Automated Global-Scale Detection and Characterization of Anthropogenic Activity using Multi-Source Satellite-Based Remote Sensing Imagery

Hirsh R. Goldberg[a], Christopher R. Ratto[a], Amit Banerjee[a], Michael T. Kelbaugh[a], Mark Giglio[b], and Eric F. Vermote[c]

[a]The Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA

[b]The MITRE Corporation, McLean, VA, USA

[c]NASA Goddard Space Flight Center, Greenbelt, MD, USA

ABSTRACT

Satellite-based remote sensing imagery is an effective means for detecting objects and structures in support of many applications. However, detecting the spatial and temporal bounds of a specific *activity* in satellite imagery is inherently more complex and research in this area is nascent. One reason for this is that describing an activity implies defining both spatial and temporal bounds and while activity is inherently continuous in nature, the geospatial (imagery) time series for any particular swath of ground provided by satellite imagery is relatively sparse and discrete in comparison. The IARPA Space-Based Machine Automated Recognition Technique (SMART)[1] program is the first large-scale research program to target advancing the state of the art for automatically detecting, characterizing, and monitoring large-scale anthropogenic activity in global, multi-spectral satellite imagery. The program has two primary research objectives: 1) the "harmonization" of multiple imagery sources and 2) automated reasoning at scale to detect, characterize, and monitor activities of interest. This paper provides details on the goals, dataset, metrics, and lessons learned of the IARPA SMART program. By releasing the annotated dataset, the program aims to foster additional research in this area by the community at large.

1. INTRODUCTION

Satellite-based remote sensing imagery has been widely demonstrated as an effective means for detecting objects and structures of interest (e.g. buildings, roads, vehicles, vegetation, etc.) as well as identifying land coverage types for a variety of applications including (but not limited to) defense, security, public health, disaster response, commercial, and environmental. However, detecting, classifying, and persistently monitoring *activities* using only temporally discrete satellite-based imagery sources is a more challenging problem, especially at global scales. This is true for many reasons, including:

1. Describing an activity implies defining both spatial and temporal bounds. While activity is inherently temporally continuous in nature, the geospatial (imagery) time series for any particular swath of ground provided by satellite imagery is often relatively temporally sparse and irregularly sampled by comparison, especially at high spatial resolutions;

2. Activities have time-varying spatial and spectral signatures that often require the development of new 3D spatio-temporal models to characterize these events. This is an inherently more complex problem than detecting and classifying objects that rely on static spatial and spectral signatures;

3. Monitoring the progress of activity over time requires spatio-temporal reasoning often outside the scope and capabilities of simple bi-temporal change detection methods.

Geospatial Informatics XIII, edited by Kannappan Palaniappan, Gunasekaran Seetharaman, Joshua D. Harguess, Proc. of SPIE Vol. 12525, 1252502 · © 2023 SPIE
0277-786X · doi: 10.1117/12.2663071

Furthermore, traditional geospatial change detection methods typically only operate on a single sensor source[2,3] and cannot easily be adapted to alternate sensor types, which often have differing spatial, temporal, and spectral characteristics.

As part of a nascent area of research, the IARPA Space-Based Machine Automated Recognition Technique (SMART)[1] program was one of the first large-scale research program to target advancing the state of the art for automatically detecting, characterizing, and monitoring large-scale anthropogenic activity in global scale, multi-source, heterogeneous satellite imagery. The program was created with a goal of leveraging and advancing the latest techniques in artificial intelligence (AI), computer vision (CV), and machine learning (ML) and apply them to geospatial applications.

To support the development of such AI/CV/ML algorithms, four multispectral satellite-based imagery sources are used, each with its own unique spectral, spatial, and temporal (revisit rate) properties: Landsat-8, Sentinel-2, multiple sensors in the WorldView constellation, and Planet Labs imagery. Section 3.5 contains more information on the imagery sources used by the SMART program. The program has focused on two primary research objectives, also referred to as Technical Areas (TAs): 1) Spatially, spectrally, and temporally "harmonize" images from the various imagery sources and 2) Develop algorithms that can perform automated reasoning at scale to detect, characterize, and monitor activities of interest. We refer to these as TA-1 and TA-2, respectively, for the remainder of this paper.

In Phase 1 of the program, the specific activity of interest was heavy construction. The ultimate goal of the program, however, was to design and build the infrastructure, algorithms, evaluation methodologies, and processes that can be readily applied to any geospatial application while being agnostic to the satellite-based imagery source and spatio-temporal activity of interest. Additional program goals include developing innovative large-scale data storage and retrieval techniques, designing and implementing computationally efficient algorithms, and building a cloud-based compute infrastructure to facilitate the primary research objectives. A custom dataset was curated to support algorithm development and quantitative evaluation. The nature of the dataset and problem formulation required both a semi-custom annotation paradigm and quantitative evaluation metrics, which presented unique challenges for data curation, algorithm evaluation, and performance analysis. This paper provides details on the goals of the IARPA SMART program and describes the nomenclature (Section 2), the formulation as an AI/CV/ML problem (Section 3), evaluation metrics (Section 4), and the dataset (Section 5) developed to support algorithm development and evaluation for this application. Section 6 discusses details of the compute framework developed for SMART and Section 7 provides lessons learned on unique challenges encountered during data curation and algorithm evaluation efforts. Finally, Section 8 provides current and proposed next steps beyond this initial dataset and research.

1.1 Related Efforts

There is a growing interest amongst consumers of Earth-observing satellite imagery in utilizing global-scale data for more than traditional detection and classification of objects, land cover identification, and observing and measuring change on the earth's surface over time. These capabilities are now relatively ubiquitous. To foster development and advancement of capabilities, several large-scale public challenges and benchmark datasets focusing on static object detection, object classification, and land cover semantic change classification from geospatial satellite imagery have been released and executed. These include SpaceNet,[4] DeepGlobe,[5] xView,[6] Urban 3D Challenge,[7] SEN12MS,[8] and OpenEarthMap[9] among others. The advanced geospatial detection capabilities enabled by these works (and others) have been remarkable; however, geospatial activity monitoring at scale, which inherently adds a temporal component, remains quite challenging.

To address this, other geospatial efforts (and accompanying datasets) have added temporal components to the respective challenges. The 2018 Functional Map of the World (fMoW)[10] dataset used a multi-temporal sequence of high resolution satellite imagery and metadata to predict the functional use of buildings and land use. DynamicEarthNet[11] is a multispectral temporal satellite dataset for semantic change segmentation and monitoring the evolution of land cover and land use. The Multi-Temporal Urban Development Dataset (SpaceNet 7)[3] aimed to detect and track individual building construction using non-video time series satellite imagery. Finally, the QFabric[12] dataset was introduced at the 2021 CVPR EarthVision Workshop with a goal of enabling the development and evaluation of algorithms to classify and track changes due to anthropogenic activities from satellite imagery. The motivation of the QFabric dataset is perhaps most closely aligned with the goals of the IARPA SMART program. However, QFabric's problem formulation only seeks to operate on a single, high spatial resolution imagery source (e.g. WorldView).

Previous efforts have undoubtedly been beneficial but most focus on a single sensing modality. Many of these efforts only make use of very high resolution imagery, which makes the problem much easier. Additionally, many do not take advantage of the diverse spectral information available across different satellite-based imaging sources. Finally, the previous efforts that do focus on activity detection using satellite imagery are typically limited to between 2 and 5 observations to describe each event[12, 13] or focus on pixel-level change detection[11] for land use classification applications.

