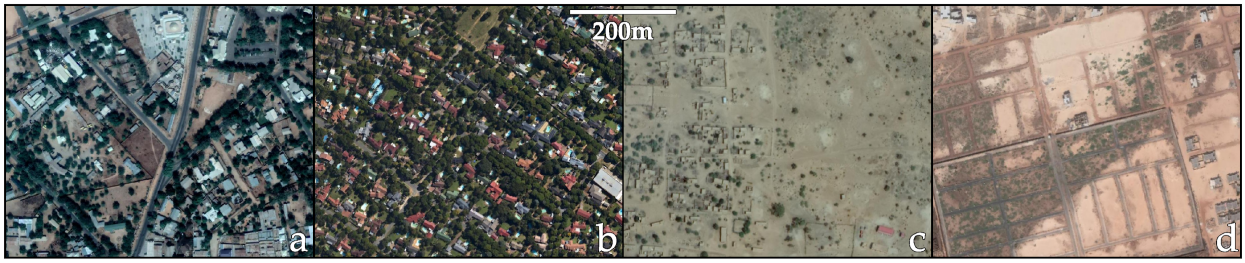

GEODOME

April 23, 2021

Project Team: Abdullah AlOthman, Sebastián Soriano Pérez, Jose Moscoso

Project Manager: Wei (Wayne) Hu

Project Partner: Dr. Kyle Bradbury

Abstract

Earth Observation (EO) data provide a collection of spatiotemporal records about the planet, usually based on satellite systems. The unique view of Earth that EO data supply reduces the time and complexity required to monitor human-Earth dynamics in diverse fields such as global transport logistics, environmental management, agriculture or public health. Recent advances in machine learning techniques make possible the automation of tasks that extract valuable information from images. Unfortunately, the shortage of available EO labeled datasets, limits the deployment at scale of current machine learning models in real-world applications.

With GEODOME, (Geographically Diverse Earth Observation Dataset with Multiple Sensor Modalities), we aim to develop a tool that simplifies the process of obtaining open source EO data. We also designed experiments to measure the generalization performance of the dataset and guide future data collection.

***Index Terms**: Earth observation, object detection, image segmentation, satellite systems, domain adaptation.*

Introduction

There are multiple data sources for EO data in the form of satellite imagery that we could use in order to develop the GEODOME dataset. These sources contain datasets of satellite images from different regions of the world with varying resolution qualities and with data from multiple sensor modalities. A sensor modality is a term that refers to the different devices that are able to capture or perceive a stimulus or signal from the environment, process this signal, and output data that can be used to measure different characteristics of a particular region of the world. Different sensor modalities can capture and measurement or pressure among others.

Sensor modalities can also be classified as active, when they emit a signal that alters the environment around them so it can then measure its effect through its receptive sensors, or passive when they only consist of sensors receptive to outside stimuli (Figure 1). For the first version of the GEODOME benchmark dataset, we focus on using passive sensor modalities that capture different ranges of the light spectrum reflected on the Earth surface on different bands available through satellite imagery. These bands include the visible light channels (red, green, and blue, or RGB), as well as bands outside the visible spectrum such as infrared and ultraviolet light.

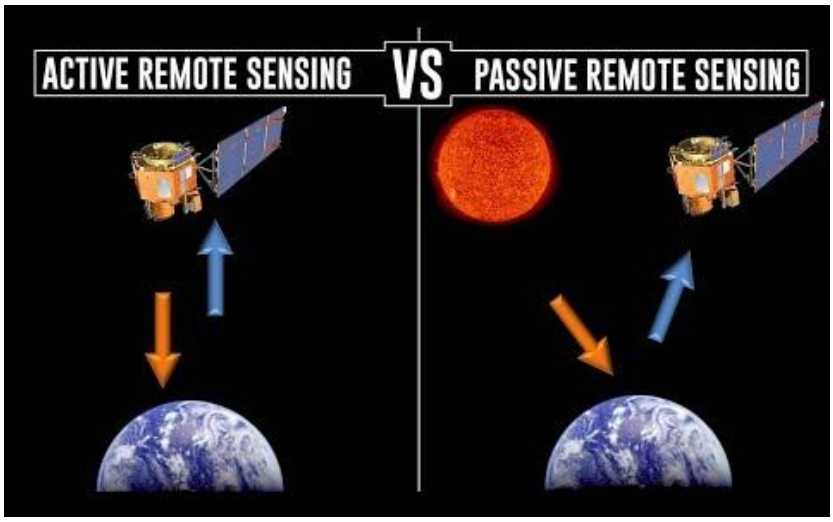

Figure 1. Active vs Passive Remote Sensing.
Image Source: https://www.geospatialworld.net/videos/active-and-passive-remote-sensing/

The evolution of EO data to reveal not only images at a large scale but also the dynamics of the human-Earth interactions in time has prompted new challenges to the EO community that require more efficient mechanisms to turn raw data into valuable insights [1]. The use of Earth observation (EO) data on different disciplines such as energy [2], environmental sciences [3], and healthcare services [4] has gained traction with the development of new distributed computer systems and machine learning algorithms that are capable of processing heavy satellite imagery data efficiently. However, in order to reach state-of-the-art results, two conditions are required. First, the deep learning model needs to be trained with a large dataset of high resolution and quality annotated images, as deep learning models are known to be data-hungry [5]. Second, the performance evaluation test should use a dataset of images with similar resolution characteristics and background context to the ones where the object of interest is annotated for the training process. This situation has been referred to in literature as the domain adaptation problem [6].

Some examples of publicly available datasets of EO data are SpaceNet [7], Inria [8], DSTL (Defence Science and Technology Laboratory) dataset and DeepGlobe [9]. These datasets provide a spatial resolution higher than 1m/pixel of aerial imagery. The SpaceNet dataset, for instance, consists of satellite imagery at 50 cm resolution from six cities around the world and contains curated annotation of images with information about the building footprints and roads located in them. Despite the unprecedented size and geographic coverage of this dataset, it fails to be a good representation of all the contrast that real-world overhead imagery contains as it lacks the geographical diversity that would make it representative of all regions of Earth.

