

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

Aysim Toker^{1,*}, Lukas Kondmann^{1,2,*}, Mark Weber¹, Marvin Eisenberger¹, Andrés Camero²,
Jingliang Hu², Ariadna Pregel Hoderlein¹, Çağlar Şenaras³, Timothy Davis³,
Daniel Cremers¹, Giovanni Marchisio^{3,†}, Xiao Xiang Zhu^{1,2,†,‡}, Laura Leal-Taixé^{1,†}

Technical University of Munich¹, German Aerospace Center², Planet Labs³



Figure 1. **Visualization of the *DynamicEarthNet* dataset.** For a specific area of interest, we show two satellite observations, 2019-08-01 and 2019-08-31, as well as the corresponding monthly ground-truth annotation (top left). The complete dataset consists of daily samples in the range from 2018-01-01 to 2019-12-31. We consider 75 separate areas of interest, spread over six continents (top right).

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

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, 46km^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

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

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 daily 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-

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

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

Table 1. **An overview of public satellite datasets.** For each dataset, we compare key characteristics like the revisit time, the number of images, data source, ground sample distance (GSD), types of annotations, and annotated objects. Most closely related are DeepGlobe [15], iSAID [43], HRSCD [14] and Hi-UCD [36] which, like ours, provide dense semantic annotations for various land cover classes. However, they either provide no time-series data or merely yearly revisit times. Closely related datasets are highlighted in blue and yellow.

eral semantic segmentation. In comparison to most common computer vision applications, segmentation of satellite images is subject to specific challenges, such as irregular sizes and shapes of segmented regions. Recent approaches show that encoder-decoder architectures [18, 23] can help to address the foreground-background imbalance of satellite data [22, 48]. Most existing algorithms focus on segmenting individual, static images. A few works leverage the additional information from time-series satellite images for the case of crop-type classification [19, 26, 34]. We believe that the *DynamicEarthNet* dataset will encourage researchers to develop specialized algorithms that can handle the particular challenges of time-series satellite imagery.

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.

class name	%	#AOIs	color
impervious surface	7.1	70	
agriculture	10.3	37	
forest & other vegetation	44.9	71	
wetlands	0.7	24	
soil	28.0	75	
water	8.0	58	
snow & ice	1.0	2	

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.

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 product¹ from Planet Labs, which provides multi-

¹<https://www.planet.com/pulse/planet-announces-powerful-new-products-at-planet-explore-2020/>



Figure 2. **An example of a changing surface.** We show four sample frames of one AOI from our dataset at different times. Two sub-regions are magnified that highlight two types of change we encounter in practice (top row). The daily nature of our data allows us to observe new buildings being built (green) or to track deforestation (yellow). Additionally, we can monitor the long-term effects of such changes over the span of multiple months, *e.g.* the changes to the forest patch here are persistent.

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.

²<https://earthdata.nasa.gov/esds/harmonized-landsat-sentinel-2>

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 AOIs. 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 $\mathbf{x} \in \mathbb{R}^{T \times H \times W \times 4}$ 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 $\mathbf{y} \in \mathcal{C}^{T \times H \times W}$ that assign each pixel in \mathbf{x} to one of the

7 LULC classes $\mathcal{C} := \{0, \dots, 6\}$ defined in Sec. 3.2. Given two consecutive frames at times t and $t + 1$, we can define the binary change $\mathbf{b} \in \{0, 1\}^{(T-1) \times H \times W}$ as a binary labeling of all pixels for which the ground-truth semantic label changes:

$$\mathbf{b}_{t,i,j} := \begin{cases} 1, & \text{if } \mathbf{y}_{t,i,j} \neq \mathbf{y}_{t-1,i,j}, \\ 0, & \text{else.} \end{cases} \quad (1)$$

When evaluating semantic change segmentation, both the binary change map $\hat{\mathbf{b}}$ and the semantic map $\hat{\mathbf{y}}$ 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 $\hat{\mathbf{b}}$ is comparing its overlap with the ground-truth change \mathbf{b} . This is commonly defined as the Jaccard index or intersection-over-union score

$$\text{BC}(\mathbf{b}, \hat{\mathbf{b}}) = \frac{|\{\mathbf{b} = 1\} \cap \{\hat{\mathbf{b}} = 1\}|}{|\{\mathbf{b} = 1\} \cup \{\hat{\mathbf{b}} = 1\}|} \quad (2)$$

where we use the short hand-notation

$$\{\mathbf{b} = 1\} := \{(t, i, j) \mid \mathbf{b}_{t,i,j} = 1\} \quad (3)$$

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.* $\mathbf{b} = 1$. On this subset of pixels, we compute the Jaccard index between the ground-truth labels \mathbf{y} and predicted labels $\hat{\mathbf{y}}$ (averaged over all classes c):

$$\text{SC}(\mathbf{y}, \hat{\mathbf{y}} | \mathbf{b}) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{|\{\mathbf{b} = 1\} \cap (\{\mathbf{y} = c\} \cap \{\hat{\mathbf{y}} = c\})|}{|\{\mathbf{b} = 1\} \cap (\{\mathbf{y} = c\} \cup \{\hat{\mathbf{y}} = c\})|}. \quad (4)$$

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

$$\text{SCS}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{2} (\text{BC}(\mathbf{b}, \hat{\mathbf{b}}) + \text{SC}(\mathbf{y}, \hat{\mathbf{y}} | \mathbf{b})). \quad (5)$$

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 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) unlabelled samples per month. Both, daily and weekly data help to improve over

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

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.

