DynamicEarthNet: Daily Multi-Spectral Satellite Dataset for Semantic Change Segmentation

Aysim Toker$^{1, *}$, Lukas Kondmann$^{1, 2, *}$, Mark Weber$^1$, Marvin Eisenberger$^1$, Andrés Camero$^2$, Jingliang Hu$^2$, Ariadna Pregel Hoderlein$^1$, Çağlar Şenaras$^3$, Timothy Davis$^3$, Daniel Cremers$^1$, Giovanni Marchisio$^{3, †}$, Xiao Xiang Zhu$^{1, 2, †, ‡}$, Laura Leal-Taixé$^1, †$

Technical University of Munich$^1$, German Aerospace Center$^2$, Planet Labs$^3$

Abstract

Earth observation is a fundamental tool for monitoring the evolution of land use in specific areas of interest. Observing and precisely defining change, in this context, requires both time-series data and pixel-wise segmentations. To that end, we propose the DynamicEarthNet dataset that consists of daily, multi-spectral satellite observations of 75 selected areas of interest distributed over the globe with imagery from Planet Labs. These observations are paired with pixel-wise monthly semantic segmentation labels of 7 land use and land cover (LULC) classes. DynamicEarthNet is the first dataset that provides this unique combination of daily measurements and high-quality labels. In our experiments, we compare several established baselines that either utilize the daily observations as additional training data (semi-supervised learning) or multiple observations at once (spatio-temporal learning) as a point of reference for future research. Finally, we propose a new evaluation metric SCS that addresses the specific challenges associated with time-series semantic change segmentation. The data is available at: https://mediatum.ub.tum.de/1650201.

Making peace with nature is the defining task of the 21st century.

António Guterres, UN Secretary General

* Authors share first authorship. † Authors share senior authorship. ‡ Corresponding author: xiaoxiang.zhu@dlr.de.

1. Introduction

Society is rapidly becoming more aware of the human footprint on the world’s climate. Overwhelming evidence shows that climate change has both short-term and long-term effects on almost every aspect of our lives [27]. Using simulations and global climate metrics, it is nowadays possible to observe changes at a global scale, like the rising sea levels or changes of the gulf stream. In contrast, precise predictions of local changes are much harder to obtain. Common examples include land use by agriculture, deforestation, flooding, wildfires, growth of urban areas, and transportation infrastructure. It is of critical importance to monitor such local changes since these are the factors that ultimately exacerbate the global climate crisis.

Satellite images are a powerful tool in this context to track local changes to the environment in specific regions. Observing change at a local scale requires two conditions: high frequency of satellite observations and pixel-precise understanding of the observed surface. Existing datasets often fail to provide these conditions. Whenever pixel-wise annotations are provided, only static images can be used [43] or the revisit frequency is limited to once a year [14, 36]. Datasets with coarser annotations have either an irregular [11] or monthly revisit frequency [38]. As an example of land changes, in 2020, 46 km$^2$ of the rainforest in Brazil were destroyed every day [29]. This suggests that if we analyze the satellite images of that area once per month, we potentially miss deforestation of the equivalent of the city of Los Angeles, California. As Brazil alone has...
millions of square kilometers of forest, automatic methods are required to detect these and other kinds of land changes. Current pixel-precise automatic methods are predominantly based on deep learning and thus require annotated data to learn.

In this work, we present DynamicEarthNet, a time-series satellite imagery dataset with daily revisits of 75 local regions across the globe. The dataset comprises consistent, occlusion-free observations with multi-spectral imagery over the span of two years (2018-2019). We further provide annotated monthly semantic segmentation labels. The main focus is to segment and detect changes in the development of general land use and land cover (LULC). Specifically, we focus on the following LULC classes: impervious surfaces, water, soil, agriculture, wetlands, snow & ice, and forest & other vegetation.

In comparison to semantic segmentation on standard computer vision benchmarks, satellite imagery is subject to various additional challenges. Most prominently, labeled areas in satellite images typically have very intricate shapes that are significantly more complex than everyday objects. We show that well-performing methods [10, 32] on standard vision benchmarks do not necessarily transfer well to this domain. Furthermore, common segmentation metrics are not optimal for quantifying the performance on the task of semantic change segmentation. We alleviate this issue by proposing a new evaluation protocol that captures the essence of semantic change segmentation. DynamicEarthNet and the proposed evaluation protocol encourage the development of more specialized algorithms that can handle the particular challenges of daily time-series satellite imagery. In summary, our contributions are as follows:

- We present a large-scale dataset of multi-spectral satellite imagery with daily observations of 75 separate areas of interest around the globe.
- We provide dense, monthly annotations of 7 land use and land cover (LULC) semantic classes.
- We propose a novel evaluation protocol that models two central properties of semantic change segmentation: binary change and semantic segmentation.
- We evaluate multiple baseline approaches on our data for the task of detecting semantic change. We show how the time-series nature of our data can be leveraged for optimal performance.

