



Technische Universität München  
TUM School of Engineering and Design

# Deep Learning for Time-Series Analysis of Optical Satellite Imagery

Lukas Johannes Kondmann

Vollständiger Abdruck der von der TUM School of Engineering and Design der Technischen Universität München zur Erlangung eines

**Doktors der Ingenieurwissenschaften (Dr.-Ing.)**

genehmigten Dissertation.

**Vorsitz:**

Prof. Dr. phil. nat. Urs Hugentobler

**Prüfer:innen der Dissertation:**

1. Prof. Dr.-Ing. habil. Xiaoxiang Zhu
2. Prof. Dr.-Ing. habil. Michael Schmitt
3. Prof. Devis Tuia, Ph.D.

Die Dissertation wurde am 20.04.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Engineering and Design am 19.08.2023 angenommen.



Dedicated to life. Arguably one of the more interesting things to  
happen in a while.





## ABSTRACT

---

Significant progress has been made in deep learning for Earth observation data in recent times. However, multi-temporal applications of satellite images such as change detection and agriculture still face a dataset bottleneck. This does not only hinder training performance but prevents standardized and transparent evaluation protocols. In this cumulative thesis, I cover four papers that aim to improve dataset availability and make use of these resources for methodological innovation in change detection. The contribution is split in two parts. The first one introduces DENETHOR and DynamicEarthNet, two landmark datasets with high-quality ground truth data for agricultural monitoring and change detection respectively. The baseline experiments on both datasets point towards a need for tailored methods since current methods seem unable to effectively use the data's high temporal and spatial resolution.

Second, I introduce SiROC and SemiSiROC, two methodological contributions to label-efficient change detection. SiROC is an unsupervised method based on distant neighborhood analysis for binary change detection in optical images. SiROC performs competitively on four datasets from a range of change detection applications. With SemiSiROC, I exploit an insightful confidence measure built-in in SiROC for pseudo labeling of unlabeled scenes. The confidence measure allows prioritization of relevant scenes for pretraining deep learning based change detection methods with pseudo labels. Then, actual labels can be reserved for finetuning the model further. Overall, our results underline that this semi-supervised pipeline boosts overall performance notably for all methods we explore. This finding is robust to various ablation studies and underlines how the advantages of traditional methods and deep learning can be combined for maximized change detection performance.

## ZUSAMMENFASSUNG

---

Der Einsatz von tiefen neuronalen Netzen in der Zeitreihenanalyse von Erdbeobachtungsdaten nimmt in den letzten Jahren stetig zu. Allerdings erfordern diese Methoden große Mengen an Referenzdaten, die aktuell nur begrenzt vorhanden sind. Diese Lücke verhindert nicht nur ein besseres Training der Methoden, sondern macht die Evaluation aktueller Methoden intransparent und nicht standardisiert. In dieser Doktorarbeit fasse ich die Beiträge vier verschiedener Aufsätze zusammen, die die Datenverfügbarkeit und Methodik für Zeitreihenanalyse optischer Satellitendaten verbessern. Strukturell ist die Arbeit in zwei Teile unterteilt. Der erste Teil präsentiert die Referenzdatensätze DENETHOR für die Landwirtschaft und DynamikEarthNet für die Veränderungsanalyse. Beide Datensätze markieren einen erheblichen Fortschritt in der Verfügbarkeit von hochqualitativen Referenzdaten für die Zeitreihenanalyse. Erste Experimente auf beiden Datensätzen zeigen, dass aktuelle Methoden der temporalen Tiefe der Daten bisher nur bedingt gerecht werden und es spezifische Methoden für diesen Einsatz benötigt.

Im zweiten Teil lege ich methodische Innovationen für die Veränderungsanalyse unter beschränkter Verfügbarkeit von Referenzdaten dar. SiROC ist eine Methode, die gänzlich ohne Referenzdaten auskommt und auf der Modellierung von Pixeln auf Basis entfernter Nachbarn beruht. SemiSiROC kombiniert SiROC mit tiefen neuronalen Netzen in einem semi-überwachten Ansatz. Zunächst werden potentielle Veränderungen in Szenen ohne verfügbare Referenzdaten mit Hilfe von SiROC vorhergesagt. Diese Vorhersagen werden für das initiale Training von neuronalen Netzen als Referenzdaten behandelt, obwohl es sich technisch gesehen um Vorhersagen handelt. Dies ermöglicht allerdings mit erheblich mehr Daten zu trainieren und die echten Referenzdaten für die Feinarbeit im Training zurückzuhalten. Somit können erhebliche Gewinne in der Genauigkeit tiefer neuronaler Netze für die Veränderungsanalyse erreicht werden. Die Verbesserungen durch SemiSiROC sind äußerst robust und bleiben auch bei der Veränderung verschiedener Parameter bestehen. Somit zeigt SemiSiROC auf wie neuronale Netze und traditionelle Methoden in der Erdbeobachtung kombiniert werden können, um die Genauigkeit der Methoden zu maximieren.