	All labelled?	<i>per class IoU</i> (\uparrow)						Val mIoU (\uparrow)	Test mIoU (\uparrow)	
		Imp. Surface	Agriculture	Forest	Wetlands	Soil	Water			
CAC [21]	monthly	✓	18.1	4.8	74.7	0.0	33.9	55.9	31.2	37.9
	weekly	✗	28.0	7.2	75.7	8.3	38.9	51.0	34.9	37.9
	daily	✗	28.9	4.0	75.5	0.5	39.0	55.6	33.9	43.6

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.

		SCS (\uparrow)	BC(\uparrow)	SC(\uparrow)	mIoU (\uparrow)
<i>mont.</i>	CAC [21]	17.7	10.7	24.7	37.9
	U-Net [32]	17.3	10.1	24.4	37.6
<i>weekly</i>	CAC [21]	17.8	10.1	25.4	37.9
	U-TAE [34]	19.1	9.5	28.7	39.7
	U-ConvLSTM [26]	19.0	10.2	27.8	39.1
	3D-Unet [26]	17.6	10.2	25.0	37.2
<i>daily</i>	CAC [21]	18.5	10.3	26.7	43.6
	U-TAE [34]	17.8	10.4	25.3	36.1
	U-ConvLSTM [26]	15.6	7.0	24.2	30.9
	3D-Unet [26]	18.8	11.5	26.1	38.8

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.

the supervised baseline. A detailed analysis of these quantitative results shows that the agriculture and wetland classes prove to be difficult for all baselines. Agricultural areas are often confused with forest or soil, see Fig. 3, whereas wetlands get confused with soil and water. This is, to a certain degree, expected due to the visual similarity of these classes. Notably, training on daily data achieves the overall best result. The obtained accuracy is 43.6% mIoU on the test set, with a considerable improvement over the monthly and weekly results of 37.9%.

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

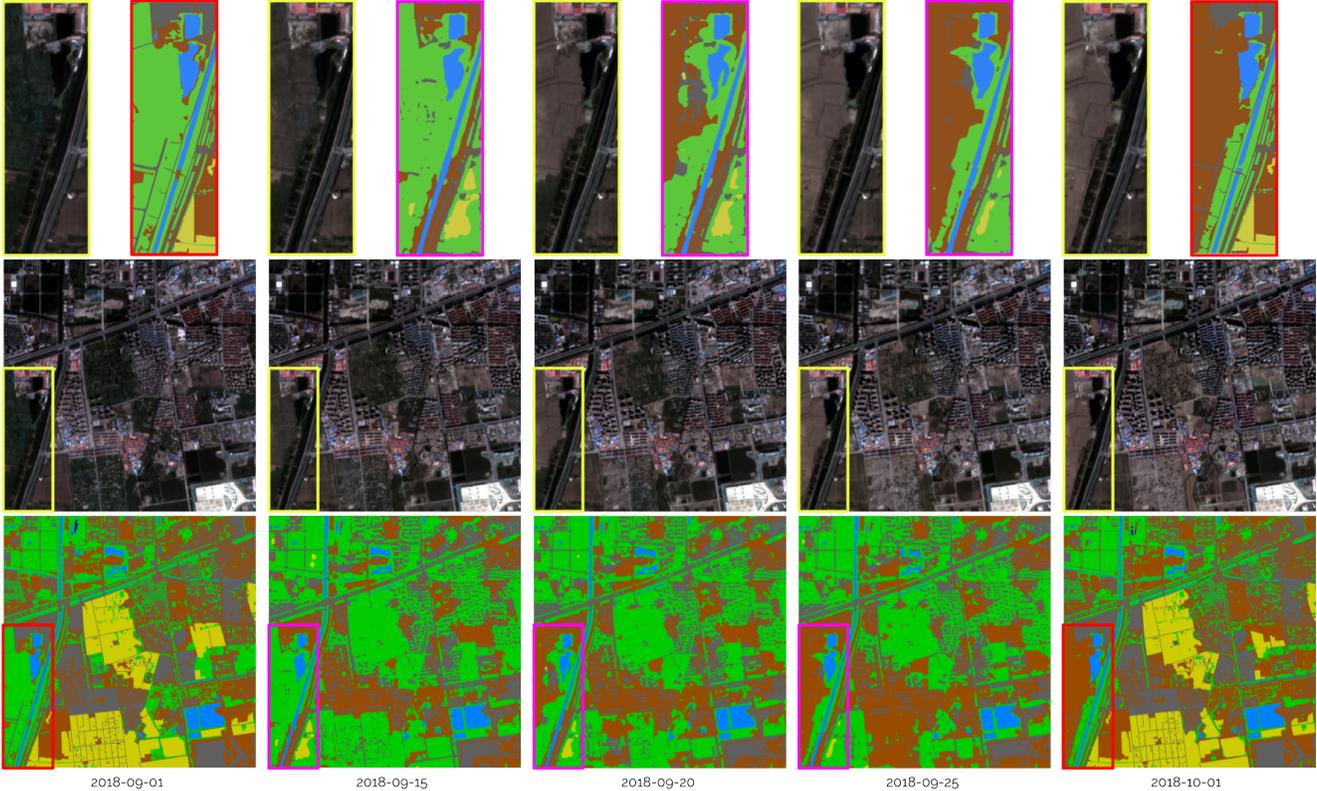


Figure 3. **Qualitative results on validation set.** Semantic maps (bottom row) of the semi-supervised baseline CAC [21] trained on daily images. The input sequence consists of 5 images (middle row) from September to October, spanning one month. For the first and last semantic map of the considered sequence, 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

