Satellite Image Time Series Semantic Change Detection: Novel Architecture and Analysis of Domain Shift

Elliot Vincent, Jean Ponce, Mathieu Aubry

Satellite Image Time Series Semantic Change Detection: Novel Architecture and Analysis of Domain Shift

Elliot Vincent[1, 2] Jean Ponce[3, 4] Mathieu Aubry[1]

[1]LIGM, Ecole des Ponts, Univ Gustave Eiffel, CNRS, France
[2]Inria Paris
[3]Department of Computer Science, Ecole normale supérieure (ENS-PSL, CNRS, Inria)
[4]Courant Institute of Mathematical Sciences and Center for Data Science, New York University

Abstract

Satellite imagery plays a crucial role in monitoring changes happening on Earth's surface and aiding in climate analysis, ecosystem assessment, and disaster response. In this paper, we tackle semantic change detection with satellite image time series (SITS-SCD) which encompasses both change detection and semantic segmentation tasks. We propose a new architecture that improves over the state of the art, scales better with the number of parameters, and leverages long-term temporal information. However, for practical use cases, models need to adapt to spatial and temporal shifts, which remains a challenge. We investigate the impact of temporal and spatial shifts separately on global, multi-year SITS datasets using DynamicEarthNet [43] and MUDS [44]. We show that the spatial domain shift represents the most complex setting and that the impact of temporal shift on performance is more pronounced on change detection than on semantic segmentation, highlighting that it is a specific issue deserving further attention. Our complete code is available at https://github.com/ElliotVincent/SitsSCD.

1. Introduction

The surface of the Earth is subject to constant changes, caused by human activity, natural disasters, and many other phenomena. As Earth observation from space has become widely accessible, it is acknowledged as "the most crucial input" [48] and "the best measure available" [16] for climate, ecosystem, and biodiversity monitoring. For example, it has proven useful in assessing flood risks in Italy [32], providing food security-related insights in South Korea [32] and responding to wildfires in Australia [5] or cyclones in New Zealand [22]. The goal of this paper is to better improve and better understand the challenges in satellite image time series semantic change detection (SITS-SCD), *i.e.* the detection of change in land use and land cover over time. We introduce an architecture that significantly boosts SITS-SCD results in the absence of any particular domain shift. However, practical monitoring necessitates online, real-time analysis, requiring models to accommodate the temporal shift between data seen during training and at inference. Additionally, due to the scarcity of annotated data [1, 2, 34], many models in practical applications are applied to images gathered from places far away from where the training data was observed. For these reasons, we conduct a comprehensive analysis of the impact of spatial and temporal domain shifts, showing their critical significance in this context.

Many works are dedicated to addressing the spatio-temporal shift through domain adaptation. While much of this work concentrates on spatial domain adaptation for single satellite images [10, 18, 19, 30, 51], recent efforts have also delved into spatial [29] or temporal [6, 33] domain adaptation for satellite image time series (SITS). However, to the best of our knowledge, no analysis of the impact of temporal or spatial domain shift on the performance of SITS-SCD and of the effects of different design choices has been performed so far. This paper is the first answer to these questions. We leverage the DynamicEarthNet [43] and MUDS [44] datasets that have both global spatial coverage and multi-year temporal coverage.

More precisely, we analyze independently the impact of these two domain shifts on both datasets for several methods, giving particular attention to the impact of model size. We evaluate state-of-the-art mono- and bi-temporal semantic segmentation approaches, which process each month or pairs of months independently. We also introduce a multi-

temporal SITS-SCD approach that jointly processes images from several months and can leverage long-term temporal information. We show it improves semantic segmentation performance in all settings on both datasets by a significant margin, and that it scales better with model size than the baselines, but that this does not always translate into better change detection performances. We also show that spatial and temporal domain shifts impact SITS-SCD approaches differently, that spatial domain shift has the most dramatic impact, and that while temporal domain shift has limited impact on semantic segmentation performance, it significantly decreases change detection accuracy. In summary, our contributions are as follows:

- We propose a new architecture to perform direct multi-temporal semantic segmentation that significantly improves SITS-SCD.
- We quantify the impact of temporal and spatial shift on the performance of SITS-SCD methods on two global and multi-year SITS datasets for different approaches.

2. Related work

In this paper, we study satellite image time series semantic change detection (SITS-SCD), considering temporal and spatial domain shifts in our evaluation settings. SITS-SCD consists in segmenting simultaneously each time stamps of a, typically monthly, SITS. We distinguish three strategies to achieve this task. First, *mono-temporal SITS-SCD* methods segment independently each image of the time series. Second, *bi-temporal SITS-SCD* methods segment the input SITS considering image pairs independently. Third, *multi-temporal SITS-SCD* approaches can segment jointly all time stamps of a given SITS. In this section, we review the literature for each of these categories, then list the existing datasets for SITS-SCD, and finally examine how related work tackles potential spatial and temporal shift between training and inference data.

Mono-temporal SITS-SCD. A first set of methods evaluated on SITS-SCD benchmarks predict the semantic segmentation for each time stamps independently. Changes are characterized by the difference of successive semantic maps. These semantic maps can be obtained from a single image using semantic segmentation models like U-Net [35], DeepLabV3 [7], Segmenter [38] or Swin-T [27]. If a monthly SITS is to be segmented and multiple images are available for each month, then methods designed for SITS segmentation into a single semantic map, such as 3D-Unet [31], UTAE [14] or TSViT [41], can be used as well, using the images from each month independently. Toker et al. [43] define 'monthly', 'weekly', and 'daily' temporal densities for these approaches, where the first image of the month, six images across the month, or all the images of the month are used respectively to form a monthly SITS.

Bi-temporal SITS-SCD. These methods perform SITS-SCD for each image pair independently, and are initially designed for the classic semantic change detection (SCD) task that requires to predict the semantic maps of a pair of satellite images at the same location but at distinct time stamps. Very early on, a series of work performed this task following a post-classification procedure [39, 40, 47], where the bi-temporal acquisitions are segmented independently. In this case, the binary change map is obtained as the difference between the predicted segmentation maps. Obviously, such method does not leverage temporal consistencies. To overcome this limitation, another approach is to classify pairs of pixels with transition labels, considering all possible pairs of semantic classes [4]. However, the number of transition labels increases as the square of the number of semantic classes, and some transitions have very few training examples because there are typically few changes and the land cover classes are very imbalanced (71%, 11% and 10% of Earth's surface is water, forest and agricultural parcels respectively). Such an approach thus faces a very challenging classification setting. Leveraging deep learning advances, most recent approaches [3, 8, 9, 11, 12, 20, 24, 26, 42, 50, 52, 53, 55, 56] tackle the SCD task with 3-branch models producing two semantic maps and a binary change detection map as output. Multi-task objectives, inner fusion modules in the architecture and/or post-processing operations help guaranty the consistency between the three outputs. In very high resolution, object-based SCD consists in detecting change on identified, often urban, semantic objects (like cars, containers or houses). Objects can be learned as bags of visual words where the dictionary is shared between time stamps [49], or as temporal correspondences between time stamps [54]. Our work focuses instead on mid-resolution satellite imagery and generic land cover classes, so object-based SCD is out of the scope of our study.

Multi-temporal SITS-SCD. Very few methods actually perform SITS-SCD in a multi-temporal manner. Saha et al. [36] propose an unsupervised framework for multi-temporal feature learning. Their model processes time stamps independently, the training loss aiming for temporal consistency. While they evaluate their model on classic bi-temporal SCD, one could imagine adapting their method to SITS-SCD but the code is not available. Very close to our approach, TSSCD [17] is a pixel-wise method extending a one-dimensional fully convolutional network for multi-temporal SITS-SCD. To the best of our knowledge, our proposed method is the first to perform multi-temporal SITS-SCD at the image level.

Semantic change detection datasets. SCD datasets [9, 23, 43–45, 52, 53] are intended for the simultaneous seman-

tic segmentation of the land cover at each time stamp and the detection of semantic changes between consecutive time stamps. Datasets like HRSCD [9] or SECOND [52] are designed for the bi-temporal task and do not exhibit SITS beyond simple image pairs. The data from the 2021 IEEE GRSS Data Fusion Contest [23] contains time series but only extreme dates have annotations, the challenge focusing more on knowledge transfer from low to high resolution rather than SITS-SCD. QFabric [45], MUDS [44] and DynamicEarthNet [43] are the SCD datasets the most relevant to our work since they include complete time series and full semantic change annotations. QFabric focuses on urban changes with partial labeling: only changing areas are annotated, with labels such as 'Prior Construction', 'Land Cleared', or 'Construction Done'. Since it is not freely available, we do not consider it. We focus on DynamicEarthNet and MUDS which both contain annotated multi-year SITS covering areas all over the world. DynamicEarthNet is designed for land-use and land-cover classification with classes such as 'impervious surface', 'forest', or 'water', and its areas of interest include a broad range of region types. MUDS, also known as SpaceNet 7, is intended for building tracking over time. We adapt its annotations to the semantic change detection task and propose a first benchmark of SITS-SCD methods on MUDS using the semantic change detection metrics defined by Toker et al. [43].

Temporal and spatial domain shifts. Domain adaptation is a well-known problem with satellite imagery [10, 18, 19, 30, 51]. In the particular case of SITS, Lucas et al. [29] attempt to adapt state-of-the-art domain adaptation methods to spatial domain shift between two regions within France for the task of generic land cover pixel-wise classification. Crop-type classification with SITS is also relevant for studying temporal domain shift because of seasonal and environmental variability. Capliez et al. [6] examine temporal domain shift in the context of crop type classification in Burkina Faso over multiple years, while Vincent et al. [46] demonstrate the challenges posed by temporal domain shift in agricultural time series pixel-wise classification with a German crop dataset [21]. Additionally, Nyborg et al. [33] propose thermal positional encoding - an encoding based on thermal time rather than calendar time - to account for varying rates of crop growth and mitigate temporal shift issues in Western Europe data. These studies are conducted at the national or continental scale, focus on classification tasks, and mainly try to bridge the performance gap due to domain shift. In contrast, we analyze the impact of temporal and spatial domain shifts independently and at a global scale for our multi-temporal SITS-SCD approach - that outperforms state of the art mono- and bi-temporal approaches - giving particular attention to the impact of model size, an important but often overlooked variable.