## ACKNOWLEDGMENTS

---

First, I would like to sincerely thank my supervisor, Xiao Xiang Zhu, for her guidance during my Ph.D. studies. I did a number of things off the beaten path during my Ph.D. which would not have been possible without her trust and support. Second, I want to thank my examiners, Michael Schmitt and Devis Tuia, for their roles on my committee and Richard Bamler for leading it (“Vorsitz”).

In recent years, I had the pleasure to learn from some exceptional teachers, leaders, and mentors. I would not be where I am today without Davide Cantoni, my economics mentor at LMU. His support even through all my data science detours means a lot to me. Sebastian Boeck and Rogerio Bonifacio offered me the fantastic opportunity to bring my research closer to practice with a fellowship at the World Food Programme. The whole Analytics R&D Team at Planet Labs Berlin played an influential role throughout my Ph.D. studies which peaked of course in my internship with them. I had the tremendous chance to work for the World Bank in the dedicated team of Parmesh Shah and Kateryna Schroeder. I learned so much already about bringing my research to policy and that Bruno Sánchez-Andrade Nuño has even more energy than characters in his name.

Throughout my Ph.D. research, I was fortunate to work with smart and dedicated people across many disciplines. Many of them have also become friends. Aysim Toker played a significant role, particularly in my early days when we could commiserate about some of the challenges of large dataset projects. Marc Rußwurm always had an open ear for any agriculture-related questions and helped me tremendously to bring Denethor to life. Yuki Asano introduced me to the depths of self-supervised learning and remains a trusted source of advice about AI and life more generally. Konstantin Klemmer introduced me among many other things to Climate Change AI which has been amazing. I could always rely on Sudipan Saha as a source of advice regarding change detection or research more generally. Beyond that, many colleagues made my time at TUM/DLR fun. Special thanks go to Anja Rösel and Andrés Camero who have put up with all my ideas and wishes over the years.

Outside of research, a lot of people make my life in Munich great. I want to thank my parents, Stefan and Dorothea, for their unconditional support during my studies and in life. Among many others, I have bothered Philipp, Sebastian, Daniel, Tassilo, Julia, and Janus the most with the perks and problems of Ph.D. life. Finally, my deepest gratitude goes to my partner Camille. Thank you for putting up with me and all my stunts across the globe for more than a decade.



# CONTENTS

---

<b>I</b>	<b>INTRODUCTION AND FOUNDATION</b>	<b>1</b>
1	INTRODUCTION	3
2	FOUNDATIONS	5
2.1	A Primer on Deep Learning	5
2.2	Multispectral Satellite Imagery	6
2.3	In-Situ Data	7
2.3.1	Agricultural Field Data	7
2.3.2	Change Detection	8
<b>II</b>	<b>CONTRIBUTIONS</b>	<b>9</b>
3	TIME-SERIES DATASETS	11
3.1	Crop Type Mapping: Methods and Datasets	11
3.1.1	Motivation	11
3.1.2	Early Methods	12
3.1.3	Deep Learning Methods	12
3.1.4	Available Datasets	13
3.2	DENETHOR: A dataset for crop-type mapping from daily data	13
3.2.1	Dataset Motivation and Description	13
3.2.2	Baselines	15
3.2.3	Baseline results	16
3.3	Change Detection	18
3.3.1	Motivation	18
3.3.2	Available Datasets	19
3.4	DynamicEarthNet: Monthly semantic change detection from daily data	20
3.4.1	Motivation	20
3.4.2	Multitemporal Semantic Segmentation Baselines	21
3.4.3	Evaluation Metric	22
3.4.4	Results	22
4	METHODOLOGICAL CONTRIBUTIONS TO CHANGE DETECTION	25
4.1	State-of-the-Art	25
4.1.1	Unsupervised Methods	25
4.1.2	Supervised Methods	26
4.1.3	Semi-Supervised Methods	27
4.2	SiROC: Sibling-Regression for Optical Change Detection	27
4.2.1	Motivation	27
4.2.2	Methodology	28
4.2.3	Data	31
4.2.4	Experiments	31

4.2.5	Results	33
4.2.6	Discussion	38
4.3	SemiSiROC: Semi-Supervised Change Detection With Optical Imagery	38
4.3.1	Motivation	38
4.3.2	Methodology	39
4.3.3	Data and Evaluation	41
4.3.4	Results	42
4.3.5	Discussion	48
III	CONCLUSION AND OUTLOOK	49
5	CONCLUSION AND OUTLOOK	51
5.1	Conclusion	51
5.2	Outlook	53
IV	APPENDIX	57
A	PUBLICATIONS	59
A.1	DENETHOR	59
A.2	DynamicEarthNet	73
A.3	SiROC	84
A.4	SemiSiROC	100
	BIBLIOGRAPHY	115

## LIST OF FIGURES

---

Figure 1	Average daily NDVI values for ten wheat and meadow fields in Brandenburg, Germany.	12
Figure 2	Three ways to operationalize the crop type mapping task with the field boundaries and satellite images as inputs	15
Figure 3	Locations of the DynamicEarthNet dataset.	20
Figure 4	Half-Sibling Regression (HSR) for Change Detection	28
Figure 5	Qualitative Comparison OSCD - Las Vegas.	34
Figure 6	Confidence-Performance Plots on four cities of the OSCD Dataset.	35
Figure 7	An overview of SemiSiROC.	39
Figure 8	DynamicEarthNet spatial split	41
Figure 9	Qualitative results of 8 sample image pairs in a rural setting	43
Figure 10	Qualitative results of 8 sample image pairs in an urban setting	45

