The Multi-Temporal Urban Development SpaceNet Dataset

Adam Van Etten[1], Daniel Hogan[1], Jesus Martinez-Manso[2], Jacob Shermeyer[3], Nicholas Weir[4,†], Ryan Lewis[4,†]

[1] In-Q-Tel CosmiQ Works, [avanetten, dhogan]@iqt.org, [2] Planet, jesus@planet.com,
[3] Capella Space, jake.shermeyer@capellaspace.com, [4] Amazon, [weirnich, rstlewis]@amazon.com

Abstract

Satellite imagery analytics have numerous human development and disaster response applications, particularly when time series methods are involved. For example, quantifying population statistics is fundamental to 67 of the 231 United Nations Sustainable Development Goals Indicators, but the World Bank estimates that over 100 countries currently lack effective Civil Registration systems. To help address this deficit and develop novel computer vision methods for time series data, we present the Multi-Temporal Urban Development SpaceNet (MUDS, also known as SpaceNet 7) dataset. This open source dataset consists of medium resolution (4.0m) satellite imagery mosaics, which includes $\approx$ *24 images (one per month) covering* $>$ *100 unique geographies, and comprises* $>$ $40,000$ *km*^{2} *of imagery and exhaustive polygon labels of building footprints therein, totaling over 11M individual annotations. Each building is assigned a unique identifier (*i.e. *address), which permits tracking of individual objects over time. Label fidelity exceeds image resolution; this "omniscient labeling" is a unique feature of the dataset, and enables surprisingly precise algorithmic models to be crafted.*

We demonstrate methods to track building footprint construction (or demolition) over time, thereby directly assessing urbanization. Performance is measured with the newly developed SpaceNet Change and Object Tracking (SCOT) metric, which quantifies both object tracking as well as change detection. We demonstrate that despite the moderate resolution of the data, we are able to track individual building identifiers over time. This task has broad implications for disaster preparedness, the environment, infrastructure development, and epidemic prevention.

1. Introduction

Time series analysis of satellite imagery poses an interesting computer vision challenge, with many human development applications. We aim to advance this field through the release of a large dataset aimed at enabling new methods in this domain. Beyond its relevance for disaster response, disease preparedness, and environmental monitoring, time series analysis of satellite imagery poses unique technical challenges often unaddressed by existing methods.

The MUDS dataset (also known as SpaceNet 7) consists of imagery and precise building footprint labels over dynamic areas for two dozen months, with each building assigned a unique identifier (see Section 3 for further details). In the algorithmic portion of this paper (Section 5), we focus on tracking building footprints to monitor construction and demolition in satellite imagery time series. We aim to identify all of the buildings in each image of the time series and assign identifiers to track the buildings over time.

Timely, high-fidelity foundational maps are critical to a great many domains. For example, high-resolution maps help identify communities at risk for natural and human-derived disasters. Furthermore, identifying new building construction in satellite imagery is an important factor in establishing population estimates in many areas (*e.g.* [7]). Population estimates are also essential for assessing burden on infrastructure, from roads[4] to medical facilities [26].

The inclusion of unique building identifiers in the MUDS dataset enable potential improvements upon existing course population estimates. Without unique identifiers building tracking is not possible; this means that over a given area one can only determine how many new buildings exist. By tracking unique building identifiers one can determine which buildings changed (whose properties such as precise location, area, *etc*. can be correlated with features such as road access, distance to hospitals, etc.), thus providing a much more granular view into population growth.

Several unusual features of satellite imagery (*e.g.* small object size, high object density, dramatic image-to-image difference compared to frame-to-frame variation in video object tracking, different color band wavelengths and counts, limited texture information, drastic changes in shadows, and repeating patterns) are relevant to other tasks and data. For example, pathology slide images or other mi-

[†]This work was completed prior to Nicholas Weir and Ryan Lewis joining Amazon

croscopy data present many of the same challenges [38]. Lessons learned from this dataset may therefore have broad-reaching relevance to the computer vision community.

2. Related Work

Past time series computer vision datasets and algorithmic advances have prepared the field to address many of the problems associated with satellite imagery analysis, allowing our dataset to explore additional computer vision problems. The challenge built around the VOT dataset [15] saw impressive results for video object tracking (*e.g.* [36]), yet this dataset differs greatly from satellite imagery, with high frame rates and a single object per frame. Other datasets such as MOT17 [17] or MOT20 [6] have multiple targets of interest, but still have relatively few (< 20) objects per frame. The Stanford Drone Dataset [23] appears similar at first glance, but has several fundamental differences that result in very different applications. That dataset contains overhead videos taken at multiple hertz from a low elevation, and typically have ≈ 20 mobile objects (cars, people, buses, bicyclists, etc.) per frame. Because of the high frame rate of these datasets, frame-to-frame variation is minimal (see the MOT17 example in Figure 1D). Furthermore, objects are larger and less abundant in these datasets than buildings are in satellite imagery. As a result, video competitions and models derived therein provide limited insight in how to manage imagery time series with substantial image-to-image variation and overly-dense instance annotations of target objects. Our data and research will address this gap.

