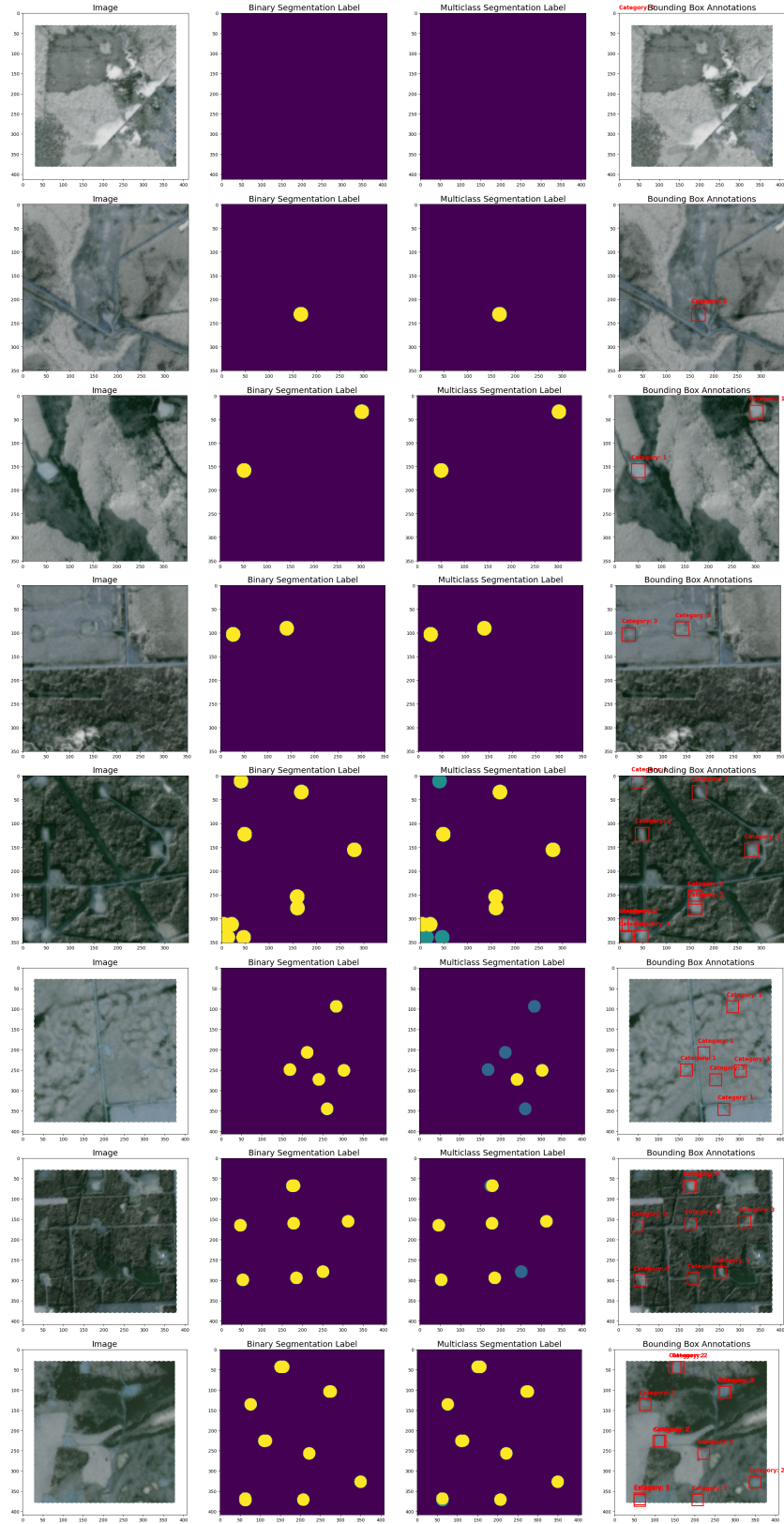


Figure 3: The qualitative results from the dataset showcase the varied distribution of wells in dataset samples, with their corresponding segmentation labels (binary and multi-class) and Bbox annotations.

151 **References**

- 152 [1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach,
153 Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Commun. ACM*, 64(12):86–92, nov
154 2021.
- 155 [2] Planet Labs PBC. Planet application program interface: In space for life on earth. [https:](https://api.planet.com)
156 [//api.planet.com](https://api.planet.com).
- 157 [3] ST37 — aer.ca. [https://www.aer.ca/providing-information/data-and-reports/](https://www.aer.ca/providing-information/data-and-reports/statistical-reports/st37)
158 [statistical-reports/st37](https://www.aer.ca/providing-information/data-and-reports/statistical-reports/st37). [Accessed 06-06-2024].