1.2 SMART Contributions

Analysis of time-series satellite data using heterogeneous imagery sources creates challenges that many change detection algorithms tend to struggle with, especially on a global scale. These include registration artifacts, varying spatial and spectral characteristics, parallax caused by different off-nadir imaging capabilities, and environmental effects such as cloud coverage, shadows, and seasonal changes. The SMART program seeks to build upon previous efforts and advance the state of the art in this field by developing AI algorithms that can operate on multi-source, heterogeneous geospatial imagery to help address many of these challenges. The effort aims to augment existing geospatial datasets and research by:

- Exploiting multi-source satellite imagery with varying spatial, spectral, and temporal resolution in the same processing pipeline, leveraging and exploiting the unique capabilities of different sensing modalities;
- Applying the latest advances in AI/CV/ML - especially those focused on temporal activity detection from ground-based, high frame rate video imagery - to spatio-temporal geospatial *activity* detection and monitoring. We emphasize that *activity detection* presents distinct challenges from object detection and scene/land cover classification using satellite imagery. This includes addressing key issues such as multi-image registration and effects caused by environmental factors;
- Building or augmenting the infrastructure required for efficient and large-scale data storage and retrieval.

To accomplish these objectives, the SMART problem formulation requires algorithms to:

- Account for differences in atmospheric effects from one region of the world to another, spatial resolution differences of different imagery sources, and a wide variety sensor-specific characteristics;
- Filter out spurious environmental changes (which are ubiquitous) to identify the spatial and temporal extents of anthropogenic activity, not simply provide semantic change detection results;
- Identify the temporal state of each activity at any given time;
- Use spatio-temporal reasoning based on past and current observations to predict when each activity will end.

As part of this effort, researchers aim to answer the following key questions, among others:

- What imagery sources and characteristics are required for this task? For example, are high resolution images (better than 1 m) required?
- How critical is spectral information in locating and monitoring activities?
- What temporal revisit rate is required?

These questions, among others, helped drive additional research and may also influence future decisions regarding satellite-based imaging capabilities.

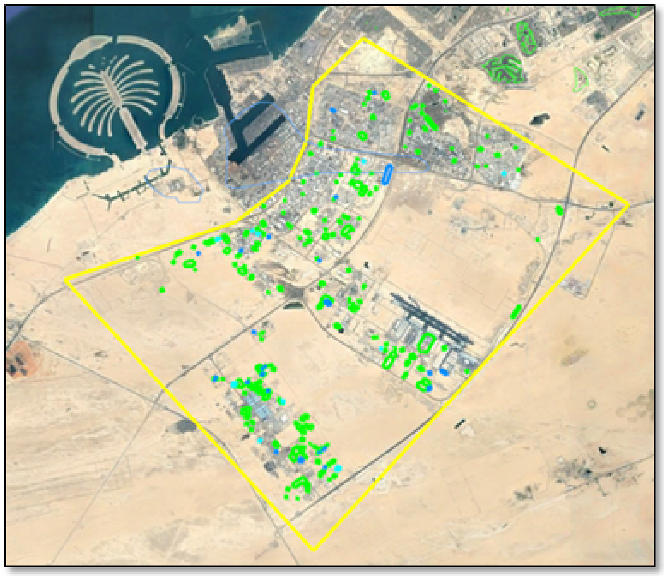


Figure 1: Example of a region of interest in the vicinity of Dubai, UAE. The region (yellow boundary) contains numerous heavy construction activities (outlined in green and blue). This region is roughly 400 km^2 in size and contains over 250 activities of interest from 2014 to 2021. (Map data: Google, Maxar Technologies)

2. RELEVANT NOMENCLATURE

The following nomenclature is used throughout this paper and in conjunction with the annotation dataset to describe the SMART application, dataset, problem formulation, and quantitative performance metrics.

Region A large geographical area defining the spatial and temporal bounds in which ground truth annotations are provided and algorithms are expected to process. In Figure 1, a sample region is identified by the large yellow polygon. Region boundaries are static across a given temporal duration and may contain any number of sites, including none. Also referred to as a region of interest (ROI).

Site A geographical area, much smaller than a region, defining the spatial and temporal bounds of large-scale change (anthropogenic or not). This is the fundamental unit of activity that SMART is focused on; algorithms are expected to detect, classify, and monitor these sites within defined regions. For the purposes of the SMART Heavy Construction dataset, sites of interest are larger than 8,000 m^2. Note that this size is in reference to the entire site area containing all change associated with that activity, not any particular object within the site boundary (e.g. a building). There can be any number of sites within a Region (including none). In Figure 1, the boundaries of sites are given by the green and blue polygons.

Object Anthropogenic structures or other natural items that exist and are visually discernible within the spatio-temporal bounds of a site that are either the reason for the construction activity (e.g. the buildings and supporting structures) or are supporting the construction activity (e.g. the construction equipment). Examples of objects include, but are not limited to, buildings, runways, roads/bridges, shipping piers/docks, and construction equipment. Annotation, localization, and classification of individual objects is not within the scope of the SMART Phase 1 Heavy Construction problem formulation. However, these objects are the basis on which sites are identified and labeled.

Observation A single image or set of images capturing the scene of interest at a discrete point in time. For the purposes of this dataset, an observation is defined to have a temporal span of a single day.

Activity A sequence of discrete states that, when considered as a whole, describe an event. The SMART Phase 1 activity of choice is anthropogenic heavy construction. Future phases of the program may consider other kinds of activity.

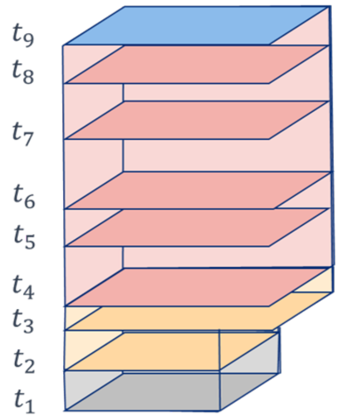


Figure 2: Illustration of a site model from time steps t_1 to t_9. Polygons represent arbitrary site boundaries at each time slice and colors represent activity phases defined in Section 3.3. Note that the site boundary may (and often does) change size from one observation to the next as shown between t_2 and t_3. This happens as the spatial extent of the activity expands over time.

Site Model A collection of polygons, each with an associated activity phase (described in Section 3.3), that defines the spatial and temporal bounds of all activity within a site for the duration of that activity. Figure 2 shows a simplified graphical representation of a simple site model with polygons representing the site boundaries at time steps t_1 through t_9. The colors represent different activity classification phases as defined in Table 1 and Section 3.3. Note that the size of the site boundary may change over time as shown between time steps t_2 and t_3. This represents the simplest case where each observation contains a single polygon/activity phase pair at each observation.

Region Model A collection of site models contained within a defined region. The region model defines both spatial and temporal bounds or the region to be processed or in which annotations are provided.

Annotations The set of site models that are used as ground truth for the purposes of algorithm training, validation, and testing.