## LIST OF TABLES

---

Table 1	Existing Datasets for Crop Type Classification.	13
Table 2	Accuracy of Benchmark Models with Planet Fusion data on the 2019 test set trained with 2018 data.	16
Table 3	Accuracy of different modalities with hand-designed features and a random forest classifier on the 2019 test set trained with 2018 data	17
Table 4	Existing Datasets for Change Detection.	21
Table 5	Monthly, weekly and daily baselines on DynamicEarthNet’s Semantic Change Segmentation test set	24
Table 6	Quantitative Results OSCD Test Set	33
Table 7	Quantitative Results Beirut Explosion	36
Table 8	Quantitative Results Agriculture Dataset	37
Table 9	Quantitative Results Alpine Dataset	37
Table 10	Quantitative Results DynamicEarthNet grouped by pseudo label use.	42
Table 11	Quantitative Results DynamicEarthNet with different pseudo labels	46
Table 12	Ablation Study: Varying the Training Set Size	46
Table 13	Ablation Study: Robustness to Finetuning Loss	47
Table 14	Ablation Study: PL Training not on Test Images with SiamUNet	47
Table 15	Quantitative Results OSCD Test Set trained on DynamicEarthNet and grouped by pseudo label use.	48



Part I

INTRODUCTION AND FOUNDATION



## INTRODUCTION

---

Remote sensing is entering a new era of time-series analysis. The paradigm shifts from mapping to monitoring our Earth thanks to the increasing capabilities of new satellites and the expansion of satellite constellations. In the eye of the climate crisis, the need to monitor the effects of natural and human activity on the Planet becomes ever more relevant. Artificial Intelligence can help to process and make sense of large amounts of satellite data and support many applications [137]. Many of these use cases are collected under the term ‘change detection’ which refers to finding differences in multitemporal satellite imagery. This includes many impactful scenarios such as damage detection after natural disasters, deforestation, or urbanization monitoring. Particularly for multi-temporal applications, however, the availability of large-scale, high-quality reference data is limited. This constrains methodological progress for AI in change detection or agricultural applications for several reasons. At first, the lack of training data simply leads to worse results since many AI methods are data-hungry. Second, missing benchmark datasets often lead to intransparent evaluation standards. This is because many researchers then evaluate their methods on different datasets which makes it difficult to compare them against each other. Therefore, this thesis aims to accomplish two main objectives:

1. Improve the availability of large-scale datasets for deep learning in multi-temporal Earth observation.
2. Push the frontier of change detection with label-efficient techniques.

These two objectives are described in two separate sections covering the work across four publications:

- L. Kondmann, A. Toker, M. Rußwurm, A. Camero Unzueta, D. Peressuti, G. Milcinski, N. Longépé, P.-P. Mathieu, T. Davis, G. Marchisio, L. Leal-Taixé, and X.X. Zhu “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, 2021
- A. Toker\*, L. Kondmann\*, M. Weber, M. Eisenberger, A. Camero, J. Hu, A. P. Hoderlein, C. Senaras, T. Davis, D. Cremers, G. Marchisio, X. X. Zhu, and L. 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, 2022

- L. Kondmann, A. Toker, S. Saha, B. Schölkopf, L. Leal-Taixé, and X. X. Zhu, “Spatial context awareness for unsupervised change detection in optical satellite images,” IEEE Transactions on Geoscience and Remote Sensing, 2021.
- L. Kondmann, S. Saha, and X. X. Zhu, “SemiSiROC: Semi-Supervised Change Detection With Optical Imagery and an Unsupervised Teacher Model” IEEE Journal of Selected Topics in Applied Earth Observation and Remote Sensing, 2023 (accepted).

Some of the ideas and Figures in this thesis have also appeared in the publications above which are appended to this thesis. After the discussion of foundations, section 3 describes the first two publications which are related to the dataset objective. Section 4 summarizes the methodological advances in the second pair of papers for change detection. Finally, section 5 concludes and describes potential for future research.

## 2.1 A PRIMER ON DEEP LEARNING

The goal of this primer is not an exhaustive review of the history of deep learning. Instead, my aim is to review recent methodological developments which are relevant for this thesis. A key moment for the popularity of deep learning in image recognition was the landmark victory of the AlexNet [62] architecture in the ImageNet competition of 2012. It represents the first instance of a deep neural network successfully trained on more than a million images and almost halved the error rate for object recognition [65]. AlexNet was based on methodological advances in convolutional neural networks (CNNs) [66] and fueled by the increasing performance of graphical processing units (GPUs). This kickstarted a variety of new convolutional architectures which became popular over the following years. The Visual Geometry Group (VGG) Net [112] further developed AlexNet's advances in 2014 into a deeper architecture and is still a popular model. ResNets [49] popularized the use of skip connections. In deep neural networks, the features can become abstract and complex which may hinder straightforward decisions and deeper networks do not necessarily lead to better performance from a certain point on. The idea behind skip connections is to allow alternative shorter paths through the network so that certain parts may be skipped and replaced with an identity mapping. Therefore, it becomes possible to add even more layers to the network compared to previous architectures.

