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

Anonymous Author(s)
Affiliation
Address
email

Abstract

Millions of abandoned oil and gas wells are scattered across the world, leaching methane into the atmosphere and toxic compounds into the groundwater. Many of these locations are unknown, preventing the wells from being plugged and their polluting effects averted. Remote sensing is a relatively unexplored tool for pinpointing abandoned wells at scale. We introduce the first large-scale dataset for this problem[1], leveraging medium-resolution multi-spectral satellite imagery from Planet Labs. Our curated dataset comprises over 213,000 wells (abandoned, suspended, and active) from Alberta, a region with especially high well density, sourced from the Alberta Energy Regulator and verified by domain experts. We evaluate baseline algorithms for well detection and segmentation, showing the promise of computer vision approaches but also significant room for improvement.

1 Introduction

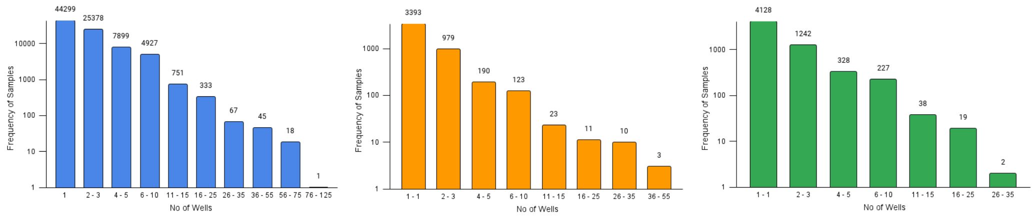
Across the world, there are millions of abandoned oil and gas wells, left to degrade by the companies or individuals that built them. No longer producing usable fossil fuels, these wells nonetheless have a significant impact on the environment, with many of them leaking significant quantities of methane, a powerful greenhouse gas, into the atmosphere. In aggregate, these emissions represent the equivalent of millions of tons of carbon dioxide per year [1]. Abandoned wells also pose health and safety concerns, in particular by leaching toxic chemicals into the groundwater of surrounding communities [2].

It is possible to plug abandoned wells to mitigate the harms associated with them (with so-called "super-emitter" wells an especially high priority [3, 4]). However, a significant fraction of abandoned wells remain unknown. In Pennsylvania, as much as 90% of abandoned wells are estimated to be unrecorded [4]. In Canada, abandoned wells have been described as the most uncertain source of methane emissions nationally, due to the poor quality of data surrounding them [1].

With the advent of large-scale remote sensing datasets and powerful machine learning tools to process them, it has become possible to label and monitor the built environment as never before [5]. Many such works have focused on opportunities to use remote sensing to accelerate climate action and environmental protection, and oil and gas infrastructure has increasingly been an object of scrutiny (see e.g. [6, 7]). In this paper, we present the first large-scale machine learning dataset for pinpointing oil and gas wells, encompassing abandoned, suspended, and active wells. Our main contributions are as follows:

[1] Dataset available at: `https://figshare.com/s/bdb097730714ee82fcb0`

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.



(a) Training set (b) Validation set (c) Test set

Figure 1: Distribution of the number of individual wells in positive samples from the dataset. We also include an equal number of images with no wells at all.

- We introduce the Alberta Wells Dataset, which includes information on over 200k abandoned, suspended, and active oil and gas wells, together with high-resolution satellite imagery.
- We frame the problem of identification of wells as a challenge for object detection and binary segmentation.
- We evaluate a wide range of deep learning algorithms commonly used for similar tasks, finding promising performance but opportunities for significant improvement.

We hope that this work will represent a step towards scalable identification of abandoned well sites and reduction of their deleterious effects upon the climate and environment.

2 Previous Work

Hundreds of satellites continuously monitor the Earth's surface, generating petabyte-scale remote sensing datasets [5]. With advancements in hardware, the quality of remote sensing images has significantly improved in terms of spatial and temporal resolution. High-quality remote sensing data are available through state-funded projects like Sentinel and Landsat, and more recently through private projects such as Planet [8]. Increasingly, machine learning has been used to parse such raw data, including in a wide range of applications for tackling climate change [6]. Benchmark datasets in this area have included tasks in land use and land cover (LULC) estimation [9], crop classification [10, 11], species distribution modeling [12], and forest monitoring [13].

Within this area of research, an increasing body of work has considered the problem of detecting artifacts associated with oil and gas operations. The detection of oil spills using a combination of remote sensing and machine learning has been widely explored [14, 15, 16]. Recently, the detection of oil and gas infrastructure has also been investigated [7, 17], with some studies focusing on the goal of estimating methane emissions [18, 19]. The dataset by [7] includes 7,066 aerial images, with 149 images of oil refineries. The METER-ML dataset [18] comprises 86,599 georeferenced images in the U.S. labeled for methane sources. The OGIM v1 dataset [19] includes 2.6 million point locations of major facilities. A dataset by [20] features 1,388 images of pipelines in the Arctic, while a dataset by [21] includes 3,266 images of heavy-polluting enterprises with 0.25 m resolution.