- [1] Josef Aschbacher. Esa's earth observation strategy and copernicus. In *Satellite earth observations and their impact on society and policy*, pages 81–86. Springer, Singapore, 2017. [4](#), [12](#)
- [2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. [2](#)
- [3] Nicolas Bourdis, Denis Marraud, and Hichem Sahbi. Constrained optical flow for aerial image change detection. In *2011 IEEE International Geoscience and Remote Sensing Symposium*, pages 4176–4179. IEEE, 2011. [2](#)
- [4] Francesca Bovolo. A multilevel parcel-based approach to change detection in very high resolution multitemporal images. *IEEE Geoscience and Remote Sensing Letters*, 6(1):33–37, 2008. [3](#)
- [5] Francesca Bovolo and Lorenzo Bruzzone. A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain. *IEEE Transactions on Geoscience and Remote Sensing*, 45(1):218–236, 2006. [3](#)
- [6] Francesca Bovolo, Silvia Marchesi, and Lorenzo Bruzzone. A framework for automatic and unsupervised detection of multiple changes in multitemporal images. *IEEE Transactions on Geoscience and Remote Sensing*, 50(6):2196–2212, 2011. [3](#)
- [7] Hongruixuan Chen, Chen Wu, Bo Du, and Liangpei Zhang. Change detection in multi-temporal vhr images based on deep siamese multi-scale convolutional networks. *arXiv preprint arXiv:1906.11479*, 2019. [3](#)
- [8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. [2](#)
- [9] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. [2](#)
- [10] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. [2](#), [6](#)
- [11] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018. [1](#), [2](#), [3](#)
- [12] Rodrigo Caye Daudt, Bertr Le Saux, and Alexandre Boulch. Fully convolutional siamese networks for change detection. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 4063–4067. IEEE, 2018. [3](#)
- [13] Rodrigo Caye Daudt, Bertr Le Saux, Alexandre Boulch, and Yann Gousseau. Urban change detection for multispectral earth observation using convolutional neural networks. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 2115–2118. IEEE, 2018. [2](#)
- [14] Rodrigo Caye Daudt, Bertrand Le Saux, Alexandre Boulch, and Yann Gousseau. Multitask learning for large-scale semantic change detection. *Computer Vision and Image Understanding*, 187:102783, 2019. [1](#), [2](#), [3](#), [5](#)
- [15] 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 Workshops*, pages 172–181, 2018. [2](#), [3](#)
- [16] Jian Ding, Nan Xue, Gui-Song Xia, Xiang Bai, Wen Yang, Micheal Ying Yang, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, et al. Object detection in aerial images: A large-scale benchmark and challenges. *arXiv preprint arXiv:2102.12219*, 2021. [3](#)
- [17] Matthias Drusch, Umberto Del Bello, Sébastien Carlier, Olivier Colin, Veronica Fernandez, Ferran Gascon, Bianca Hoersch, Claudia Isola, Paolo Laberinti, Philippe Martimort, et al. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote sensing of Environment*, 120:25–36, 2012. [4](#), [12](#)
- [18] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6399–6408, 2019. [3](#)
- [19] Lukas Kondmann, Aysim Toker, Marc Rußwurm, Andrés Camero, Devis Peressuti, Grega Milcinski, Pierre-Philippe Mathieu, Nicolas Longépé, Timothy Davis, Giovanni Marchisio, et al. Denethor: The dynamicearthnet dataset for harmonized, inter-operable, analysis-ready, daily crop monitoring from space. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. [3](#)
- [20] Lukas Kondmann, Aysim Toker, Sudipan Saha, Bernhard Schölkopf, Laura Leal-Taixé, and Xiao Xiang Zhu. Spatial context awareness for unsupervised change detection in optical satellite images. *arXiv preprint arXiv:2110.02068*, 2021. [3](#)
- [21] Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Semi-supervised semantic segmentation with directional context-aware consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1205–1214, 2021. [6](#), [7](#), [8](#), [14](#), [15](#), [17](#), [18](#)
- [22] Xiangtai Li, Hao He, Xia Li, Duo Li, Guangliang Cheng, Jianping Shi, Lubin Weng, Yunhai Tong, and Zhouchen Lin. Pointflow: Flowing semantics through points for aerial image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4217–4226, 2021. [3](#)