Empirical studies have shown that deep learning models trained on imagery datasets without sufficient geographic diversity have limitations in their generalization performance [10]. When satellite imagery of different geographies is used in object detection tasks using models trained on a dataset which was collected within a specific geographic region (i.e. cities) at a specific time of the day, the accuracy of the model is considerably lower. In consequence, the inability to transfer knowledge from models trained on datasets with limited geographical diversity restricts its applicability at scale. Besides natural differences between urban and rural locations, terrain types or idiosyncratic architectural designs particular to a region, technical and atmospheric conditions also affect the visual characteristics of images.

Different approaches have been explored to overcome the domain adaptation problem of publicly available datasets. One approach to remediate the lack of background geographic diversity in these datasets is to use synthetic imagery [11]. Artificial 3D models of objects of interest (i.e. wind turbines) are placed over a more diverse real background imagery. Since the location of the objects of interest is known from design, there is no need to manually annotate them. The major challenge with this approach is that it requires a skilled software designer to invest time recreating features in the virtual world that mimic characteristics of the objects of interest in the real world. For illustrative purposes, the Synthinel-1 Dataset paper reproduces small representations of the world with a focus on high fidelity visual features and randomness of small objects (vegetation, building, road, people). As we approximate the level of variation and details present in a real-world scenario, the amount of work required in the development of large-scale virtual scenarios would become cumbersome.

In this paper, we build upon previous efforts to expand the availability of high-quality and geographically diverse datasets using crowd-sourced label acquisition [12] [13]. One clear benefit of this approach is that public crowd-platforms, such as OpenStreetMap (OSM), provide inexpensive and useful sources of information at scale. GEODOME incorporates by design, a collection of diverse EO geographies representing the real-world variability.

Project Objectives

- The first goal of the GEODOME project is the creation of a tool that enables the continuous generation of annotated EO data. We will use existing sources of labeled geodata to retrieve labels of buildings and other structures on the Earth surface and match them with the corresponding satellite images over multiple sensor modalities from publicly available EO datasets.

- The second goal of the project is to create a geographically diverse dataset of annotated data that can become a new benchmark dataset for the earth observation computer vision community.
- Stretch goal: explore how to improve the performance of models on images from geographic regions that are not similar to the ones that it has been trained on.

Pipeline Process

In this section, we describe the overall structure of the pipeline process we have devised to create the GEODOME dataset (outlined in figure 2). These steps are data agnostic and thus will have only a general overview here, but a more detailed description of the work done throughout this process will be provided in the relevant sections

We first survey publicly available data sources of infrastructure labels and satellite imagery and choose a candidate source. It is important that this source be open to the public to align with our goal of having our resulting GEODOME dataset be open for use and expansion by everyone. After picking the data source we perform the prototypical exploratory data analysis to get more familiar with the source beyond what is available in the documentation and description. The EDA step is usually accompanied by quality assurance as we will inspect the data as we explore it, which allows for discovering any potential issues inherent in the data source (such as inconsistent labels or limited availability) in the early stages of the pipeline.

What follows is what we call the 'annotation' step where we match the infrastructure labels with the satellite imagery based on location and perform all the preprocessing steps required to make the dataset usable as a machine learning training set, this includes steps such as combining data from multiple satellite bands as a single image, normalization, and creating a pixel mask for the locations of the labels (rasterization). Finally, we run a domain adaptation experiment that measures how well the dataset generalizes, the quality of our domain designation, and highlights which domains need more samples. The results of the domain adaptation experiment are then used to guide us in the iterative process of continuously developing and expanding GEODOME and informs the steps of the next iteration through the pipeline.

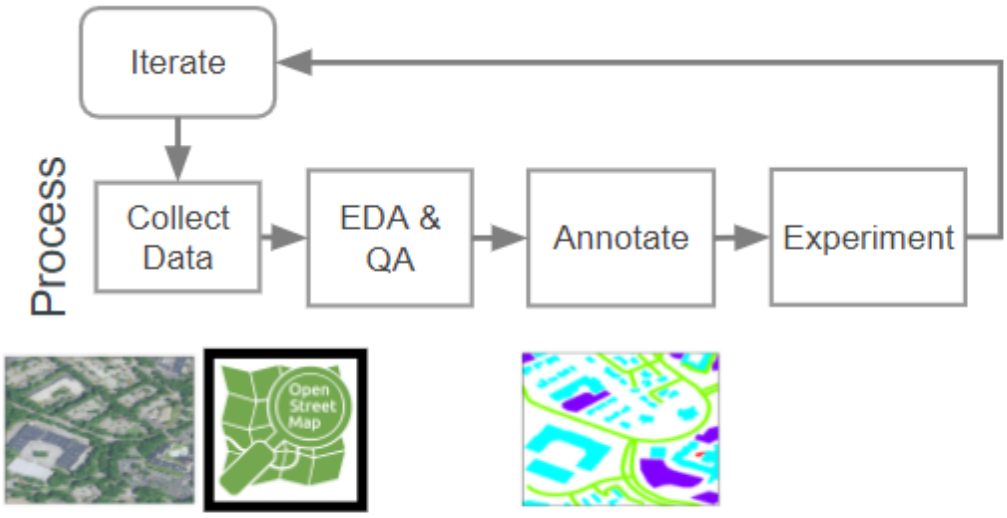

Figure 2. Flow chart of the GEODOME pipeline.

Data Sources

Data Source: OpenStreetMaps

To collect data that we can use to annotate the GEODOME dataset we use the open-source OpenStreetMap (OSM) project. OSM is a collaborative project for the creation of an open and editable map of the world for which thousands of collaborators around the world contribute. OSM was created in 2004, inspired by Wikipedia and its collaborative nature, to provide an alternative to the restrictions of the proprietary map data currently available at the time.