Convolutional networks can transfer to other tasks and domains given the right conditions. This means that a model that was trained for one task may still be used as a starting point for another application. It may need to be finetuned for this new use but the previous model can still be utilized. Therefore, the rise of CNNs for image classification also fueled advances in related computer vision tasks such as semantic segmentation. Fully convolutional networks (FCNs) relieve the restriction of standard CNNs that are bound to a fixed input size and exploit this for semantic segmentation. The UNet architecture [98] was originally proposed for biomedical image segmentation but has become widely popular in other segmentation tasks as well. It contains a contracting path that is comparable to a CNN for image classification and an expansion path where the learned representations are gradually upsampled to the original image size for a segmentation output. This upsampling procedure is based on an up-convolution operation.

While convolutional networks have drastically improved computer vision performance across a variety of tasks they do not natively handle sequential data. This, however, is a crucial element of time-series analysis. One way to deal with this limitation in a bi-temporal setting is a siamese structure. Both inputs are processed separately with shared weights and the extracted features are further processed and compared afterwards to reach a final decision. This could be determining a change or if the same person is visible in both input images.

As longer sequences are common in Earth observation, however, a class of networks with high relevance is recurrent neural networks (RNNs). RNNs can process sequential data as they have built-in loops on top of a feed-forward architecture. However, standard RNNs have fairly short memory. This means that it becomes ever harder for the model to remember early inputs of the sequence the longer it is. Long short-term memory networks [51] are designed to counter this weakness with specific cells that include a constant error carousel. This connection across the sequence makes it easier for information to persist across long sequences within the model.

In recent years, transformers have surpassed LSTMs for many sequential applications [119]. They are purely based on self-attention entirely without the use of convolutions or recurrence. While initially pioneered in natural language processing, transformers are fairly common also in image recognition applications by now. Vision transformers [33] achieve new heights in image recognition performance without using convolutions which laid the groundwork for a lot of current research. While transformers can be more efficient than CNNs, they are typically data hungry which means a lot of pretraining is essential. Therefore, CNNs can often still be practical alternatives when data availability is limited. In the absence of labels, vision transformers require a lot of compute for self-supervised pretraining. Current research with transformers under limited labels, for example, successfully combines vision transformers with masked autoencoders [48]. Alternatively, better data availability can guide methodological progress which is a primary goal of this thesis.

## 2.2 MULTISPECTRAL SATELLITE IMAGERY

Multispectral images are the result of optical instruments on board of satellites that orbit the Earth while imaging it. The collected data is transferred to ground stations which allows us to get a perspective on Earth from above. The on-board instrument is able to separate reflections in different wavelengths of the electromagnetic spectrum. In many cases, this includes more channels than the red, green, and blue (RGB) wavelengths human eyes capture. For example, near-infrared (NIR) emissions increase with the heat of on an object which can be helpful to distinguish objects on the ground. In this thesis,

two different kinds of multispectral satellite data sources are used: Sentinel-2 and Planet Fusion. Sentinel-2 is operated by the European Space Agency (ESA) and is part of the Copernicus Programme [2]. The Copernicus Programme has an open data policy which implies that the collected data is shared for free for scientific research or commercial purposes. Sentinel-2 collects 13 different spectral bands with a resolution varying from 10-60m per pixel. RGB bands and one NIR band are, for example, captured at 10m spatial resolution. Two Sentinel-2 satellites are in sun-synchronous orbit which means that they visit the same spots on Earth during the same time of day. Every larger landmass on Earth is reimaged by one of the satellites at least every five days. The closer to the poles the area is, the shorter the revisit intervals. This is because the satellites pass over or close to the poles on every orbit but can capture only a fraction of the Earth at the equator at each pass [34]. Planet Fusion is a commercial product offered by Planet Labs based on the Planetscope constellation. It comes in 3m resolution and with four spectral channels (RGB+NIR). One image a day is available which makes its spatial and temporal resolution significantly higher compared to Sentinel-2. The data is shipped heavily preprocessed and is advertised as an ‘analysis-ready’ product. If clouds obstruct observations on certain days, the missing areas are gap-filled from the closest available date. The data is harmonized across time for temporal consistency and shadows are removed. Additionally, the spectral channels are calibrated to be consistent with Landsat and Sentinel data. This is important since there are small differences in the precise wavelength each of the sensors collects. For example, the wavelength of the red channel may slightly differ among the sensors. Fusion data is a harmonized Landsat Sentinel (HLS) product that aims to resolve these differences and make the data interoperable with these public sources of optical imagery.

## 2.3 IN-SITU DATA

### 2.3.1 *Agricultural Field Data*

The Common Agricultural Policy (CAP) of the European Union requires farmers to self-report the crops they plant in order to receive subsidies [43]. Some regions make the collected geodata around field boundaries and planted crops openly available which can be used for the creation of crop type datasets in conjunction with Earth observation data.