3. Method

3.1. Proposed architecture

We propose to modify UTAE [14] by changing the core temporal attention mechanism to output one segmentation map per input image instead of aggregating temporal information, in order to better leverage temporal knowledge. The overall pipeline is illustrated by Figure 1 where we show the encoder branch, the temporal attention block, and the decoder branch.

Encoder. Our encoder takes as input a SITS $\mathbf{x}$ in $\mathbb{R}^{T\times H\times W\times C}$ of T satellite images of spatial dimensions $H \times W$ with C spectral bands. The encoder branch of our model strictly mirrors UTAE and produces a series of feature maps $\mathbf{z}^1$, ..., $\mathbf{z}^L$ using L successive down-sampled convolutions. Positional encoding is added to the feature map at the last level $\mathbf{z}^L$ in $\mathbb{R}^{T\times H'\times W'\times D}$, with $H' \times W'$ the spatial resolution and D the feature size at level L. Similar to Garnot and Landrieu [14], sections of size D/h of $\mathbf{z}^L$ are processed independently in a h-head manner. For the sake of conciseness, we ignore positional encoding and multi-head processing in our notations in the following sections.

Attention mechanism. Our temporal attention mechanism outputs multi-temporal attention maps $\mathbf{a}^L$ in $[0,1]^{T\times T\times H'\times W'}$ at the lowest resolution. Its role is to combine the different temporal feature maps while maintaining a temporal dimension. For each time stamp t in the range $\{1,\dots,T\}$, we aim to incorporate information from all dates into the prediction for t, with the contributions of each date t' in the range $\{1,\dots,T\}$ being specific to t. Our proposed attention mechanism builds on TAE [15], which predicts the queries as a function of the feature maps at the lowest level $\mathbf{z}^L$. However, instead of computing the attention weights as a scalar product of the keys and queries, we define the weights as their matrix multiplication in order to keep the temporal dimension. Note that UTAE, on which our overall architecture is built upon, uses a lightweight temporal attention encoder (LTAE) [13] at its core to aggregate the temporal feature maps. LTAE is a lightweight version of TAE where the queries are free parameters of the model and are the same for all time stamps. We illustrate in Figure 2 the differences between TAE, LTAE and our proposed attention mechanism. For a given spatial location (i,j) in $[1,H'] \times [1,W']$, we compute the attention weights $\mathbf{a}^L_{i,j}$ from queries

$$\mathbf{q} = \text{FC}^q(\mathbf{z}^L_{i,j}) \ \in \mathbb{R}^{T\times d}, \tag{1}$$

and keys

$$\mathbf{k} = \text{FC}^k(\mathbf{z}^L_{i,j}) \ \in \mathbb{R}^{T\times d}, \tag{2}$$

as

$$\mathbf{a}^L_{i,j} = \mathbf{k}\mathbf{q}^\top \ \in \mathbb{R}^{T\times T}, \tag{3}$$

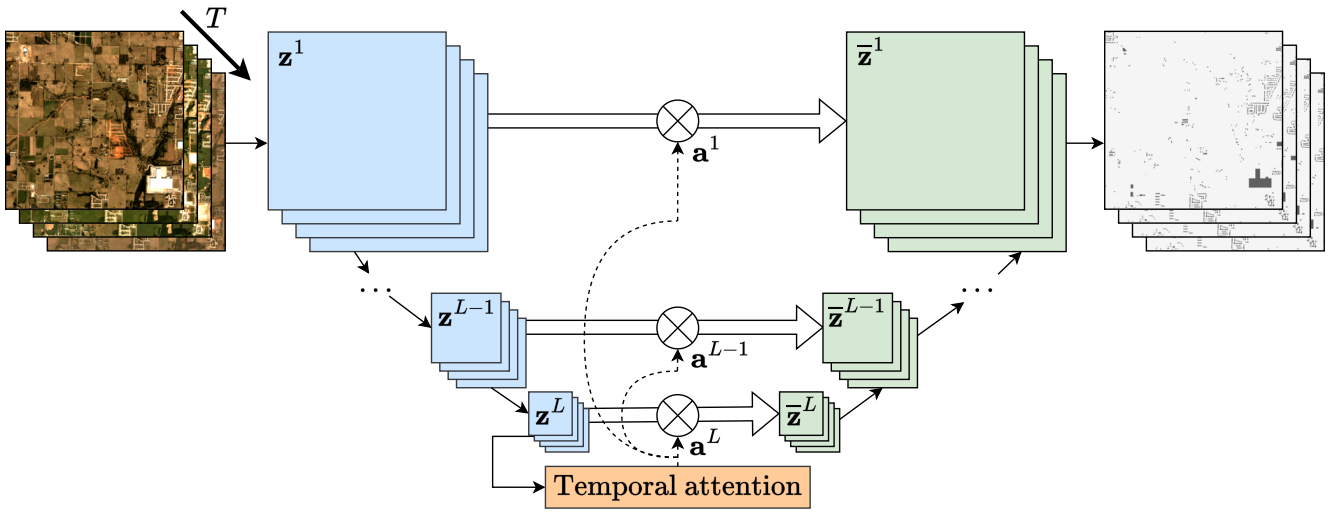


Figure 1. **Overall architecture.** Given an input SITS, we compute feature maps at various scales. Our contribution is the temporal attention mechanism that allows to account for long-term temporal information. The decoder branch up-scales the feature maps for all time stamps in parallel, while propagating the attention maps at all levels.

where FC^q and FC^k denote fully-connected layers. The attention maps are up-sampled at all levels l in $\{1, \dots, L\}$ into attention maps $\mathbf{a}^l$, so that the combined feature map is obtained for all time stamps t as:

$$\overline{\mathbf{z}}_t^l = \sum_{t'=1}^{T} \mathbf{a}_{t,t'}^l \odot \mathbf{z}_{t'}^l, \tag{4}$$

where $\odot$ denotes the element-wise multiplication.

Decoder. The decoder uses strided transposed convolutions to up-sample the feature maps $\overline{\mathbf{z}}^l$ to the upper level. Following Garnot and Landrieu [14], we propagate the up-sampled attention maps $\mathbf{a}^1, \dots, \mathbf{a}^{L-1}$ at all levels with skip connections: before each up-sampling convolution, the obtained feature map $\overline{\mathbf{z}}_t^l$ is concatenated to the up-sampled feature map of the lower level $\overline{\mathbf{z}}_t^{l+1}$. All time steps are processed in parallel using the same decoder branch. Outputs are the segmentation maps $\mathbf{s}$ in $\mathbb{R}^{T \times H \times W \times K}$ where K is the number of semantic classes.

3.2. Analysis methodology

We perform our analysis on two global and multi-year SITS datasets: DynamicEarthNet [43] and MUDS [44]. We evaluate all methods in three different settings: a setting without domain shift between train and test sets, a setting with temporal shift, and a setting with spatial domain shift. These different settings are visualized in Figure 3.

No domain shift. We split the datasets into 4 subsets as visualized in Figure 3a: we split each SITS into four SITS of equal size. We keep two for training, one for validation and one for test purposes. We follow a 4-fold cross validation scheme that we detail in Section A of the appendix. Though folds cover distinct areas, they all share common regions so

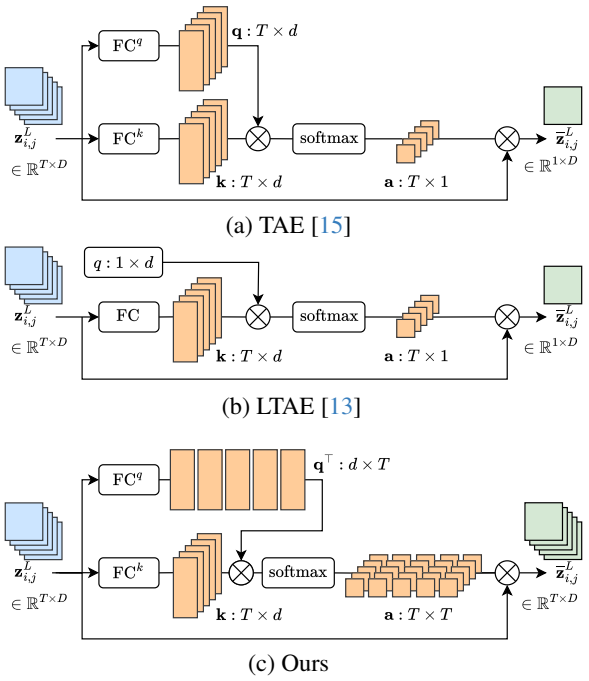
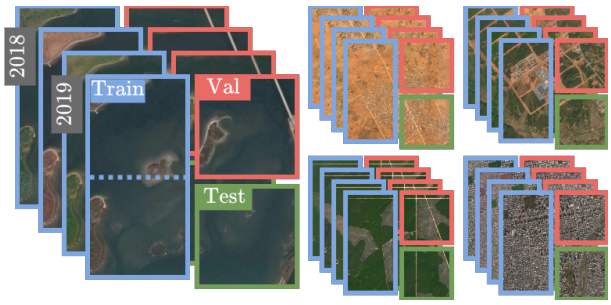
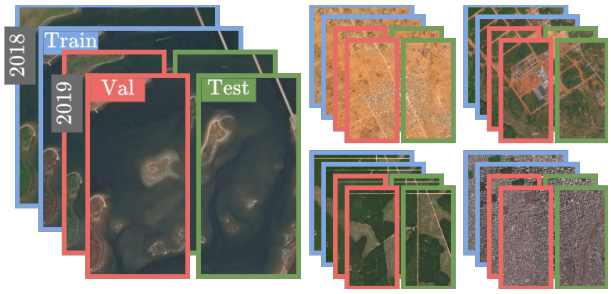


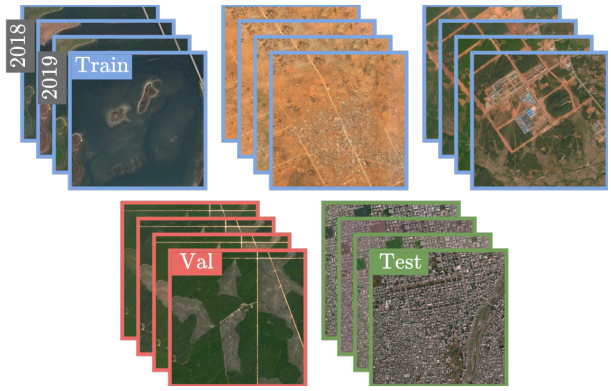
Figure 2. **Attention mechanism of TAE, LTAE and our method.** We show the temporal attention mechanism of TAE [15], LTAE [13] and our method for a given patch $\mathbf{z}_{i,j}^L$ of the feature map $\mathbf{z}^L$. Here, d is the dimension of the key and query vectors.