OSM has thousands of registered users from all over the world who contribute to the project by editing maps, labeling buildings, roads, sites, structures, and other land features that can be observed from satellite imagery. The main output of the OSM project is the free and publicly available geodata that its contributors generate and keeps growing on a daily basis. The information generated for a particular geographical region can be accessed and retrieved in the form of GeoJSON structured data that includes the labels of every building, road, site, and other relevant landmarks in the area as well as their exact location as indicated by a polygon drawn with their coordinates. These polygons and their respective labels are the data we use to annotate the GEODOME dataset. There are several APIs publicly available to retrieve OSM data that can be used to identify and label objects in satellite imagery.

Being a crowdsourced project with so many people working on it at the same time, handling the OSM data also presents a few challenges to consider. The labeling of places,

buildings, and roads depends entirely on the OSM collaborators and their knowledge of a particular region, so the data being available or up-to-date depends on the availability and interest of people to work on it. We have encountered that the low coverage of the labels (i.e. the proportion of visible objects or structures in satellite imagery that have a corresponding label in OSM data) is a very important challenge when retrieving OSM data. The labels that are used to describe similar objects may not always be consistent, especially across different regions since consensus for a standard label is sometimes difficult to reach and enforce. This leads to, for example, church buildings being labeled with either the '*amenity=place_of_worship*' label or the '*building=church*' label in different locations.

Moreover, there are some instances of objects that are mislabeled (incorrect labels are used to tag them). Collaborators also label services or amenities that are not visible from satellite imagery and therefore are not useful for computer vision tasks that utilize satellite imagery (such as labels that indicate whether there are toilets inside a building). Finally, there are some labels that are not widely used and exist in very small and non-representative numbers. All of these problems represent an important challenge that must be addressed by analyzing and processing the OSM geodata after it is collected, so it can be used to annotate the GEODOME dataset.

Data Source: Google Earth Engine (GEE)

We use the Google Earth Engine (GEE) API as our primary source of acquiring satellite imagery and geospatial data. Google Earth Engine is a service provided by Google with the goal of providing an easy tool for academics and researchers to conduct scientific analysis and visualizations on geospatial datasets. It also provides easy access to a massive archive of non-proprietary satellite imagery and geospatial data. Through the GEE API, it is possible to access satellite imagery with multiple sensor modalities or 'bands' from a vast number of sources or satellites.

A major drawback for this data source is that it was made with the goal of easing the analysis and exploration of the data, not for the massive acquisition of it. This means that while the API provides ways for the user to manipulate the data and analyze it on the GEE servers, it is more complicated and challenging to download the data as we intend to do in the GEODOME project. Luckily, there is a vibrant and helpful community of GEE users that have provided tools that enabled us to work around that hurdle, making GEE a feasible source for our project.

Data Source: National Agriculture Imagery Program (NAIP) aerial imagery Dataset

For the first version of the GEODOME, we decided to use the NAIP dataset as our main source of satellite imagery. NAIP is a USDA Farm Service Agency program that captures aerial imagery of the continental United States during the agricultural growing season. The imagery provided has a resolution ranging between 0.6 and 1-meter gsd (ground sample distance) and has four major bands that cover the visible light spectrum

(RGB) and near-infrared wavelength (N). NAIP is an excellent source for the pilot version of GEODOME because of the high quality of the imagery and the diverse geographies and terrain types contained within the United States borders. Those diverse terrain types would be used as domain designations to both measure the generalizability of each terrain type and guide the development of a universally applicable dataset.

Data Source: National Land Cover Database (NLCD 2016)

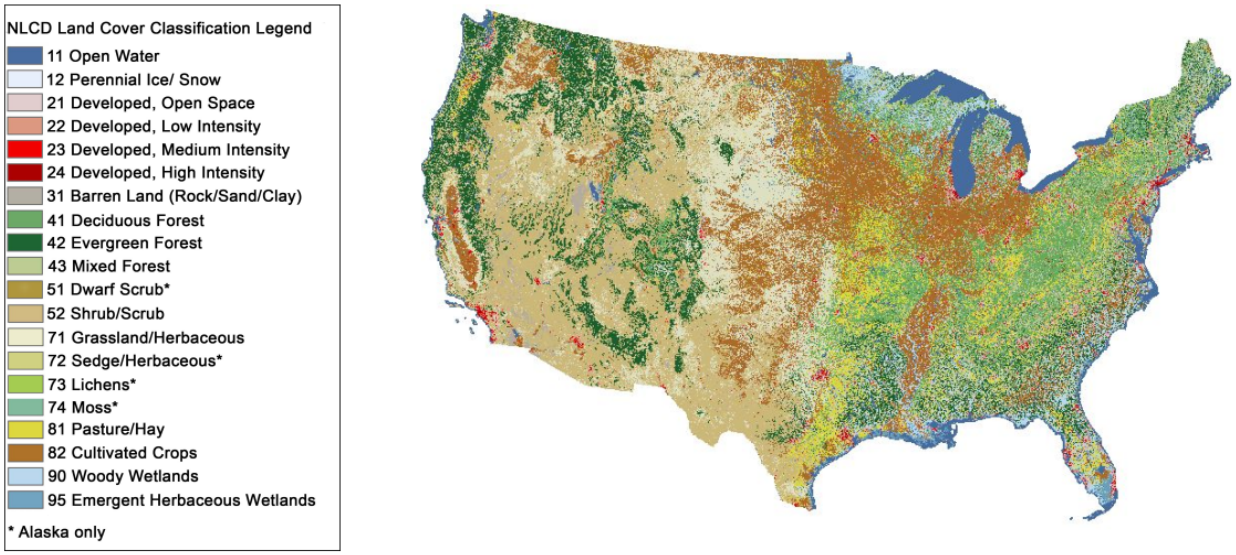

Figure 3. National Land Cover Database Land Cover Classification of the United States.

The US Geological Survey (USGS) has developed the NLCD which provides 30-meter resolution data on the type of land cover visible in satellite imagery across the US. Areas in mainland USA are labeled with one of 16 classes (as seen in figure 3) using a decision-tree-based classifier with accuracies ranging between 71% and 91% between land covers. We plan to use this data as a domain identifier for GEODOME and thus have elected to collapse these 16 classes into their 9 main classes (barren, crops, developed, forest, grassland, open water, pasture, scrub, and wetland) in order to minimize cross-domain leakage and have a more robust domain designation.