2. Related work

For our discussion of related work, we provide an overview of publicly available satellite imagery datasets, see also Tab. 1. Furthermore, we summarize existing work on the tasks of semantic segmentation and change detection.

2.1. Earth observation datasets

Segmentation and detection. Semantic segmentation of land cover classes for satellite imagery was originally pioneered by the ISPRS project [30, 37]. Similarly, the DeepGlobe [15] and SpaceNet [39] challenges provide datasets for building detection, road extraction, and land cover classification. In contrast to ours, such early works have a relatively small number of areas of interest.

Subsequently, the main focus started to shift towards large-scale aerial imagery [43, 46]. To that end, DOTA [46] proposes to detect objects on a large collection of images cropped from Google Earth. iSAID [43] extends this concept to the task of instance segmentation. Along the same lines, SpaceNet MVOI [44] proposes a benchmark on building detection for multi-view satellite imagery. Our benchmark, on the other hand, provides semantic annotations that are dense, i.e. defined for every single pixel.

Change detection. Several works aim at predicting change between observations of the same area of interest at different times. Most relevant datasets focus on binary change detection which is agnostic to specific types of change [3, 13]. HRSCD [14] and Hi-UCD [36] propose a multi-class semantic change detection datasets. In comparison to time-series data, these benchmarks show only one observation per year, for 2-3 years in total, rather than a full sequence. Moreover, the diversity is limited – HRSCD [14] and Hi-UCD [36] cover specific regions of France and Tallinn, Estonia, respectively. More recently, QFabric [41] presented a large-scale multi-temporal dataset, with polygonal annotations for change regions. In contrast, our dataset contains daily observations and pixel-wise LULC classes.

Time-series analysis. In recent times, time-series satellite datasets gained increasing attention [11, 31, 38]. For instance, Earthnet2021 [31] presents a surface forecasting dataset based on public Sentinel-2 imagery with a revisit rate of 5 days. Since the intended applications are quite dissimilar to ours, no land cover annotations are provided. fMoW [11] provides temporal satellite imagery with bounding box annotations. Similarly, MUDS [38] aims at monitoring urbanization by tracking buildings for several areas of interest that are annotated with polygons. Varying acquisition conditions make it challenging to consistently collect data over an extended period of time. Consequently, existing datasets often contain irregular revisit frequencies [11] or infrequent (monthly) observation intervals [38]. In contrast, our DynamicEarthNet dataset provides high-quality, consistent daily observations.

2.2. Considered tasks

Semantic segmentation. There are countless recent deep learning methods [2, 8–10, 24, 32, 42] that address gen-
Change detection. Change detection is an extensively studied topic in earth observation. Classical approaches define axiomatic, pixel-based [4–6, 20, 35] algorithms to obtain change whereas many recent approaches are data-driven [7, 12, 33, 47]. The development of new algorithms is often inhibited by a lack of high-quality data and expert annotations. Most methods focus on binary change and are usually limited to two distinct observations in time (bitemporal) [4–7, 20, 35, 47]. Moreover, datasets and metrics used for evaluation differ widely and are often not public.

These considerations underline the necessity for a standardized benchmark with a consistent evaluation protocol. Up to now, there are few approaches suitable for multi-class change detection. Most of them typically consider two snapshots, often years apart. Among these works, [25, 28] directly predict the multi-class change map whereas, [36] define change as the difference between two semantic maps. We follow the latter approach in our evaluations since existing work on multi-class change detection is not primarily designed to handle high temporal frequencies. Therefore, we benchmark state-of-the-art semantic segmentation algorithms on our dataset and compare differences in the predicted multi-class semantic masks over time.