The EIONET Action Group on Land Monitoring (EAGLE) defines a common classification standard for agricultural activity. This thesis uses an aggregated version of this with nine classes: Wheat, rye, barley, oats, corn, oil seeds, root crops, meadows, and forage crops. Most of these are fairly distinct but oil seeds contain, for example,

sunflower or soy. Potatoes or sugar canes belong to root crops and forage crops are used for animal food.

### 2.3.2 *Change Detection*

Based on the Corine Land Cover (CLC) [13] classification scheme, five different types of land cover can be distinguished at the highest level: Impervious surfaces, agriculture, solid natural areas, wetlands, and water. In this thesis, I split solid natural areas into its subclasses forests and soil and additionally treat snow and ice as a separate class to create a more meaningful classification task with 7 categories. For the purpose of this thesis, a change is defined as a transition from one of the land cover classes to the other if not mentioned otherwise. Changes are generally rare but are often events of interest. This is particularly the case if the change of land cover is unexpected or unintended with events such as natural disasters.



Part II

CONTRIBUTIONS



## TIME-SERIES DATASETS

---

Multitemporal analysis is at the core of many relevant applications of Earth observation data. The time-series can be operationalized in different ways, however. In change detection, the explicit task is to determine what happened between two images from the same location at different times. In crop type mapping, a time-series of images from a growing season is typically used as an input to determine what was planted in the field. Here, the multitemporal analysis serves as means to an end in a classification task rather than being the explicit target. Both areas have been known fields for decades but they have been heavily influenced by recent advances in deep learning [137]. However, deep learning methods require suitable large-scale benchmark datasets which are scarce in both of these applications as outlined below.

### 3.1 CROP TYPE MAPPING: METHODS AND DATASETS

#### 3.1.1 *Motivation*

Agricultural analysis is among the premier uses of remote sensing data [84]. Satellites can shed light on a variety of aspects starting from identifying agricultural areas [121], the planted crops [95], or yield potential [126]. Further, they can inform about soil moisture [44], vegetation cycle indicators [56] and also sustainable farming practices [134]. In crop type mapping, the boundaries of a field are typically assumed as given and the task is to determine its crop type based on remote sensing imagery. Often, the input data will be multitemporal as the evolution of a crop over time is a key element for the prediction [85]. In its early steps, the amount of data to be processed was often a challenge so many approaches relied on feature extraction.

Tucker (1979) [93] was among the first to develop a feature-based approach to study vegetation with satellite data based on the Normalized Difference Vegetation Index (NDVI). NDVI is commonly defined as

$$\text{NDVI} = (\text{NIR} - \text{RED}) / (\text{NIR} + \text{RED}) \quad (1)$$

where NIR stands for the reflectance value in the near-infrared spectrum and red for the red spectrum respectively. Since plants absorb red but barely any infrared light during photosynthesis, a high NDVI close to one is indicative of vegetation activity on the ground. If the NDVI is close to zero, the ratio of reflected NIR and red light are similar and therefore not indicative of photosynthetic activity.

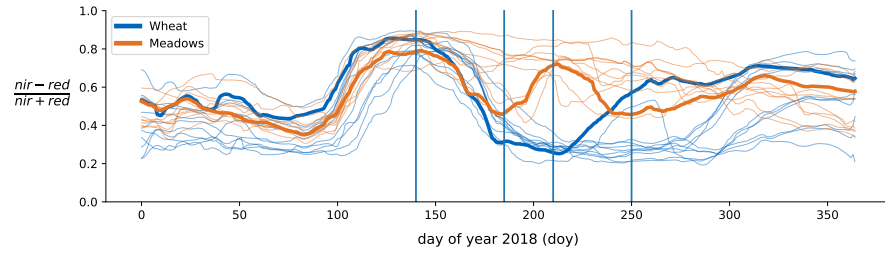


Figure 1: Average daily NDVI values for ten wheat and meadow fields in Brandenburg, Germany.

Figure 1 gives an intuition how temporal features of the NDVI can be used to differentiate crop types. It shows an NDVI time-series based on Planet Fusion for wheat and meadow fields in Brandenburg, Germany. The temporal patterns are distinct particularly between days of year 150 and 250 which helps algorithms to distinguish the two crop types. Many crop types therefore are not only distinct in their visual appearance in satellite images. Their temporal vegetation activity during the growing season is a critical input to classification models.

### 3.1.2 Early Methods

Many approaches make use of a version of vegetation indices for crop classification [25, 26, 36, 47, 89]. One popular approach for crop type mapping with vegetation indices is the combination with random forests [87, 117]. The time-series can, for example, be aggregated by determining temporal features of the average field pixel such as its mean, median or also extreme values [60, 102]. One advantage of this approach is its ability to scale [59]. Support Vector Machines are also frequently used in combination with vegetation index features [31, 63, 135]. Additionally, Dynamic Time Warping (DTW) [81] is a popular tool in the analysis of phenological stages with remote sensing data. [8, 27, 78].