GEODOME Version 1.0

In order to use the OSM data appropriately in a way that would be useful for the purpose of the GEODOME dataset, we limited the number of labels obtained from the platform that we actually utilize for the annotation of the dataset (see Data Sources section above). For this purpose, we decided to sample and analyze several locations of Earth that

we deemed to be representative, so we could retrieve the OSM labels at these regions and execute exploratory data analysis to decide which labels were the most significant for our project considering our goals. The OSM labels consist of key-value pairs (i.e. 'key=value') where the key is used to describe a topic, category, or type of feature, and the value provides detail for the key-specified feature [15].

We handpicked a total of 200 coordinates across all continents of places that had buildings or other structures labeled in the OSM data. Additionally, we randomly selected 925 locations, 5 from each country plus an additional 100 in the United States since we plan to use NAIP as the initial satellite source. We assume these 1125 locations to be representative of the totality of the surface of the Earth. We then retrieved the OSM labels at these 1125 locations by taking a square area of 1000 m per side centered at the location's coordinates. We aggregated and counted the number of occurrences of each label by the number of images they appeared on and by the total number of times they appeared on all of the images. We collected a total of 682 labels that we used as a basis to analyze and create our final list of labels for the GEODOME project.

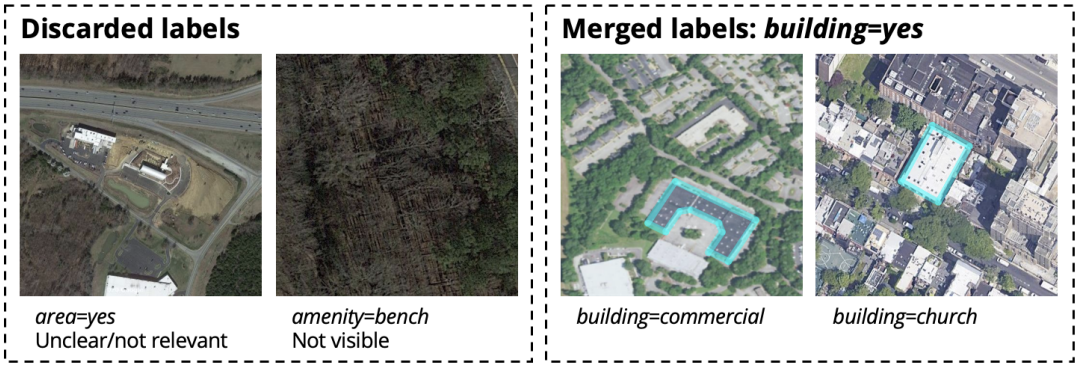

Figure 4. Left: Examples of discarded OSM labels. Right: Examples of merged OSM labels (polygons are highlighted in blue).

The next step in our process was to analyze each of the OSM labels we collected individually by looking at their frequency and a few sample satellite images of the locations where they are found to determine if they were visible and relevant both in terms of the number of times they are used and the significance to the GEODOME project (the labels represented structures that can be used for computer vision tasks our dataset is intended for). Figure 4 shows two examples of labels that were discarded for representing objects that are not visible with satellite imagery, were unclear and not relevant for our project goals, or covered regions that are too big to be useful for our purposes. Also, we show an example of two labels that were merged into a single new 'taxonomy' of labels that consists of two proposed hierarchies: a primary label and an optional secondary label.

The output of our OSM analysis consists of a dictionary of labels that can be used when retrieving OSM data for the annotation of the GEODOME dataset. The dictionary maps the original 682 OSM labels into the primary and optional secondary labels we assigned through our hierarchical taxonomy. It also indicates which labels are to be discarded and not used in the annotation of the GEODOME dataset. The grouping label classes are also assigned a label priority to determine its depth when annotating the dataset. After our initial experiments, we decided to assign all visible and useful label classes to one of the following three grouping label classes (with no secondary label):

- ***building=yes*** [priority 1]: Includes 272 label classes such as *building=commercial*, *building=residential*, or *building=church*. All of them represent visible building structures in satellite imagery.
- ***amenity=parking*** [priority 2]: Includes 5 label classes, all of which are used to label distinctly shaped parking lots visible in the satellite imagery.
- ***highway=road*** [priority 3]: Includes 37 label classes used to label different types of roads.

Using the NLCD classifications we randomly sampled 8145 coordinates to cover all 8 major land cover types and used the tools we developed to download NAIP imagery of at least 550 meters per side square area centered on each of these coordinates. We also obtained the shapefiles and infrastructure labels that correspond to each of these locations.

The angle of an aerial image is not preset and therefore could affect how well the labels match up to the image. Thankfully the exact transformations that affect the image are available in the metadata of the tif file. This information can be used to apply those transformations to the polygons and lines in the label shapefile and ensure that each label holds the same position as the object that it corresponds to. After transforming the labels we then convert them from vectors into pixel bitmaps or 'rasters' that mark the location of each label and can be used as targets in training deep learning models for the domain adaptation experiment.

In this way, the output of the GEODOME data collection pipeline are:

- Satellite imagery download

A total of 7834 RGB + Near-infrared images with a resolution of 1m/pixel aerial imagery from NAIP were downloaded. The image's width size ranged between 607 and 845 pixels, and the image's height size was 550 pixels. These images come from a set of coordinates sampled from the 9 mainland cover types across the U.S. The set of tools developed for this purpose is expandable to other satellite systems.