### Table 1. An overview of public satellite datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Temporal</th>
<th>Revisit Time</th>
<th># Images</th>
<th>Sources</th>
<th>GSD (m)</th>
<th>Annotation</th>
<th>Objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>SpaceNet [39]</td>
<td>✗</td>
<td>✗</td>
<td>&gt;24,586</td>
<td>Maxar</td>
<td>0.31</td>
<td>Polygon</td>
<td>Buildings and Roads</td>
</tr>
<tr>
<td>DOTA [46]</td>
<td>✗</td>
<td>✗</td>
<td>2,806</td>
<td>Google Earth</td>
<td>0.15-12†</td>
<td>Oriented Bbox</td>
<td>Various</td>
</tr>
<tr>
<td>fMoW [11]</td>
<td>✓</td>
<td>irregular</td>
<td>&gt;1,000,000</td>
<td>Maxar</td>
<td>0.31-1.60</td>
<td>BBox</td>
<td>Various</td>
</tr>
<tr>
<td>SpaceNet MVOI [44]</td>
<td>✗</td>
<td>✗</td>
<td>60,000</td>
<td>Maxar</td>
<td>0.46-1.67</td>
<td>Polygon</td>
<td>Buildings</td>
</tr>
<tr>
<td>MUDS [38]</td>
<td>✓</td>
<td>monthly</td>
<td>2,389</td>
<td>Planet</td>
<td>4.0</td>
<td>Polygon</td>
<td>Buildings</td>
</tr>
<tr>
<td>DOTA-v2.0 [16]</td>
<td>✓</td>
<td>✗</td>
<td>11,268</td>
<td>Google Earth</td>
<td>0.15-12†</td>
<td>Oriented Bbox</td>
<td>Various</td>
</tr>
<tr>
<td>DeepGlobe [15]</td>
<td>✓</td>
<td>✓</td>
<td>1,146</td>
<td>Maxar</td>
<td>0.3</td>
<td>Seg. Mask</td>
<td>Various LULC</td>
</tr>
<tr>
<td>iSAID [43]</td>
<td>✗</td>
<td>✗</td>
<td>2,806</td>
<td>DOTA</td>
<td>0.15-12†</td>
<td>I. Seg Mask</td>
<td>Various</td>
</tr>
<tr>
<td>HRSCD [14]</td>
<td>✓</td>
<td>yearly</td>
<td>582</td>
<td>BD ORTHO</td>
<td>0.5</td>
<td>Seg. Mask</td>
<td>Various LULC</td>
</tr>
<tr>
<td>Hi-UCD [36]</td>
<td>✓</td>
<td>yearly</td>
<td>2,586</td>
<td>ELB†</td>
<td>0.1</td>
<td>Seg. Mask</td>
<td>Various LULC</td>
</tr>
<tr>
<td>DynamicEarthNet</td>
<td>✓</td>
<td>daily</td>
<td>54,750</td>
<td>PlanetFusion</td>
<td>3.0</td>
<td>Seg. Mask</td>
<td>Various LULC</td>
</tr>
</tbody>
</table>

† Estonian Land Board. † Google Earth gathers information from various sensors, so the resolution is diverse [44].

Table 2. LULC class distribution. The distribution of LULC classes averaged over all $24 \times 75 = 1800$ semantic maps in the dataset. Additionally, we report the absolute number of AOIs with any occurrences of a given LULC class. We visualize the colors we use for each class throughout the paper.

<table>
<thead>
<tr>
<th>class name</th>
<th>%</th>
<th>#AOIs</th>
<th>color</th>
</tr>
</thead>
<tbody>
<tr>
<td>impervious surface</td>
<td>7.1</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>agriculture</td>
<td>10.3</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>forest &amp; other vegetation</td>
<td>44.9</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td>wetlands</td>
<td>0.7</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>soil</td>
<td>28.0</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>water</td>
<td>8.0</td>
<td>58</td>
<td></td>
</tr>
<tr>
<td>snow &amp; ice</td>
<td>1.0</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

3. The DynamicEarthNet dataset

We present the DynamicEarthNet dataset that contains daily, cloud-free satellite data acquired from January 2018 to December 2019. It consists of images from 75 areas of interest (AOIs) across the globe, as illustrated by the world map in Fig. 1. The dataset covers a wide variety of environments with diverse types of land cover changes. For each region, we provide a sequence of images with daily revisits. Furthermore, we present pixel-wise semantic labels for the first day of each month. These serve as ground-truth to define land cover changes over the span of two observed years. In the remainder of this section, we provide details on the imagery, semantic labels, and statistics of the dataset.