- [23] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. **3**
- [24] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. **2**
- [25] Haobo Lyu, Hui Lu, and Lichao Mou. Learning a transferable change rule from a recurrent neural network for land cover change detection. *Remote Sensing*, 8(6):506, 2016. **3, 14**
- [26] Rose M Rustowicz, Robin Cheong, Lijing Wang, Stefano Ermon, Marshall Burke, and David Lobell. Semantic segmentation of crop type in africa: A novel dataset and analysis of deep learning methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 75–82, 2019. **3, 6, 7, 14, 15, 16, 18**
- [27] V. Masson-Delmotte, P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, L. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekci, R. Yu, and B. Zhou. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2021. **1**
- [28] Lichao Mou, Lorenzo Bruzzone, and Xiao Xiang Zhu. Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 57(2):924–935, 2019. **3, 14**
- [29] A.H. Pickens, M.C. Hansen, B. Adusei, and Potapov P. Sentinel-2 forest loss alert. www.globalforestwatch.org, 2020. Accessed through Global Forest Watch on 11/09/2021. **1**
- [30] ISPRS Potsdam. 2d semantic labeling dataset, 2018. **2**
- [31] Christian Requeena-Mesa, Vitus Benson, Markus Reichstein, Jakob Runge, and Joachim Denzler. Earthnet2021: A large-scale dataset and challenge for earth surface forecasting as a guided video prediction task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1132–1142, 2021. **2**
- [32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. **2, 6, 7**
- [33] Sudipan Saha, Francesca Bovolo, and Lorenzo Bruzzone. Unsupervised deep change vector analysis for multiple-change detection in vhr images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(6):3677–3693, 2019. **3**
- [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, 14, 16, 18**
- [35] Frank Thonfeld, Hannes Feilhauer, Matthias Braun, and Gunter Menz. Robust change vector analysis (rcva) for multi-sensor very high resolution optical satellite data. *International Journal of Applied Earth Observation and Geoinformation*, 50:131–140, 2016. **3**
- [36] Shiqi Tian, Ailong Ma, Zhuo Zheng, and Yanfei Zhong. Hi-ucd: A large-scale dataset for urban semantic change detection in remote sensing imagery. *arXiv preprint arXiv:2011.03247*, 2020. **1, 2, 3, 5, 14**
- [37] ISPRS Vaihingen. 2d semantic labeling dataset, 2018. **2**
- [38] Adam Van Etten, Daniel Hogan, Jesus Martinez Manso, Jacob Shermeyer, Nicholas Weir, and Ryan Lewis. The multi-temporal urban development spacenet dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6407, 2021. **1, 2, 3, 5**
- [39] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018. **2, 3**
- [40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. **6**
- [41] Sagar Verma, Akash Panigrahi, and Siddharth Gupta. Qfabric: Multi-task change detection dataset. In *Earthvision Workshop Computer Vision and Pattern Recognition (CVPR 2021)*, page 10, 2021. **2, 5**
- [42] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020. **2**
- [43] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–37, 2019. **1, 2, 3**
- [44] 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*, pages 992–1001, 2019. **2, 3**
- [45] Curtis E Woodcock, Richard Allen, Martha Anderson, Alan Belward, Robert Bindschadler, Warren Cohen, Feng Gao, Samuel N Goward, Dennis Helder, Eileen Helmer, et al. Free access to landsat imagery. *SCIENCE VOL 320: 1011*, 2008. **4**
- [46] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018. **2, 3**

- [47] Yang Zhan, Kun Fu, Menglong Yan, Xian Sun, Hongqi Wang, and Xiaosong Qiu. Change detection based on deep siamese convolutional network for optical aerial images. *IEEE Geoscience and Remote Sensing Letters*, 14(10):1845–1849, 2017. [3](#)
- [48] Zhuo Zheng, Yanfei Zhong, Junjue Wang, and Ailong Ma. Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4096–4105, 2020. [3](#)

A. Dataset details

In the following, we provide additional details on our dataset. For once, we summarize the metadata complementing our raw sensory data in Appendix A.1. In Appendix A.2 we describe the auxiliary Sentinel-2 [1, 17] images. Finally, we provide additional information on the different sampling densities of our dataset in Appendix A.3.

A.1. Planet metadata

The commercial Planet Fusion data constitutes the core part of the *DynamicEarthNet* dataset. In addition to the surface reflectance values (RGB+near-infrared) that we use in the main paper, Planet provides additional quality assurance (QA) information. The purpose of this is to denote which parts of the data are raw observations and which parts are gap-filled with temporally close observations. For every pixel, the QA product gives the distance and direction to the day of the observation. For example, a pixel value of -1 implies that the pixel has been filled from the previous day.

A.2. Sentinel 2 auxiliary images

Sentinel-2 (S2) images are publicly available through the open data policy of the European Space Agency’s (ESA) Copernicus Program. The mission collects images of all landmasses every 5 days at a resolution of 10m per pixel [17]. While the temporal and spatial resolution of S2 time-series imagery is smaller than the Planet data, S2 collects 13 channels compared to 4 channels of Planet Fusion. In certain scenarios, the additional channels, particularly in the short-wave infrared spectrum, may provide useful auxiliary information about changes on the ground.

In order to encourage cross-research between Planet Fusion and S2 data, we accompany our dataset with monthly images of S2 data from the same locations. The Sentinel-2 images are composite images which means they have been created from multiple S2 images throughout the month. This allows for a direct comparison of the effectiveness of different sources of satellite imagery.

Our Sentinel-2 data is provided as a so-called Bottom-Of-Atmosphere product which includes the correction of distortions to the surface reflectance values caused by atmospheric interference. The S2 pre-processing quality is relatively low compared to the analysis-ready Planet Fusion product. For some areas of interest (AOIs), the collected S2 data suffer from occlusions through cloud coverage for all S2 images in a month. This naturally compromises the quality of the monthly composites. We have collected affected months for all AOIs manually in a designated S2 quality assessment spreadsheet that we provide, together with the dataset. 26% of monthly S2 composites suffer from minor quality issues and around 5% have major quality issues. When the community explores applications of S2 data with

DynamicEarthNet, we advise to investigate whether considered cubes or months are potentially impacted.

A.3. Temporal densities

In our experiments in Sec. 5.2 and Sec. 5.3, we use three different temporal sampling densities for both the spatio-temporal and semi-supervised baselines:

- The monthly setting (fully supervised) shows the first day of each month, resulting in a one-to-one correspondence between input images and labels.
- For the weekly setting, we feed the architectures with samples from the 1st, 5th, 10th, 15th, 20th and 25th days of each month.
- The daily setting uses all the available images in a considered month, as well as the corresponding monthly label.

In Fig. 4, we show the images of 5 time-series with a weekly sampling density.