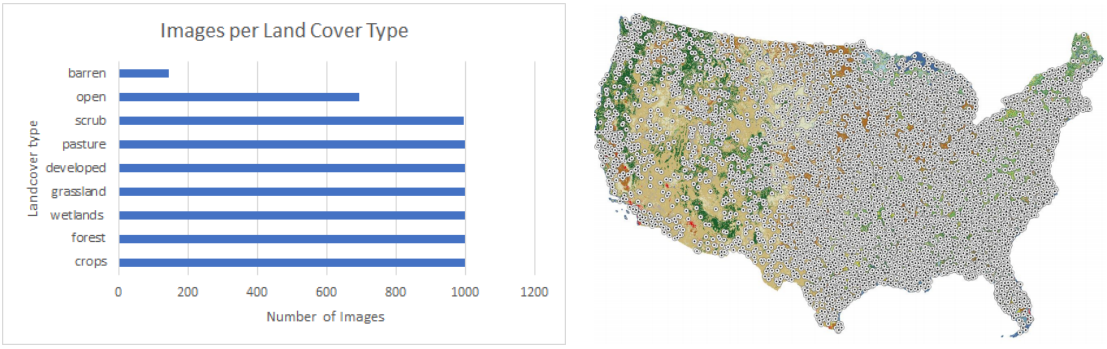

Figure 5. Left: Distribution of Images per land cover type. Right: visualization of the location of the samples over the U.S. map

- OSM label pipeline

Using the OSM tools module we developed, we retrieved OSM data asynchronously from the same locations where we collected the satellite imagery dataset. We then transformed OSM GeoJson into a shapefile for each location. Based on an iterative process, and considering the frequency and consistency of labels across different regions, we created our own hierarchical taxonomy of labels that translates the original OSM labels into one of 3 classes: building, parking or roads.

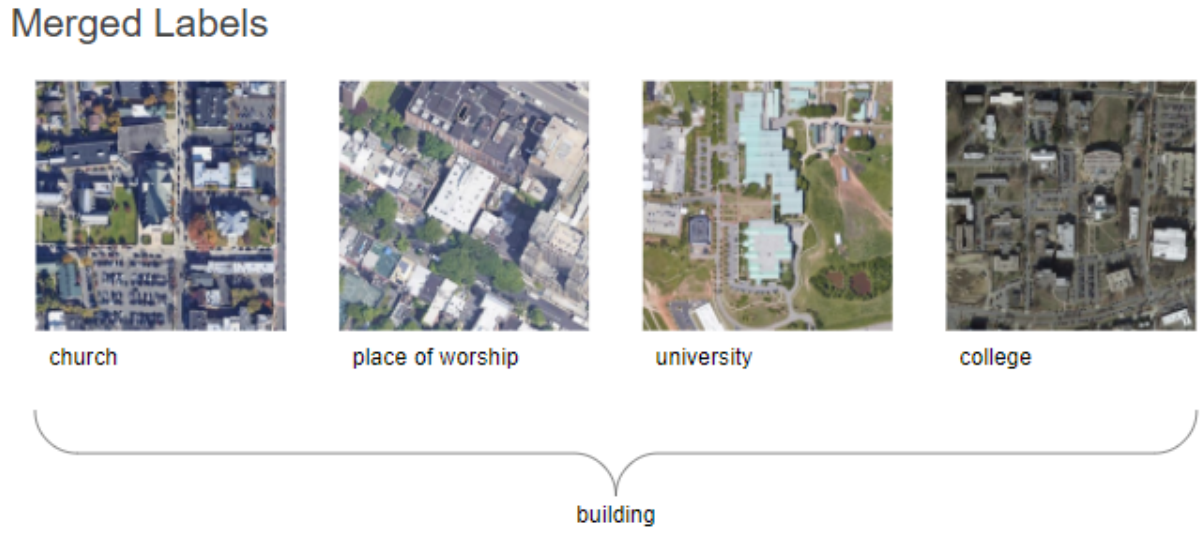

Figure 6. Example of OSM tags for similar structures merged into building category

- Rasterization pipeline

Take the satellite imagery and its corresponding shapefile, and align them using the information stored in the metadata to ensure that OSM polygons match the viewing angle in the image. We also assign a “depth” value for each polygon to handle the overlapping

infrastructure. Finally, we obtain a set of bitmaps where each pixel is assigned to one of our three grouping label classes or class 0 (background).