To our knowledge, no existing dataset has offered a deep time series of satellite imagery. A number of previous works have studied building extraction from satellite imagery ([8], [5], [39], [27]), yet these datasets were static. The closest comparison is the xView2 challenge and dataset [10], which examined building damage in satellite image pairs acquired before and after natural disasters (*i.e.* only two timestamps) in < 20 locations; however, this task fails to address the complexities and opportunities posed by analysis of deep time series data such as seasonal vegetation and lighting changes, or consistent object tracking on a global scale. Other competitions have explored time series data in the form of natural scene video, *e.g.* object detection [6] and segmentation [2] tasks. There are several meaningful dissimilarities between these challenges and the task described here. Firstly, frame-to-frame variation is very small in video datasets (see Figure 1D). By contrast, the appearance of satellite images can change dramatically from month to month due to differences in weather, illumination, and seasonal effects on the ground, as shown in Figure 1C. Other time series competitions have used non-imagery data spaced regularly over longer time intervals [9], but none focused on computer vision tasks.

The size and density of target objects are very different in this dataset than past computer vision challenges. When comparing the size of annotated instances in the COCO dataset [18], there's a clear difference in object size distributions (see Figure 1A). These smaller objects intrinsically provide less information as they comprise fewer pixels, making their identification a more difficult task. Finally, the number of instances per image is markedly different in satellite imagery from the average natural scene dataset (see Section 3 and Figure 1B). Other data science competitions have explored datasets with similar object size and density, particularly in the microscopy domain [21, 11]; however, those competitions did not address time series applications. Taken together, these differences highlight substantial novelty for this dataset.

3. Data

The Multi-Temporal Urban Development SpaceNet (MUDS) dataset consists of 101 labelled sequences of satellite imagery collected by Planet Labs' Dove constellation between 2017 and 2020, coupled with building footprint labels for every image. The image sequences are sampled at the 101 distinct areas of interest (AOIs) across the globe, covering six continents (Figure 2). These locations were selected to be both geographically diverse and display dramatic changes in urbanization across a two-year timespan.

The MUDS dataset is open sourced under a CC-BY-4.0 ShareAlike International license[‡] to encourage broad use. This dataset can potentially be useful for many other geospatial computer vision tasks: it can be easily fused or augmented with any other data layers that are available through web tile servers. The labels themselves can also be applied to any other remote sensing image tiles, such us high resolution optical or synthetic aperture radar.

3.1. Imagery

Images are sourced from Planet's global monthly basemaps, an archive of on-nadir imagery containing visual RGB bands with a ground sample distance (GSD) (*i.e.* pixel size) of ≈ 4 meters. A basemap is a reduction of all individual satellite captures (also called scenes) into a spatial grid. These basemaps are created by mosaicing the best scenes over a calendar month, selected according to quality metrics like image sharpness and cloud coverage. Scenes are stack-ranked with best on top, and spatially harmonized to smoothen scene boundary discontinuities. Monthly basemaps are particularly well suited for the computer vision analysis of urban growth, since they are relatively cloud-free, homogeneous, and represented in a consistent spatio-temporal grid. The monthly cadence is also a good match to the typical timescale of urban developments.

[‡]https://registry.opendata.aws/spacenet/

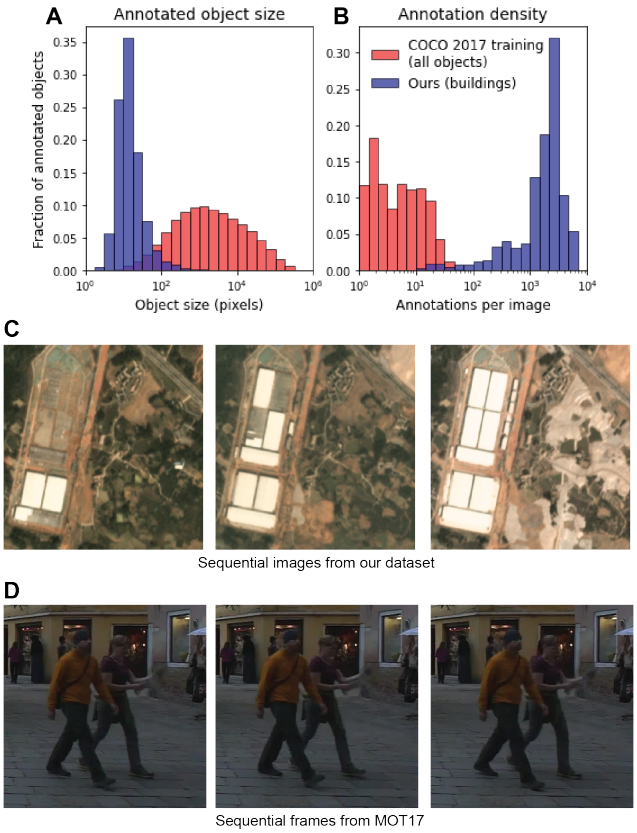


Figure 1: A comparison between our dataset and related datasets. **A.** Annotated objects are very small in this dataset. Plot represents normalized histograms of object size in pixels. Blue is our dataset, red represents all annotations in the COCO 2017 training dataset [18]. **B.** The density of annotations is very high in our dataset. In each 1024×1024 image, our dataset has between 10 and over 20,000 objects (mean: 4,600). By contrast, the COCO 2017 training dataset has at most 50 objects per image. **C.** Three sequential time points from one geography in our dataset, spanning 3 months of development. Compare to **D.**, which displays three sequential frames in the MOT17 video dataset [17].

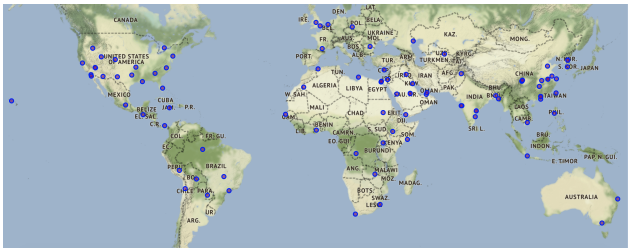


Figure 2: Location of MUDS data cubes.