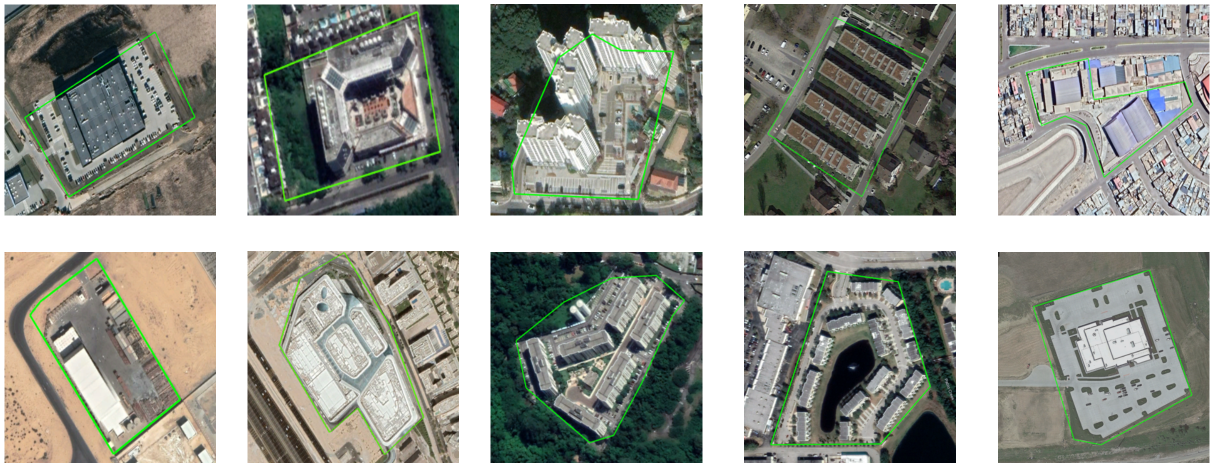
Datacube A collection of images from one or more sensor platforms covering the same spatial area of interest. The set of images will span some time period T = $[T_{start}, T_{end}]$. Often used to describe all images needed for a single site, a group of sites, or an entire region.

3. PROBLEM FORMULATION

3.1 Positive Classes

Visual examples of positive activity types can be found in Figure 3. Examples of large-scale heavy construction activity types in the positive class include the following:

- **Industrial**: Factories, power plants, manufacturing facilities, warehouses, distribution centers, shipping infrastructure (shipping ports, industrial piers), etc;
- **Commercial**: Malls, grocery stores, strip malls, gas stations, hospitals, stadiums, arenas, office buildings, hotels, storage units, etc.
- **Heavy Residential**: Large apartments or condominium highirise buildings that are five stories or more;
- **Medium Residential**: Low-rise apartments/condominium buildings, townhouse/row homes typically with four stories or fewer;
- **Miscellaneous**: Heavy construction that does not fall into one of the categories above, including schools, parking garages, religious buildings, power substations, fire stations, etc.



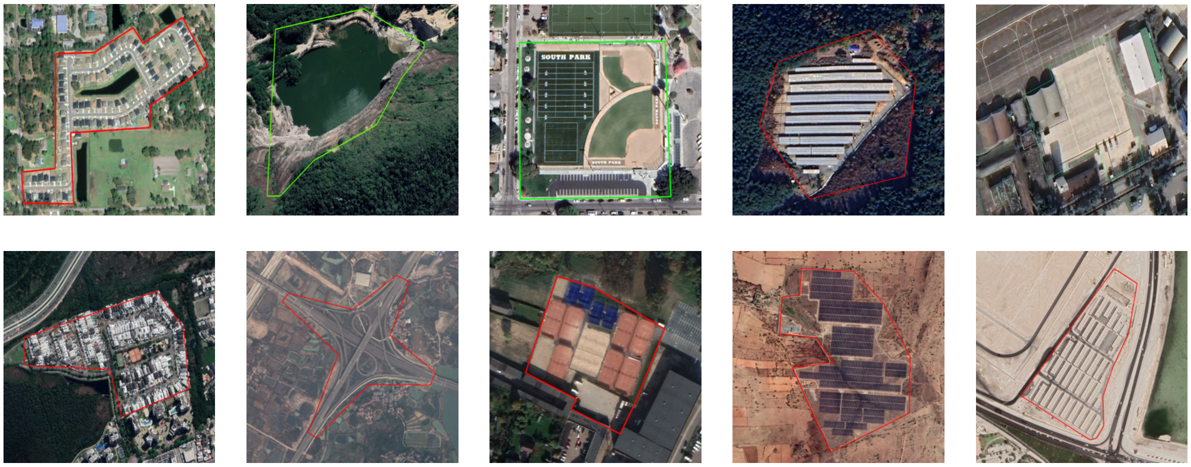
Industrial Commercial Heavy Residential Medium Residential Other

Figure 3: Examples of heavy construction activity types (Positive Class). In the right-most column ('Other'), the top image is a school and the bottom image is a church. (Map data: Google, Maxar Technologies)

3.2 Negative Classes

Types of large-scale change that are not considered to be of interest for this problem are included in the negative class set. Visual examples of negative activity types can be found in Figure 4. Negative classes include the following:

- **Light residential**: e.g. A collection of one or more detached single family homes, including large neighborhoods where the only activity is the construction of single family or detached homes;
- **Sports fields** not associated with the construction of a large building other than the (often small) clubhouse associated with the sports complex;
- **Golf courses;**
- **Standalone surface parking lots** not associated with the construction of at least one other 'Positive' example (e.g. a park and ride, an airport parking lot);
- **Standalone road infrastructure** not otherwise associated with the construction of buildings including interchanges, bridges, tunnels, etc;
- Change that is the result of a **natural disaster** (e.g. tornado, hurricane, flood, etc.) which tends to take place over the course of a handful of days;
- The general **clearing of land or destruction** of a building/structure without the explicit or immediate purpose of continued construction (e.g. the destruction of a stadium or factory on land that is then abandoned) or left as greenspace;
- The creation of **water retention ponds**, unless directly associated with the construction of a 'Positive' example;
- Change associated with the **resurfacing** of a parking lot or building roof;
- The installation of **solar panel "fields".**



Light Residential Infrastructure Recreation Area Solar Panel Field Parking Lot

Figure 4: Examples of large scale change that is not considered a heavy construction activity type (Negative Class). (Map data: Google, Maxar Technologies)

3.3 Activity Phase Categories

The SMART Heavy Construction dataset uses four classes of activity phases to characterize heavy construction activity: *No activity*, *site preparation*, *active construction*, and *post construction*. Algorithms are expected to classify each observation of part or all of an identified site as one of the latter three phase categories; the *No Activity* phase represents a 'null' class for the purposes of algorithm development, as algorithms are not required to classify observations in that phases. We also define an 'Unknown' label to represent observations in which the exact activity phase cannot be reliably determined from the available imagery or other relevant external information sources. Table 1 provides information about these four categories. Figure 5 shows examples of different sites in each phase.

3.4 Site Boundaries

The SMART problem formulation revolves around identifying and monitoring the spatial and temporal extents of *all change* associated with heavy construction activities, not simply localizing the buildings or other infrastructure being constructed; this includes all land used to support the construction activity from the beginning of the activity to the end. Because the concept of a 'site' is more nebulous than a clearly defined object (e.g. a building, a vehicle, a road, etc.), we must clearly define the concept of a site boundary in order to describe the spatial bounds of a site and support the desired system functionality. In addition, to account for the fact that different areas of a single site might be in different phases of activity in any given observation, we also introduced the concept of a *subsite*. Our goal was to create a rule set that was explainable, repeatable, and meaningful. The specific definitions of site and subsite boundaries are outside the scope of this publication but are provided alongside the dataset.[14]

3.5 Imagery Sources

The SMART problem formulation incorporates three primary satellite-based imagery sources: Landsat 8, Sentinel 2, and the Maxar WorldView constellation (specifically WorldView-1, WorldView-2, and WorldView-3). These three sources were selected for their complementary set of spatio-temporal and spectral characteristics. A fourth imagery source, Planet Labs, was added later in the program as a supplemental source to fill gaps in the spatio-temporal trade space. Table 2 describes the four data sources used by the SMART program. Other sources may be added in the future to encourage the development of systems that are agnostic to data source, or that can at least be easily augmented to operate on any future data source.