(a) No domain shift setting



(b) Temporal domain shift setting



(c) Spatial domain shift setting

Figure 3. **Domain shift settings.** We organize the dataset splits in three different manners such that there is respectively (a) no domain shift, (b) a temporal domain shift, and (c) a spatial domain shift between train and val/test sets. DynamicEarthNet [43] images are shown here for visualization, and we use the same settings for MUDS [44].

that there is no significant spatial domain shift. Additionally, since all splits cover the same two years, there is thus no temporal shift.

Temporal shift. DynamicEarthNet and MUDS contain 2-year image time series from January 2018 to December 2019. We split all time series in half, keeping 2018 for training and 2019 for validation and test purposes as visualized in Figure 3b. Note that more annual data would be necessary to create multiple folds for cross-validation in the temporal setting. In this setting, there is no spatial domain shift.

Spatial shift. We split the 55 (resp. 60) areas of interest of DynamicEarthNet (resp. MUDS) into five subsets of 11 (resp. 12) image time series. As illustrated in Figure 3c, we keep three sets for training while the remaining two are kept for validation and test purposes respectively. We then follow a 5-fold cross validation scheme described in Section A of the appendix. Here, all subsets cover regions that are significantly different and far from each other so that there is a spatial domain shift. Folds are random on MUDS and selected so that the class distribution is approximately similar in each fold on DynamicEarthNet. More details to reproduce this setting can be found in Section A of the appendix. Note that this spatial shift setting is the one used in the DynamicEarthNet[1] and the SpaceNet 7[2] challenges.

3.3. Training and implementation details

Following Tarasiou et al. [41], we trained all methods using focal loss [25] and the AdamW optimizer [28] with a learning rate starting from 0 and gradually reaching 10^{-4} as a warmup after 5000 iterations. We train our model for 500 000 iterations and keep the checkpoint that achieves the best semantic change segmentation score on the validation set. For data augmentation, we randomly crop image patches of size 128×128 out of the 1024×1024 images and additionally do random horizontal and vertical flips as well as random rotations.

Our model can take image sequences of variable length as input. During training, we sample 12 random monthly images out of the 24 available in the setting without domain shift and the spatial shift setting. In the temporal shift setting, we sample 6 out of the 12 available images. At inference, we take the full monthly time series as input, *i.e.* a sequence of length 24 or 12 depending on the case. In Section 4.3.1, we study alternative inference schemes.

We set as default values for the number of levels $L = 4$, for the spatial feature size $D = 512$ and for the dimension of keys and queries $d = 4$. We investigate the impact of changing D in Section 4.3.1 and d in Section C of the appendix.

4. Experiments

4.1. Datasets and metrics

We evaluate our method along with baselines on DynamicEarthNet [43] and MUDS [44] datasets and more precisely on their training set for which ground truth annotations are available. Images for both these datasets were acquired by Planet Labs with a ground sample distance (GSD) of approximately 3 meters.

DynamicEarthNet [43]. This dataset contains 55 daily SITS from January, 1st 2018 to December, 31st 2019 dis-

[1] https://codalab.lisn.upsaclay.fr/competitions/2882
[2] https://spacenet.ai/sn7-challenge/

	Method	Input type	Strategy	No domain shift				Temporal domain shift				Spatial domain shift			
				SCS↑	SC↑	BC↑	mIoU↑	SCS↑	SC↑	BC↑	mIoU↑	SCS↑	SC↑	BC↑	mIoU↑
DynamicEarthNet	Random	—	—	5.9	6.9	4.9	7.3	5.9	6.8	5.0	7.3	5.7	6.6	4.9	7.1
	TSViT monthly	Single image	Mono	23.0	34.1	11.8	50.5	19.3	28.6	9.9	47.3	13.3	18.6	7.9	31.2
	UTAE monthly	Single image	Mono	25.9	38.0	13.8	53.7	20.8	30.7	10.9	53.7	15.5	21.9	9.0	36.9
	TSViT weekly	SITS	Mono	25.0	37.6	12.5	50.9	23.4	36.0	10.9	51.4	13.2	19.1	7.4	32.2
	UTAE weekly	SITS	Mono	26.7	39.1	14.3	54.4	22.4	33.6	11.3	54.7	**16.1**	**23.6**	8.7	37.8
	A2Net	Image pair	Bi	22.2	32.9	11.5	47.2	21.6	32.2	11.0	46.7	15.4	22.5	8.2	37.9
	SCanNet	Image pair	Bi	24.8	35.8	13.9	53.0	24.8	**36.4**	13.1	55.6	15.4	21.5	9.3	37.3
	TSSCD	Pixel-wise SITS	Multi	12.0	19.4	4.7	33.9	10.0	14.8	5.2	29.4	9.1	13.0	5.2	22.9
	Ours	SITS	Multi	**31.7**	**41.0**	**22.4**	**60.5**	**25.6**	36.0	**15.3**	**61.7**	15.8	21.5	**10.1**	**38.5**
MUDS	Random	—	—	15.0	29.9	0.1	28.1	**14.7**	**29.2**	0.1	28.1	**15.0**	**30.0**	0.1	28.1
	TSViT monthly	Single image	Mono	11.8	23.1	0.5	60.2	9.2	17.8	0.5	56.8	7.0	13.7	0.4	49.8
	UTAE monthly	Single image	Mono	**15.4**	**30.2**	0.6	67.1	12.4	24.1	0.6	66.0	14.2	27.9	0.6	63.0
	A2Net	Image pair	Bi	11.1	21.7	0.5	61.5	8.7	16.7	0.6	56.1	8.6	16.7	0.5	53.0
	SCanNet	Image pair	Bi	13.2	25.7	0.7	64.9	10.1	19.5	0.8	62.8	11.9	23.4	0.4	58.8
	TSSCD	Pixel-wise SITS	Multi	11.7	23.3	0.1	47.7	12.8	25.4	0.2	49.6	9.1	18.0	0.1	43.6
	Ours	SITS	Multi	13.7	25.7	**1.7**	**72.0**	11.0	20.1	**1.9**	**71.1**	10.8	20.8	**0.7**	**66.2**

Table 1. **Results for all three settings.** We report for our method and competing methods the semantic change segmentation score (SCS), the binary change score (BC), the semantic change score (SC) and the mean intersection-over-union (mIoU) in all settings on DynamicEarthNet [43] and MUDS [44]. 'Random' refers to a baseline predicting a random label for each pixel.

Seq. len.	Inference time series splitting	SCS↑	SC↑	BC↑	mIoU↑
6		27.9	38.0	17.8	56.9
		29.0	**41.5**	16.6	59.5
8		29.7	39.6	19.8	58.3
		29.3	41.2	17.3	59.9
12		30.6	40.1	21.1	59.6
		29.8	39.7	19.9	59.0
		30.9	41.2	20.6	59.9
		30.4	40.8	20.0	59.8
		30.4	41.3	19.5	60.1
		29.8	41.3	18.3	60.2
24		**31.7**	41.0	**22.4**	**60.5**

1st inference 2nd inference 3rd inference 4th inference

Table 2. **Inference time series size.** We report the SCS, SC, BC and mIoU of our model for various input sequence length and different splitting configurations. The splitting is best viewed in color, where cells of the same color were gathered together as a SITS to produce their corresponding predictions simultaneously.

tributed over the globe. The first day of each month is annotated, leading to 24 ground truth segmentation maps per area of interest (AoI). Images are of size 1024×1024 and multi-spectral with 4 channels (RGB + near-infrared). Annotations are general land-use and land-cover classes: 'impervious surface', 'agriculture', 'forest', 'wetlands', 'soil' and 'water'. The 'snow' class is only present on very few AoIs of the dataset and is discarded for this study.

MUDS [44]. The Multi-temporal Urban Development SpaceNet (MUDS) dataset consists of 60 monthly SITS collected between 2017 and 2020 all over the globe. MUDS contains few images acquired in 2017 and 2020: they are discarded for this study. Due to an excessive amount of clouds or haze some images were excluded from the dataset, causing a few gaps in some of the time series, with length ranging from 18 to 24 images per SITS. Images are of size 1024×1024 with 3 channels (RGB). Default annotations for this dataset are polygons indicating buildings from which we generate binary segmentation maps with classes 'building' and 'not building'.

Metrics. We use the four metrics defined by Toker et al. [43] to assess the performance of evaluated methods. The mean intersection-over-union (mIoU) on the semantic labels indicates the ability of the method to predict correct semantic segmentations, irrespective of the change. The binary change (BC) score depicts how well a method can predict a semantic change while the semantic change (SC) score focuses on the semantic prediction for pixels where a change actually occurs. The semantic change segmentation (SCS) score is the average of both previous scores.

4.2. Baselines

4.2.1 Mono-temporal

We evaluate two state-of-the-art semantic segmentation methods designed for SITS: UTAE [14] and TSViT [41].

UTAE [14]. The U-Net with Temporal Attention Encoder (UTAE) consists of a U-Net architecture where a temporal attention mechanism squeezes the temporal dimension before the decoding branch. Thus, the model outputs a single segmentation map for the whole input time series. This method shows competitive performance on recent segmentation benchmarks [41, 43].

TSViT [41]. In contrast to UTAE, the Temporo-Spatial Vision Transformer (TSViT) has a fully-attentional architec-

ture, processing the tokens first temporally then spatially. It was shown to improve semantic segmentation on several datasets [41].

Both these methods output a single prediction map for a given SITS as input. In order to evaluate them on the SITS-SCD task, we follow the setting of Toker et al. [43] where the monthly segmentation maps are predicted independently from one another by using as input signal one or several images in the month. In the *monthly* setting, only the first image of each month is used, and the input time series is actually composed of a single image. In the *weekly* setting, a SITS of six images - corresponding to an image every 5 days through the month - serves as input to obtain the monthly prediction. Toker et al. [43] show that a *daily* setting - where all images of a month are used as a SITS to predict a monthly segmentation map - does not improve over the weekly setting, thus we do not consider it in this work. We set same values of L, D and d for UTAE-based methods as with our method for fair comparison. For TSViT, we set the feature dimension to 512 so that the number of trainable parameters is of the same order of magnitude as other evaluated methods. Additional details on our UTAE and TSViT implementations are provided in Section B of the appendix.