The size of each image is 1024×1024 pixels, corresponding to $\approx 18\,\mathrm{km}^2$, and the total area of the images in the dataset is $41,250\,\mathrm{km}^2$. See Table 1 or spacenet.ai for additional statistics. The time series contain imagery of $18 - 26$ months, depending on AOI (median of 24). This lengthy time span captures multiple seasons and atmospheric conditions, as well as the commencement and completion of multiple construction projects. See Figure 3 for examples. Images containing an excessive amount of clouds or haze were fully excluded from the dataset, thus causing minor temporal gaps in some of the time series.

3.2. Label Statistics

Each image in the dataset is accompanied by two sets of manually created annotations. The first set of labels are building footprint polygons defining the outline of each building. Each building is assigned a unique identifier (*i.e.* address) that persists throughout the time series. The second set of annotations are "unusable data masks" (UDMs) denoting areas of images that are obscured by clouds (see Figure 4) or that suffer from image geo-reference errors greater than 1 pixel. Geo-referencing is the process of mapping pixels in sensor space to geographic coordinates, performed via an empirical fitting procedure that is never exact. In rare cases, the scenes that compose the basemaps have spatial offsets of 5-10 meters. Accounting for such spatial displacements in the time series would make the modeling task significantly harder. Therefore, we decided to eliminate this complexity by including these regions in the UDM.

Each image has between 10 and $\approx 20,000$ building annotations, with a mean of $\approx 4,600$ (the earliest timepoints in some geographies have very few buildings completed). This represents much higher label density than natural scene datasets like COCO [18] (Figure 1B), or even overhead drone video datasets [34]. As the dataset comprises ≈ 24 time points at 101 geographic areas, the final dataset includes $>$ 11M annotations, representing $> 500,000$ unique buildings. (Compare the training data quantities shown for other datasets in Table 1.) The building areas vary between approximately 0.25 and 13,000 pixels (median building area of 193 m^2 or 12.1 pix^2), markedly smaller than most labels in natural scene imagery datasets (Figure 1A).

Seasonal effects and weather (i.e. background variation) pervade our dataset given the low frame rate of 4×10^{-7} Hz (Figure 1C). This "background" change adds to the change detection task's difficulty. This frame-by-frame background variation is particularly unique and difficult to recreate via simulation or video re-sampling.

3.3. Labeling Procedure

We define buildings as static man-made structures where an individual could take shelter, with no minimum footprint size. The uniqueness of the dataset presents distinct labeling challenges. First, small buildings can be under-resolved to the human eye in a given image, making it difficult to locate and discern from other non-building structures. Second, in locations undergoing building construction, it can

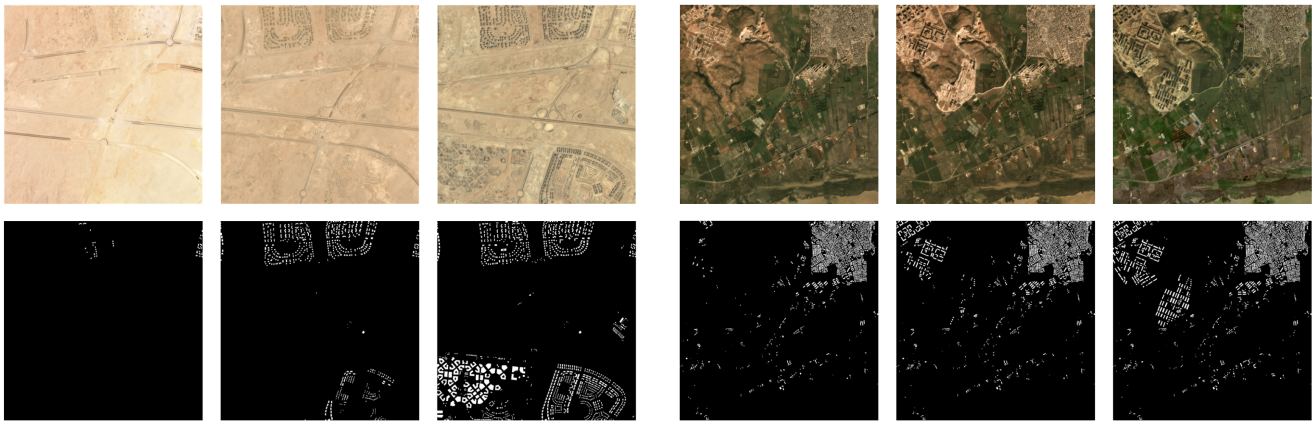


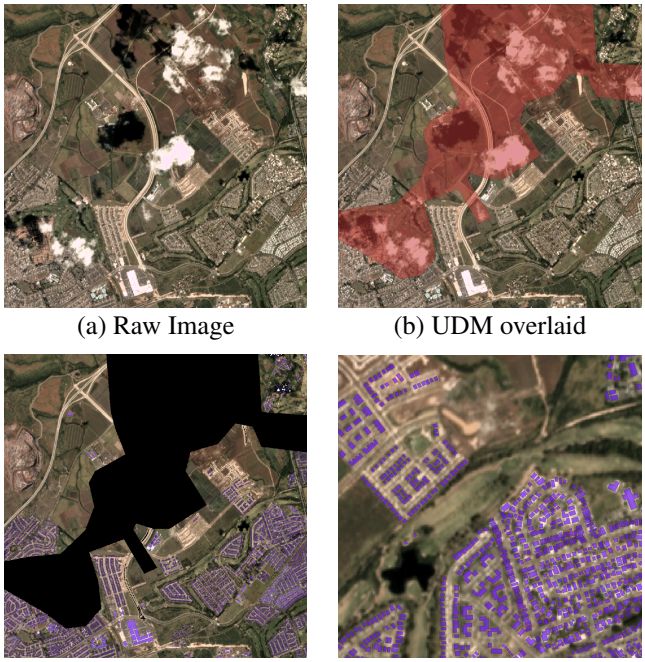
Figure 3: Time series of two data cubes. Left column (*e.g.* 1a) denotes the start of the times series, the middle column (*e.g.* 1b) the approximate midpoint, and the right column (*e.g.* 1c) shows the final image. The top row displays imagery, while the bottom row illustrates the labeled building footprints.