Phase	**Description**
No Activity	• Status quo for any given site. What the scene looks like before activity started. • Default class for un-annotated areas before any activity begins. • An algorithm is not expected to detect this phase or anticipate transitions out of this phase. Observations of sites in this phase are included in site models for algorithm training purposes only.
Site Preparation	• Includes activity such as ground clearing, ground shaping, and other activity related to preparing the site for construction. • Earliest possible annotated activity phase class for a site.
Active Construction	• This phase defines when objects in the scene are actively being built, including building foundations and supporting infrastructure. • Includes activity such as pre-foundation and foundation building, transient construction, building of intermediate structures, etc.
Post Construction	• Phase in which apparent completion of all construction activity within the site or sub-site bounds has occurred. • At least one view of post-construction is required to be annotated for each site model, assuming the construction has been completed within the temporal bounds of the datacube. More may be provided if available.
Unknown	• No image or external information source is available to reliably classify that observation as being in a specific phase. • Will always occur in the temporal gap between phase transitions (e.g. between 'Site Prep' and 'Active Construction')

Table 1: Activity phase classification categories used in the SMART Heavy Construction dataset.

Source	**Nominal Spatial Resolution (m)**	**Nominal Temporal Resolution**	**Spectral Bands**	**Notes**
Landsat 8[15]	30 (spectral) 15 (panchromatic)	15 days	8 spectral 1 panchromatic	
Sentinel 2[8,16]	10 (RGB and NIR) 20 (spectral) 60 (spectral)	5-10 days	4 (RGB/NIR) 9 (VNIR/SWIR)	
WorldView 1[17]	0.5	1-6 months	1 (panchromatic)	
WorldView 2[18]	0.5 (panchromatic) 1.8 (VNIR)	1-6 months	1 (panchromatic) 8 (VNIR)	
WorldView 3[19]	0.3 (panchromatic) 1.24 (VNIR) 3.7 (SWIR)	1-6 months	1 (panchromatic) 8 (VNIR) 8 (SWIR)	Not annotated but available for algorithm development and testing
Planet Imagery[20]	3-5	1 day	4 (RBG and NIR)	Not annotated but available for algorithm development and testing

Table 2: Imagery sources used for the SMART program with relevant information about each source.

4. PERFORMANCE METRICS

Quantitative metrics were identified for both technical areas (TAs) to establish a performance baseline, set near- and long-term goals, and track the progress of algorithms over time. The following subsections describe the metrics for TA-1 (Section 4.1) and TA-2 (Section 4.2).

4.1 TA-1: Data Harmonization

TA-1 algorithms are designed to estimate and normalize (”harmonize”) physical quantities from multi-source, heterogeneous imagery and then format the image sequence in the form of a “datacube”. Corrections to be applied include:

- **Geometric corrections**: Align, co-register, georeference, and orthorectify all imagery;
- **Radiometric corrections:** Estimate surface reflectance by accounting for atmospheric and surface scattering effects;

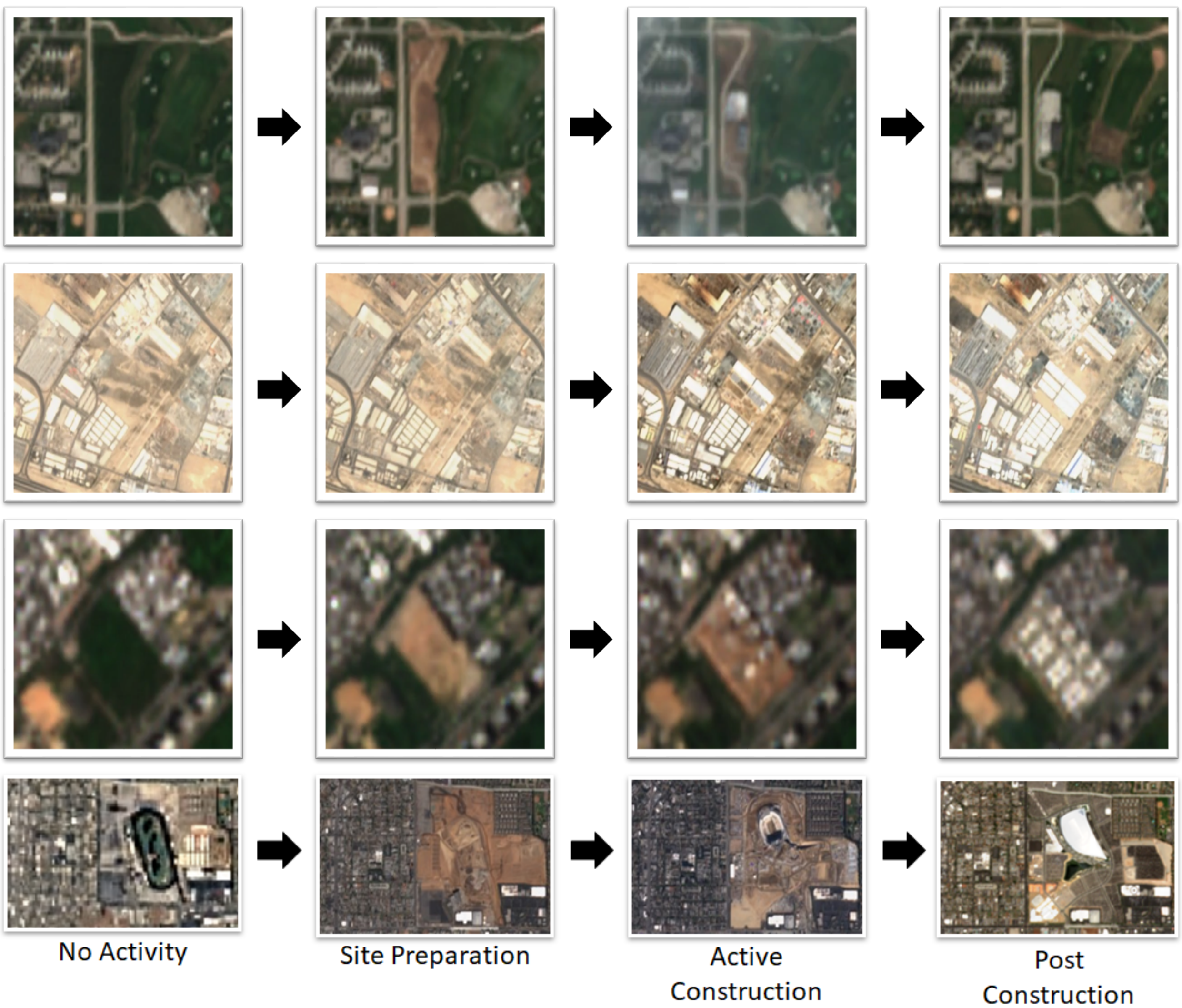


Figure 5: Example sequences showing a construction site transitioning through each of the four phases defined in Table 1. All images shown are from Sentinel-2.

- **Sensor-to-Sensor Harmonization:** Normalize spatial and spectral information across all available sources.

Table 3 summarizes the performance metrics that are used to evaluate TA-1 algorithms. Each of the desired capabilities described above is measured using a separate set of metrics. Each capability also has one or more relative and absolute metrics.

4.2 TA-2: Activity Detection and Characterization

TA-2 algorithms use the TA-1 output product to *detect*, *classify*, and *predict* anthropogenic activity in a region with pre-defined spatial and temporal bounds. Detected activity is formatted using the site model convention described in Section 2, which contains the phase classification and prediction information. The desired capabilities include:

1. **Broad area search (BAS)**: Identify and localize (spatially and temporally) activities in a sequence of images;

2. **Activity classification (AC)**: Label the detected activity at each observation (time step) as one of a specified set of activity phases (See Table 1);

3. **Activity prediction (AP)**: For a given site that is not yet observed in an end-state, predict when the end-state transition will occur.