4.2.2 Bi-temporal

We evaluate two state-of-the-art bi-temporal SCD methods: A2Net [24] and SCanNet [12].

A2Net [24]. A2Net first extracts multi-stage feature maps from a pair of images with a shared-weight MobileNetV2 [37]. The difference of the two feature maps at all stages are combined and decoded into a binary change mask and two semantic segmentation maps.

SCanNet [12]. The Semantic Change Network (SCanNet) has a three-branch encoder-decoder architecture. The image pairs and the concatenation of intermediate feature representations are used to learn two sets of semantic tokens (one for each time stamps) and a set of change tokens. All tokens are concatenated and processed by an inner transformer. The output is decoded into a binary change mask and two semantic segmentation maps.

There are multiple manners to adapt bi-temporal methods to SITS-SCD. We train both methods with all possible ordered image pairs. At inference, a SITS is divided in consecutive image pairs that are segmented independently. We only consider the semantic outputs and disregard the predicted binary change maps at inference. Additional details on our A2Net and SCanNet implementations are provided in Section B of the appendix.

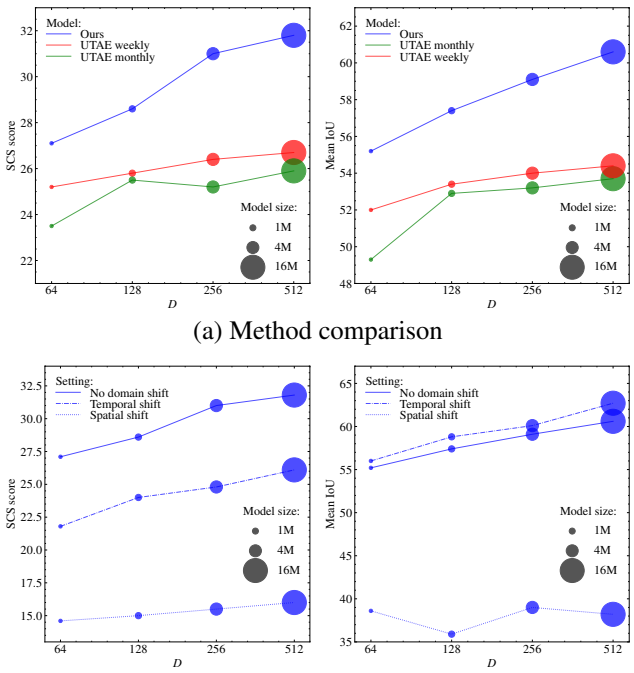


Figure 4. **Impact of D on performance.** We compare the impact of the spatial feature size D on performance (a) for our model and UTAE-based methods in the setting without domain shift and (b) for our model in the three domain shift settings. In each case, we report the SCS score (left) and the mean IoU (right).

4.2.3 Multi-temporal

We evaluated TSSCD [17], a state-of-the-art pixel-wise multi-temporal approach, adapting a one-dimensional fully convolutional network to the SITS-SCD task. Its training requires a particular sampling of pixel time series that we detail in Section B of the appendix.

4.3. Results

4.3.1 Leveraging long-term temporal information

In Table 1, we report the performance obtained on DynamicEarthNet and MUDS in all three settings. Additional quantitative and qualitative results can be found in Sections D & E of the appendix. Our architecture performs better than other evaluated baselines in all settings and on both datasets in terms of binary classification and semantic segmentation. We discuss SC and SCS performance in Section 4.3.3. We explain the better scores obtained by our model by its ability to extract temporal knowledge from long-term images of the time series.

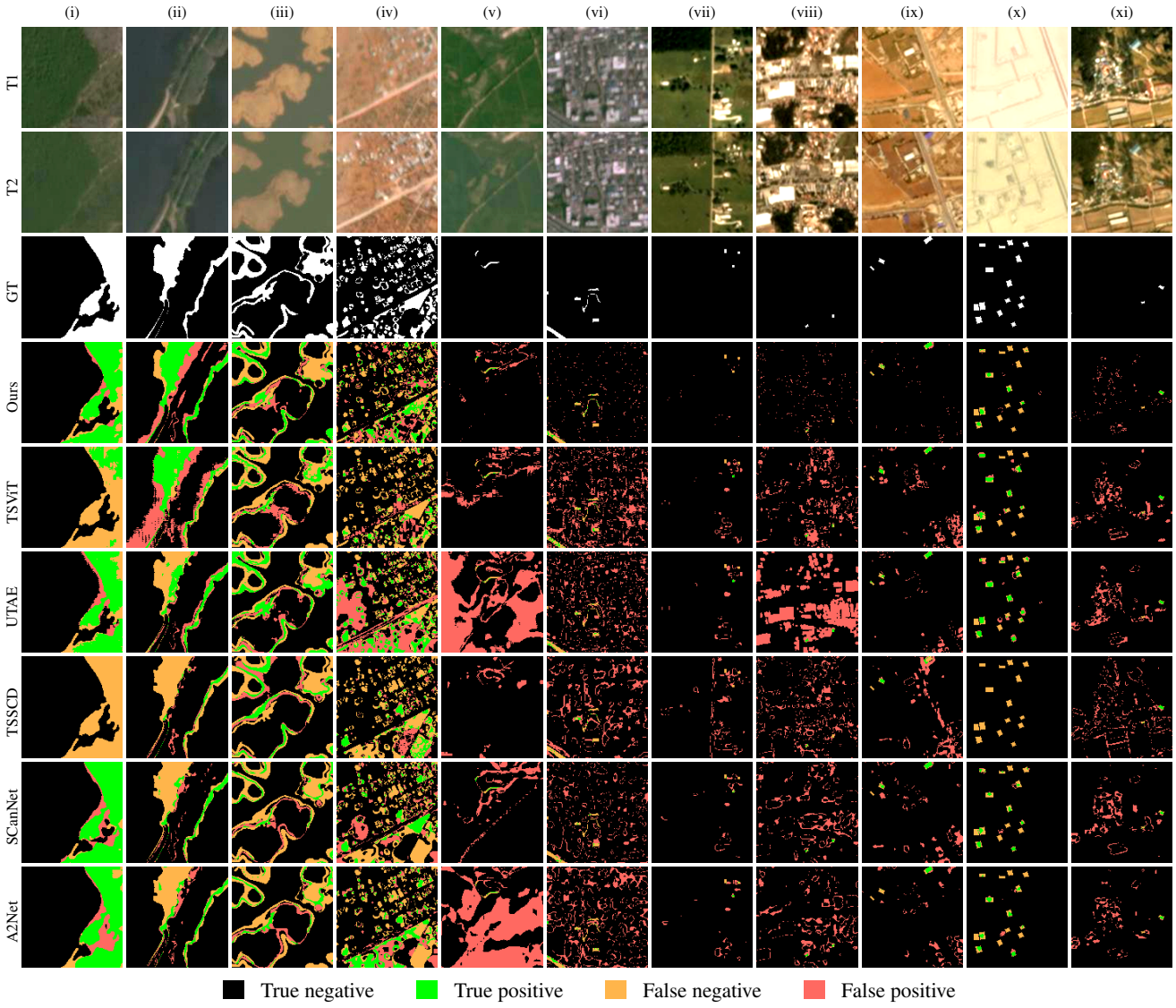


Figure 5. **Qualitative change detection results.** We show the binary change detection maps predicted by our model and competing methods in the setting without domain shift for randomly selected input images. From top to bottom, we show the input pairs at time T1 (01/09/2018) and T2 (01/09/2019), the ground truth binary change map and the predictions of different methods for DynamicEarthNet (i-vi) and MUDS (vii-xi). For TSViT and UTAE we use the weekly setting for DynamicEarthNet and the monthly setting for MUDS. Best viewed in color.

To confirm this intuition, we investigate various inference schemes in the setting without domain shift and report the results in Table 2. We evaluate the performance of our method when performing inference on sub-sequences of the full 24-month image sequence of sizes 6, 8, 12 and 24. We also explore various ways to sample these sub-sequences, as visualized by the colors on the left of the Table. The results clearly show that the performance improves when longer series are used at inference, validating our hypothesis that our model leverages temporal information over a long temporal range. Interestingly, one can also see that for a given input sequence length, having sub-sequences that span the 24-month period is better than using successive slices of the full time series. This seems particularly important when using short sequences.

Another indicator of our method's ability to learn more informative features than UTAE-based methods is that its performance improves as the spatial feature size D increases, as illustrated on Figure 4a. While the SCS score and the mIoU of our architecture increases as the model gets bigger, there is no significant increase in UTAE performance for feature size higher than $D = 128$. This shows our architecture can better leverage spatial information.