The problem of detection of oil and gas wells has also been proposed by a number of authors. Existing datasets, however, are quite small (500-5,000 samples), and typically are limited to a small region and contain only active wells, limiting their applicability in the context of identifying abandoned or suspended wells. The NEPU-OWOD V1.0 dataset [22] includes 432 0.41m/px resolution Google Earth Imagery-based high-resolution images from Daqing City, China, containing 1,192 oil wells. The NEPU-OWS V1.0 dataset [23] consists of 1,200 10m/px resolution Sentinel-2 images from Russia with a resolution of 10 m per pixel, covering 1192 oil wells and V2.0 [24] includes 120 multispectral images from Austin, USA. NEPU-OWOD 3.0 [25] contains 722 images with 3749 oil wells from various locations in China & California, with resolutions of 0.48 m/px. A dataset with 5,895 images from Daqing City, each containing 1–5 oil wells at 0.26 m per pixel, was proposed in [26] Another dataset of 930 images from the Permian Basin, USA, was introduced in [27], with resolutions ranging from 15 cm to 1 m per pixel. These various works have largely considered

Table 1: Statistics of wells represented across the Alberta Wells Dataset.

Split	Count Total	Count Wells	Count Non-Wells	Count of Well Type in Wells Patches of Split		
				Abandoned	Suspended	Active
Train	167436	83718	83718	46342	47595	100294
Validation	9463	4731	4731	3166	2671	2406
Test	11789	5894	5894	4024	3609	3340

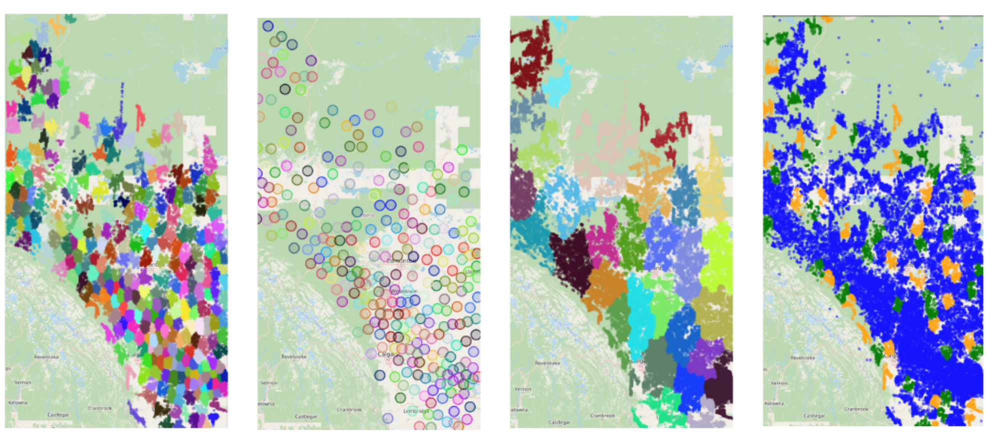
only simple machine learning algorithms for well detection, without evaluating the more complex approaches which have proven useful in other remote sensing contexts.

3 Alberta Wells Dataset

In this paper, we introduce the benchmark **Alberta Wells Dataset** for oil and gas well detection.The dataset is drawn from the province of Alberta, Canada, a region with a substantial number of oil and gas wells and infrastructure present for over a century, including over 94,000 patches of satellite imagery acquired from Planet Labs [8], covering more than 213,000 individual wells. Each patch is annotated with labels for both segmentation and bounding box localization. The annotations are based on data from the Alberta Energy Regulator, quality-controlled by domain experts.

Our dataset attempts to maximize the amount of data available for learning by including a mixture of active and suspended wells alongside abandoned wells. These types of wells appear overall similar in satellite imagery. In contrast to abandoned wells, "suspended" refers to wells that have merely paused operations temporarily, though this designation can be inaccurate, and some wells are classified as suspended for long enough that they are truly abandoned. Active wells are those that are currently in operation.

To simulate real-world conditions, we ensure a varied density of wells per image, as highlighted in Figure 1. We also include satellite imagery patches with no wells present from areas nearby to areas with wells, ensuring no overlap between the samples. This balanced dataset maintains an equal distribution of well and non-well images. Table 1 details the total sample count in each dataset split, alongside the number of well and non-well patches.



(a) Step-1 k_{1i} clusters (b) c_{2i} cluster centroids (c) Step-2 k_{2i} clusters (d) Final Dataset Split

Figure 2: Illustration of the outcome of applying our dataset splitting algorithm: In Figures (a) to (c), different colors represent various cluster IDs. In Figure (d), blue refers to the training set, orange to the validation set, and green to the test set.