Table 4 summarizes the performance metrics that are used to evaluate TA-2 algorithms. Each algorithm capability (BAS, AC, and AP) is measured using a separate set of metrics.

4.3 Current Status of Metrics

At the time of this paper's publication, all metrics have been implemented and are used to evaluate TA-1 and TA-2 algorithms using the processing framework described in Section 6. The SMART test and evaluation (T&E) team continues to revise the metric calculations as lessons are learned throughout the program about what implicit assumptions have been made in their formulation. A detailed summary of how the various TA-1 and TA-2 metrics are calculated will be provided in a future publication along with preliminary quantitative and qualitative analyses.

5. DATASET DESCRIPTION

The SMART Heavy Construction dataset consists of site model annotations that support TA-2 algorithm development and quantitative assessment of Broad Area Search (BAS, detection), Activity Classification (AC, phase classification), and Activity Prediction (AP, temporal reasoning). The dataset contains different types of annotations, each described uniquely by certain characteristics. Table 5 provides a description of these site types and each type has a specific utility. For example, while all types can be used for BAS algorithm development and evaluation, only Types 1 and 2 can be used for AC and AP. Type 3 sites are intended to be the negative class (Section 3.2).

The SMART Heavy Construction dataset can be further categorized by the annotation process and the purpose for which annotations were generated. These broad categories, discussed in Table 6, notionally describe a tradeoff between annotation quality and quantity. This table also notes which site types are included in each dataset. The following sections provide additional information on each of these two dataset categories, which we generically refer to here as the *Primary Dataset* and the *Secondary Dataset.* Both datasets, as well as more information about each, can be obtained on JHU/APL's Pubgeo Github page.[14]

Capability	Metric Name	Symbol	Description
Geospatial Corrections	Relative geospatial error (easting)	μ_E, σ_E	Average and standard deviation of easting error
	Relative geospatial error (northing)	μ_N, σ_N	Average and standard deviation of northing error
	Relative circular error	$CE_{90}^{(rel)}$	90% circular error, relative between image features
	Absolute circular error	$CE_{90}^{(rel)}$	90% circular error, relative to ground control points
Radiometric Corrections	Reflectance accuracy, precision, and uncertainty	$A_r(\lambda)$ $P_r(\lambda)$ $U_r(\lambda)$	Statistics of linear regression residuals per pixel, per band
Sensor-to-sensor harmonization (SSH)	Coefficient of variation	$CV(\lambda)$	Relative uncertainty in harmonized surface reflectance per band
	Temporal smoothness index	$TSI(\lambda)$	De-trended moving average of harmonized surface reflectance

Table 3: Summary of TA-1 performance metrics.

Capability	Metric Name	Symbol	Description
Broad area search (BAS)	F1 score for BAS	$F_1^{(BAS)}$	Overall measure of BAS performance, measured as the harmonic mean of precision and recall.
	Probability of detection	P_D	Equivalent to recall, measured as the fraction of known activities detected successfully.
	False alarm rate	FAR	The rate at which false positives are detected per unit area.
	Fractional false positive area	$FFPA$	The fraction of the ROI's negative area that is covered by site model polygons scored as false positives.
Activity classification (AC)	F1 score for AC	$F_1^{(c_i)}$	Harmonic mean of precision and recall for activity phase c_i
	Average temporal error	$TE_{(c_i)}$	Average absolute difference between the true and estimated dates an activity entered phase c_i
Activity prediction (AP)	Average temporal error	$TE_{(end)}$	Average absolute difference between the true and predicted dates an activity ends.

Table 4: Summary of TA-2 performance metrics.

5.1 Primary Dataset

The program initially aimed to produce 1,000 positive (Type 1 and Type 2) and 500 negative (Type 3) site models for the *Primary Dataset*; to date, 1,174 positive and 516 negative sites have been generated. An additional 214 positive sites (Type 4) have been identified and annotated but are provided without phase labels at this time. These sites are still useful for BAS, but cannot be used for AC or AP. As of this publication, the dataset is still being updated and these numbers are expected to change as modifications and corrections are made.

As discussed in Section 7.2, the annotation process for Type 1 and Type 2 sites proved to be cumbersome, time-consuming, and costly. As a result, the number of site model annotations produced in the *Primary Dataset* is relatively limited, especially for supporting many machine learning-based algorithmic approaches which often require a significant number of data for algorithm training. Nevertheless, while the number of annotated *activities* may be relatively low, the number of annotated *observations* is quite large. Table 7 shows the number of annotated observations across all three primary image sources while Figures 6 shows how those observations are distributed between the five phase labels described in Table 1. The dataset contains over 59,000 observations annotated with phase labels across nearly 13,000 unique satellite images. The median number of observations per site model is 35 while the average temporal duration of an activity is roughly two years. Roughly 32,000 annotated phase labels are of the *Active Construction* label, nearly four times greater than all other phase labels which are roughly equivalent in magnitude. This class imbalance, while not ideal for machine learning applications, is representative of the real world prevalence of that activity phase as compared to *Site Preparation*; that is, this class imbalance is expected and is a challenge that must be handled appropriately.

Figure 7 shows three examples of regions and the sites within them. The spatial density of activity within each

Type	No. of Annotated Observations Per Site	Has Phase Labels?	Completed Activity?	Notes
1	Many	Yes	Yes	Positive Sites
2	Many	Yes	No	Positive Sites
3	2 (Start and End)	No	Yes	Negative Sites
4	2 (Start and End)	No	Yes or No	Positive sites pending phase labels

Table 5: Annotation types found in the SMART Heavy Construction annotation dataset

Dataset	No. of Annotated Site Models (All Types)	Relative Quality	Used For	Site Types Included
Primary	Lower	Higher	BAS, AC, and AP	1, 2, 3, 4
Secondary	Much Higher	Lower	BAS Only	4

Table 6: Summary of the two types of annotations in the SMART Heavy Construction dataset