Qualitatively, we report the predicted binary change detec-

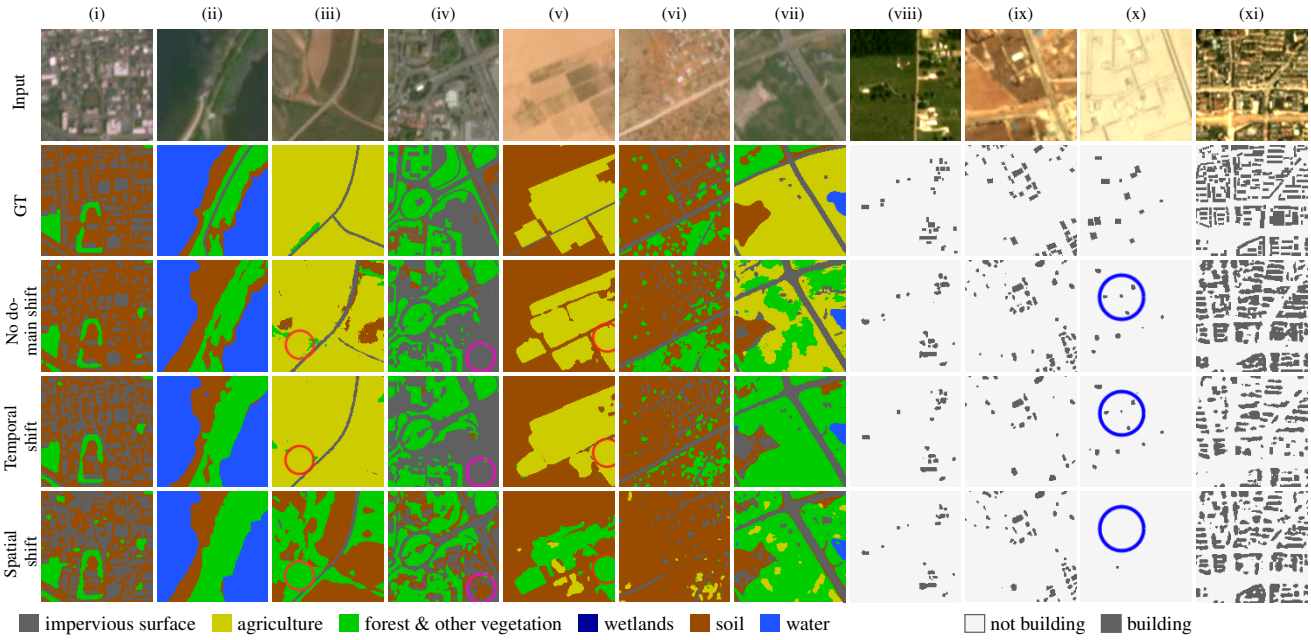


Figure 6. **Qualitative segmentation results in different settings.** We show the segmentation maps predicted by our model in different settings for randomly selected input images. From top to bottom, we show the input image (on 01/07/2019), the corresponding ground truth, the predictions without domain shift, the predictions with temporal shift and the predictions with spatial shift. Images from (i-vii) are taken from DynamicEarthNet and (viii-xi) from MUDS. We highlight areas where our method fails in the spatial domain shift setting on the 'agriculture' ◯ and the 'impervious surface' ◯ classes for DynamicEarthNet and the 'building class' ◯ for MUDS. Best viewed in color.

tion maps for pairs of images one year apart without domain shift in Figure 5. Our predictions show significantly fewer false positive than competing baselines.

4.3.2 Comparing domain shift settings

The results from Table 1 show that spatial domain shift has the most significant impact on performance, both regarding semantic and change detection scores. For all evaluated methods, we observe an average relative drop of the mIoU from the setting without domain shift to the spatial shift setting of 31.9% on DynamicEarthNet and of 10.5% on MUDS. The SCS score similarly decreases by 39.0% on DynamicEarthNet and by 20.7% on MUDS. This drop is explained by the diversity of geographies contained in these two global datasets. The fact the performance drop is smaller on MUDS than on DynamicEarthNet is likely related to the fact that the geographic variability is less pronounced on buildings than on other land-cover types. On DynamicEarthNet, the drop of IoU for the 'impervious surface' class (*i.e.* artificial land) is only of 23.0%, similar to the drop observed on MUDS, while it is 49.7% on average for the other land-cover classes.

The impact of the domain shift can be seen qualitatively in Figure 6, where we show segmentation results of our method trained in each setting on the same images. We highlight in red circles areas where 'agriculture' is classified as 'forest' and in pink circles, areas where 'impervious surface' is classified as 'soil' in the spatial setting on DynamicEarthNet. On MUDS, though some buildings are not detected in the spatial setting as highlighted by the blue circles, our method seems to rarely classify 'not building' as 'building' in any of the settings.

In figure 4b, we analyze the relation between the model size and performance for our method in the different domain shift settings. Two effects are striking. First, while performance exhibits gradual improvement as the number of parameters increases in the absence of domain shift and under temporal shift conditions, there is no significant improvement in mIoU in the spatial shift setting, and only a slight increase in the SCS score. This again highlights the importance of spatial domain shift. Second, this graph confirms that the semantic segmentation results are similar without domain shift and in the temporal setting (which can also be seen qualitatively in Figure 6), but change detection performance is clearly impacted by temporal domain shift. We believe we are the first to highlight this very specific impact of temporal domain shift for change detection.

4.3.3 Limits of current methods and future work

Binary change detection on DynamicEarthNet and MUDS is a challenging task. The BC score in all settings and for all methods is relatively low, below 23%. This clearly is an obstacle for the application of current methods for SITS-

SCD. To understand this low performance, it is important to note that there are very few changes occurring in these two datasets: the proportion of pixels that semantically change from one month to the next is of 1.28% in DynamicEarthNet and of 0.03% in MUDS. This makes the SC and SCS scores, for which our method is not always better than competing baselines, hard to interpret since they focus on very few pixels on which change is accurately detected. However, this is consistent with practical use cases where semantic changes, *e.g.* the construction/destruction of a building, droughts/floods or deforestation, are rare events at a global scale. Thus, while we believe that adapting SITS-SCD methods to address temporal and spatial domain shifts for multi-year and global applications is an important challenge, we also argue that improving performance in the setting without any domain shift is important, and will likely require better addressing the scarcity of change data. On MUDS especially, no method significantly outperforms a random baseline giving random labels to each pixel (see the 'Random' line in Table 1).

5. Conclusion

This paper introduces a novel architecture for semantic change detection in satellite image time series (SITS-SCD) and a detailed analysis of the impact of temporal and spatial domain shifts on this task. Our method outperforms existing baselines for semantic segmentation and binary change segmentation across various evaluation settings, demonstrating its effectiveness in extracting temporal knowledge. Our analysis outlines that spatial domain shift has a significant impact on overall performances, and that temporal domain shift impacts more specifically change detection. However, it also suggests that even in the absence of domain shift, the performance of current methods for SITS-SCD is limited. We believe this is due to the rarity of significant changes, underlining the importance of addressing data scarcity. Overall, our study contributes to advancing SITS-SCD methods and highlights avenues for future research in this area.

Acknowledgments. The work of MA was partly supported by the European Research Council (ERC project DISCOVER, number 101076028). JP was supported in part by the Louis Vuitton/ENS chair on artificial intelligence and the French government under management of Agence Nationale de la Recherche as part of the *Investissements d'avenir* program, reference ANR19-P3IA0001 (PRAIRIE 3IA Institute). This work was granted access to the HPC resources of IDRIS under the allocation 2021-AD011013067 made by GENCI. We thank Antoine Guédon, Loïc Landrieu and Ioannis Siglidis for valuable feedbacks, and Zeynep Sonat Baltaci and Syrine Kalleli for their careful proofreading.

References

[1] Ting Bai, Le Wang, Dameng Yin, Kaimin Sun, Yepei Chen, Wenzhuo Li, and Deren Li. Deep learning for change detection in remote sensing: a review. *Geo-spatial Information Science*, 26(3):262–288, 2023. 1

[2] John Ball, Derek Anderson, and Chee Seng Chan. Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community. *Journal of applied remote sensing*, 11(4):042609–042609, 2017. 1

[3] Maximilian Bernhard, Niklas Strauß, and Matthias Schubert. Mapformer: Boosting change detection by using pre-change information. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16837–16846, 2023. 2

[4] Lorenzo Bruzzone and Sebastiano B Serpico. An iterative technique for the detection of land-cover transitions in multitemporal remote-sensing images. *IEEE transactions on geoscience and remote sensing*, 35(4):858–867, 1997. 2

[5] Bushfire Earth Observation Taskforce. Report on the role of space based earth observations to support planning, response and recovery for bushfires. In *Australian Space Agency*, pages 1–33, 2020. 1

[6] Emmanuel Capliez, Dino Ienco, Raffaele Gaetano, Nicolas Baghdadi, and Adrien Hadj Salah. Temporal-domain adaptation for satellite image time-series land-cover mapping with adversarial learning and spatially aware self-training. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:3645–3675, 2023. 1, 3

[7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 2

[8] Fengzhi Cui and Jie Jiang. Mtscd-net: A network based on multi-task learning for semantic change detection of bitemporal remote sensing images. *International Journal of Applied Earth Observation and Geoinformation*, 118:103294, 2023. 2

[9] 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. 2, 3

[10] Xueqing Deng, Hsiuhan Lexie Yang, Nikhil Makkar, and Dalton Lunga. Large scale unsupervised domain adaptation of segmentation networks with adversarial learning. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 4955–4958. IEEE, 2019. 1, 3

[11] Lei Ding, Haitao Guo, Sicong Liu, Lichao Mou, Jing Zhang, and Lorenzo Bruzzone. Bi-temporal semantic reasoning for the semantic change detection in hr remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–14, 2022. 2

[12] Lei Ding, Jing Zhang, Haitao Guo, Kai Zhang, Bing Liu, and Lorenzo Bruzzone. Joint spatio-temporal modeling for semantic change detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 2, 7

[13] Vivien Sainte Fare Garnot and Loic Landrieu. Lightweight temporal self-attention for classifying satellite images time series. In *Advanced Analytics and Learning on Temporal*

Data: 5th ECML PKDD Workshop, AALTD 2020, Ghent, Belgium, September 18, 2020, Revised Selected Papers 6, pages 171–181. Springer, 2020. 3, 4

[14] Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4872–4881, 2021. 2, 3, 4, 6, 13

[15] Vivien Sainte Fare Garnot, Loic Landrieu, Sebastien Giordano, and Nesrine Chehata. Satellite image time series classification with pixel-set encoders and temporal self-attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12325–12334, 2020. 3, 4

[16] Ivan Haščič and Alexander Mackie. Land cover change and conversions. In *OECD iLibrary*, 2018. 1

[17] Haixu He, Jining Yan, Dong Liang, Zhongchang Sun, Jun Li, and Lizhe Wang. Time-series land cover change detection using deep learning-based temporal semantic segmentation. *Remote Sensing of Environment*, 305:114101, 2024. 2, 7, 15

[18] Yi Huang, Jiangtao Peng, Na Chen, Weiwei Sun, Qian Du, Kai Ren, and Ke Huang. Cross-scene wetland mapping on hyperspectral remote sensing images using adversarial domain adaptation network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 203:37–54, 2023. 1, 3

[19] Javed Iqbal and Mohsen Ali. Weakly-supervised domain adaptation for built-up region segmentation in aerial and satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 167:263–275, 2020. 1, 3