Table 2: Information on the numbers of wells represented in the dataset across different states (suspended, abandoned, and active), including domain-specific technical details such as the mode and the types of fossil fuel reserves represented.

Well State	Count	License Status	Mode Short Description	Fluid Short Description
Suspended	55007	Suspension Issued Amended	All Suspended	Gas, Crude oil, Crude bitumen, Liquid petroleum gas, Coalbed methane-coals and other Lith, Coalbed methane-coals only, Shale gas only, Acid gas, CBM and shale and other sources, Shale gas and other sources.
Abandoned	54947	Abandoned Issued Amended	All Abandoned, Abandoned Zone, Junked and Abandoned.	
Active	107139	Issued Amended Re-Entered	Flowing, Pumping,Gas Lift. Abandoned and Re-Entered	

3.1 Well Data Collection, Quality Control & Patch Creation

The Alberta Energy Regulator (AER) oversees the energy industry in the province, ensuring companies adhere to regulations as they develop oil and gas resources. AER publishes AER ST37[28], a monthly list of all wells reported in Alberta, detailing their geographic location, mode of operation, license status, and type of product being extracted, among other attributes. This data is provided in shapefile format along with metadata. However, this data cannot be used directly because the license status or mode of operation does not always correlate with the actual status of the well. Therefore, we work with domain experts to perform quality control on the dataset.

First, we remove duplicate entries from the well metadata, which often contain multiple instances of the same well identified by duplicate license numbers. We resolve these duplicates by retaining the most recent update.A similar approach is applied to the shapefile, where duplicates are resolved using the license date. Afterward, we merge both datasets and filter the data as shown in Table 2, categorizing the wells as active, abandoned, or suspended based on specific criteria developed in consultation with domain experts. We check for duplicate location coordinates in the dataset and resolve them by retaining the instance with the latest drill date. Finally, we ensure all the well instances in the dataset are indeed within the boundaries of Alberta.

After filtering and performing quality control on the datasets with domain experts, we calculate the geographical bounds covered by the well instances across the province and divide the region into nonoverlapping square image patches, each covering an area of 1.1025 sq km (with sides of 1050m). These images include various numbers of individual wells (see Fig. 1), and we ensure that an approximately equal number of patches exist with and without wells.

3.2 Dataset Splitting

To create a well-distributed dataset that represents various geographical regions and offers a diverse benchmark for evaluation and testing, we developed a splitting algorithm (see Algorithm 1). This method involves forming small clusters k_{1i} of nearby well patches based on their centroids as illustrated in Figure 1(a). These small clusters are then grouped into larger, non-intersecting super-clusters k_{2i}, with each super-cluster representing a city or larger geographical area. The formation of super-clusters involves calculating a centroid for each k_{1i} cluster based on the centroids of the well patches it contains as illustrated in Figure 1(b). By clustering wells in this manner, we ensure that k_{1i} clusters group wells from nearby localities together, while k_{2i} clusters group wells from the same geographic region as illustrated in Figure 1(c). Thus, each k_{2i} cluster represents a geographic distribution, with each k_{1i} cluster within it representing a sample of that distribution. To ensure a diverse and well-distributed evaluation and testing of our machine learning model, we select the k_{1i} clusters with the two fewest well instances from each k_{2i} super-cluster for inclusion in the evaluation and test sets. This approach ensures a diverse representation of the dataset as observed in Figure 1(d). Moreover, we maintain an equal distribution of well and non-well patches. In cases of imbalance in non-well images, we exclude such patches from the contributing k_{1i} clusters as specified in Algorithm 1. For imbalances in well images, we sample non-well patches that are not part of any other clusters. The parameters used in constructing the dataset are $M = 300$ and $N = 30$.