Table 1: Comparison of Selected Time Series Datasets

Property	**MUDS**	**VOT-ST2020** [14]	**MOT20** [6]	**Stanford Drone** [23]	**DAVIS 2017** [3]	**YouTube-VOS** [41]
Scenes	101	60	4	60	90	4,453
Total Frames	2,389	19,945	8,931	522,497	6,208	∼603,000
Unique Tracks	538,073	60	2,332	10,300	216	7,755
Total Labels	11,079,262	19,945	1,652,040	10,616,256	13,543	197,272
Median Frames/Scene	24	257.5	1,544	11,008	70.5	∼135 (mean)
Ground Sample Dist.	4.0m	n/a	n/a	∼2cm	n/a	n/a
Frame Rate	1/month	30fps	25fps	30fps	20fps	30fps (6fps labels)
Annotation	Polygon	Seg. Mask	BBox	BBox	Seg. Mask	Seg. Mask
Objects	Buildings	Various	Pedestrians, etc.	Pedestrians & Vehicles	Various	Various

be difficult to determine what point in time the structure becomes a building per our definition. Third, variability in image quality, atmospheric conditions, shadows, and seasonal phenology can introduce additional confusion. Mitigating these complexities and minimizing label noise was of paramount importance, especially along the temporal dimension. Even though the dataset AOIs were selected to contain urban change, construction events are still highly imbalanced compared to the full spatio-temporal volume. Thus, temporal consistency was a fundamental area of focus in the labeling strategy. In cases of high uncertainty with a particular building candidate, annotators examined the full time series to gain temporal and contextual information of the precise location. For example, a shadow from a neighboring structure might be confused as a building, but this becomes evident when inspecting the full data cube. Temporal context can also help identify groups of objects. Some regions have structures that resemble buildings in a given image, but are highly variable in time. Objects that appear and disappear multiple times are unlikely to be buildings. Once one type of such ephemeral structures is identified as a confusion source, all other similar structures are also excluded (Figure 5). Labeling took 7 months by a team of 5; each data cube was annotated by one person, reviewed and corrected by another, with final validation by the team lead.

Annotators also used a privately-licensed high resolution imagery map to help discriminate uncertain cases. This high resolution map is useful to gain contextual information of the region and to guide the precise building outlines that are unclear from the dataset imagery alone. Once a building candidate was identified in the MUDS imagery, the high resolution map was used to confirm the building geometry. In other words, labels were not created on the high resolution imagery first. While the option of labeling on high resolution might seem attractive, it poses labeling risks such as capturing buildings that are not visible at all in the MUDS imagery. In addition, the high resolution map is static and composed of imagery acquired over a long range of dates, thus making it difficult to perform temporal comparisons between this map and the dataset imagery.



(a) Raw Image (b) UDM overlaid

(c) Masked image + labels (d) Zoomed labels

Figure 4: Single image in a data cube. (a) Image with cloud cover. (b) Image with UDM overlaid. (c) Masked image with building labels overlaid. (d) Zoom showing the high fidelity of building labels.

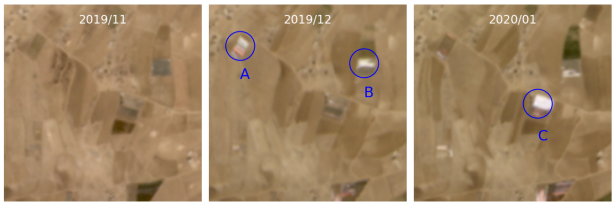


Figure 5: Example of how temporal context can help with object identification. If the middle image were to be labeled in isolation, objects A and B could be annotated as buildings. However, taking into account the adjacent images, these objects exist only for one month and therefore are unlikely to be buildings. Object C is also unlikely to be a building, just by group association.

The procedure to annotate each time series can be summarized as follows:

1. Start with the first image in the series. Identify the location of all visible structures. If the building location and outline are clear, draw a polygon around it. Otherwise, overlay a high resolution optical map to help confirm the presence of the building and draw the outline. Assign a unique integer identifier to each building. In addition, identify any regions in the image with impaired ground visibility or defects and add their polygons to the UDM layer of this image.

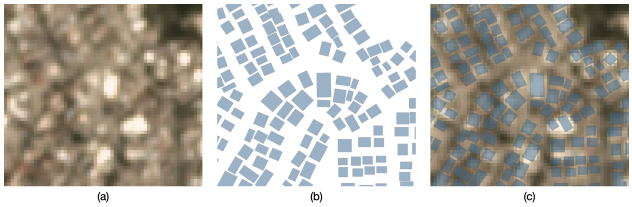


Figure 6: Zoom in of one particularly dense region illustrating the very high fidelity of labels. (a) Raw image. (b) Footprint polygon labels. (c) Footprints overlaid on imagery.