### 3.1.3 Deep Learning Methods

The increasing popularity of deep learning [65] has inspired many innovations in crop type classification. Convolutional neural networks (CNNs) [66] can also be modified to incorporate temporal dimension in crop type mapping [88]. The resulting TempCNN [88] applies convolutions also to the temporal dimension. Recurrent neural networks have the capacity to operate on sequential data which makes them particularly suitable for time-series classification of crops [99, 100, 111]. Self-attention based mechanisms have surpassed the performance of CNNs and RNNs for many vision tasks in recent years [119].

Table 1: Existing Datasets for Crop Type Classification.

	Inputs	GSD	RT	#Fields	Size[GB]
Breizhcrops (FR) [102]	S2	10m	5 days	768,000	17.4
TimeSenzCrop (AUT) [122]	S2	10m	5 days	1,200,000	2.1
CV4A Kenya [91]	S2	10m	5 days	4,700	3.5
Crop Type Uganda [10]	S2	10m	5 days	52	59.4
Spot the Crop Challenge (SA) [92]	S1+S2	10m	5 days	35,300	52.1
DENETHOR (Contribution)	PF+S2+S1	3m	Daily	4,500	254.5

Rußwurm et al. [101] study the application of self-attention to raw optical time-series data since the attention mechanism helps to prioritize relevant scenes. A pixel set encoder with temporal self-attention (Pse-Tae) has shown promising results for temporal applications including crop type mapping [108]. In this work, a small window is randomly sampled from a given field to increase spatial variability.

#### 3.1.4 Available Datasets

Deep learning methods are generally data-hungry [65]. Therefore, they require large-scale datasets to be trained effectively which can often be an issue in remote sensing applications where reference data is typically scarce [137]. For Europe, EuroCrops [109] aims to collect and harmonize publicly available data through the CAP policy. Still, benchmark datasets that combine crop type data with remote sensing imagery are scarce. BreizhCrops [102] is a Sentinel-2 based dataset from the Brittany region in France with around 800,000 parcels from one growing season. TimeSenzCrop [122] covers large parts of Austria but contains only average pixel information per field. For Africa, some small competition datasets exist [10, 91, 92] mainly through the work of the Radiant Earth Foundation. Overall, however, the capacity of deep learning for crop type classification is limited by the availability of large-scale benchmark datasets. This becomes even more critical when looking at geographic or temporal generalization which is often a challenge in crop type mapping [83].

## 3.2 DENETHOR: A DATASET FOR CROP-TYPE MAPPING FROM DAILY DATA

### 3.2.1 Dataset Motivation and Description

As part of this thesis, the DENETHOR dataset for crop type mapping from daily data is presented. It stands for the **D**ynamic**E**arth**N**ET dataset for **H**armonized, **i**nter-**O**perable, **a**nalysis-**R**eady, **d**aily crop monitoring from space. It provides the first opportunity to explore analysis-ready data (ARD) from Planet Fusion for open scientific research. The area of interest is in Brandenburg, Germany and the

dataset contains Earth observation data, field boundaries and crop types for about 4500 fields in the years 2018 and 2019. The fields are from two spatially distinct areas where one is used for training with 2018 data and the other for testing with 2019 data. This aims to incentivize spatial as well as temporal generalization. Additional to the Planet Fusion data, the dataset contains Sentinel-1 and 2 imagery for the respective fields as well. Sentinel-2 is included at preprocessing level L2A. This implies that images have, among other steps, undergone atmospheric correction, orthorectification and spatial registration. 12 bands are given at 10m resolution which means that lower resolution bands have been resampled to 10m. No observations are excluded because of cloud coverage. For Sentinel-1 radar data that is not obstructed by clouds, we use the Ground Range Detected (GRD) product. Since a share of the radar waves may be repolarized when they interact with the surface, we include vertical-vertical (VV) and vertical-horizontal (VH) polarization values.

To understand why DENETHOR can extend the scope of methodological research in crop type mapping, Table 1 presents specifications of current datasets for crop type mapping together with DENETHOR. Current datasets, such as BreizhCrops or TimeSen2Crop are based primarily on Sentinel-2 data which has a resolution of 10m and a revisit time of 5 days. However, next-generation Earth observation products such as Planet Fusion can deliver data daily in higher resolution. Therefore, the opportunities for dense time-series data in conjunction with deep learning for agriculture are underexplored because of missing datasets. Denethor aims to bridge this gap between high-cadence time-series and deep learning. For additional experiments with Copernicus data, the dataset also contains Sentinel-1 and 2 data. This can shed light on the practical usefulness of Fusion imagery compared to common data sources for crop type mapping. While particularly TimeSen2Crop covers more fields than DENETHOR, their dataset provides only an average pixel per field. On the other hand, one goal of DENETHOR is to provide the full field every day to allow the community to explore methodological approaches for this. This results, however, in a notably larger dataset size. In summary, DENETHOR has the following advantages compared to previous datasets:

- Higher spatial resolution with 3m vs 10m
- Higher processing level as a harmonized and gap-filled product
- Higher temporal resolution with daily data.
- Data from two seasons and two locations which allows testing for spatial and temporal generalization