Algorithm 1 Clustering Algorithm for Dataset Splitting

W: Set of image patches ids containing wells ; NW: Set of image patches ids not containing wells
Input: x_i represents the i-th patch with centroid coordinates c_i, where $i \in W$ or $i \in NW$;
Output: T_s: Test Set ; T_r: Train Set ; E_v: Eval Set ;
Step 1: Clustering into M Clusters
Perform K-Means Clustering $k_1(*)$ with M clusters using all centroid coordinates c_i, where $i \in W$.
Assign each i-th patch into the m-th cluster where $m \in${1,...,M} and $i \in W$: cluster $k_{1i} = k_1(c_i) = m$ and update patches (x_i, c_i, k_{1i})
for $z \in \{1, \ldots, M\}$ **do**
  $W_{cz} = \{j \in W \mid k_{1j} = z\}$
  Calculate cluster centroids c_{2j} based on values of c_i and update patch: $(x_i, c_i, k_{1i}, c_{2j})$, where $i \in W_{cz}$.
end for
Step 2: Clustering into N Super Clusters
Let W_{cc} be the set of unique c_{2j} for $j \in W$
Perform K-Means clustering $k_2(*)$ with N clusters using all $c_{2i} \in W_{cc}$.
Assign each $c_{2i} \in W_{cc}$ to n-th cluster, where $n \in${1,..,N} & $k_{2i} = k_2(c_{2i}) = n$.
Update patches $(x_j, c_j, k_{1j}, c_{2j}, k_{2j})$ where $c_{2j} = c_{2i}$ and $j \in W$.
Step 3: Assigning Patches to Sets
for $z \in \{1, \ldots, N\}$ **do**
  Find all j with $k_{2j} = z$, where $j \in W$ as W_{fz}.
  Find unique k_{1j} and count o_j associated with it for j in W_{fz}. The, assign k_{1j} with minimum counts as $\min_1$ and $\min_2$.
  For each i in W_{fz}, append i to E_v if $k_{1i} = \min_1$, to T_s if $k_{1i} = \min_2$, otherwise to T_r.
end for
Step 4: Assigning Non-Well Patches
for each set_counter in $\{E_v, T_s, T_r\}$ **do**
  for each unique k_{1i} as $z_i \in set\_counter$ **do**
    Find convex hull radius $r(z_i)$ of area occupied by c_j , where $j \in set\_counter$ & $k_{1j} = z_i$.
    Locate non-well patches $f \in NW$ within radius $r(z_i)$ not in any other cluster; Assign f to cluster z_i: $(x_f, c_f, k_{1f}) : k_{1f} = z_i$.
  end for
end for
Step 5: Imbalance Correction
T_w refers to Count of Well Instances & T_{nw} refers to Count of Non-Well Instances in a Dataset Split
if $T_{nw} > T_w$ **then**
  Identify clusters k_{1j} in data split contributing to the imbalance of excess non-well patches, assign to W_{ic}
  for each i in W_{ic} **do**
    $R(i) = (T_{nw} - T_w) \cdot \frac{\text{Count\_Non\_Wells}(k_{1i})}{\sum \text{Count\_Non\_Wells}(k_{1l}) \text{ where } l \in W_{ic}}$; where $R(i)$ is the no. of Samples to be Removed from i-th Cluster.
  end for
else
  Sample non-well patches $x_j : j \in NW$ & $j \notin k_{1j}$.
end if

3.3 Satellite Imagery Acquisition & Label Creation

We used PlanetScope-4-Band imagery [8] featuring RGB and Near Infrared bands to represent satellite images of the region with a medium resolution of about 3 meters per pixel. PlanetScope, a product of Planet Labs, consists of approximately 130 satellites that can image the entire Earth's land surface daily, collecting up to 200 million sq. km of data each day. We obtained Surface Reflectance imagery, which is offset-corrected, flat-field-corrected, ortho-rectified, visually processed, and radiometrically corrected. These processes ensure consistency across varying atmospheric conditions and minimize uncertainty in spectral response over time and location, making the data ideal for temporal analysis and monitoring applications. To ensure the highest quality, we selected images with no

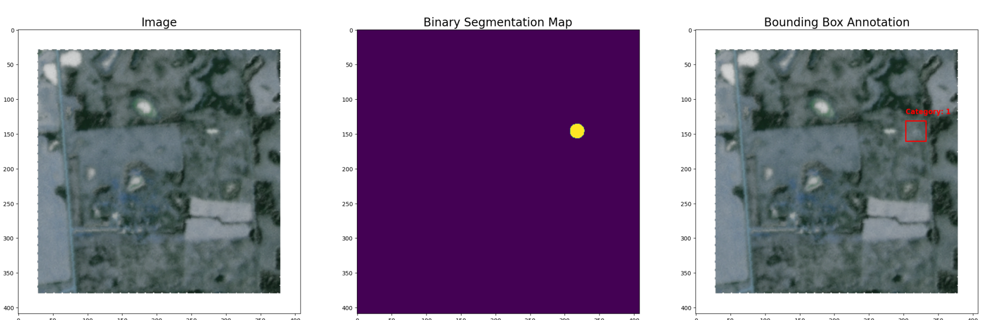


Figure 3: A sample image patch from our dataset, including the infrastructure comprising a single well (visible as a lighter region against the darker gray background), alongside target images from the binary segmentation and object detection tasks.

cloud cover. The images were acquired by Planet satellites within a timeframe that aligns with the well location data from AER. We obtained satellite images for each sample based on geographical coordinates, ensuring an intersection between the actual area of interest and the acquired imagery.

We frame the task of identifying wells as both an object detection and segmentation task, since related remote sensing tasks have found both framings to be constructive. For each image patch as shown in Figure 3, we generated corresponding segmentation maps and object detection annotations for all known wells in the image based on the point labels provided in the AER data. For binary segmentation, we annotated each well site with a circle to match the teardrop shape typical for well sites. We standardized the diameter of a well site to a value of 90 meters (such sites typically range from 70 to 120 meters in diameter). We used the same scale to define bounding boxes in the object detection task, following the COCO [29] format for annotations. Additionally, we created multi-class segmentation maps, where each class represents a different state of the well (active, suspended, or abandoned), and included this information in the object detection annotations. (We do not perform multi-class segmentation experiments here, but it is possible that future researchers may find this task useful.)