2. Copy all the building labels onto the next image (not the UDM). Examine carefully all buildings in the new image, and edit the labels with any changes. Edits are only be made when there is significant confidence that a building appeared or disappeared. If a new building appeared, assign a new unique identifier. Toggle through multiple images in the time series to ensure: (a) there is a true building change and (b) that it is applied to the correct time point. Also, create a UDM.

3. Repeat step 2 for the remaining time points.

This process attempts to enforce temporal consistency and reduce object confusion. While label noise is appreciable in small objects, the use of high resolution imagery to label results in labels of significantly higher fidelity that would be achievable from the Planet data alone, as illustrated in Figure 6. This "omniscient labeling" is one of the key features of the MUDS dataset. We will show in Section 5 that the baseline algorithm does a surprisingly good job of extracting high-resolution features from the medium-resolution imagery. In effect, the labels are encoding information that is not visible to humans in the imagery, which the baseline algorithm is able to capitalize upon.

4. Evaluation Metrics

To evaluate model performance on a time series of identifier-tagged footprints such as MUDS, we introduce a new evaluation metric: the SpaceNet Change and Object Tracking (SCOT) metric [13]. As discussed later, existing metrics have a number of shortcomings that are addressed by SCOT. The SCOT metric combines two terms: a tracking term and a change detection term. The tracking term evaluates how often a proposal correctly tracks the same buildings from month to month with consistent identifier numbers. In other words, it measures the model's ability to characterize what stays the same as time goes by. The change detection term evaluates how often a proposal correctly picks up on the construction of new buildings. In other words, it measures the model's ability to characterize what changes as time goes by.

For both terms, the calculation starts the same way: find-

ing "matches" between ground truth building footprints and proposal building footprints for each month. A pair of footprints (one ground truth and one proposal) are eligible to be matched if their intersection over union (IOU) exceeds 0.25, and no footprint may be matched more than once. We select an IOU of 0.25 to mimic Equation 5 of ImageNet [25]), which sets IOU < 0.5 for small objects. A set of matches is chosen that maximizes the number of matches. If there is more than one way to achieve that maximum, then as a tie-breaker the set with the largest sum of IOUs is used. This is an example of the unbalanced linear assignment problem in combinatorics.

If model performance were being evaluated for a single image (instead of a time series), a customary next step might be calculating an F1 score, where matches are considered true positives (tp) and unmatched ground truth and proposal footprints are considered false negatives (fn) and false positives (fp) respectively.

$$F_1 = \frac{tp}{tp + \frac{1}{2}(fp + fn)} \tag{1}$$

The tracking term and change detection term both generalize this to a time series, each in a different way.

The tracking term penalizes inconsistent identifiers across time steps. A match is considered a "mismatch" if the ground truth footprint's identifier was most recently matched to a different proposal ID, or vice versa. For the purpose of the tracking term, mismatches (mm) are not counted as true positives. So each mismatch decreases the number of true positives by one. This effectively divorces the ground truth footprint from its mismatched proposal footprint, creating an additional false negative and an additional false positive. That amounts to the following transformations:

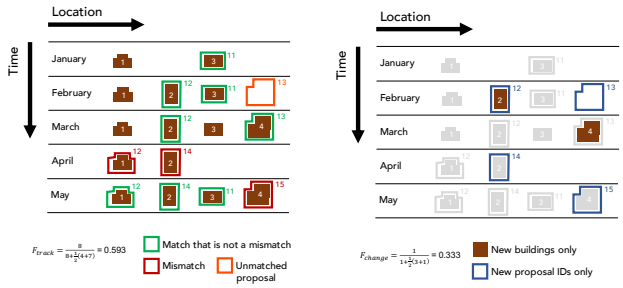
$$\begin{aligned} tp &\rightarrow tp - mm \\ fp &\rightarrow fp + mm \\ fn &\rightarrow fn + mm \end{aligned} \tag{2}$$

Applying these to the F1 expression above gives the formula for the tracking term:

$$F_{\text{track}} = \frac{tp - mm}{tp + \frac{1}{2}(fp + fn)} \tag{3}$$

The second term in the SCOT metric, the change detection term, incorporates only new footprints. That is, ground truth or proposal footprints with identifier numbers making their first chronological appearance. Letting the subscript new indicate the count of tp's, fp's, and fn's that persist after dropping non-new footprints:

$$F_{\text{change}} = \frac{tp_{\text{new}}}{tp_{\text{new}} + \frac{1}{2}(fp_{\text{new}} + fn_{\text{new}})} \tag{4}$$



(a) Tracking Term (b) Change Term

Figure 7: (a) Example of SCOT metric tracking term. Solid brown polygons are ground truth building footprints, and outlines are proposal footprints. Each footprint's corresponding identifier number is shown. (b) Example of SCOT metric change detection term, using the same set of ground truth and proposal footprints. This term ignores all ground truth and proposal footprints with previously-seen identifiers, which are indicated in a faded-out gray color.

One important property of this term is that a set of static proposals that do not vary from one month to another will receive a change detection term of 0, even for a time series with very little new construction. (In the MUDS dataset, the construction of new buildings is by far the most common change; the metric could be generalized to accommodate building demolition or other changes by any of several straightforward generalizations.)

To compute the final score, the two terms are combined with a weighted harmonic mean:

$$F_{\text{scot}} = (1 + \beta^2)\frac{F_{\text{change}} \cdot F_{\text{track}}}{\beta^2 F_{\text{change}} + F_{\text{track}}} \tag{5}$$

We use a value of $\beta = 2$ to emphasize the part of the task (tracking) that has been less commonly explored in an overhead imagery context. For a dataset like MUDS with multiple AOIs, the overall SCOT score is the arithmetic mean of the scores of the individual AOIs.