3.1. Multi-spectral imagery

The primary source of our dataset is the Fusion Monitoring product from Planet Labs, which provides multi-
spectral time-series satellite imagery. Each snapshot contains four channels (RGB + near-infrared) with a ground sample distance (GSD), i.e., pixel granularity, of 3 meters and a resolution of 1024x1024.

Beyond the raw observational data, Planet applies a combination of post-processing techniques to ensure data quality and consistency: For once, all images are processed to remove occlusions by weather, overcast and related visual artifacts. The data is gap-filled, which means that missing information due to cloud coverage is filled with suitable observations from the closest available point in time. Moreover, the Fusion bands are calibrated to the Harmonized Landsat-Sentinel (HLS)\(^2\) spectrum to make them compatible with other publicly available datasets such as Landsat 8 [45] or Sentinel 2 [1, 17].

To encourage the exploration of data fusion, we provide monthly Sentinel-2 (S2) imagery from the same 75 AOIs for reference. The main idea of this auxiliary set of images is to allow for comparisons with publicly available data. Moreover, the additional data potentially gives rise to interesting multi-modal settings in future experiments. For more details, we refer the reader to our supplementary material.

3.2. Pixel-wise labels

Having described the raw satellite imagery, we now provide more details on the monthly ground-truth annotations. They comprise a collection of pixel-wise semantic segmentation labels corresponding to the first day of each month. These labels are defined as the common LULC classes, i.e., impervious surfaces, agriculture, forest & other vegetation, wetlands, soil, water, snow & ice. The resolution of each annotation is 1024x1024 with a pixel granularity of 3 meters, just like the corresponding satellite images.

The annotation procedure was rigorous with an emphasis on the temporal consistency of the labels. The first image was manually annotated for each AOI and used as a basis for the following months. Subsequent maps are updated if there is a perceptible change in a certain region that is evident to the human annotator. Three quality control gates, each with a different annotator, ensure accurate annotations, topological correctness, and format correctness, respectively.

3.3. Dataset statistics

The DynamicEarthNet dataset contains 75 different AOIs across the globe, each of which consists of a sequence of 730 images covering two years from January 2018 to December 2019. We provide semantic LULC classes for the first day of each month, 24 per sequence in total. In total, this amounts to 54750 satellite images and 1800 ground-truth annotations.

We illustrate the distribution of LULC classes over the whole dataset in Tab. 2. Due to the nature of the data, occurrences of certain semantic classes are imbalanced with forest & other vegetation and soil dominating less frequent classes like wetlands. Such general ambient classes often take up large portions of a considered region, see the bottom third of the images in Fig. 2.

We split our data into train, validation, and test sets with 55, 10, and 10 AOIs, respectively. The number of distinct classes per AOI ranges from 2 to 6. For instance, some AOIs from the dataset contain only forest & other vegetation and soil, whereas others include impervious surfaces, water, soil, agriculture, wetlands, and forest & other vegetation. No single AOI contains all 7 classes. For an optimal balance, we ensure that the splits’ classes are distributed as equally as possible. We refrain from providing more fine-grained statistics on the class distribution to avoid disclos-
ing any additional information on the (currently concealed) test set. Since the snow & ice class occurs in only 2 cubes, see Tab. 2, we have no such examples in the validation or test sets. Consequently, we also do not consider this class in our quantitative evaluations presented in Sec. 5.

3.4. Advantages over existing benchmarks

In comparison to other publicly available, annotated satellite datasets, DynamicEarthNet has a number of crucial distinguishing features, see Tab. 1. First and foremost, it is the first to provide daily observations from a large diversity of AOs. The closest work to ours in terms of revisit rates is [38] with monthly observations. Yet, they have a narrower focus with the main objective of tracking buildings to monitor urbanization. Other related change detection datasets [14, 36, 41] show merely one observation per year, see Tab. 1. In our dataset, we provide consistent daily observations for two years allowing the study of both short-term and long-term change. Fig. 2 highlights the potential of such data: We can observe the change of new buildings being built day by day. At the same time, we can pin down exact dates of deforestation, and successively observe long-term effects over the span of multiple months.

4. Semantic change segmentation

One key application of our dataset is to measure how a given local region changes over time. For the standard task of binary change detection, we classify each pixel into change or no-change. This definition, however, disregards semantic information. We, therefore, generalize this classical notion to a multi-class segmentation task, which we refer to as semantic change segmentation.