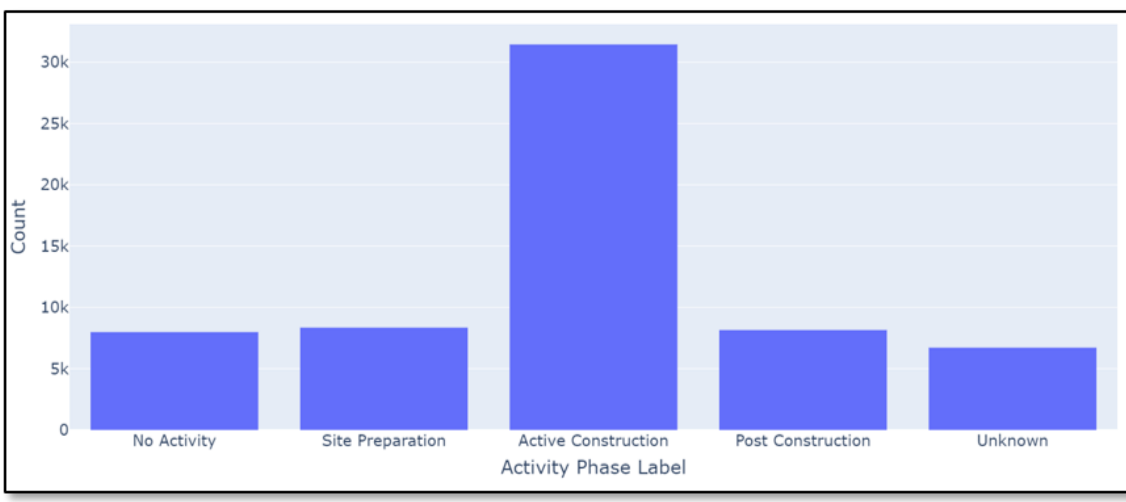
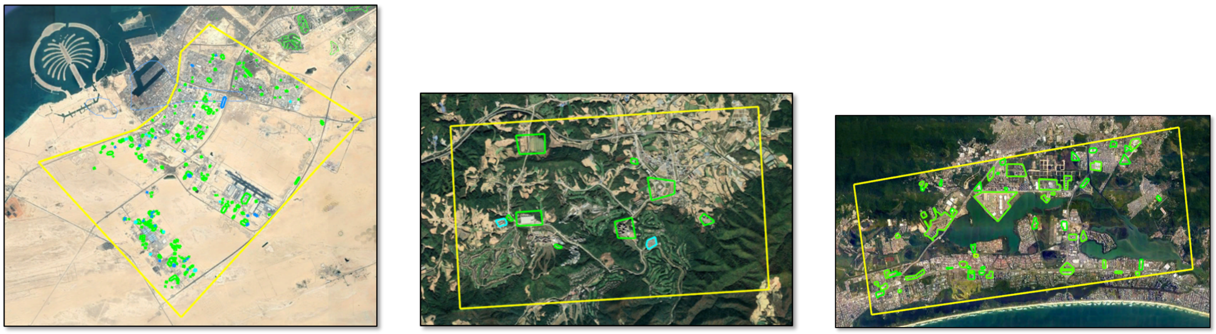


Figure 6: Number of annotated observations by activity phase

Sensor	Observations Annotated	Unique Images Annotated
Landsat 8	12,916	2,199
Sentinel 2	25,473	5,619
WorldView (all)	19,971	5,145
Total	**58,360**	**12,963**

Table 7: Number of observations and unique images annotated in the SMART primary dataset



a) Dubai, UAE b) Pyeongchang, South Korea c) Rio de Janeiro, Brazil

Figure 7: Examples of annotated regions (yellow polygons) showing all heavy construction sites within them (green and blue polygons). Spatial and temporal density of activities (sites) varies by region. (Map data: Google, Maxar Technologies)

region varies, with some regions being extremely dense and others being sparse - both spatially and temporally. Overall, the *Primary Dataset* contains annotations across 25 regions, 17 of which are in the training/validation set while 8 are withheld for independent sequestered test and evaluation by the program's T&E team. Overall, the annotated regions cover more than 4,600 km^2. Figure 8 shows the locations of the regions in which annotations are provided.

All regions include annotated heavy construction activities over more than 7.5 years from at least January

Figure 8: Location of regions annotated in the SMART *Primary Dataset* showing reasonable geographic diversity of the dataset.

1, 2014 through August 31, 2021. In many cases, sites outside of these temporal bounds are also included for algorithm training and validation purposes. However, evaluation is limited to the dates identified above due to increased reliability and availability of sufficient information to support annotation of site boundaries and phase labels. More information about the annotations and the regions in which they are contained can be found with the dataset.

5.1.1 Inter-Annotator Agreement (IAA) Analysis

In order to quantitatively assess the trustworthiness and consistency of the TA-2 *Primary Dataset*, an inter-annotator agreement (IAA) analysis was performed. This analysis is often accomplished by having *several* independent annotators annotate *many* objects and comparing the annotation products generated by each annotator for consistency. However, the complexity of the SMART annotation task as well as program timelines limited this effort. Our approach involved two sets of independent annotators annotating a representative set of 50 site models. Sites were categorized into three categories (Easy, Medium, and Hard) based on the complexity and consistency of the site boundaries and the presence (or not) of subsites. Using those categories, 20 easy, 15 medium, and 15 hard sites were randomly selected for this study.

The analysis aimed to assess the quality of three aspects of the annotations: site boundary (polygon shape, size, and location), accuracy of activity phase labels, and subsite usage. Agreement was measured using Cohen's Kappa coefficient. Figure 9 shows the IAA results for all three annotation characteristics along with a widely accepted qualitative interpretation of the quantitative results. These results indicate that on average, substantial agreement (or better) is seen across all three annotation characteristics. Nevertheless, there is room for improvement with the dataset and updates are warranted to further increase the annotation accuracy. However, the SMART metrics and associated thresholds were explicitly designed to accommodate some of this uncertainty and potential errors identified by this analysis.

5.2 Secondary Dataset

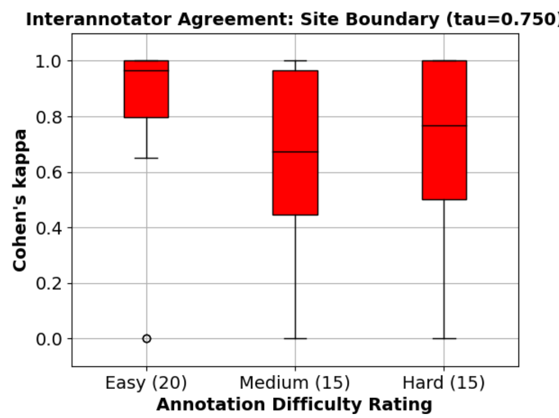
A secondary annotation methodology was created to complement the *Primary Dataset*. The purpose of this dataset, which we refer to here as the *Secondary Dataset*, was to significantly increase the quantity of site model annotations to better facilitate broad area search algorithm development. This methodology produced more than 28,500 labeled sites in 84 regions spanning more than 183,000 sq. km; this represents a significant increase in the number of site models available for algorithm development and testing. Figure 10 shows the locations of all regions annotated in the SMART *Secondary Dataset*. While the number of site models in this dataset is significantly higher, there are no phase labels in this dataset, only activity start and end dates (they are all Type 4 annotations); therefore, the annotations in this dataset cannot be used for evaluating activity classification (AC) or prediction (AP). Furthermore, the higher quantity of sites in this dataset may come at the expense of annotation accuracy and slightly higher label noise compared to the *Primary Dataset*. Nevertheless, this dataset has proven to be quite useful in the development and evaluation of broad area search algorithms. More information about the *Secondary Dataset* can be found alongside the dataset.[14]

6. DATA PROCESSING SYSTEM

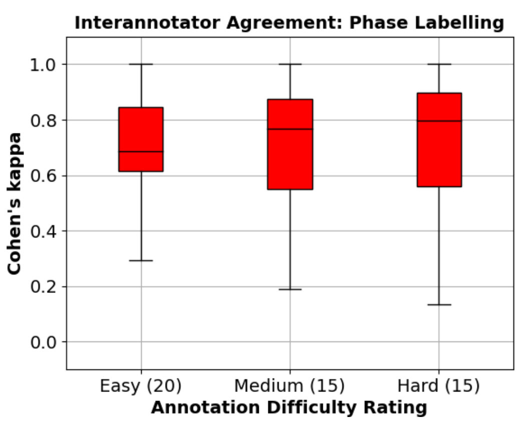
The goals of CV/ML research, especially that involving large amounts of data, often conflict with the need to develop robust software systems and scalable solutions. To this end, the SMART program focused on the development of a software framework in parallel with the algorithm research in Phase 1. Specifically, one of the key program goals in Phase 1 was to build a common framework responsible for connecting algorithms to data and abstracting the management of scalable compute resources from the research teams. This facilitates the use of MLOps in an isolated T&E environment and is expected to also facilitate the transition of research to operations.