B. Evaluation protocol details

In the following, we motivate our design choices for the metric proposed in Sec. 4 and compare it to other existing metrics.

B.1. Semantic change

In contrast to semantic segmentation, semantic change segmentation focuses on the changed parts of a given semantic map. Similar to how boundary segmentation restricts evaluation to the boundary pixels, our proposed metric is restricted to changed pixels. We consider several options on how to restrict this subset. In the following, we refer to pixels that have changed their semantic class from one timestep to the next as changed pixels:

- R1.** We restrict the evaluation to the set of changed pixels, as predicted by the considered method.
- R2.** We restrict the evaluation to the set of changed pixels defined by the ground-truth semantic maps.
- R3.** We restrict the evaluation to the intersection of R1 and R2, which is the set of true positives.

Using the set of R1 or R3 has the disadvantage that it couples the semantic change performance with the binary change performance. Only the pixels that are predicted as changed are potentially also evaluated for the semantic change score. Hence, errors in the binary change influence the semantic change score, which potentially opens the metric to misconduct. One can easily imagine a method that reduces the set of predicted change artificially to a single pixel for which the semantic class is predicted with very high confidence. Then the SC score would be perfect (1.0), while the BC score would be close to 0. The resulting overall SCS score would be around 0.5, which is much higher than the scores reported in Tab. 5. Such behavior is completely undesired and leads to a metric that is not aligned

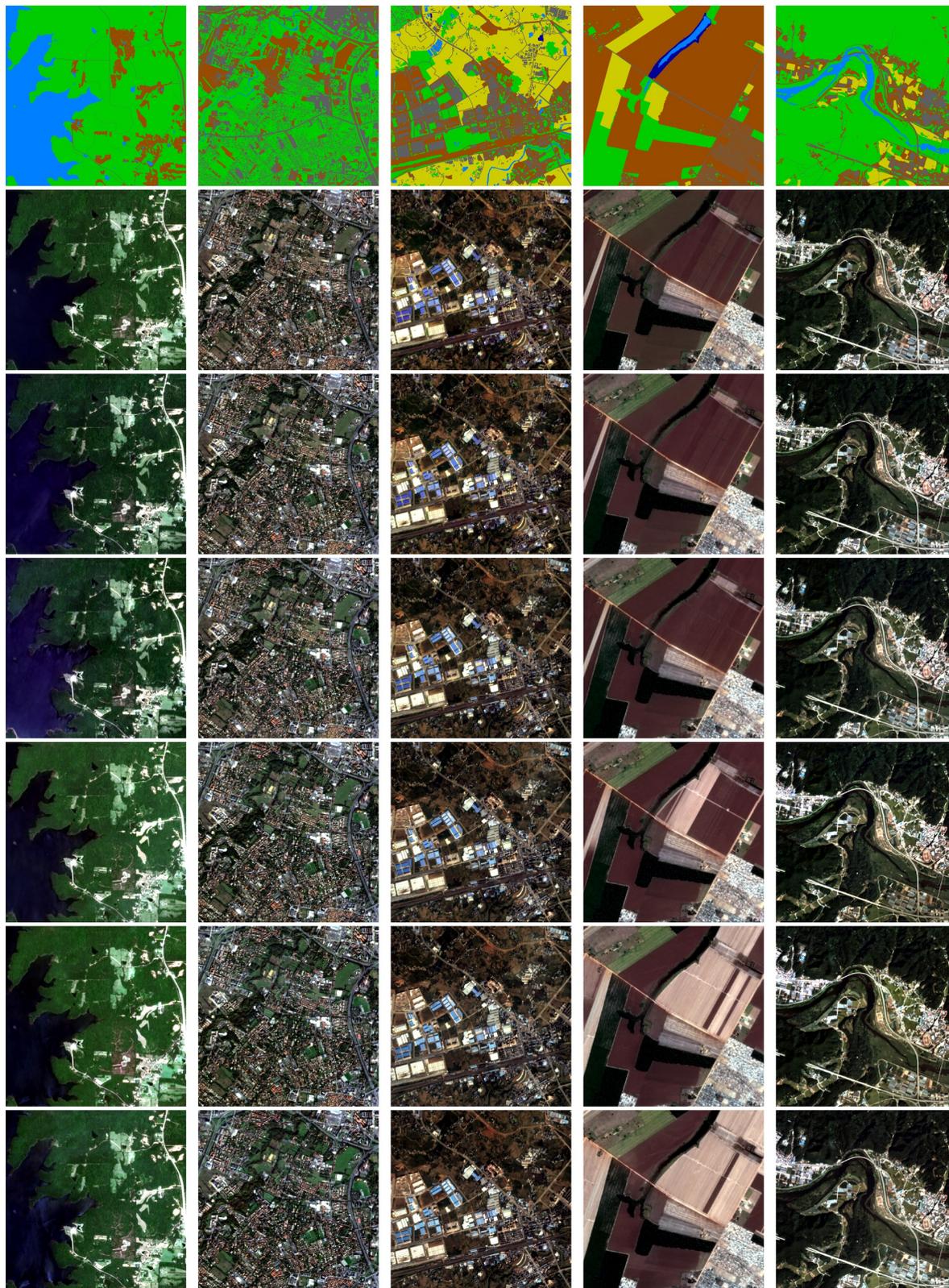


Figure 4. **Training set samples.** We visualize 5 sample time series (one per column) from the training set of the presented *DynamicEarthNet* dataset. Each sequence illustrates weekly samples (row 2-7) and the corresponding annotated monthly labels (1st row).

with human intuition, with results that are hard to interpret. Thus, we use the second option R2 to compute the SC metric. This makes the errors decoupled and the scores easy and intuitive to interpret.