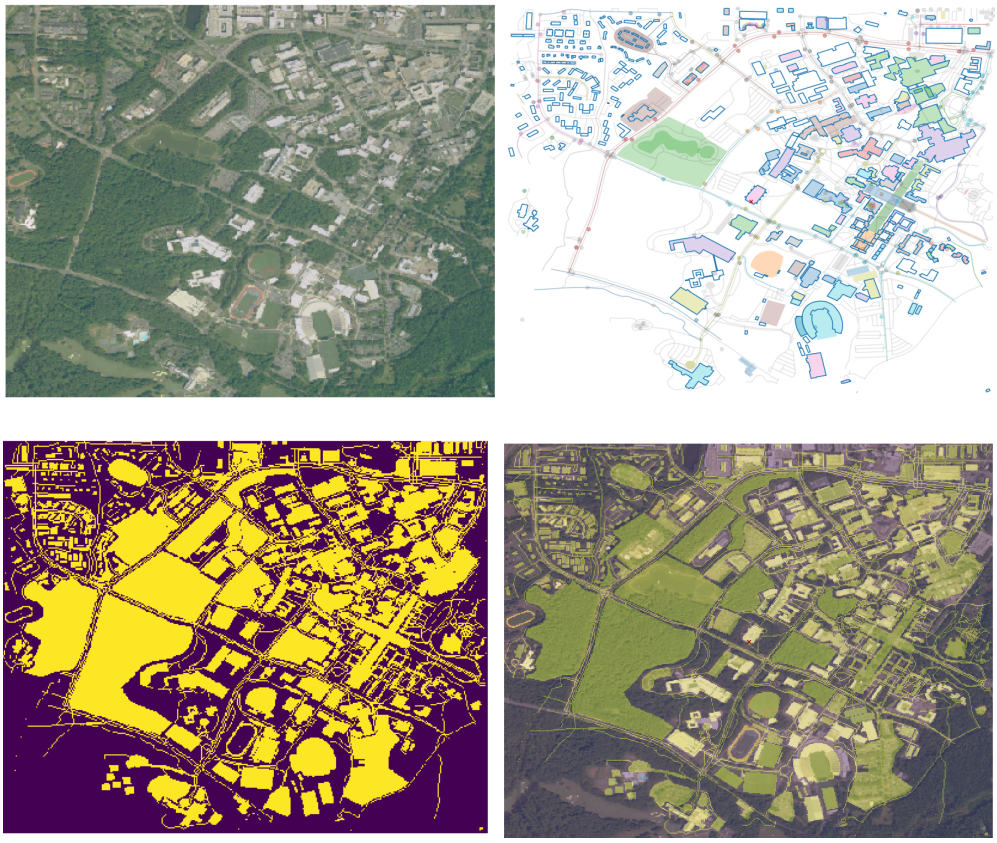

Figure 7. OSM labels and satellite imagery around Gross Hall.
Top left: Satellite image. Top right: OSM labels and Geometries.
Bottom left: OSM labels rasterized and transformed to match the viewing angle of the satellite.
Bottom right: Transformed labels superimposed on the satellite image.

Image Segmentation Experiments

In order to measure the quality of our collected dataset, we executed a series of image segmentation experiments on the GEODOME dataset and analyzed the results. Our assumption is that a high-quality, well-labeled dataset would yield a high-performing model. Our image segmentation model is trained on a dataset that consists of 3-channel 512 by 512 RGB images and their corresponding targets (the final output from our rasterization pipeline). The model learns to assign each of the pixels in the input image to one of the 4 classes present in the dataset (0 for the background class that contains no OSM labels, 1 for *building*, 2 for *parking*, and 3 for a *road*). Our chosen model architecture consists of a ResNet-50 encoder with a U-Net decoder. This architecture was first developed and implemented by Wayne Hu and Bohao Huang at Duke University [14]. This

architecture has a validation intersection over union (IoU) score of 78.78% on the Inria dataset and 79.43% on the DeepGlobe dataset.

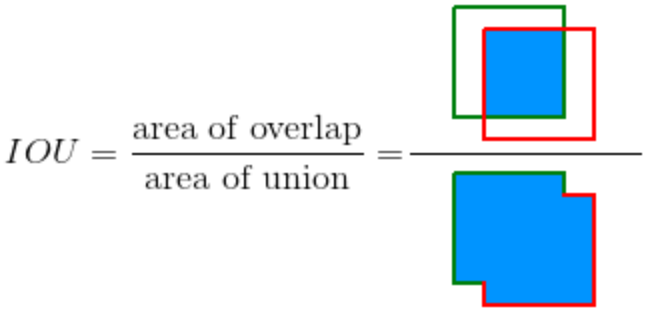

Figure 8. Visual representation of the intersection over union (IoU) metric. The area of the intersection between the raster (green) and the prediction (red) is divided by the union of both. Source: https://supervise.ly/explore/plugins/precision-and-recall-75278/overview

The IoU score is the principal metric when evaluating an image segmentation model's performance. It is a measure of the similarity between a predicted class bitmap (prediction) and the actual labels bitmap (raster). For each class, it is the ratio of the area of the intersection or overlap between the prediction and the raster, and the area of their union. See Figure 8 for a visual representation of this calculation. An overall IoU is obtained by getting a weighted average for each class according to their incidence in the number of pixels. The overall IoU in itself however is not enough to assess a model's performance. For this purpose, we also obtained class-wise confusion matrices to count the number of pixels that were correctly and incorrectly predicted when compared to the rasters in the training and validation datasets (this is the number of true positives, false positives, false negatives, and true negatives). From here, we also considered the class-wise precision, negative predictive value (NVP), sensitivity, specificity, and accuracy to evaluate how well the model is performing.

The first step for the experiment process was to create a dataset preprocessing pipeline that would transform the output of our data collection and rasterization pipelines into a format that was usable by the deep learning model architecture. We considered each RGB and raster pair in the GEODOME dataset as a pair of "tiles" that our preprocess pipeline transformed into 512 by 512 pixel "patches", being able to produce up to four of them located at each corner of the 550 by 550 tile. The preprocess saves all the patches into the same directory and produces training and a validation list file, with the list of all RGB and raster patches pairs to be used for training and validation according to the specified domain or land cover type.

The initial runs of our experiments highlighted important challenges when collecting data from OSM. We encountered a big class imbalance in the collected dataset as 96.75%

of the pixels in the entire dataset belonged to the background class 0. After manually inspecting many of the RGB-raster pairs, we observed many of the locations did not have all the labels for the buildings and objects visible in the RGB satellite images. Initial models were run using 59 classes, which made it more challenging for the model to differentiate among them. Finally, we sought to increase the thickness of the road labels, which were initially only one pixel thick, and thus the models could not predict them at all. Our latest experiments were able to overcome these challenges by training with a manually selected dataset of 100 RGB-raster pairs that we ensured were fully labeled (every visible object in the RGB image had a corresponding polygon label in the raster).

Experiment Results

Run **	Train IoU	Class 0 Precision	Class 0 Specificity	Road* Precision	Road* Specificity	Building* Precision	Building* Specificity
1	0.947	0.9108	0.0194	-	1.0000	0.2977	1.0000
2	0.676 ↓	0.9138 ↑	0.6847 ↑	-	1.0000	0.4117 ↑	0.8670 ↓
3	0.770 ↑	0.9439 ↑	0.8316 ↑	0.8442 ↑	0.9817 ↓	0.8514 ↑	0.9812 ↑

*The classes used for comparison in runs 1 and 2 for road and building are classes 35 and 47 respectively, while for run 3 we used classes 1 and 3 for road and building respectively.

**All of the metrics displayed in this table should tend to 1 in order to indicate a good performance.

Figure 9. Performance results for experiment runs 1 to 3.