Kappa	Agreement
< 0	Less than chance
0.01 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 0.99	Almost perfect

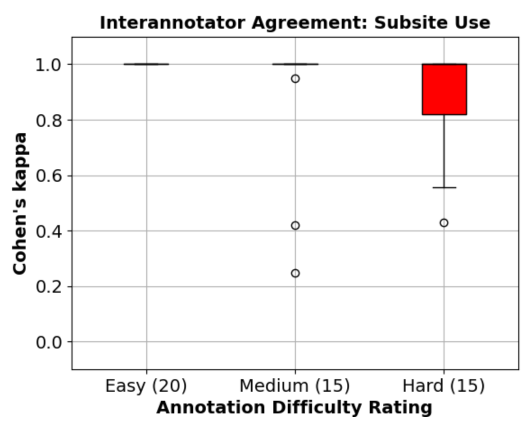
(a) Cohen's Kappa Interpretation



(b) Site Boundaries



(c) Activity Phase



(d) Subsite Usage

Figure 9: Inter-Annotator Agreement (IAA) analysis results for the Heavy Construction *Primary Dataset*. On average, analysis indicates substantial agreement of the dataset, indicating sufficient dataset accuracy.

6.1 Processing Architecture

The processing framework designed for SMART uses Apache Airflow on a Kubernetes back end for task orchestration. The system is deployed on Amazon Web Services (AWS) using CloudFormation and the managed Elastic Kubernetes Service (EKS). The framework includes several common functions, such as satellite image search, geographic region splitting, and recombination. SMART also uses the Spatio-Temporal Asset Catalog (STAC)[21] as a common metadata format and API across multiple satellite platforms to facilitate data search and retrieval; this is essential given the number of large satellite images needed by algorithms for development and testing. All satellite imagery is stored in one of many STAC catalogs. Landsat and Sentinel imagery is found in publicly available catalogs while WorldView and Planet imagery was procured by the program and made available via self-managed STACs. Information on how to access satellite imagery data used for this program can be found alongside the dataset.[14] Algorithms are deployed in one or more Docker[22] containers and linked together using an airflow directed acyclic graph (DAG).

The most significant unsolved challenge is efficient satellite image retrieval and container-to-container storage solutions. The satellite collections are hosted on AWS S3 storage and need to be downloaded to EKS pods. Using instances designed for efficient processing (e.g., GPU instances) to download data decreases computational efficiency and increases cost. Likewise, information can be passed from container to container via S3 or EFS, but this also has cost implications. Future goals include developing a solution that more efficiently pairs data search, retrieval, and storage with processing of that data.

6.2 Processing Workflow

Figure 11 contains a high level representation of the SMART algorithm, test and evaluation, and analytics workflow. Algorithm processing is initiated with an empty *region model* file that contains the spatial and temporal bounds over which the algorithm is expected to process. Algorithms first make STAC queries to search and retrieve all relevant imagery required for processing that region model. After processing is complete, algorithms output one or more *site models* indicating the spatial and temporal extents of each detected activity. These site models are then processed by the SMART evaluation test harness. The test harness compares algorithm outputs with ground truth annotations to provide a quantitative assessment of performance. These metrics results are stored in a database and are made available for viewing and analysis in a custom user interface.

Figure 10: Location of regions annotated in the SMART Heavy Construction *Secondary Dataset* showing increased geographic diversity of the dataset. Markers in blue indicate overlap with *Primary Dataset*.

7. LESSONS LEARNED

Producing a dataset that enables researchers to sufficiently address the problem of anthropogenic activity detection from satellite imagery was challenging. During the process, we identified several issues that will likely present similar challenges for organizations producing similar datasets for AI/ML research, and share them here.

7.1 Complexity of Activity Phase Label Hierarchy

The original goal of the program was to utilize a more complex activity phase hierarchy with more activity phase labels. However, the program discovered that readily available imagery was not suitable for reliable annotation (at scale) of many of the finer grained activity phases in all locations around the world. Consequently, the phase labels as originally envisioned were consolidated into the four broader categories described in Section 3.3. The four categories were selected based on what annotators could reliably determine at scale from the available imagery and also with consideration of what would be most useful to an end-user of the SMART system. Even with this consolidation into fewer categories, preliminary results (to be shared in a forthcoming publication) suggest that Activity Classification and Activity Prediction both remain as challenging problems. Future work could re-expand this hierarchy in order to increase the capability of the SMART system, but we expect obtaining a sufficient number of samples of each sub-phase to be a challenge. This may drive the need for more self-supervised or unsupervised learning approaches.

7.2 Data Annotation Challenges

The annotation task for this effort proved to be uniquely challenging, manually intensive, and quite time-consuming. For most geospatial object detection tasks involving annotation of static objects, many open-source and established utilities exist to aid annotation. The process also often involves a semi-supervised approach where an existing algorithm is used to aid in the initial detection of the objects. There are also utilities available for annotating continuous activities in high frame rate video sequences. However, most existing open source annotation utilities do not natively work with large satellite images at scale, let alone sparse sequences with (potentially) different spatial resolutions from frame to frame. At the outset of the annotation activity the program had neither utilities to easily locate activities of interest nor the capability to track activity phase over time. Even with a semi-custom annotation utility to help ease the annotation burden, the overall effort was still quite labor intensive. Specific challenges presented by the annotation effort include:

- Finding inherently continuous activity in sparse, irregularly-sampled video derived from satellite imagery;

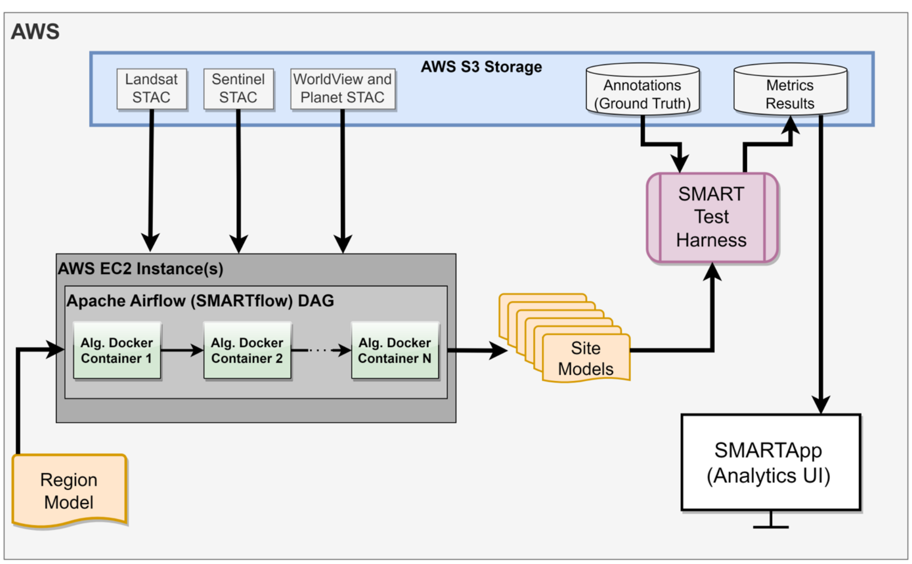


Figure 11: Generalized SMART processing architecture.

- Locating *all* activities across a wide swath of land. While many countries have databases with information about construction projects, we found that these databases were often incomplete, unreliable, or inaccessible and/or unavailable worldwide. Therefore, we were forced to use openly available imagery alone to find the sites. In our workflow we used historical data found in Google Earth;
- The requirement for high resolution imagery to ensure correctness and completeness of the labels. Google Earth contains high resolution imagery but with relatively low temporal resolution. In some regions of the world, the temporal resolution of this high spatial resolution imagery is lower due to higher prevalence of clouds or other atmospheric effects, making this process even more challenging;
- Searching across more than 7.5 years of imagery, which made the process quite tedious.