[20] Liangcun Jiang, Feng Li, Li Huang, Feifei Peng, and Lei Hu. Ttnet: A temporal-transform network for semantic change detection based on bi-temporal remote sensing images. *Remote Sensing*, 15(18):4555, 2023. 2

[21] 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

[22] Land Information New Zealand. High resolution imagery of flood-hit areas supports cyclone recovery. -, 2023. https://www.linz.govt.nz/news/2023-03/high-resolution-imagery-flood-hit-areas-supports-cyclone-recovery. 1

[23] Zhuohong Li, Fangxiao Lu, Hongyan Zhang, Lilin Tu, Jiayi Li, Xin Huang, Caleb Robinson, Nikolay Malkin, Nebojsa Jojic, Pedram Ghamisi, et al. The outcome of the 2021 ieee grss data fusion contest—track msd: Multitemporal semantic change detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:1643–1655, 2022. 2, 3

[24] Zhenglai Li, Chang Tang, Xinwang Liu, Wei Zhang, Jie Dou, Lizhe Wang, and Albert Y Zomaya. Lightweight remote sensing change detection with progressive feature aggregation and supervised attention. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–12, 2023. 2, 7

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 5

[26] Xuanguang Liu, Chenguang Dai, Zhenchao Zhang, Mengmeng Li, Hanyun Wang, Hongliang Ji, and Yujie Li. Tbscdnet: A siamese multi-task network integrating transformers and boundary regularization for semantic change detection from vhr satellite images. *IEEE Geoscience and Remote Sensing Letters*, 2024. 2

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[29] Benjamin Lucas, Charlotte Pelletier, Daniel Schmidt, Geoffrey I. Webb, and François Petitjean. Unsupervised domain adaptation techniques for classification of satellite image time series. In *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, pages 1074–1077, 2020. 1, 3

[30] Muying Luo and Shunping Ji. Cross-spatiotemporal land-cover classification from vhr remote sensing images with deep learning based domain adaptation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 191:105–128, 2022. 1, 3

[31] 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. 2

[32] Anastasia Moumtzidou, Stelios Andreadis, Ilias Gialampoukidis, Stefanos Vrochidis, Vasileios Sitokonstantinou, Dennis Hoppe, Michael Gienger, and Li Zhong. Change detection techniques in earth observation. In *EOPEN: opEn interOperable Platform for unified access and analysis of Earth observatioN data - European Commission*, pages 13–14, 2020. 1

[33] Joachim Nyborg, Charlotte Pelletier, and Ira Assent. Generalized classification of satellite image time series with thermal positional encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1392–1402, 2022. 1, 3

[34] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and fnm Prabhat. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, 2019. 1

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2

[36] Sudipan Saha, Lichao Mou, Chunping Qiu, Xiao Xiang Zhu, Francesca Bovolo, and Lorenzo Bruzzone. Unsupervised deep

joint segmentation of multitemporal high-resolution images. *IEEE Transactions on Geoscience and Remote Sensing*, 58 (12):8780–8792, 2020. 2

[37] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 7

[38] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021. 2

[39] PH Swain. Bayesian classification in a time-varying environment. *IEEE Transactions on Systems, Man and Cybernetics*, 8:879–883, 1978. 2

[40] Philip H Swain and Hans Hauska. The decision tree classifier: Design and potential. *IEEE Transactions on Geoscience Electronics*, 15(3):142–147, 1977. 2

[41] Michail Tarasiou, Erik Chavez, and Stefanos Zafeiriou. Vits for sits: Vision transformers for satellite image time series. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10418–10428, 2023. 2, 5, 6, 7, 14

[42] Shiqi Tian, Yanfei Zhong, Zhuo Zheng, Ailong Ma, Xicheng Tan, and Liangpei Zhang. Large-scale deep learning based binary and semantic change detection in ultra high resolution remote sensing imagery: From benchmark datasets to urban application. *ISPRS Journal of Photogrammetry and Remote Sensing*, 193:164–186, 2022. 2

[43] Aysim Toker, Lukas Kondmann, Mark Weber, Marvin Eisenberger, Andrés Camero, Jingliang Hu, Ariadna Pregel Hoderlein, Çağlar Şenaras, Timothy Davis, Daniel Cremers, Giovanni Marchisio, Xiao Xiang Zhu, and Laura Leal-Taixé. Dynamicearthnet: Daily multi-spectral satellite dataset for semantic change segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21158–21167, 2022. 1, 2, 3, 4, 5, 6, 7, 14, 15

[44] 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, 3, 4, 5, 6, 15

[45] Sagar Verma, Akash Panigrahi, and Siddharth Gupta. Qfabric: Multi-task change detection dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1052–1061, 2021. 2, 3

[46] Elliot Vincent, Jean Ponce, and Mathieu Aubry. Pixel-wise agricultural image time series classification: Comparisons and a deformable prototype-based approach. *arXiv preprint arXiv:2303.12533*, 2023. 3, 15

[47] RA Weismiller, SJ Kristof, DK Scholz, PE Anuta, and SA Momin. Change detection in coastal zone environments. *Photogrammetric Engineering and Remote Sensing*, 43(12):1533–1539, 1977. 2

[48] World Meteorological Organization. Earth observation satellites. -, 2024. https://wmo.int/topics/earth-observation-satellites,. 1

[49] Chen Wu, Lefei Zhang, and Liangpei Zhang. A scene change detection framework for multi-temporal very high resolution remote sensing images. *Signal Processing*, 124:184–197, 2016. 2

[50] Hao Xia, Yugang Tian, Lihao Zhang, and Shuangliang Li. A deep siamese postclassification fusion network for semantic change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022. 2

[51] Mengqiu Xu, Ming Wu, Kaixin Chen, Chuang Zhang, and Jun Guo. The eyes of the gods: A survey of unsupervised domain adaptation methods based on remote sensing data. *Remote Sensing*, 14(17):4380, 2022. 1, 3

[52] Kunping Yang, Gui-Song Xia, Zicheng Liu, Bo Du, Wen Yang, Marcello Pelillo, and Liangpei Zhang. Asymmetric siamese networks for semantic change detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2021. 2, 3

[53] Panli Yuan, Qingzhan Zhao, Xingbiao Zhao, Xuewen Wang, Xuefeng Long, and Yuchen Zheng. A transformer-based siamese network and an open optical dataset for semantic change detection of remote sensing images. *International Journal of Digital Earth*, 15(1):1506–1525, 2022. 2

[54] Xueliang Zhang, Pengfeng Xiao, Xuezhi Feng, and Min Yuan. Separate segmentation of multi-temporal high-resolution remote sensing images for object-based change detection in urban area. *Remote Sensing of Environment*, 201:243–255, 2017. 2

[55] Manqi Zhao, Zifei Zhao, Shuai Gong, Yunfei Liu, Jian Yang, Xiong Xiong, and Shengyang Li. Spatially and semantically enhanced siamese network for semantic change detection in high-resolution remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:2563–2573, 2022. 2

[56] Zhuo Zheng, Yanfei Zhong, Shiqi Tian, Ailong Ma, and Liangpei Zhang. Changemask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 183:228–239, 2022. 2

Appendix A - Dataset details

No domain shift setting. As illustrated in Figure 3a, SITS are spatially divided in 4 sub-SITS in this setting. We number all 512×512 quarters of the 1024×1024 SITS from 1 to 4. We follow the 4-fold validation scheme described in Table A2a, where a distinct quarter is kept for test purposes for each fold.

Spatial shift setting. We detail here the organization of DynamicEarthNet and MUDS's areas of interest into 5 distinct subsets for the spatial shift setting for reproducibility. The no domain shift and temporal shift settings are fully explained in Section 3.2. In Table A1, we report the compositions of each subset for both datasets. Areas of interest are designated in the table by their unique ID. Note that these subsets have been formed randomly for MUDS (two-class dataset) and in order to have similar class distributions for DynamicEarthNet (multi-class dataset). We show the class distribution for each fold in this setting in Figures A1 and A2. We follow the 5-fold validation scheme described in Table A2b inspired by [14].

DynamicEarthNet				
Set 1	Set 2	Set 3	Set 4	Set 5
2235_3403_13	2528_4620_13	1417_3281_13	1311_3077_13	1700_3100_13
4254_2915_13	2850_4139_13	1487_3335_13	2470_5030_13	2006_3280_13
4421_3800_13	4240_3972_13	2415_3082_13	2832_4366_13	2029_3764_13
4768_4131_13	4426_3835_13	2459_4406_13	4223_3246_13	2065_3647_13
5111_4560_13	4780_3377_13	2624_4314_13	4622_3159_13	2697_3715_13
5989_3554_13	4856_4087_13	3002_4273_13	4806_3588_13	4791_3920_13
6730_3430_13	5926_3715_13	3998_3016_13	5863_3800_13	4881_3344_13
6752_3115_13	6381_3681_13	4127_2991_13	6204_3495_13	5125_4049_13
6810_3478_13	6813_3313_13	4169_3944_13	6466_3380_13	6468_3360_13
6824_4117_13	7026_3201_13	4397_4302_13	7367_5050_13	6475_3361_13
8077_5007_13	7312_3008_13	4838_3506_13	7513_4968_13	6688_3456_13
MUDS				
Set 1	Set 2	Set 3	Set 4	Set 5
1446_2989_13	1549_3087_13	1736_3318_13	1327_3160_13	1429_3296_13
1474_3210_13	2345_3680_13	2027_3374_13	1433_3310_13	1950_3207_13
1831_3648_13	4056_2688_13	2176_3279_13	2265_3451_13	2287_3888_13
3041_4643_13	4102_2726_13	2383_3079_13	2528_4620_13	2309_3217_13
4061_3941_13	4553_3325_13	2459_4406_13	3911_3441_13	2732_4164_13
5184_3399_13	4742_4450_13	4802_4803_13	4838_3737_13	3699_3757_13
5342_3524_13	4815_3378_13	4816_3380_13	5753_3655_13	4196_2710_13
6460_3366_13	4819_3372_13	5105_3761_13	5927_3715_13	4688_2967_13
6679_3549_13	5156_3514_13	5193_2903_13	6460_3370_13	4840_4088_13
6813_3313_13	5916_3785_13	5759_3655_13	6468_3360_13	5557_3054_13
6993_3202_13	6678_3579_13	6154_3539_13	6678_3548_13	6691_3363_13
7394_5018_13	6838_3742_13	6864_3345_13	6764_3347_13	6763_3346_13

Table A1. **Composition of subsets in the spatial shift setting.** For reproducibility, we share the composition of our subsets in the spatial shift setting. Areas of interest are designated by their unique ID.