B.2. Comparison

Even though there exists no unified evaluation protocol for semantic change segmentation, there are a few metrics that focus on certain aspects of the task. In the following, we discuss the different options and compare their efficacy for the task of semantic change segmentation.

Pixel accuracy. Pixel accuracy, also referred to as overall accuracy, is one of the simplest measures for (binary) segmentation problems. It is defined as the ratio of correctly classified pixels to all pixels. In settings like ours, in which there are 2 classes for binary change and a high imbalance between them, the pixel accuracy is not able to report meaningful insights. In our setting, 95% of all pixels do not change. Thus, a score of 95% can be obtained by predicting no change all the time. Therefore, we refrain from using pixel accuracy as a metric.

mIoU. The standard mean intersection-over-union addresses the immediate shortcomings of the vanilla pixel accuracy metric. It is possible to use it for both, binary change and semantic change. However, using the mIoU metric for binary change directly, *i.e.* computing the mean IoU of the 2 classes, suffers also from the imbalance issues discussed for the pixel accuracy. Thus, the proposed BC metric computes the IoU of only the change class, rather than both the change and no-change class. For the semantic change, we however apply mIoU, *i.e.* computing the mean over all semantic classes. As explained in the previous subsection, an insightful change metric should focus on the changed regions. We, therefore, refrain from using mIoU on the whole image but compute the scores solely on the changed pixels.

Cohen’s kappa. Previous works [25, 28, 36] have used Cohen’s kappa to measure the performance in similar settings. Cohen’s kappa is a statistical measure of the agreement between the predictions and ground-truth. It is more robust compared to pixel accuracy as it takes the agreement occurring by pure chance into account. However, this measure is not as informative as mIoU. It does not offer insights into the performance of individual classes. Moreover, since scores are not aggregated per class, the performance of classes with high appearance rates will dominate the score and therefore lead to an overall higher score. For more details about the dataset imbalance, we refer to Tab. 2. We thus choose to adapt the well-established IoU measure for our needs.

		SCS (↑)	BC(↑)	SC(↑)
<i>bi-temp</i>	CAC [21]	17.8	10.1	25.4
	U-TAE [34]	19.1	9.5	28.7
	U-ConvLSTM [26]	19.0	10.2	27.8
	3D-Unet [26]	17.6	10.2	25.0
<i>multi-temp</i>	CAC [21]	27.7	23.6	31.8
	U-TAE [34]	27.6	23.4	31.8
	U-ConvLSTM [26]	27.5	24.2	30.7
	3D-Unet [26]	25.3	21.2	29.4

Table 6. **Quantitative results of our metric variant on our test set.** The first row shows the bi-temporal, and the second row shows the multi-temporal results on weekly data. The first row results are identical to the weekly results in Tab. 5.

B.3. Correcting wrong predictions

Our proposed metric requires a separate binary change map $\hat{\mathbf{b}}$ and semantic map $\hat{\mathbf{y}}$. It is therefore not limited to the special case of computing the binary change $\hat{\mathbf{b}}$ directly from the predicted semantic maps for two consecutive timesteps $\hat{\mathbf{y}}_{t-1}$ and $\hat{\mathbf{y}}_t$. This provides additional flexibility, as it is often preferable to decouple the semantic maps from the change predictions [25, 28]. Moreover, it allows for the correction of previous mistakes in online methods that obtain predictions frame-by-frame for an input time-series. As an example, suppose that a semantic class for a certain pixel is predicted wrong at a given timestep. If that pixel does not change in the next timestep, its prediction would either need to keep the wrong semantic class or predict a different semantic class. However, predicting a different semantic class would automatically be recognized as a predicted change, resulting in an error in the binary change. Thus, there is no way to correct previous mistakes without introducing another one. This also holds for other types of errors. By requiring each method to pass explicitly a binary change map $\hat{\mathbf{b}}$ and semantic map $\hat{\mathbf{y}}$, this issue can be avoided. In our setting and the above example, the semantic class can be corrected without predicting a binary change for this pixel. This is especially important for methods that are used for both semantic segmentation as well as semantic change segmentation.

B.4. Discussion on bi-temporal change

In Sec. 4.1, we define the problem as a bi-temporal semantic change segmentation that measures the SCS, SC, and BC scores for a given ground truth \mathbf{y}_t and \mathbf{y}_{t+1} . Given that our dataset contains consistent multi-temporal land use and land cover ground-truth information, it allows us to extend the bi-temporal metric definition and calculate the scores on time intervals of varying lengths. Specifically, we investigate a variant of our bi-temporal semantic change segmentation metrics by measuring the change between all viable pairs of months (t to $t + 1$, $t + 2$, $t + 3$, ...) at each

area of interest ($24 \times 23 = 552$ pairs in total).

We report the resulting accuracies in Tab. 6. For the most part, the modified metric yields slightly higher values than our bi-temporal metric. We attribute this to the fact that the modified metric has a certain smoothing effect, *i.e.* less emphasis is placed on pinpointing the exact frame where change occurs. Throughout our dataset, we notice that different types of changes happen over different time periods (daily, weekly, monthly quarterly, or even seasonally/yearly). On the other hand, the smoothing effect of longer time intervals potentially under-penalizes prediction errors on small time intervals, which goes against one of the main motivations of having daily time-series observations. In our work, we ultimately prefer the bi-temporal setting and leave the detailed multi-temporal discussion as future work.