4 Benchmark Experiments

We train benchmark deep learning models for both the binary segmentation and object detection tasks. Our focus includes all oil and gas wells, regardless of their operational status, since they exhibit similar footprints and consistent features, making them detectable in satellite imagery.

For both tasks, we augment images by randomly resizing images to 256×256, ensuring all bounding boxes remain intact for object detection. We then apply horizontal and vertical flipping with a probability of 0.25 each, followed by normalization using channel-wise mean and standard deviation calculated from the training split of the dataset. The hyperparameters we use in these various models represent standard performant settings and are not intended to represent the outcome of hyperparameter optimization.

4.1 Binary Segmentation

We selected well-known baseline models for binary segmentation, encompassing the deep CNN-based approaches U-Net [30] and DeepLabV3+ [31] as well as the Transformer-based architectures Segformer[32] and UperNet[33]. U-Net [30] was chosen for its widespread use as a baseline, offering an effective encoder-decoder architecture for multi-scale feature extraction. DeepLabV3+[31] was selected for its popularity in remote sensing tasks with its Atrous Convolution and ASPP module for capturing contextual information at various scales. SegFormer [32] is a transformer-based architecture designed for semantic segmentation, utilizing self-attention mechanisms for capturing long-range dependencies. UperNet [33] combines UNet [30] and PSPNet [34] architectures, featuring a UNet-like structure for multi-scale feature fusion and PSPNet's pyramid pooling module integrated with a Swin Transformer [35] backbone for efficient multi-scale processing.

We train all CNN-based models with a ResNet50 [36] backbone, batch size of 128, and BCELogits loss function. A cosine annealing scheduler [37] adjusts the learning rate smoothly in a cyclical manner, aiding in fine-tuning the model by gradually decreasing the learning rate. For transformer-based models, while both Segformer and UperNet use a Dice loss function and a polynomial learning rate scheduler, Segformer utilizes a mit-b0-ade [32] backbone with a batch size of 128, and UperNet employs a Swin Small Transformer with a batch size of 64. All models are optimized using AdamW [38] for 50 epochs, with the learning rate specified in Table 3.

We evaluate the binary segmentation task with respect to IoU, Precision, Recall, and F1-Score. High Precision corresponds to reducing false positives, while high Recall corresponds to reducing false negatives. IoU measures the overlap between predicted and ground truth masks, offering further insight into segmentation accuracy. F1-Score, the harmonic mean of precision and recall, provides a balanced measure considering both false positives and false negatives.

Table 3: Results for the binary segmentation task for a variety of models evaluated over the test set.We report the Intersection over Union (IoU), precision, recall, and F1-score.

Architecture	Backbone	Learning Rate	IoU	F1 Score	Precision	Recall
U-Net	ResNet50	10^{-3}	56±0.4	59.3±0.2	78.5±2.8	68±1.8
DeepLabV3+	ResNet50	10^{-4}	55.1±0.6	58.5±0.5	77.8±1.7	67.3±1.2
Segformer	mit-b0-ade	6.10^{-4}	51.3±0.7	54.1±0.6	74.8±2.4	69.8±0.2
UperNet	swin small	10^{-4}	51.4± 0.5	54.8±0.5	69.3±0.2	75.3±0.3

4.2 Object Detection

For binary object detection, we consider the CNN-based architectures RetinaNet [39] and Faster R-CNN [40] and the transformer-based architecture DETR [41]. RetinaNet is a one-stage architecture trained using focal loss, which helps to address class imbalance. It uses a Feature Pyramid Network (FPN) for multi-scale feature extraction and efficient object detection across different scales. Faster R-CNN is a two-stage model recognized for its high accuracy. It employs a Region Proposal Network (RPN) for generating region proposals and a separate network for predicting class labels and refining bounding box coordinates. DETR (DEtection TRansformers) is a transformer-based model that treats object detection as a set prediction problem. It eliminates the need for specialized components such as anchor boxes and NMS, using transformers to directly predict the final set of detections.

All object detection models are trained with a ResNet50 backbone. The batch size is 256 for Faster R-CNN and DETR and 512 for RetinaNet. For RetinaNet and Faster R-CNN, we use a cosine annealing scheduler [37]. DETR uses a step-wise learning rate scheduler, reducing the learning rate by a factor of 50 epochs. We train Faster R-CNN and RetinaNet for 120 epochs, and DETR for 150 epochs. All models are optimized using AdamW [38].