Additional details for MUDS. MUDS dataset is not originally designed for semantic segmentation (where annotations are semantic masks) but rather for object detection with polygons labeling all buildings contained in the images of the dataset. To adapt the annotations for our task, we color the interior of polygons in white on a black background to obtain binary semantic mask where 0 corresponds to the class 'not building' and 1 to the class 'building'. Note that some images of MUDS are not exactly of size 1024×1024 but have a side of size 1023. These images are resized alongside with their semantic mask to the size 1024×1024. Finally, some images from MUDS were acquired on 2017 and 2020: they are discarded for our study. Finally, MUDS comes with 'unusable data mask' annotations, *i.e.* binary masks indicating zones where the data cannot realistically be labeled because of cloud, shadows or geo-reference errors. 3.35% of total pixels are concerned: they are discarded from our metrics.

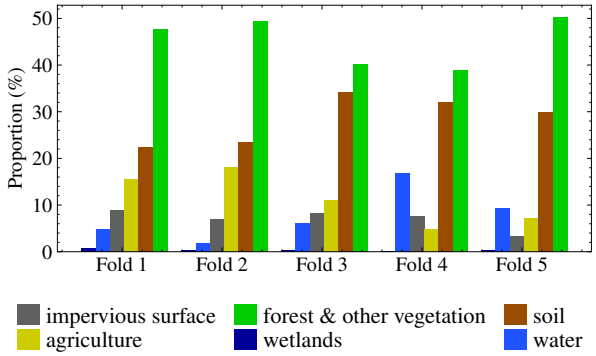


Figure A1. Class distribution per fold on DynamicEarthNet in the spatial shift setting.

Fold	Train	Val	Test
I	1-2	3	4
II	2-3	2	3
III	3-4	1	2
IV	4-1	4	1

(a) No domain shift setting

Fold	Train	Val	Test
I	1-3	4	5
II	2-4	5	1
III	3-5	1	2
IV	4-1	2	3
V	5-2	3	4

(b) Spatial shift setting

Table A2. 4- and 5-fold cross validation schemes for the setting without domain shift and the spatial shift setting. Each line gives the organization of the splits into train, validation and test set for each fold. The temporal domain shift scheme follows the usual train, validation, test single fold procedure.

Appendix B - Implementation details

We train all models on up to 4 NVIDIA GeForce RTX 2080 Ti or NVIDIA V100 GPUs in a data parallel fashion.

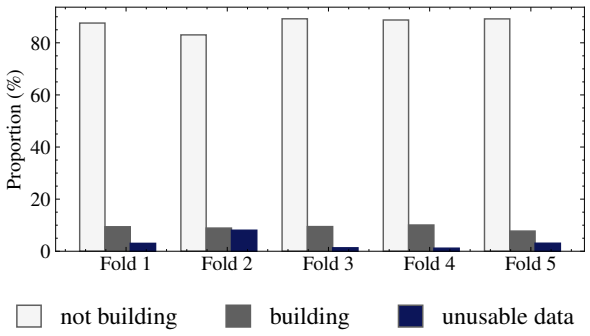


Figure A2. Class distribution per fold on MUDS in the spatial shift setting.

UTAE. We used UTAE official Pytorch implementation[3] with default parameters, except for the spatial feature size D that we change according to our different experiments. For positional encoding, we use the number of days after 01/01/2018 at the time of acquisition of the image. For example, an input time series for UTAE weekly corresponding to September 2018 will have [243, 247, 252, 257, 262, 267] as positional encoding vector, since 01/09/2018 is the 244th day since 01/01/2018, and the 5th, 10th, 15th, 20th and 25th days of each months are additionally selected in the weekly setting, following [43].

TSViT. We used TSViT official Pytorch implementation[4] with default parameters except for the feature dimension that we set to 512 for fair comparison with other evaluated methods, in terms of trainable parameters, even though [41] shows in their supplementary material that the feature size has only a limited impact on performance. We use an input image size of 64×64, since a size 128×128 is exceeding GPU memory capacity. We use the same positional encoding as we do with UTAE.

A2Net and SCanNet. We use A2Net and SCanNet Pytorch implementation available in the Unified Framework for Change Detection[5]. We train over all possible ordered pairs of images. We investigate several inference pairing strategies in Table B1 where the image pairs are defined as follows:

- Consecutive: $(\mathbf{x}_{2t}, \mathbf{x}_{2t+1})$ with $t \in \{1, ..., 12\}$;
- 12 months apart: $(\mathbf{x}_t, \mathbf{x}_{t+12})$ with $t \in \{1, ..., 12\}$;
- 6 months apart: $(\mathbf{x}_t, \mathbf{x}_{t+6})$ with $t \in \{1, ..., 6\} \cup \{13, ...18\}$;
- Random: $(\mathbf{x}_{t_1}, \mathbf{x}_{t_2})$ with $[t_1, t_2] \in \{[0, 15], [1, 13], [2, 9], [3, 23], [4, 10], [5, 14], [6, 18], [7, 19], [8, 21], [11, 12], [16, 17], [20, 22]\}$.

It is not clear what is the best strategy to divide a SITS into pairs of image for bi-temporal SITS-SCD approaches since the conclusion drawn depends on what is the considered score and is not always consistent across methods. For these reasons, we stick to the 'consecutive' strategy that we deemed is the most natural.

[3]https://github.com/VSainteuf/utae-paps
[4]https://github.com/michaeltrs/DeepSatModels
[5]https://github.com/guanyuezhen/UFCD

Pairing	A2Net				SCanNet			
	SCS↑	SC↑	BC↑	mIoU↑	SCS↑	SC↑	BC↑	mIoU↑
Consecutive	**22.2**	32.9	**11.5**	47.2	24.8	35.8	**13.9**	53.0
12 months apart	22.0	32.9	11.2	48.2	**24.9**	**36.6**	13.2	54.6
6 months apart	21.0	31.4	10.5	48.0	23.0	33.9	12.1	53.5
Random	21.9	**33.0**	10.8	**48.4**	24.5	36.4	12.6	**54.7**

Table B1. **Inference pairing for bi-temporal SITS-SCD.** We report the SCS, SC, BC and mIoU of A2Net and SCanNet for various manners of pairing time stamps at inference.

	SCS↑	SC↑	BC↑	mIoU↑
$d = 4$	**31.7**	**41.0**	**22.4**	**60.5**
$d = 16$	31.5	40.8	22.2	60.4

Table C1. **Impact of d on performance.** We report the SCS, SC, BC and mIoU of our method in the setting without domain shift on DynamicEarthNet for two values of d.

Appendix C - Impact of d on performance

In the paper, we investigate the impact of the spatial feature size D. While the feature maps contain spatial representations, the keys and queries of the attention mechanism store temporal information. Thus, we also analyze the impact of the dimension of keys and queries d. We train and evaluate our model on DynamicEarthNet in the setting without domain shift with $d = 16$ instead of $d = 4$ and report the results in Table C1. We see little difference if not a very slight decrease for all metrics.

Appendix D - Additional quantitative results

We report in Table D1 the per-class mean intersection-over-union (IoU) in the setting without domain shift on DynamicEarthNet and MUDS. Our method not only achieves best performance in terms of mean IoU but also on each class in all seetings on MUDS and in the setting without domain shift and in the temporal shift setting on DynamicEarthNet. Note

[6]https://github.com/CUG-BEODL/TSSCD

TSSCD. We use TSSCD official Pytorch implementation[6]. Since the method requires pixel time series as input, we flatten 128×128 image patches. At inference all 16,384 resulting pixel time series are used in a batch. During training, we randomly sample 64 pixel time series out of them.

	Method	Input type	Strategy	DynamicEarthNet							MUDS		
				imp. surf.	agr.	forest	wetlands	soil	water	mean	not build.	build.	mean
No domain shift	TSViT monthly	Single image	Single	26.7	46.5	72.6	13.3	56.7	87.1	50.5	91.8	28.6	60.2
	UTAE monthly	Single image	Single	33.7	53.7	75.9	11.8	59.6	87.7	53.7	92.7	41.5	67.1
	TSViT weekly	SITS	Single	29.5	49.4	73.9	9.4	56.8	86.6	50.9	—	—	—
	UTAE weekly	SITS	Single	33.9	56.4	76.6	10.6	60.8	88.2	54.4	—	—	—
	A2Net	Image pair	Bi	26.2	38.2	71.1	7.6	54.2	86.1	47.2	91.5	31.5	61.5
	SCanNet	Image pair	Bi	30.7	53.7	74.6	12.8	58.6	87.9	53.0	92.1	37.6	64.9
	TSSCD	Pixel-wise SITS	Multi	0.2	11.5	65.6	0	44.9	81.3	33.9	80.5	14.9	47.7
	Ours	SITS	Multi	**41.6**	**67.3**	**80.3**	**17.9**	**65.5**	**90.5**	**60.5**	**93.8**	**50.2**	**72.0**
Temporal domain shift	TSViT monthly	Single image	Single	24.4	34.5	71.1	15.7	52.5	86.1	47.3	91.4	22.2	56.8
	UTAE monthly	Single image	Single	37.0	50.5	74.9	17.4	55.5	86.8	53.7	92.9	39.0	66.0
	TSViT weekly	SITS	Single	28.4	42.4	72.1	22.4	54.2	88.0	51.4	—	—	—
	UTAE weekly	SITS	Single	37.4	52.8	75.1	15.8	57.9	89.0	54.7	—	—	—
	A2Net	Image pair	Bi	32.8	33.2	69.5	7.1	50.6	86.8	46.7	91.5	20.8	56.1
	SCanNet	Image pair	Bi	31.3	55.9	76.0	23.3	58.3	88.7	55.6	92.3	33.4	62.8
	TSSCD	Pixel-wise SITS	Multi	0	0	62.2	0	39.0	75.2	29.4	83.3	15.9	49.6
	Ours	SITS	Multi	**42.5**	**66.3**	**79.6**	**30.0**	**61.8**	**90.2**	**61.7**	**94.0**	**48.2**	**71.1**
Spatial domain shift	TSViT monthly	Single image	Single	12.2	9.6	55.8	0	40.4	69.4	31.2	90.8	8.9	49.8
	UTAE monthly	Single image	Single	27.1	11.1	**64.2**	0.1	43.0	76.1	36.9	92.1	33.9	63.0
	TSViT weekly	SITS	Single	15.1	13.4	55.7	**0.7**	40.7	67.6	32.2	—	—	—
	UTAE weekly	SITS	Single	29.1	15.1	63.1	0.1	43.4	76.3	37.8	—	—	—
	A2Net	Image pair	Bi	24.6	**19.9**	63.4	0.1	41.7	**77.7**	37.9	90.9	15.1	53.0
	SCanNet	Image pair	Bi	22.0	14.9	64.1	0.1	**46.5**	75.9	37.3	90.1	27.6	58.8
	TSSCD	Pixel-wise SITS	Multi	0.1	1.7	61.7	0	39.2	34.9	22.9	78.4	8.7	43.6
	Ours	SITS	Multi	**30.2**	16.8	62.5	0.3	45.2	76.2	**38.5**	**92.5**	**39.9**	**66.2**