The first run of the model was done with the full training dataset of 7,827 RGB-raster pairs, which yielded a total of 15,654 patch pairs after preprocessing, trained for 50 epochs. The results of all model runs can be observed in Figure 9. Although the training IoU appeared to be high, it was solely due to the class imbalance described previously where 96.75% of the pixels in the rasters belonged to class 0. When analyzing the class-specific metrics, we can see that the class 0 specificity, which measures the proportion of actual pixels belonging to a class other than 0 that were correctly identified as such (true negatives / actual negatives), is extremely low. The road precision could not be computed because the model was not making any predictions of this class. Finally, the building precision signals that the proportion of predicted positives that were actual positives in the rasters was really low too. These class-wise metrics indicated that the model, and thus the GEODOME dataset too, was flawed. Some prediction examples

showed this to be true, as the model was predicting mostly zeros. See Figure 10 for a prediction sample of the experiment run 1.

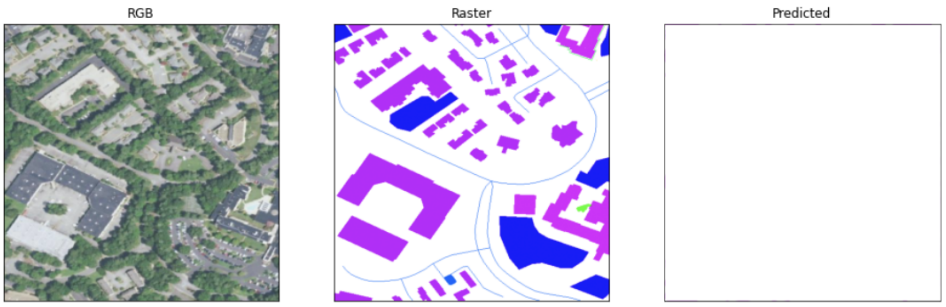

Figure 10. Experiment 1 RGB, raster, and prediction samples.

For the second run, we trained the model with the handpicked 100 patches of well-labelled RGB-raster pairs to see how much the performance would improve just by improving the quality of the data. This time we trained the model for 100 epochs and obtained the results shown in Figure 9. We observe that although the training IoU decreased, every problematic class metric from the first run showed significant improvement. The class 0 specificity went up to 0.6847 (a 3,429% improvement), while the building precision went up to 0.4117 (a 38% improvement). This meant that not all of the pixels were being predicted to class 0 anymore. However, the quality of the data alone did not fix all of the challenges. As we can observe in Figure 11, the predictions are still far from the actual rasters, and the road class is still not being learned by the model.

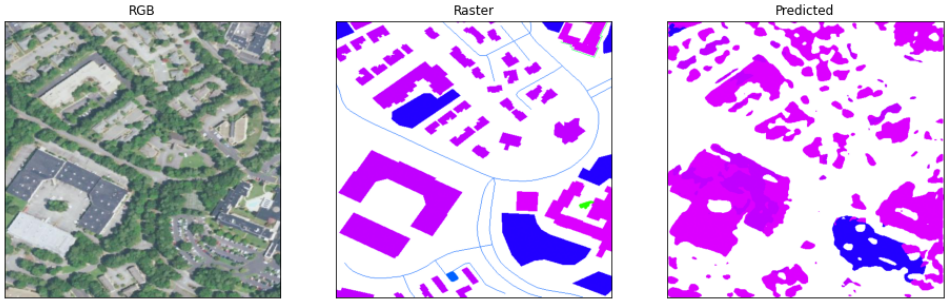

Figure 11. Experiment 2 RGB, raster, and prediction samples.

For the third run, we implemented more changes. First, we reduced the number of classes from 59 to 4, as described previously. This class reduction made it possible for the model to learn to distinguish between them and abstract the appropriate features from the

RGB images to predict them better. Additionally, we implemented a change in the rasterization pipeline to increase the thickness of the road class in the rasters so that it would be thicker and would closely resemble the average thickness of a 3-lane US road of ~11.11m (configurable value). When comparing the rasters from Figure 11 and Figure 12, one can see the difference in road thickness before and after the change was implemented. It can be observed in Figure 12 that effectively all of the problematic class metrics showed an important improvement, and all of them are above 0.8316 now. These experiments demonstrated that the measures we took help improve the quality of the GEODOME dataset, so we can take them into account when collecting data iteratively again.

Figure 12. Experiment 3 RGB, raster, and prediction samples.

Conclusion

In this work, we documented the process of researching, designing and implementing an end to end pipeline that generates a training dataset on EO data, specifically designed for scalability. The set of tools produced by the project are well documented, and designed for adaptation on user specific purposes. These tools include:

- A command-line interface to download large-scale aerial imagery from GEE given a list of coordinates and an area size. The tool currently retrieves data from the NAIP satellite, but is easily configurable for other satellites such as LANSAT and SENTINEL.
- A command-line interface to acquire and OSM data and structure it as valid geoJSON.
- A command-line interface that transforms OSM labels into bitmaps to be used as the annotations of the objects of interest. The labels of interest can be easily configured by editing a simple csv label mapper.
- A web-based quality assurance tool to help in validating the truthiness of the rasterized labels without having to download the dataset locally.

One major challenge we faced with the OSM data, is that even though it is suitable for capturing large sets of labels based on geographic coordinates, it also introduces

label-noise because of inconsistent object labeling or areas where object annotations are incomplete. So a human validation is still required in this iterative process.

Limitations and Future Work

Open crowded source labeled platforms, such as OSM, do not enforce rigid annotation rules. This situation leads to multiple overlapping in the tagging categories. Also, some label categories could be treated as outliers because the infrastructure's tags are only used in some regions. In order to generate a training dataset of EO data that can be globally applied, significant effort was required to ensure a minimum consistency level across tagging schemes. For GEODOME, we iterated over the general set of labels that OSM provides to develop a compact hierarchical taxonomy of labels. Labels for objects of interest can be aggregated into higher label categories (i.e. shop & offices labels can be sub-categories for general buildings)

We have already tried out a couple different automatic approaches to ensure the quality of the labels in the rasters and filter out the RGB-raster pairs that were not fully labelled. For the first approach, we filtered out all the rasters containing less than a specified number of label classes, and for the second one, we trained a classifier that took the mean and standard deviation for each of the RGB channels and the raster as features to try to determine if the pair was well labelled or not. To train the classifiers, we handpicked 100 fully and correctly labelled patches that were marked as "well labelled", and mixed them with another 200 incorrectly labelled patches that were marked as "poorly labelled". To further improve the quality of the dataset, we can use the QA web tool to filter out the poorly labelled images even further.