In evaluating binary object detection, we compute IoU with various thresholds ($\text{IoU}_{0.1}$, $\text{IoU}_{0.3}$, $\text{IoU}_{0.5}$), indicating how well the model distinguishes between predicted and actual well locations across different overlap levels. We also assess Mean Average Precision (mAP) metrics, including mAP_{50} and $\text{mAP}_{50:95}$, measuring the model's precision-recall trade-off and detection accuracy at various IoU thresholds.

4.3 Results & Analysis

Our tasks involve identifying a roughly circular well region with a 90m diameter in real life, which translates to less than 30 pixels in satellite imagery due to resizing and other augmentations. This poses a challenge for machine learning models given the heterogeneous nature of the background, including various similarly shaped and sized features of the natural and built environment. Additionally, vegetation can occlude wells in RGB channels, highlighting the importance of near-infrared imagery for guiding the model. The wells themselves also vary somewhat in shape, and can be in various states of disrepair as a result of differing ages and maintenance.

For the binary segmentation task framing, we train both CNN-based and Transformer-based backbones, considering the prevalent imbalance in the image data due to the small size of wells. Among our models, as shown in Table 3, the traditional U-Net performs the best, with CNN-based models showing higher IOU, precision, and F1 scores, indicating more accurate predictions of well instances compared to other models. Precision, which reflects the accuracy of our positive detections compared to the ground truth, is crucial. However, a high recall value ensures the model captures most actual well instances, reducing the risk of missing important information. Thus, the Uper-Net model with the highest recall value of 75.3± 0.3, which excels at capturing global context information, appears well-suited for this task.

For the object detection task framing, the IoU metrics measure how accurately the model identifies predicted well locations compared to the actual locations, at different levels of overlap. A higher IoU indicates better alignment between predicted and ground truth bounding boxes. Mean Average Precision (mAP) metrics, including mAP_{50} and $\text{mAP}_{50:95}$, provide a comprehensive assessment of

Table 4: Results for the object detection task for a variety of models evaluated over the test set. We report the intersection over union (IoU) over thresholds $0.1, 0.3, 0.5$ and the mean average precision (mAP) for both IoU$= 0.5$ and IoU$\in [0.5, 0.95]$ thresholds.

Architecture	Learning Rate	$\text{IoU}_{0.1}$	$\text{IoU}_{0.3}$	$\text{IoU}_{0.5}$	mAP_{50}	$\text{mAP}_{50:95}$
RetinaNet	10^{-4}	24.58±0.11	43.07±0.8	59.79±0.36	0.72±1.12	0.18±0.28
FasterRCNN	10^{-3}	36.79±1.07	46.95±0.66	61.29±0.35	19.12±3.41	5.2±1.0
DETR	10^{-4}	21.6±0.25	42.1±1.38	60±2.64	$24.1\times10^{-5} \pm 7.75\times10^{-5}$	$6.8\times10^{-5} \pm 4.09\times10^{-5}$

the model's precision-recall trade-off. mAP_{50} considers precision at a single IoU threshold of 0.5, giving an overall measure of the model's accuracy in detecting well instances. On the other hand, $\text{mAP}_{50:95}$ evaluates the model's performance across a range of IoU thresholds from 0.5 to 0.95, providing a detailed understanding of its precision-recall behavior across different levels of detail in the predictions.

Our evaluation, as shown in Table 4 indicates that while all models perform reasonably well in terms of aligning predicted and actual well locations, Faster R-CNN stands out with the highest $\text{IoU}_{0.5}$ score of 61.29 ± 0.35. However, all models perform poorly in terms of mean average precision, with Faster R-CNN achieving the highest score of only 19.12 ± 3.41. DETR and RetinaNet perform particularly poorly, with near-zero scores indicating their inability to identify well-bounding box locations accurately. This could be attributed to the fact that these models might not produce region proposals confidently enough, especially considering instances with a large number of wells. While IoU scores are decent with increasing thresholds, the mAP scores indicate that a more complex model may be required for this task.

5 Conclusion

In this paper, we have introduced the first large-scale dataset for identifying oil and gas wells, in particular abandoned wells, which represent a major source of greenhouse gases and other pollutants. We combine high-resolution imagery, an extensive database of well locations, and expert verification to create the Alberta Wells Dataset. We frame well identification both in terms of object detection and binary segmentation, and evaluate the performance of a wide range of popular deep learning methods on these tasks. We find that the Uper-Net model in particular represents the most promising baseline for the binary segmentation task, while for object detection all models demonstrate more mixed results, with relatively strong IoU scores but weak mAP. These results show that the Alberta Wells Dataset represents both a challenging as well as a societally impactful set of tasks.

We do not envision any significant negative uses of our work. Localization of wells is primarily of interest to the climate change mitigation community and is not, for example, a primary means whereby fossil fuel companies select new locations for drilling. Therefore, we do not believe this dataset is susceptible to dual use.