For time-series satellite data, changes are usually caused by external forces, such as weather and climate effects or human destruction and creation. Compared to standard vision benchmarks, they often appear gradually over time and with a limited spatial extent. When predicting semantic labels for a whole observed region, such rare changes between frames have a low influence on the overall segmentation score. In our dataset, only 5% of all pixels change from month to month on average. Hence, standard evaluation metrics defined on the full image like the Jaccard index (IoU) are not suitable to express how accurately semantic classes of changed areas are predicted. We, therefore, propose a new metric to quantify the performance of methods in semantic change segmentation of satellite images.

4.1. Problem definition

Let \( x \in \mathbb{R}^{T \times H \times W \times 4} \) be an input time-series of satellite images consisting of \( T \) frames with a spatial size of \( H \times W \) and 4 input channels (RGB + near-infrared). For each such time-series, we further provide semantic annotations \( y \in C^{T \times H \times W} \) that assign each pixel in \( x \) to one of the 7 LULC classes \( C := \{0, \ldots, 6\} \) defined in Sec. 3.2. Given two consecutive frames at times \( t \) and \( t + 1 \), we can define the binary change \( b \in \{0, 1\}^{(T-1) \times H \times W} \) as a binary labeling of all pixels for which the ground-truth semantic label changes:

\[
b_{t,i,j} := \begin{cases} 1, & \text{if } y_{t,i,j} \neq y_{t-1,i,j}, \\ 0, & \text{else.} \end{cases}
\]  

(1)

When evaluating semantic change segmentation, both the binary change map \( b \) and the semantic map \( y \) need to be predicted. This requires methods to answer which pixels change and what class do these pixels change to.

4.2. Evaluation protocol

There are two distinct types of errors that are common in the context of semantic change segmentation: failing to detect the binary change and predicting the wrong semantic class for a changed pixel. Our goal is to design an evaluation protocol that captures both of these errors in a single signal. Thus, the resulting semantic change segmentation (SCS) metric consists of two components, a class-agnostic binary change score (BC) and a semantic segmentation score among changed pixels (SC).

Binary change (BC). The standard approach to measure the quality of a predicted change map \( \hat{b} \) is comparing its overlap with the ground-truth change \( b \). This is commonly defined as the Jaccard index or intersection-over-union score:

\[
\text{BC}(b, \hat{b}) = \frac{|\{b = 1\} \cap \{\hat{b} = 1\}|}{|\{b = 1\} \cup \{\hat{b} = 1\}|}
\]  

(2)

where we use the short hand-notation

\[
\{c = 1\} := \{(t, i, j) | b_{t,i,j} = 1\}
\]  

(3)

for the indicator set of indices with binary change.

Semantic change (SC). The second component of our metric measures semantic change accuracy. It is defined as the segmentation score, conditioned on the set of pixels where any change occurs in the ground-truth maps, i.e., \( b = 1 \). On this subset of pixels, we compute the Jaccard index between the ground-truth labels \( y \) and predicted labels \( \hat{y} \) (averaged over all classes \( c \)):

\[
\text{SC}(y, \hat{y} | b) = \frac{1}{|C|} \sum_{c \in C} \frac{|\{b = 1\} \cap (\{y = c\} \cap \{\hat{y} = c\})|}{|\{b = 1\} \cap (\{y = c\} \cup \{\hat{y} = c\})|}
\]  

(4)

Semantic change segmentation (SCS). The total SCS score is the arithmetic mean of the binary change and the semantic change:

\[
\text{SCS}(y, \hat{y}) = \frac{1}{2} \left( \text{BC}(b, \hat{b}) + \text{SC}(y, \hat{y} | b) \right).
\]

(5)
In practice, we first accumulate confusion matrices of all time-series before computing the final SCS score.

**Metric properties.** In the following, we summarize a few distinguishing features of the proposed SCS metric.

i. **Focus on change.** In comparison to standard metrics, like the Jaccard index, the SCS metric specifically emphasizes accurate change predictions.

ii. **Separation of errors.** It separates the problems of detecting areas where change occurs (BC) and predicting the correct semantic labels for changed areas (SC).

iii. **Single output signal.** Both signals contribute equally to the final SCS score.

5. Experiments

In this section, we demonstrate the utility of our dataset with various experiments on land cover segmentation and semantic change segmentation. We first give an overview of considered baseline methods in Sec. 5.1 and then present corresponding results in Sec. 5.2 and Sec. 5.3.