Figure 7a is a cartoon example of calculating the tracking term on a row of four buildings imaged over five months (during which time two of the four are newly-constructed, and two are temporarily occluded by clouds). Figure 7b illustrates the change detection term for the same case.

For geospatial work, the SCOT metric has a number of advantages over evaluation metrics developed for object tracking in video, such as the Multiple Object Tracking Accuracy (MOTA) metric [1]. MOTA scores are mathematically unbounded, making them less intuitively interpretable for challenging low-score scenarios, and sometimes even yielding negative scores. More critically, for scenes with only a small amount of new construction, it's possible to achieve a high MOTA score with a set of proposal footprints

that shows no time-dependence whatsoever. Since understanding time-dependence is usually a primary purpose of time series data, this is a serious drawback. SCOT's change detection term prevents this. In fact, many such approaches to "gaming" the SCOT metric by artificially increasing one term will decrease the other term, leaving no obvious alternative to intuitively-better model performance as a way to raise scores.

5. Experiments

For object tracking, one could in theory leverage the results of previous challenges (*e.g.* MOT20 [6]), yet the significant differences between MUDS and previous datasets such as high density and small object size (see Figure 1) render previous approaches unsuitable. For example, approaches such as TrackR-CNN [35] are untrainable as each instance requires a separate channel resulting in a memory explosion for images with many thousands of objects. Other approaches such as Joint Detection and Embedding (JDE) [37] are trainable; however inference results are ultimately incoherent due to the tiny object size and density overwhelming the YOLOv3 [22] detection grid. Despite these challenges, the spatially static nature of our objects of interest somewhat simplifies tracking objects between each observation. Consequently, this dataset should incentivize the development of new object tracking algorithms that can cope with a lack of resolution, spatial stasis, minimal size, and dense clustering of objects.

As a result of the challenges listed above, we choose to experiment with semantic segmentation based approaches to detect and track buildings over time. These methods are adapted from prize winning approaches for the SpaceNet 4 and 6 Building Footprint Extraction Challenges [40, 28]. Our architecture comprises a U-Net [24] with different encoders. The first "baseline" approach uses a VGG16 [30] encoder and a custom loss function of $\mathcal{L} = \mathcal{J} + 4 \cdot BCE$, where $\mathcal{J}$ is Jaccard distance and BCE denotes binary cross entropy. The second approach uses a more advanced EfficientNet-B5 [32] encoder with a loss of $\mathcal{L} = \mathcal{F} + \mathcal{D}$ where $\mathcal{F}$ is Focal loss [19] and $\mathcal{D}$ is Dice loss.

To ensure robust testing statistics, we train the model on 60 data cubes, testing on the remaining 41 data cubes. We train the segmentation models with an Adam optimizer on the 1424 images of the training set for 300 epochs and a learning rate of 10^{-4} (baseline) or 100 epochs and a learning rate of 2×10^{-4} (EfficentNet).

At inference time binary building prediction masks are converted to instance segmentations of building footprints. Each footprint at $t = 0$ is assigned a unique identifier. For each subsequent time step building footprints polygons are compared to the positions of the previous time step. Building identifier matching is achieved by an optimized matching of polygons with a minimum IOU overlap of 0.25.

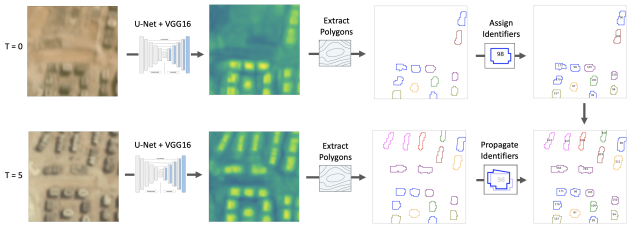


Figure 8: Baseline algorithm for building footprint extraction and identifier tracking showing evolution from T = 0 (top row) to T = 5 (bottom row). The input image is fed into our segmentation model, yielding a building mask (second column). This mask is refined into building footprints (third column), and unique identifiers are allocated (right column).

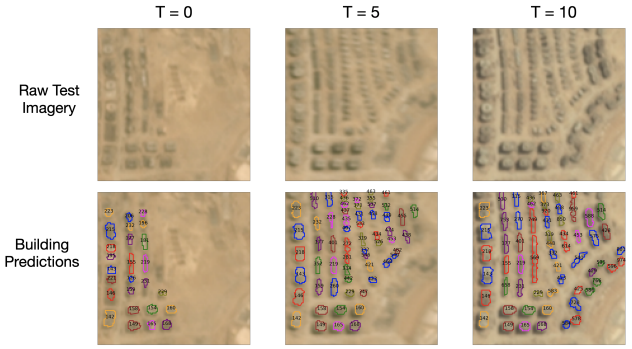


Figure 9: Example tracking performance of the baseline algorithm. Note that larger, well-separated buildings are tracked well between epochs, while denser regions are more challenging for tracking.

Table 2: Building Tracking Performance

Metric	Approach	
	VGG-16	**EfficentNet**
F1 (IOU ≥ 0.25)	0.45 ± 0.13	0.42 ± 0.12
Tracking Score	0.40 ± 0.10	0.39 ± 0.10
Change Score	0.06 ± 0.05	0.07 ± 0.05
SCOT	0.17 ± 0.10	0.18 ± 0.09

Matched footprintes are assigned the same identifier as the previous timestep, while footprints without significant overlap with preceding geometries are assigned a new unique identifier. The baseline algorithm is illustrated in Figure 8; note that building identifiers are well matched between epochs. Performance is summarized in Table 2. For scoring we assess only buildings with area $\geq 4\,\mathrm{px}^2$.