One potential limitation of our work is that we rely on well locations listed by the Alberta Energy Regulator. It is likely that many true well locations are missing in this data, leading to the potential for false negatives in the ground-truth data for this problem. However, it is to be expected that this will not significantly affect the training of algorithms since these labels represent a small fraction of the negative locations in the dataset, and deep learning algorithms are known to be robust to moderate amounts of label noise (see e.g. [42]). Instead the effect may simply be that the reported test accuracy is actually lower than the true value (due to certain correctly predicted well locations being evaluated as false). We hope to investigate such effects further in future work.

Another noteworthy limitation is the exclusive focus on Alberta, which we selected because there is a large amount of labeled data available for this region. Another promising direction for future work will be to assess the capacity for few- or zero-shot transfer learning from the region of Alberta to other regions with a high expected concentration of abandoned wells, including the Appalachian and Mountain West regions of the United States, as well as a number of former Soviet states.

We hope that our work may be of use to policymakers and other stakeholders involved in climate action and environmental protection, according to the following envisioned steps:

- Use the Alberta Wells Dataset to train algorithms for pinpointing well locations.
- Run these algorithms at scale across a broader region of interest, comparing against any existing databases to identify those wells which may be undocumented.
- Flag abandoned wells for plugging, prioritizing those identified as super-emitters.

We believe that the scalability of machine learning tools for remote sensing will make them an invaluable tool in pinpointing and mitigating the global environmental impact of abandoned oil and gas wells.

References

[1] James P Williams, Amara Regehr, and Mary Kang. Methane emissions from abandoned oil and gas wells in Canada and the United States. *Environmental science & technology*, 55(1):563–570, 2020.

[2] Aaron G Cahill, Roger Beckie, Bethany Ladd, Elyse Sandl, Maximillian Goetz, Jessie Chao, Julia Soares, Cara Manning, Chitra Chopra, Niko Finke, et al. Advancing knowledge of gas migration and fugitive gas from energy wells in northeast british columbia, canada. *Greenhouse Gases: Science and Technology*, 9(2):134–151, 2019.

[3] Stuart N Riddick, Mercy Mbua, Arthur Santos, Ethan W Emerson, Fancy Cheptonui, Cade Houlihan, Anna L Hodshire, Abhinav Anand, Wendy Hartzell, and Daniel J Zimmerle. Methane emissions from abandoned oil and gas wells in colorado. *Science of The Total Environment*, 922:170990, 2024.

[4] Mary Kang, Shanna Christian, Michael A Celia, Denise L Mauzerall, Markus Bill, Alana R Miller, Yuheng Chen, Mark E Conrad, Thomas H Darrah, and Robert B Jackson. Identification and characterization of high methane-emitting abandoned oil and gas wells. *Proceedings of the National Academy of Sciences*, 113(48):13636–13641, 2016.

[5] Esther Rolf, Konstantin Klemmer, Caleb Robinson, and Hannah Kerner. Mission critical – satellite data is a distinct modality in machine learning. In *International Conference in Machine Learning (ICML)*, 2024.

[6] Jun Yang, Peng Gong, Rong Fu, Minghua Zhang, Jing Chen, Shunlin Liang, Bing Xu, Jiancheng Shi, and Robert Dickinson. The role of satellite remote sensing in climate change studies. *Nature Climate Change*, 3:875–883, 09 2013.

[7] Hao Sheng, Jeremy A. Irvin, Sasankh Munukutla, Shenmin Zhang, Christopher Cross, Kyle T. Story, Rose Rustowicz, Cooper W. Elsworth, Zutao Yang, Mark Omara, Ritesh Gautam, Robert B. Jackson, and A. Ng. OGNet: Towards a global oil and gas infrastructure database using deep learning on remotely sensed imagery. *ArXiv*, abs/2011.07227, 2020.

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

[9] Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. BigEarthNet: A large-scale benchmark archive for remote sensing image understanding. *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 5901–5904, 2019.

[10] Dimitrios Sykas, Maria Sdraka, Dimitrios Zografakis, and Ioannis Papoutsis. A Sentinel-2 multi-year, multi-country benchmark dataset for crop classification and segmentation with deep learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2022.

[11] Gabriel Tseng, Ivan Zvonkov, Catherine Lilian Nakalembe, and Hannah Kerner. CropHarvest: A global dataset for crop-type classification. In *Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

[12] Mélisande Teng, Amna Elmustafa, Benjamin Akera, Yoshua Bengio, Hager Radi, Hugo Larochelle, and David Rolnick. Satbird: a dataset for bird species distribution modeling using remote sensing and citizen science data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[13] Nikolaos Ioannis Bountos, Arthur Ouaknine, and David Rolnick. FoMo-Bench: a multi-modal, multi-scale and multi-task Forest Monitoring Benchmark for remote sensing foundation models. *arXiv e-prints*, page arXiv:2312.10114, December 2023.

[14] Xiaodao Chen, Dongmei Zhang, Yuewei Wang, Lizhe Wang, Albert Y. Zomaya, and Shiyan Hu. Offshore oil spill monitoring and detection: Improving risk management for offshore petroleum cyber-physical systems. *2017 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 841–846, 2017.