5.1. Baselines

*DynamicEarthNet* contains daily images and dense semantic annotations for the first day of each month. This raises the question of how one can leverage additional unlabelled examples to improve the results when training on the labeled data. We study two separate approaches in this work: spatio-temporal and semi-supervised semantic segmentation. The former addresses the time-series nature of our data by combining spatial information with temporal architectures. The latter uses the annotated images (first day of each month) as supervision while taking advantage of the additional unlabeled samples in an unsupervised manner.

**Spatio-temporal baselines.** The first class of baselines we consider are spatio-temporal methods. The main idea is to fuse individual observations of an input time series and produce a single output prediction – the monthly semantic map. As a backbone, we use the U-Net feature extractor [32]. Following [26, 34], we compare different temporal architectures. First, we apply a U-ConvLSTM network [26]. As a second method, we utilize 3D convolutions that process spatial and temporal information at once [26]. Finally, we employ U-TAE [34] that encodes temporal features in the latent space via self-attention [40].

**Semi-supervised baselines.** As an alternative to modeling the input images as sequences, we can interpret them as an unordered collection of training samples. Analogous to standard supervised learning, the labeled examples are used directly as training data. To extract information from the remaining set of unlabeled training examples, we employ the recent state-of-the-art consistency-based semi-supervised segmentation method by Lai *et al.* [21]. The main idea is to randomly crop unlabeled images into pairs of patches and enforce consistent outputs for the overlap of both sub-regions. Robustness to varying contexts is crucial for our data since the surrounding of an overlapping region is generally an unreliable predictor for its class label. For example, water occurs in quite different environmental contexts in our dataset, like forests, agriculture, or impervious surfaces. We evaluate this method [21] with the segmentation backbone DeepLabv3+ [10].

5.2. Land cover and land use segmentation

The first task we consider is semantic segmentation of land cover classes. Specifically, the goal is to predict one of the LULC labels described in Sec. 3.2. We compare the performance of the two classes of baseline methods discussed in the previous section. For each setting, we evaluate the intersection-over-union score averaged over all 6 evaluation LULC classes (mIoU). Due to its overall scarcity, we exclude the snow & ice class from the evaluations, see Sec. 3.3 for more details.

**Spatio-temporal results.** Results of spatio-temporal methods are summarized in Tab. 3. As a first reference point, we consider the purely supervised setting. Here, we train a standard U-Net architecture only on the monthly labeled samples. It achieves 33.5% mIoU on the validation and 37.6% mIoU on the test set.

We further assess whether existing spatio-temporal architectures benefit from the time-series nature of our data. All three considered architectures improve the performance over the supervised baseline for weekly temporal inputs on the validation set. U-TAE and U-ConvLSTM show the strongest generalization performance on the test set.

On the other hand, when using daily sequences of 28–31 images, the performance drops considerably. This suggests that generic spatio-temporal techniques are not necessarily optimal for extracting information from daily satellite data. The individual images of such daily time series are often highly correlated. Consequently, when labeled data is limited, increasing the length of a sequence at some point leads to unstable training. For our benchmark, using weekly samples is optimal for the considered baselines. We conclude that more specialized techniques are needed to allow for robust learning on daily time-series satellite imagery.

**Semi-supervised results.** We report the performances of our the baseline [21] in combination with DeepLabv3+ [10] in Tab. 4. Similar to the spatio-temporal experiments, we consider different temporal densities. For the purely supervised setting, all unlabeled images are discarded (monthly). Additionally, we compare different semi-supervised settings with 6 (weekly), 28–31 (daily) unlabeled samples per month. Both, daily and weekly data help to improve over
Table 3. **Quantitative results of spatio-temporal methods.** We compare the performance of different spatio-temporal architectures on the task of LULC segmentation. Individual values denote the intersection-over-union score for individual classes (cols. 3-8), as well as the averaged scores over the whole validation set (9th col.) and test set (10th col.). The monthly U-Net baseline is generally less accurate than the considered temporal architectures.