Localizing and tracking buildings in medium resolution ($\approx$ 4m) imagery is quite challenging, but surprisingly achievable in our experiments. For well separated buildings, building localization and tracking performs fairly well; for example in Figure 9) we find a localization F1

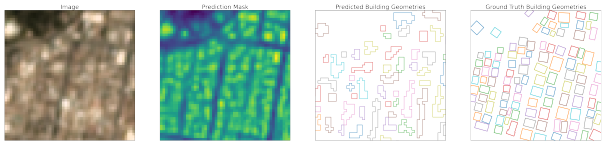


Figure 10: Prediction in a difficult, crowded region. Despite the inherent difficulties in separating nearby buildings at medium resolution, for this image F1 = 0.40.

score of 0.55, and a SCOT score of 0.31. For dense regions, building tracking is far more difficult; in Figure 10 we still see decent performance in building localization (F1 = 0.40), yet building tracking and change detection is very challenging (SCOT = 0.07) since inter-epoch footprints overlap poorly. The change term of SCOT is particularly challenging, as correctly identifying the origin epoch of each building is non-trivial, and spurious proposals are also penalized.

In an attempt to raise the scores of Table 2, we also endeavor to incorporate the time dimension into training. As previously mentioned, existing approaches transfer poorly to this dataset, so we attempt a simple approach of stacking multiple images at training time. For each date we train on the imagery for that date plus the four chronologically adjacent future observations $[t = 0, t + 1, t + 2, t + 3, t + 4]$ for five total dates of imagery. When the number of remaining observations in the time series becomes less than five, we repeatedly append the final image for each area of interest. We find no improvement with this approach (SCOT $= 0.17 \pm 0.08$).

We also note no significant difference in scores between the VGG-16 and EfficentNet architectures (Table 2), implying that older architectures are essentially as adept as state-of-the-art architectures when it comes to extracting information from the small objects in this dataset.

While not fully explored here, we also anticipate that researchers may improve upon the baseline using models specifically intended for time series analysis (*e.g.* Recurrent Neural Networks (RNNs) [20] and Long-Short Term Memory networks (LSTMs) [12]. In addition, numerous "classical" geospatial time series methods exist (*e.g.* [42]) which researchers may find valuable to incorporate into their analysis pipelines as well.

6. Discussion

Intriguingly, the score of $F1 = 0.45$ for our baseline mode parallels previous results observed in overhead imagery. [29] studied object detection performance in xView [16] satellite imagery for various resolutions and five different object classes. These authors used the YOLT [33] object detection framework, which uses a custom network based on the Googlenet [31] architecture. The mean extent of the objects in this paper was 5.3 meters; at a resolution of 1.2 meters objects have an average extent of 4.4 pixels.

The average building area for the MUDS dataset is 332 m^2, implying an extent of 18.2 m for a square object. For a 4 meter resolution, this gives an average extent of 4.5 pixels, comparable to the 4.4 pixel extent of xView. The observed MUDS F1 score of 0.45 is within error bars of the results of the xView results, see Table 3. Of particular note is that while the F1 scores and object pixel sizes of Table 3 are comparable, the datasets stem from vastly different sensors, and the techniques are wildly different as well (a Googlenet-based object detection architecture versus a VGG16-based segmentation architecture). Apparently, object detection performance holds across sensors and algorithms as long as object pixel sizes are comparable.

Table 3: F1 Performance Across Datasets

Dataset	GSD (m)	Object Size (pix)	F1
xView	1.2	4.4	0.41 ± 0.03
MUDS	4.0	4.5	0.45 ± 0.13

7. Conclusions

The Multi-temporal Urban Development SpaceNet (MUDS, also known as SpaceNet 7) dataset is a newly developed corpus of imagery and precise labels designed for tracking building footprints and unique identifiers. The dataset covers over 100 locations across 6 continents, with a deep temporal stack of 24 monthly images and over 11,000,000 labeled objects. The significant scene-to-scene variation of the monthly images poses a challenge for computer vision algorithms, but also raises the prospect of developing algorithms that are robust to seasonal change and atmospheric conditions. One of the key characteristics of the MUDS dataset is exhaustive "omniscient labeling" with labels precision far exceeding the base imagery resolution of 4 meters. Such dense labels present significant challenges in crowded urban environments, though we demonstrate surprisingly good building extraction, tracking, and change detection performance with our baseline algorithm. Intriguingly, our object detection performance of $F1 = 0.45$ for objects averaging 4-5 pixels in extent is consistent with previous object detection studies, even though these studies used far different algorithmic techniques and datasets. There are numerous avenues of research beyond the scope of this paper that we hope the community will tackle with this dataset: the efficacy of super-resolution, adapting video time-series techniques to the unique features of MUDS, experimenting with RNNs, Siamese networks, LSTMs, etc. Furthermore, the dataset has the potential to aid a number of humanitarian efforts connected with population dynamics and UN sustainable development goals.