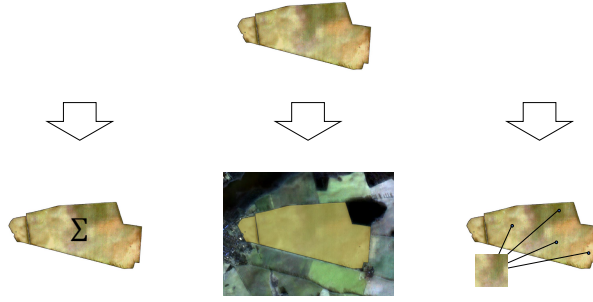


Figure 2: Three ways to operationalize the crop type mapping task with the field boundaries and satellite images as inputs

### 3.2.2 Baselines

To understand how current methods in crop type mapping perform on this novel dataset, a variety of baselines are tested in three categories shown in Figure 2. As fields vary in size and shape, we explore different ways to break down this issue. At first, one can average the pixels in space and only use the mean pixel per field over time as an input. This scales comparably well since the input size per field is minimized in space. However, large parts of spatial information are discarded.

Second, one can take images of the fields in a constant pixel size (here  $32 \times 32$ ) which requires downsampling of large fields and zero-padding of smaller ones. This increases data input size but allows to process spatial as well as temporal characteristics of a field. We use separate encoders for the spatial and temporal dimension. As spatial encoders we rely on small models such as ResNet18 [49], MobileNetv3 [52], and SqueezeNet [53]. One of the temporal encoders we use is TempCNN [88] which is a competitive model for temporal learning for crop type mapping discussed above. Further, we compare against a transformer [101] and an LSTM baseline [99] for the temporal encoding. Finally, MultiScale ResNet (MSResNet) has also shown promising results for crop type mapping in [102] and is included as a temporal encoder as well.

Third, PseTae [108] samples an equal number of pixels from each field which combines some spatial variation with the temporal component. The pixel set encoder transforms the field window into a representation with a multi-layer perceptron which is further processed by a temporal attention encoder. Additional to PseTae we also test a lightweight version of the model which is called PseLTae [39].

Besides the deep learning baselines, we further evaluate attempts based on random forests. This gives a reference point based on more traditional methods. For this, we use explicit temporal features of the mean pixel per field and not the time-series itself. We obtain the

Table 2: Accuracy of Benchmark Models with Planet Fusion data on the 2019 test set trained with 2018 data.

Spatial Encoder	Temporal Encoder			
	TempCNN [88]	MSResNet [120]	LSTM [100]	Transformer [101]
ResNet18 [49]	52.22%	49.53%	44.64%	43.61%
SqueezeNet [53]	53.94%	49.78%	35.89%	42.58%
MobileNetv3 [52]	53.20%	54.33%	43.46%	48.06%
Pixel Average [100]	64.46%	58.83%	48.40%	52.56%
Pixel-Set Encoding + Self-Attention				
PseLTae [39]	67.25%			
PseTae [108]	64.95%			
Ablation Scores				
PseLTae (2018)	78.77%			
PseLTae (Val)	88.02%			

minimum, maximum, argmin, argmax, mean, median, and the standard deviation for all available bands and additionally the respective NDVI. For example, as Planet Fusion has only 4 channels, this results in 35 features. With Sentinel-2, the number of features is 91. Since this approach is more focused on spectral depth rather than temporal or spatial resolution, it should benefit the Sentinel data more. However, one intention is also to explore data fusion of the different sources in this approach which may be fruitful because of their different modalities.

All models are trained until convergence with a cross-entropy loss and with imagenet weights as a starting point. All spatial and temporal encoders use defaults from their Breizhcrops or torchvision implementation respectively. We evaluate results based on accuracy and macro-averaged F1 score. Accuracy is defined as the number of correctly classified samples over all samples. An F1 score balances precision and recall. For the macro-average, the individual F1 scores per class are calculated separately and then averaged with equal weights. Alternatively, one could also use a weighted average by the number of samples per class but this score is easily distorted by the majority classes.

### 3.2.3 Baseline results

Table 2 contains scores of the deep learning based experiments with Planet Fusion data. In the first panel, results for spatial (vertical) and temporal (horizontal) are shown. Pixel average is the first approach from Figure 2 in which the field is spatially averaged instead of using an additional spatial encoder. Even though the pixel average strategy is simple, it performs better than all three other spatial encoders explored here which is somewhat surprising. It seems that it is not straightforward for the spatial encoders to extract meaningful infor-



Table 3: Accuracy of different modalities with hand-designed features and a random forest classifier on the 2019 test set trained with 2018 data

Data Type	# Features	Accuracy	Macro F1-Score
Sentinel-1 (S1)	42	0.58	0.43
Sentinel-2 (S2)	91	0.59	0.42
Planet (PL)	35	0.37	0.12
S1 + S2	133	0.62	<b>0.46</b>
S1 + PL	77	0.60	0.42
S2 + PL	126	0.59	0.41
S1 + S2 + PL	168	<b>0.63</b>	<b>0.46</b>

mation from the  $32 \times 32$  field images. Between the three encoders, it seems that MobileNet v3 is on average slightly better than ResNet18 or SqueezeNet but performance is similar.