C. Implementation details

All the experiments are implemented in PyTorch. Our dataset contains 4 spectral bands (RGB + near-infrared). The theoretical valid range for all 4 channels is 1-32,767; however, in practice, the maximum value for the type of data contained in our dataset is 10,000. For data normalization, we calculate the mean and standard deviation per band, averaged over the whole dataset. The exact obtained values are

$$\begin{aligned} \text{mean} &= [1042.59, 915.62, 671.26, 2605.21] & \text{and} \\ \text{std} &= [957.96, 715.55, 596.94, 1059.90], \end{aligned}$$

respectively. For data augmentation, we randomly resize the images with a ratio between $[0.5, 2]$ and crop them to half the input resolution (512, 512). Additionally, we apply random horizontal flips. As we specified in Sec. 3.3, due to the scarcity of the snow & ice class, we do not include them in the test and validation set. For the spatio-temporal architectures, we use the Adam optimizer with a learning rate of $1e - 4$. The batch size is set to 4. We generally train our networks for up to 100 epochs. For the spatio-temporal experiment with daily samples, we use 200 epochs to ensure convergence. The reported results are taken from the epoch that achieves the highest validation accuracy. For the semi-supervised architecture, analogous to [21], we use the SGD optimizer with the poly learning rate decay policy. For both the supervised and unsupervised samples, we use a batch size of 8.

D. Additional qualitative results

Additional visualizations. We present additional qualitative visualizations corresponding to the results in Sec. 5.2. In Fig. 5, we depict a comparison of the different spatio-temporal baselines described in Sec. 5.1. Furthermore, we

compare the effect of different temporal densities for the semi-supervised baseline CAC [21] in Fig. 6. The weekly training achieves the best results on the validation set, as indicated by the results in Tab. 4. This is mostly due to the fact that monthly and daily settings struggle to predict uncommon classes like wetlands (first example in Fig. 6) and agriculture (second example in Fig. 6). Note that these observations are consistent with the confusion matrices shown in Fig. 7.

Confusion matrices. We provide confusion matrices to complement our results on LULC segmentation in Sec. 5.2. The main idea is to allow for a more fine-grained analysis in terms of the 6 semantic classes, see Fig. 7. We show results for both spatio-temporal methods and semi-supervised learning. Each confusion matrix depicts which classes typically get misclassified as certain other classes. For example, the overall uncommon class wetland frequently gets mislabeled as soil, see *e.g.* the first example in Fig. 6. Beyond that, one can also directly read the relative segmentation accuracy of each class in the diagonal entries. As can be expected, the predictions are overall more stable for the more common classes like forest & other vegetation and soil, see Tab. 2 for reference. Among the spatio-temporal methods, the 3D-Unet [26] setting yields the best results for the challenging impervious surface class, see *e.g.* the first example in Fig. 5. All in all, these results indicate that future approaches might benefit from reweighting the individual class labels for a more balanced training that can account for rare LULC classes.