References

[1] Keni Bernardin, Alexander Elbs, and Rainer Stiefelhagen. Multiple object tracking performance metrics and evaluation in a smart room environment. *Sixth IEEE International Workshop on Visual Surveillance, in conjunction with ECCV*, 2006. 6

[2] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool. The 2019 davis challenge on vos: Unsupervised multi-object segmentation. *arXiv:1905.00737*, 2019. 2

[3] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool. The 2019 DAVIS challenge on VOS: Unsupervised multi-object segmentation. *arXiv:1905.00737*, 2019. 4

[4] Simiao Chen, Michael Kuhn, Klaus Prettner, and David E. Bloom. The global macroeconomic burden of road injuries: estimates and projections for 166 countries. 2019. 1

[5] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 2

[6] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. MOT20: A benchmark for multi object tracking in crowded scenes, 2020. 2, 4, 7

[7] R. Engstrom, D. Newhouse, and V. Soundararajan. Estimating small area population density using survey data and satellite imagery: An application to sri lanka. *Urban Economics & Regional Studies eJournal*, 2019. 1

[8] Adam Van Etten, Dave Lindenbaum, and Todd M. Bacastow. Spacenet: A remote sensing dataset and challenge series. *CoRR*, abs/1807.01232, 2018. 2

[9] Google. Web traffic time series forecasting: Forecast future traffic to wikipedia pages. 2

[10] Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston. Creating xbd: A dataset for assessing building damage from satellite imagery. In *Proceedings of the 2019 CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 2

[11] Booz Allen Hamilton and Kaggle. Data science bowl 2018: Spot nuclei. speed cures. 2

[12] Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. *Neural Computation*, 9, 1997. 8

[13] Daniel Hogan and Adam Van Etten. The SpaceNet change and object tracking (SCOT) metric, August 2020. 5

[14] Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kamarainen, Luka Čehovin Zajc, Martin Danelljan, Alan Lukezic, Ondrej Drbohlav, Linbo He, Yushan Zhang, Song Yan, Jinyu Yang, Gustavo Fernandez, et al. The eighth visual object tracking VOT2020 challenge results, 2020. 4

[15] Matej Kristan, Jiri Matas, Aleš Leonardis, Tomas Vojir, Roman Pflugfelder, Gustavo Fernandez, Georg Nebehay, Fatih Porikli, and Luka Čehovin. A novel performance evaluation methodology for single-target trackers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2137–2155, Nov 2016. 2

[16] Darius Lam, Richard Kuzma, Kevin McGee, Samuel Dooley, Michael Laielli, Matthew Klaric, Yaroslav Bulatov, and Brendan McCord. xview: Objects in context in overhead imagery. *CoRR*, abs/1802.07856, 2018. 8

[17] Laura Leal-Taixé, Anton Milan, Konrad Schindler, Daniel Cremers, Ian D. Reid, and Stefan Roth. Tracking the trackers: An analysis of the state of the art in multiple object tracking. *CoRR*, abs/1704.02781, 2017. 2, 3

[18] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 2, 3

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 7

[20] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*, 2010. 8

[21] Recursion Pharmaceuticals. Cellsignal: Disentangling biological signal from experimental noise in cellular images. 2

[22] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 7

[23] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *ECCV*, 2016. 2, 4

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the 2015 International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015. 7

[25] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014. 6

[26] Nadine Schuurman, Robert S. Fiedler, Stefan C.W. Grzybowski, and Darrin Grund. Defining rational hospital catchments for non-urban areas based on travel time. 5, 2006. 1

[27] Jacob Shermeyer, Daniel Hogan, Jason Brown, Adam Van Etten, Nicholas Weir, Fabio Pacifici, Ronny Hansch, Alexei Bastidas, Scott Soenen, Todd Bacastow, and Ryan Lewis. Spacenet 6: Multi-sensor all weather mapping dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2

[28] Jacob Shermeyer, Daniel Hogan, Jason Brown, Adam Van Etten, Nicholas Weir, Fabio Pacifici, Ronny Hansch, Alexei Bastidas, Scott Soenen, Todd Bacastow, and Ryan

Lewis. Spacenet 6: Multi-sensor all weather mapping dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 7

[29] Jacob Shermeyer and Adam Van Etten. The effects of super-resolution on object detection performance in satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 8

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceeings of the 2015 International Conference on Learning Representations*, 2015. 7

[31] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. 8

[32] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020. 7

[33] A. Van Etten. Satellite imagery multiscale rapid detection with windowed networks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 735–743, 2019. 8

[34] Stanford Computational Vision and Geometry Lab. Stanford drone dataset. 3

[35] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. MOTS: Multi-object tracking and segmentation, 2019. 7

[36] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip H.S. Torr. Fast online object tracking and segmentation: A unifying approach. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[37] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking, 2020. 7

[38] Nicholas Weir, JJ Ben-Joseph, and Dylan George. Viewing the world through a straw: How lessons from computer vision applications in geo will impact bio image analysis, Jan 2020. 2

[39] Nicholas Weir, David Lindenbaum, Alexei Bastidas, Adam Van Etten, Sean McPherson, Jacob Shermeyer, Varun Kumar, and Hanlin Tang. Spacenet mvoi: A multi-view overhead imagery dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[40] Nicholas Weir, David Lindenbaum, Alexei Bastidas, Adam Van Etten, Sean McPherson, Jacob Shermeyer, Varun Kumar Vijay, and Hanlin Tang. Spacenet MVOI: a multi-view overhead imagery dataset. In *Proceedings of the 2019 International Conference on Computer Vision*, volume abs/1903.12239, 2019. 7

[41] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. YouTube-VOS: A large-scale video object segmentation benchmark, 2018. 4

[42] Zhe Zhu. Change detection using landsat time series: A review of frequencies, preprocessing, algorithms, and applications. *ISPRS Journal of Photogrammetry and Remote Sensing*, 130:370 – 384, 2017. 8