[15] Yuewei Wang, Xiaodao Chen, and Lizhe Wang. Cyber-physical oil spill monitoring and detection for offshore petroleum risk management service. *Scientific Reports*, 13, 2023.

[16] Junfang Yang, Yi Ma, Yabin Hu, Zongchen Jiang, J. Zhang, Jianhua Wan, and Zhongwei Li. Decision fusion of deep learning and shallow learning for marine oil spill detection. *Remote. Sens.*, 14:666, 2022.

[17] Samyak Prajapati, Amrit Raj, Yash Chaudhari, Akhilesh Nandwal, and Japman Singh Monga. OGInfra: Geolocating oil & gas infrastructure using remote sensing based active fire data. *ArXiv*, abs/2210.16924, 2022.

[18] Bryan Zhu, Nicholas Lui, Jeremy Irvin, Jimmy Le, Sahil Tadwalkar, Chenghao Wang, Zutao Ouyang, Frankie Y. Liu, Andrew Y. Ng, and Robert B. Jackson. METER-ML: A multi-sensor Earth observation benchmark for automated methane source mapping, 2022.

[19] Mark Omara, Ritesh Gautam, Madeleine A. O'Brien, Anthony Himmelberger, Alexandre Puglisi Barbosa Franco, Kelsey Meisenhelder, Grace Hauser, David R. Lyon, Apisada Chulakadabba, C. Chan Miller, Jonathan E. Franklin, Steven C. Wofsy, and Steven P. Hamburg. Developing a spatially explicit global oil and gas infrastructure database for characterizing methane emission sources at high resolution. *Earth System Science Data*, 2023.

[20] Huan Chang, Lu Bai, Zhibao Wang, Mei Wang, Ying Zhang, Jinhua Tao, and Liangfu Chen. Detection of over-ground petroleum and gas pipelines from optical remote sensing images. In *Remote Sensing*, 2023.

[21] Zhibao Wang, Xi Zhao, Lu Bai, Mei Wang, Man Zhao, Meng Fan, Jinhua Tao, and Liangfu Chen. Detection of heavy-polluting enterprises from optical satellite remote sensing images. In *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 6474–6477, 2023.

[22] Zhibao Wang, Lu Bai, Guangfu Song, Jie Zhang, Jinhua Tao, Maurice D. Mulvenna, Raymond R. Bond, and Liangfu Chen. An oil well dataset derived from satellite-based remote sensing. *Remote Sensing*, 13(6), 2021.

[23] Hao Wu, Hongli Dong, Zhibao Wang, Lu Bai, Fengcai Huo, Jinhua Tao, and Liangfu Chen. Semantic segmentation of oil well sites using Sentinel-2 imagery. In *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 6901–6904, 2023.

[24] Hao Wu, Hongli Dong, Zhibao Wang, Lu Bai, Fengcai Huo, Jinhua Tao, and Liangfu Chen. Spatial information extraction of oil well sites based on medium-resolution satellite imagery. In *Image and Signal Processing for Remote Sensing XXIX*, volume 12733, page 127330K. International Society for Optics and Photonics, SPIE, 2023.

[25] Yu Zhang, Lu Bai, Zhibao Wang, Meng Fan, Anna Jurek-Loughrey, Yuqi Zhang, Ying Zhang, Man Zhao, and Liangfu Chen. Oil well detection under occlusion in remote sensing images using the improved YOLOv5 model. *Remote Sensing*, 15(24), 2023.

[26] Pengfei Shi, Qigang Jiang, Chao Shi, Jing Xi, Guo Tao, Sen Zhang, Zhenchao Zhang, B. Liu, Xin Gao, and Qian Wu. Oil well detection via large-scale and high-resolution remote sensing images based on improved YOLO v4. *Remote. Sens.*, 13:3243, 2021.

[27] Jade Eva Guisiano, Éric Moulines, Thomas Lauvaux, and Jérémie Sublime. Oil and Gas Automatic Infrastructure Mapping: Leveraging High-Resolution Satellite Imagery through fine-tuning of object detection models. In *International Conference on Neural Information Processing (ICONIP)*, Changsha, China, November 2023.

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

[29] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 2014.

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, abs/1505.04597, 2015.

[31] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision*, 2018.

[32] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, José Manuel Álvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Neural Information Processing Systems*, 2021.

[33] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. *ArXiv*, abs/1807.10221, 2018.

[34] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6230–6239, 2016.

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021.

[36] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015.

[37] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *arXiv: Learning*, 2016.

[38] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in Adam. *ArXiv*, abs/1711.05101, 2017.

[39] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:318–327, 2017.

[40] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015.

[41] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *ArXiv*, abs/2005.12872, 2020.

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

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] We do not provide exact numbers, instead providing qualitative estimates; compute used is low overall.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]