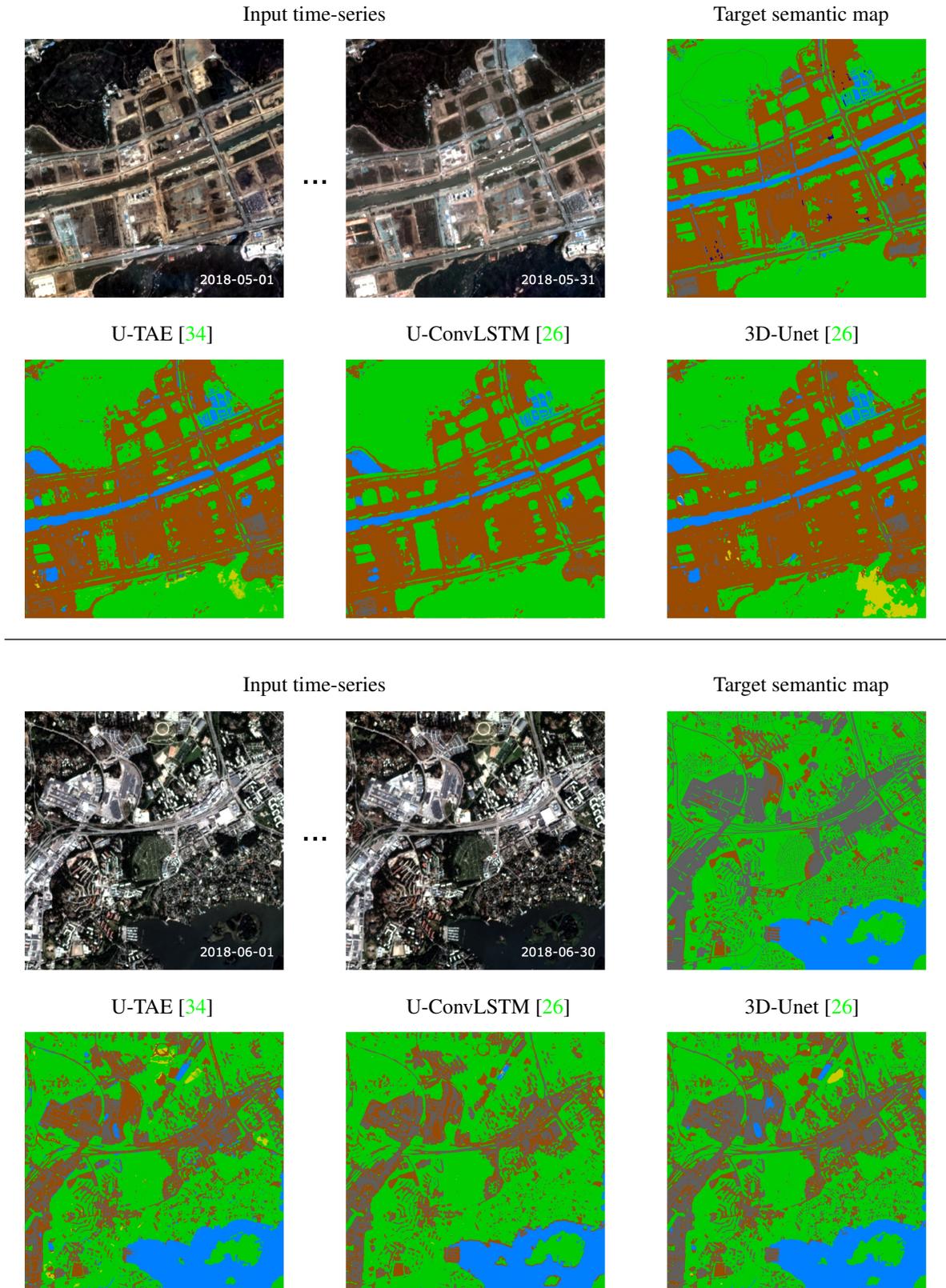


Figure 5. **Spatio-temporal predictions.** We show two qualitative comparisons of the spatio-temporal methods discussed in Sec. 5.1. Both examples are taken from our validation set. The methods take a sequence of 31 and 30 daily samples as inputs (top left) and predict a single semantic map for the whole month (bottom row). We furthermore show the ground-truth annotated map for comparison (top right).

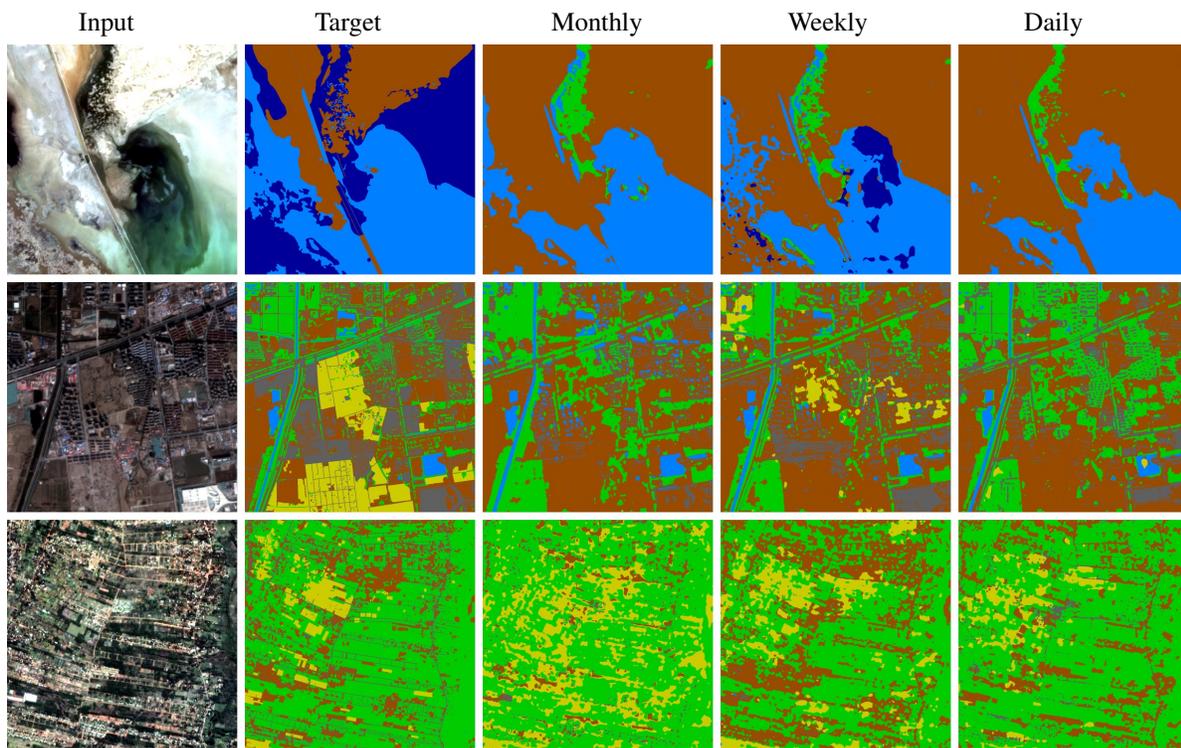


Figure 6. **CAC [21] predictions.** We show sample predictions by the semi-supervised baseline CAC [21] for three different examples from our validation set. For each example, we depict the input sample (1st column), the ground-truth semantic map (2nd column), as well as the predictions of [21] for the monthly, weekly, and daily training setup (3rd-5th column) respectively.

Daily U-TAE [34]



Monthly CAC [21]



Daily U-ConvLSTM [26]



Weekly CAC [21]



Daily 3D-Unet [26]



Daily CAC [21]



Figure 7. **Confusion matrices.** We show confusion matrices corresponding to the LULC segmentation results in Sec. 5.2 on the validation set. The goal is to provide a fine-grained analysis of which classes frequently get misclassified as certain other classes. Each column of an individual confusion matrix is normalized, meaning that it shows the relative distribution of predictions (in percent) for a given, true class. Results are shown for both spatio-temporal (left column) and semi-supervised baselines (right column) with three different settings each.