7.3 Metrics formulation: Pixel-level or Instance-level?

Traditional approaches for evaluating the performance of change detection algorithms often rely on pixel-level metrics that compare the change detection pixel maps generated by an algorithm with the ground truth. However, activities such as large-scale construction evolve in time in more subtle and complicated ways. For example, the spectral and temporal change signatures often vary spatially (e.g. changes due to pouring the foundation are different than changes due to landscaping). Furthermore, the spatial bounds of the activity change over time. Observing such characteristics of construction activity drove the team to consider each activity as a 3D spatio-temporal volume (hence the derivation of the site model). The resulting site model construct using time-varying polygons that each have an associated activity phase label allows for characterization of activities that have varying spatial, spectral, and temporal characteristics as a single detectable entity. The problem of detecting and monitoring these sites can then be treated as an AI problem that leverages more traditional object- or instance-level detection and classification metrics. One downside of this decision is that it adds to the complexity of polygon spatial and temporal association logic. For this, we partially leveraged SpaceNet-7[3] for our metric formulation but also made minor modifications to these more traditional approaches to better fit our problem construct and desired program goals.

7.4 Spatial Boundary Complexity: Sites and Subsites

Custom rules created for determining site and subsite boundaries (Section 3.4) added an extra layer of complexity for both annotation and algorithm development. This added complexity slowed the annotation process down considerably and it may have also hindered the algorithm development process unnecessarily. Annotator training was more difficult than anticipated since there was a steep learning curve required to become familiar with the concepts. In the end, it is not clear that the decision to include subsites in the problem formulation was worthwhile. A simpler annotation approach may have helped considerably, while not sacrificing the goals of the program.

7.5 Importance of the FFPA metric

One of the key limitations of F1 score for BAS is that it only considers the number of detections and false alarms but not the spatial area of the detections. Using only the F1 metric, an algorithm that produces large, undersegmented polygons will score as well as another algorithm that better localizes its detections. This is because the F1 considers only the *number* of detections and not the size of the proposed polygons. By pairing the F1 with the FFPA metric (See Table 4), one can gain deeper insight into algorithm performance and visualize the trade-off between the two metrics. The FFPA metric provides a spatial measure of how much of the area under test is covered by false alarm proposals.

8. NEXT STEPS

At the time of this publication, the SMART program is currently in the middle of Phase 2. Current objectives include:

- Increase quantitative evaluation (test harness) automation to facilitate development of algorithms using MLOps concepts and techniques;

- Improve heavy construction detection, classification, and prediction (temporal reasoning) algorithms;
- Improve accuracy of heavy construction annotation datasets;
- Curate data for a new TA-2 activity detection task using the same approaches developed for heavy construction. The program has identified **Transient Activities** as the activity of choice. These are activities that are *significantly* shorter in temporal duration than heavy construction, which creates more unique challenges and imagery requirements;
- Quantitatively assess the value of Planet imagery, which has both high spatial and temporal resolution;
- Continue building custom user interfaces and analytic tools to facilitate development and end-user performance analysis;
- Develop capabilities and evaluation metrics to support "online" or "incremental" processing of imagery, which is closer to the expected operational use case of receiving smaller batches of new imagery over time as opposed to having all imagery up-front.

Several of these tasks have been completed while others are currently in progress. The program hopes that publicly releasing the annotation dataset[14] will encourage more research in this area and help further advance the state of the art in *activity detection* using satellite-based imagery. Future publications will provide additional details on metrics formulation, algorithm development, and quantitative and qualitative assessment of performance.

9. ACKNOWLEDGMENTS

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. The authors would specifically like to thank the IARPA Program Manager, Dr. Jack Cooper. We would also like to acknowledge other key members of the test and evaluation team who have made significant contributions to this work. These include: Christopher Barber and Kelcy Smith (United States Geological Survey (USGS); Geoff Emmer, Colin Lee, and Sujit Wariyar (The MITRE Corporation); and Matt Dinmore, Jay Huang, Dave Jansing, Isolde Moyer, Eric Nguyen, May Palace, Christine Piatko, Trevor Stout, and Sean Wang (JHU/APL). Finally, a special thank you to iMERIT and CrowdAI for supporting a challenging annotation effort.

REFERENCES

[1] https://www.iarpa.gov/research-programs/smart. Accessed: April 21, 2023.

[2] Zhu, Z. and Woodcock, C. E., "Continuous change detection and classification of land cover using all available landsat data," *Remote sensing of Environment* **144**, 152–171 (2014).

[3] Van Etten, A., Hogan, D., Manso, J. M., Shermeyer, J., Weir, N., and Lewis, R., "The multi-temporal urban development spacenet dataset," in [*2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 6394–6403 (2021).

[4] https://spacenet.ai/. Accessed: April 6, 2023.

[5] Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., and Raskar, R., "Deepglobe 2018: A challenge to parse the earth through satellite images," in [*2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*], 172–17209 (2018).

[6] https://www.diu.mil/ai-xview-challenge. Accessed: April 6, 2023.

[7] Goldberg, H. R., Wang, S., Christie, G. A., and Brown, M. Z., "Urban 3d challenge: building footprint detection using orthorectified imagery and digital surface models from commercial satellites," in [*Geospatial informatics, motion imagery, and network analytics viii*], **10645**, 12–31, SPIE (2018).

[8] https://sentinel.esa.int/web/sentinel/missions/sentinel-2. Accessed: March 28, 2023.

[9] Xia, J., Yokoya, N., Adriano, B., and Broni-Bediako, C., "Openearthmap: A benchmark dataset for global high-resolution land cover mapping," in [*2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*], 6243–6253 (2023).

[10] Christie, G., Fendley, N., Wilson, J., and Mukherjee, R., "Functional map of the world," in [*2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 6172–6180 (2018).

[11] Toker, A., Kondmann, L., Weber, M., Eisenberger, M., Camero, A., Hu, J., Hoderlein, A. P., Şenaras, , Davis, T., Cremers, D., Marchisio, G., Zhu, X. X., and Leal-Taixé, L., "Dynamicearthnet: Daily multi-spectral satellite dataset for semantic change segmentation," in [*2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 21126–21135 (2022).

[12] Verma, S., Panigrahi, A., and Gupta, S., "Qfabric: Multi-task change detection dataset," in [*2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*], 1052–1061 (2021).

[13] Papadomanolaki, M., Verma, S., Vakalopoulou, M., Gupta, S., and Karantzalos, K., "Detecting urban changes with recurrent neural networks from multitemporal sentinel-2 data," in [*IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*], 214–217 (2019).

[14] https://github.com/pubgeo/datasets#iarpa-smart-2023.

[15] https://earth.esa.int/eogateway/missions/landsat-8. Accessed: March 28, 2023.

[16] "Sentinel 2 user handbook." https://sentinel.esa.int/documents/247904/685211/Sentinel-2_User_Handbook. Accessed: March 28, 2023.

[17] https://resources.maxar.com/data-sheets/worldview-1. Accessed: March 28, 2023.

[18] https://resources.maxar.com/data-sheets/worldview-2. Accessed: March 28, 2023.

[19] https://worldview3.digitalglobe.com/. Accessed: March 28, 2023.

[20] https://earth.esa.int/eogateway/documents/20142/37627/Planet-combined-imagery-product-specs-2020.pdf?text=Combined-Imagery-Product-Spec.pdf. Accessed: March 28, 2023.

[21] https://stacspec.org/en. Accessed: April 20, 2023.

[22] https://www.docker.com/. Accessed: April 20, 2023.