<table>
<thead>
<tr>
<th>Method</th>
<th>Frequency</th>
<th>Imp. Surface</th>
<th>Agriculture</th>
<th>Forest</th>
<th>Wetlands</th>
<th>Soil</th>
<th>Water</th>
<th>Val mIoU (↑)</th>
<th>Test mIoU (↑)</th>
</tr>
</thead>
<tbody>
<tr>
<td>U-Net [32]</td>
<td>monthly</td>
<td>28.6</td>
<td>6.9</td>
<td>76.4</td>
<td>0.0</td>
<td>38.4</td>
<td>50.5</td>
<td>33.5</td>
<td>37.6</td>
</tr>
<tr>
<td>U-TAE [34]</td>
<td>weekly</td>
<td>31.8</td>
<td>8.0</td>
<td>77.3</td>
<td>0.0</td>
<td>39.1</td>
<td>58.1</td>
<td>35.7</td>
<td>39.7</td>
</tr>
<tr>
<td></td>
<td>daily</td>
<td>26.3</td>
<td>6.5</td>
<td>73.7</td>
<td>0.0</td>
<td>35.7</td>
<td>51.2</td>
<td>32.2</td>
<td>36.1</td>
</tr>
<tr>
<td>U-ConvLSTM [26]</td>
<td>weekly</td>
<td>31.4</td>
<td>2.2</td>
<td>77.7</td>
<td>0.0</td>
<td>36.1</td>
<td>58.6</td>
<td>34.3</td>
<td>39.1</td>
</tr>
<tr>
<td></td>
<td>daily</td>
<td>14.4</td>
<td>0.6</td>
<td>72.1</td>
<td>0.0</td>
<td>32.0</td>
<td>58.8</td>
<td>29.7</td>
<td>30.9</td>
</tr>
<tr>
<td>3D-Unet [26]</td>
<td>weekly</td>
<td>32.4</td>
<td>2.1</td>
<td>77.4</td>
<td>0.0</td>
<td>35.3</td>
<td>65.5</td>
<td>35.5</td>
<td>37.2</td>
</tr>
<tr>
<td></td>
<td>daily</td>
<td>31.1</td>
<td>1.8</td>
<td>75.8</td>
<td>0.0</td>
<td>34.1</td>
<td>66.0</td>
<td>34.8</td>
<td>38.8</td>
</tr>
</tbody>
</table>

Table 4. **Quantitative results of semi-supervised methods.** The table shows the semantic segmentation results of using the context-aware consistency-based semi-supervised approach [21] on our DynamicEarthNet dataset. We further present the IoU scores per class for the validation set. ‘Monthly’ indicates that the architecture is trained in a supervised manner. Using unlabelled satellite images improves the results over the fully supervised baseline.

<table>
<thead>
<tr>
<th>Method</th>
<th>All labelled?</th>
<th>Imp. Surface</th>
<th>Agriculture</th>
<th>Forest</th>
<th>Wetlands</th>
<th>Soil</th>
<th>Water</th>
<th>Val mIoU (↑)</th>
<th>Test mIoU (↑)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAC [21]</td>
<td>monthly ✓</td>
<td>18.1</td>
<td>4.8</td>
<td>74.7</td>
<td>0.0</td>
<td>33.9</td>
<td>55.9</td>
<td>31.2</td>
<td>37.9</td>
</tr>
<tr>
<td></td>
<td>weekly ✗</td>
<td>28.0</td>
<td>7.2</td>
<td>75.7</td>
<td>8.3</td>
<td>38.9</td>
<td>51.0</td>
<td>34.9</td>
<td>37.9</td>
</tr>
<tr>
<td></td>
<td>daily ✗</td>
<td>28.9</td>
<td>4.0</td>
<td>75.5</td>
<td>0.5</td>
<td>39.0</td>
<td>55.6</td>
<td>33.9</td>
<td>43.6</td>
</tr>
</tbody>
</table>

Table 5. **Quantitative results of semantic change segmentation on our test set.** This table shows semantic change segmentation results of all methods on our DynamicEarthNet dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>SCS (↑)</th>
<th>BC (↑)</th>
<th>SC (↑)</th>
<th>mIoU (↑)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAC [21]</td>
<td>17.7</td>
<td>10.7</td>
<td>24.7</td>
<td>37.9</td>
</tr>
<tr>
<td>U-Net [32]</td>
<td>17.3</td>
<td>10.1</td>
<td>24.4</td>
<td>37.6</td>
</tr>
<tr>
<td>U-TAE [34]</td>
<td>19.1</td>
<td>9.5</td>
<td>28.7</td>
<td>39.7</td>
</tr>
<tr>
<td>U-ConvLSTM [26]</td>
<td>19.0</td>
<td>10.2</td>
<td>27.8</td>
<td>39.1</td>
</tr>
<tr>
<td>3D-Unet [26]</td>
<td>17.6</td>
<td>10.2</td>
<td>25.0</td>
<td>37.2</td>
</tr>
<tr>
<td>CAC [21]</td>
<td>18.5</td>
<td>10.3</td>
<td>26.7</td>
<td><strong>43.6</strong></td>
</tr>
<tr>
<td>U-TAE [34]</td>
<td>17.8</td>
<td>10.4</td>
<td>25.3</td>
<td>36.1</td>
</tr>
<tr>
<td>U-ConvLSTM [26]</td>
<td>15.6</td>
<td>7.0</td>
<td>24.2</td>
<td>30.9</td>
</tr>
<tr>
<td>3D-Unet [26]</td>
<td>18.8</td>
<td>11.5</td>
<td>26.1</td>
<td>38.8</td>
</tr>
</tbody>
</table>