An important takeaway from our experiments is that utilizing the land cover types as domains is not feasible because of the inconsistency and unreliability of the OSM tags in most land cover type regions other than the developed type. After analyzing the causes for our initial poor results, and how a large percentage of our rasters contained mostly class 0 pixels, we identified that the "well labelled" data came mostly from the developed type. In fact, when running the filters described above, filtering for 2 or more classes returned RGB-raster pairs of which 62% belonged to the developed type; filtering for 3 or more returned 74%. In order to run the domain adaptation experiments effectively, one must choose an alternative definition of domains when using the OSM data to annotate the satellite imagery. One alternative we looked into is the use of geographical coordinates to define clusters of locations that are physically close to each other. These clusters could then act as the domains for the domain adaptation experiments.

After choosing a domain designation, a training/validation split of the dataset can be done using the preprocess pipeline that we have created for this project. A first step would include running the image segmentation experiments with training and validation

data from the same domain, for each of the domains (see the Within Domain case in Figure 13). Then, for each domain one would create a validation split while all of the data in the training set would come from a different domain to measure the cross-domain performance compared to the baseline and measure the drop in performance when the validation domain is not seen when training the model (see the Cross Domain case in Figure 13).

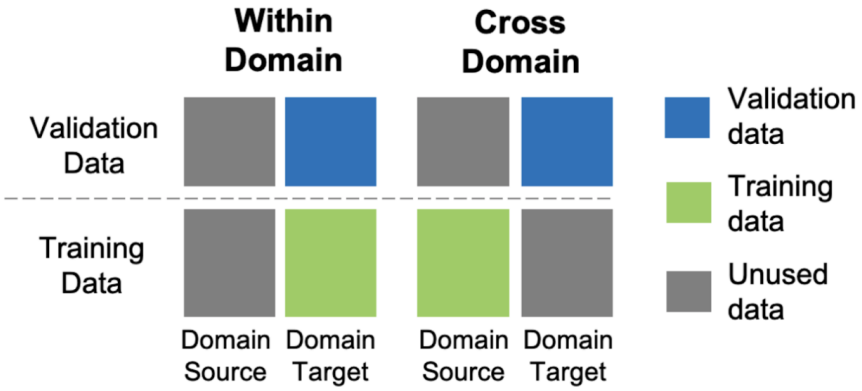

Figure 13. Domain adaptation experiments design; training/validation splits.

References

[1] Martin Sudmanns et al., “Big Earth Data: Disruptive Changes in Earth Observation Data Management and Analysis?,” International Journal of Digital Earth 13, no. 7 (July 2, 2020): 832–50, https://doi.org/10.1080/17538947.2019.1585976.

[2] Amy Leibrand et al., “Using Earth Observations to Help Developing Countries Improve Access to Reliable, Sustainable, and Modern Energy,” Frontiers in Environmental Science 7 (2019), https://doi.org/10.3389/fenvs.2019.00123.

[3] “Earth Observation Big Data for Climate Change Research | Elsevier Enhanced Reader,” accessed September 26, 2020, https://doi.org/10.1016/j.accre.2015.09.007.

[4] Lyn D Wigbels, “Using Earth Observation Data to Improve Health in the United States,” n.d., 34.

[5] Aiswarya Munappy et al., “Data Management Challenges for Deep Learning,” in 2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA) (2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), Kallithea-Chalkidiki, Greece: IEEE, 2019), 140–47, https://doi.org/10.1109/SEAA.2019.00030.

[6] Nadir Bengana and Janne Heikkilä, “Improving Land Cover Segmentation across Satellites Using Domain Adaptation,” ArXiv:1912.05000 [Cs], April 1, 2020, http://arxiv.org/abs/1912.05000.

[7] Adam Van Etten, Dave Lindenbaum, and Todd M. Bacastow, “SpaceNet: A Remote Sensing Dataset and Challenge Series,” ArXiv:1807.01232 [Cs], July 14, 2019, http://arxiv.org/abs/1807.01232.

[8] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat and Pierre Alliez. “Can Semantic Labeling Methods Generalize to Any City? The Inria Aerial Image Labeling Benchmark”. IEEE International Geoscience and Remote Sensing Symposium (IGARSS). 2017.

[9] Demir et al. DeepGlobe 2018: A Challenge to Parse the Earth through Satellite Images arXiv:1805.06561

[10] R. Wang et al., “The Poor Generalization of Deep Convolutional Networks to Aerial Imagery from New Geographic Locations: An Empirical Study with Solar Array Detection,” in 2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), 2017, 1–8, https://doi.org/10.1109/AIPR.2017.8457960.

[11] Fanjie Kong et al., “The Synthinel-1 Dataset: A Collection of High Resolution Synthetic Overhead Imagery for Building Segmentation,” 2020 IEEE Winter Conference on

Applications of Computer Vision (WACV), March 2020, 1803–12, https://doi.org/10.1109/wacv45572.2020.9093339.

[12] Bradbury, Kyle; Brigman, Benjamin; Collins, Leslie; Johnson, Timothy; Lin, Sebastian; Newell, Richard; et al. (2016): Aerial imagery object identification dataset for building and road detection, and building height estimation. figshare. Collection. https://doi.org/10.6084/m9.figshare.c.3290519.v1

[13] Hamid, Raffay & O'Hara, Stephen & Gueguen, Lionel. (2015). Towards Automatically Generating Object-Aware Maps Of Earth Using Satellite Imagery.

[14] Huang, Bohao; Hu, Wayne (2020): Models for Remote Sensing. https://github.com/bohaohuang/mrs

[15] OpenStreetMap Wiki contributors. Tags [Internet]. OpenStreetMap Wiki, ; 2020 Jun 21, 16:35 UTC [cited 2021 Apr 23]. Available from: https://wiki.openstreetmap.org/w/index.php?title=Tags&oldid=2002791.