Table D1. **Per-class IoU.** We report for our method and competing methods the per-class mean intersection-over-union (IoU) in the setting without domain shift on DynamicEarthNet [43] and MUDS [44].

that performance are consistent across all methods, with the hardest class being also the less represented in the datasets ('wetlands' on DynamicEarthNet and 'building' on MUDS). An additional comment to be made is that the pixel-wise method TSSCD [17] performs much worse than all others, once again showing that spatial context-related information leveraged by whole-image methods is crucial for segmentation tasks with SITS. This is a well-known fact in remote sensing, already discussed in [46] for example.

Appendix E - Additional qualitative results

We show in Figures E1 and E2 the semantic predictions of our model for 12-month time series on DynamicEarthNet and MUDS respectively in the setting without domain shift. Our architecture is able to capture changes while jointly processing the whole time series at once. Note how our model can classify accurately buildings masked as 'unusable data' in MUDS dataset.

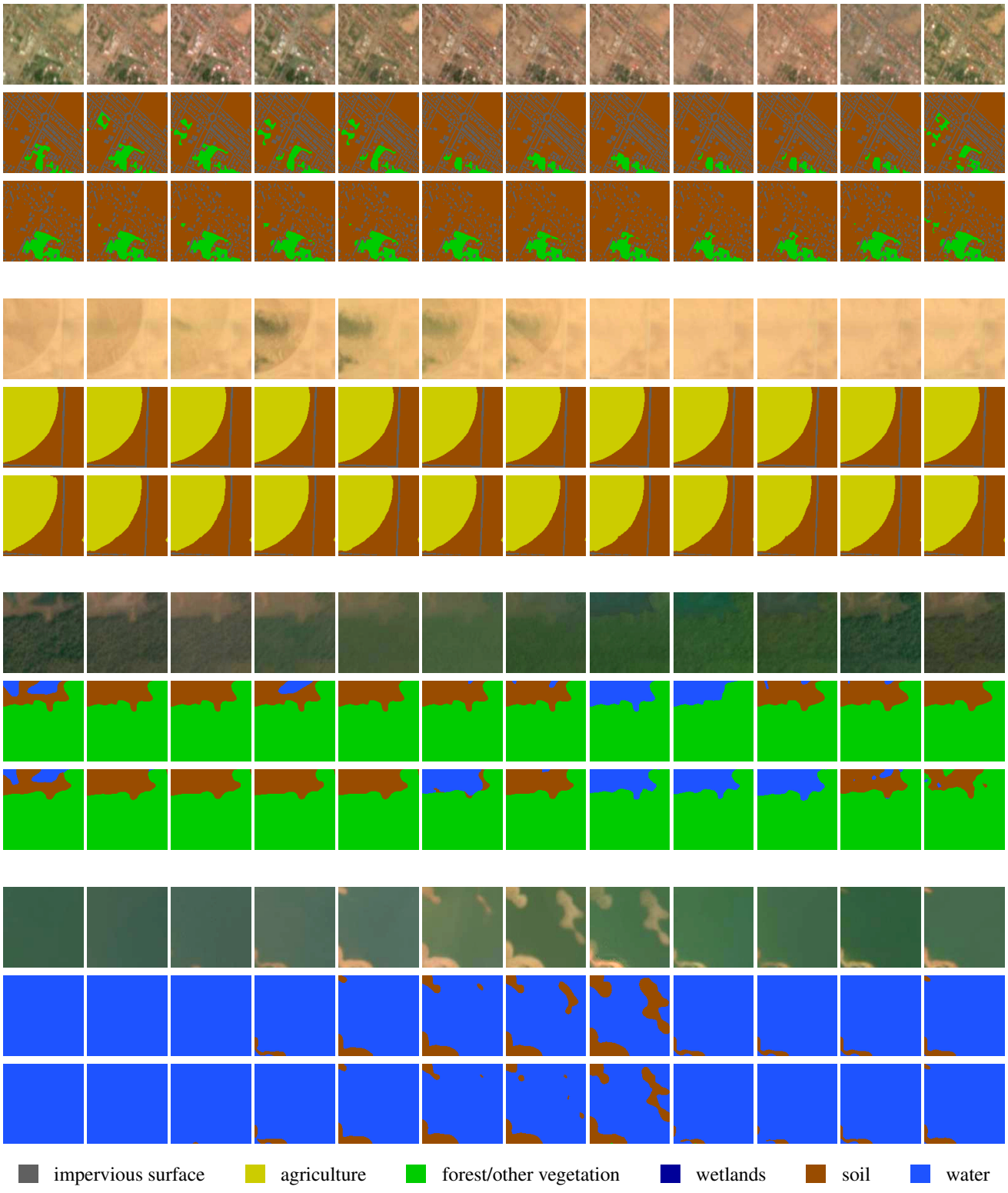


Figure E1. **Qualitative segmentation results on DynamicEarthNet.** We show the segmentation maps predicted by our model in the setting without domain shift for four randomly selected input SITS from DynamicEarthNet. For each SITS, we show the monthly input time series from January to December 2019 (top row), the corresponding ground truth (middle row) and the predictions (bottom row).

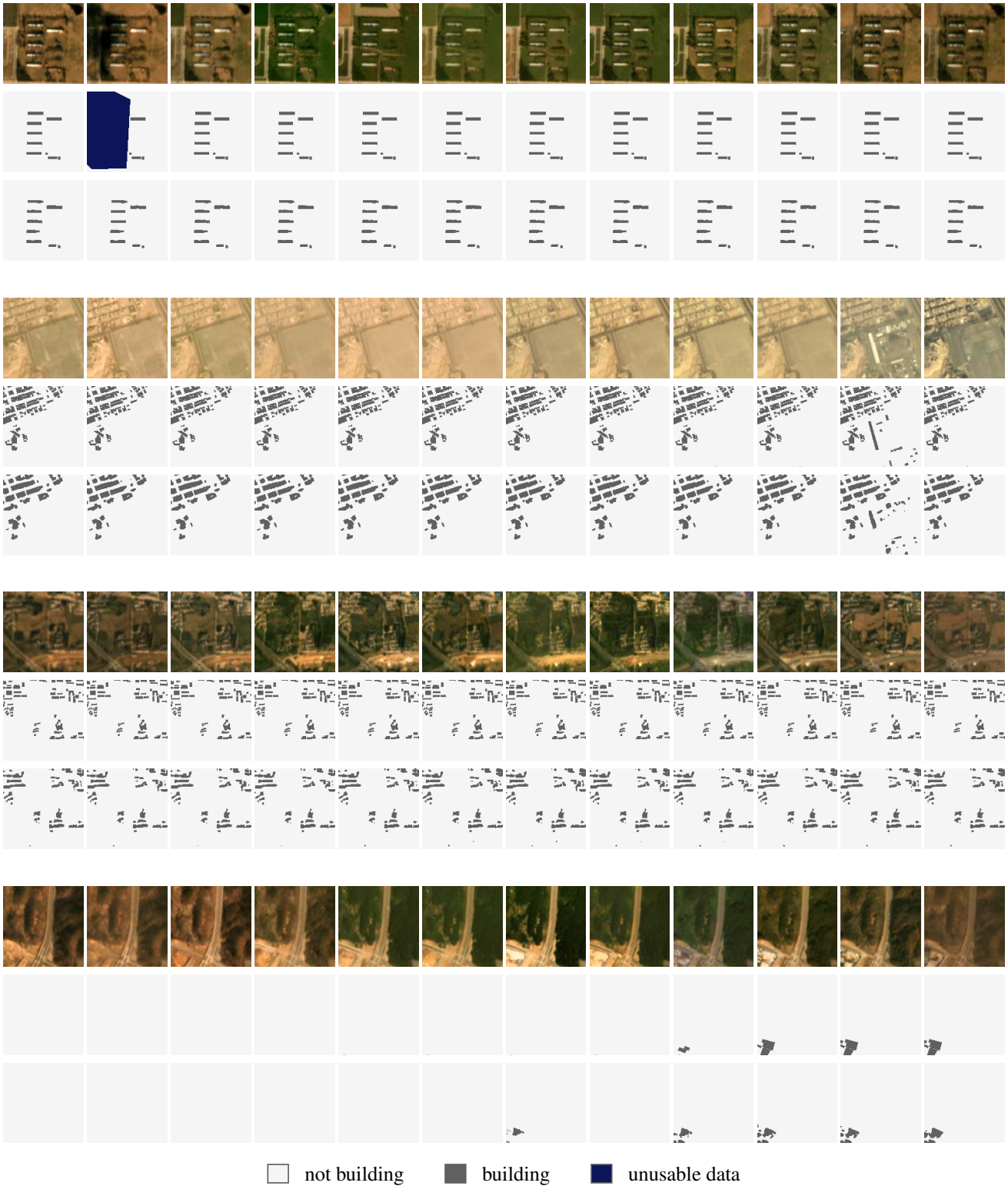


Figure E2. **Qualitative segmentation results on MUDS.** We show the segmentation maps predicted by our model in the setting without domain shift for four randomly selected input SITS from MUDS. For each SITS, we show the monthly input time series from January to December 2019 (top row), the corresponding ground truth (middle row) and the predictions (bottom row).