5.3. **Semantic change segmentation**

In the following, we compare the performance of our considered baseline methods on the metrics that we introduced in Sec. 4, see Tab. 5 for results. Similar to Sec. 5.2, we use different degrees of temporal densities with monthly, weekly, and daily observations. As a general trend, the additional weekly observations improve the performance over the purely supervised, monthly baselines. For the semi-supervised approach [21] the performance on the test set further improves with daily samples. On the other hand, the benefits from additional daily observations are less consistent for spatio-temporal baselines. In this case, increasing the sequence length is inherently subject to a trade-off between providing more information and decreasing the training stability. Since daily observations are highly correlated, optimal results are achieved for a weekly sampling.

Overall, our results suggest that detecting change (BC) is particularly challenging for our considered baselines. Most obtained accuracies are around 10%. Considering that the ground-truth change maps cover only 5% of all pixels on average, there exist a high number of potential false positives. Oftentimes, change occurs between two classes that are visually very similar, like forest & other vegetation to soil. The results further confirm that the mIoU metric alone is not sufficient to measure the performance of semantic change segmentation. A high LULC segmentation score
Figure 3. **Qualitative results on validation set.** Semantic maps (bottom row) of the semi-supervised baseline CAC [21] trained on daily images. The input sequence consists of 5 images (middle row) from September to October, spanning one month. For the first and last semantic map of the considered sequence, we show ground-truth labels (bottom right, bottom left). The three middle columns show predictions of [21]. For each sample, we magnify a specific area to highlight the temporal transition from forest & other vegetation to soil, marked red for ground-truth and pink for baseline predictions [21]. Notably, this development is captured with high fidelity by our baseline [21]. On the other hand, in certain areas, it is not able to distinguish between the generic forest & vegetation class and the ground-truth label agriculture. For the color representation of segmentation maps see Tab. 2.

(mIoU) does not guarantee optimal performance in terms of the change segmentation score (SCS). When compared directly, the semantic change and binary change performance are somewhat decoupled which warrants the split of our SCS metric into binary change BC and semantic change SC.

### 6. Conclusion

We presented DynamicEarthNet, a novel dataset that provides daily, multi-spectral satellite imagery for a broad range of areas of interest. Beyond the raw imagery, it comprises monthly semantic annotations of 7 common LULC classes. This unique combination of dense time-series data and high-quality annotations distinguishes DynamicEarthNet from existing benchmarks, see Tab. 1, which are either temporally sparse or do not provide comparable ground-truth labels. We showed that this gives rise to previously unexplored settings like semi-supervised learning, as well as spatio-temporal methods with an unprecedented temporal resolution. We further devised a new evaluation protocol for semantic change segmentation. It involves several metrics that focus on distinct, common errors in the context of multi-class change prediction. We believe that our benchmark has the potential to spark the development of more specialized techniques that can take full advantage of daily, multi-spectral data. Finally, we highlight in several compelling case-studies how high frequency satellite data can be used to track land cover evolution, e.g. due to deforestation, and assess both its short and long-term effects.

### Acknowledgements

This work is supported by the Humboldt Foundation through the Sofja Kovalevskaja Award, the framework of Helmholtz AI [grant number: ZT-I-PF-5-01] - Local Unit “Munich Unit @Aeronautics, Space and Transport (MASTr)”, the Helmholtz Association under the joint research school “Munich School for Data Science - MUDS”, and the German Federal Ministry for Economic Affairs and Energy (BMWi) under the grant DynamicEarthNet (grant number: 50EE2005).
References


[30] ISPRS Potsdam. 2d semantic labeling dataset, 2018. 2


[34] Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. ICCV, 2021. 3, 6, 7, 13, 15, 17


[37] ISPRS Vaihingen. 2d semantic labeling dataset, 2018. 2