Looking at temporal encoders, TempCNN reaches the highest score in the first panel with an accuracy of 64.46% followed by MSResNet. The LSTM as well as the transformer model do not seem competitive here with scores ranging from about 36% to 53%.

The best model overall is PseLTae with 67.25% which is visible in the second panel. It performs even slightly better than its heavier sibling model PseTae with 64.95%. It appears that the temporal self-attention mechanism can exploit discriminatory temporal features much better if fields are not resized or padded but pixels are sampled. Even if the whole field is averaged, however, the transformer model is not competitive to PseTae. The combination of some spatial variation together with temporal self-attention in the PseTae framework seems to make the initial difference here.

As the test results are obtained on 2019 data, we further test the effects of spatial and temporal shifts on accuracy in the third panel. PseLTae achieves 88% accuracy on validation data which is from the same year and the same tile. If we move to the test tile which is in the same geographic area the performance drops by 9 percentage points (p.p.). A key factor here is that the distribution of crop types shifts slightly between the two tiles even though they are geographically close. This is likely a factor in the observed performance drop. Interestingly, the performance drop becomes even larger by another 10.52 p.p. when we use the test data for 2019. This outlines the challenge of temporal generalization in crop type mapping. Through differences such as the weather, the vegetation cycles can differ significantly from year to year. This makes comparisons such as Figure 1 more difficult when comparing different years.

Table 3 contains the scores for feature-based experiments with random forests. At first, we compare data sources separately in the first

part. Temporal features from Sentinel-1 and 2 perform similarly with 58% and 59% in accuracy and 43% and 42% in macro F1 score respectively. Expectedly, temporal feature extraction with random forests seems to have few advantages with Planet Fusion on its own with comparatively low scores. This is because this approach prioritizes spectral depth over temporal and spatial resolution which is a relative weakness of the Planet Fusion data here.

In the second part, results from fusion experiments are reported where we use the same features from all respective input sources. Combining S<sub>1</sub> and S<sub>2</sub> yields an improvement of 3-4 p.p. accuracy over only using one of them. Combining PL with S<sub>1</sub> improves S<sub>1</sub> scores by 2 p.p. but combining PL with S<sub>2</sub> yields no effect. This seems intuitive since the modalities of S<sub>1</sub> and PL are different but similar for PL and S<sub>2</sub> since they are both multispectral satellites operating in similar spectra. Finally, combining all three satellite types has again little effect compared to only using S<sub>1</sub> and S<sub>2</sub> since spectral features from PL are likely somewhat redundant if S<sub>2</sub> is already included.

In comparison to the deep learning results of Table 2, the margin is fairly small. Only the best deep learning models surpass a random forest baseline even if only S<sub>2</sub> or S<sub>1</sub> is used. This suggests that current methods may not be cut out for this kind of data and there is ample potential for improvements to exploit the spatial and temporal depth better.

To summarize, DENETHOR provides the community with the first opportunity to explore next generation Earth observation products with daily inputs with deep learning. We show that current baselines are suboptimal to deal with the spatial and temporal depth of the data. This is because their edge over Sentinel-based random forest models is small even though these baselines disregard large parts of the spatial and temporal information. Additionally, we outline the hurdles of spatial and temporal generalization in crop type mapping as a starting point for future research. Beyond crop type mapping, DENETHOR could also be used, for example, for declouding or super-resolution experiments given the different kinds of EO data provided. It will be exciting to see what the scientific community builds on top of this dataset. Additional details on the dataset and baselines are provided in Appendix A.1.

### 3.3 CHANGE DETECTION

#### 3.3.1 *Motivation*

Change detection commonly refers to the task of identifying changes between images of the same location in multitemporal satellite images. This task supports monitoring of natural disasters [46, 73, 74, 80, 82, 128], forests [14, 17, 110], urban [54, 71] or mountain areas [23,

57] as well as sea and ice [38, 96]. The task is typically performed either as a classification task or segmentation task. The goal of a classification task is to determine if something between the two points in time has changed that fulfills the respective criteria. It is not necessarily determined how many changes happened or where in the image a specific event occurred. This is different with change segmentation where for every pixel in the time-series a decision is made if the pixel has changed since the last observation.

What makes change detection challenging is that differences between two images over time do not necessarily indicate a change. Varying acquisition conditions of the images can play a role because of different viewpoints, illumination conditions or clouds and shadows. Additionally, it is not necessarily straightforward what constitutes a change for a certain use case. Some applications only try to identify changes of a certain kind such as urban changes [16] or forest changes [30, 58] and may disregard other kinds of changes. Further, some changes on the ground do not always constitute a change in the corresponding land cover class. An example can be an agricultural field where the visual appearance changes significantly over the growing season but the blooming and harvest of crops are not unexpected. In some applications, this is not seen as a change. However, in other cases, the harvesting of crops may be the event of interest and therefore the primary goal of the change analysis.

### 3.3.2 Available Datasets

Available datasets for change detection are still small and lack geographic diversity [103]. Among them is the Onera Satellite Change Detection Dataset (OSCD) [28]. It contains Sentinel-2 image pairs from 24 locations across the globe with manually annotated changes in varying image sizes. HRSCD [29] is based on aerial imagery but labels are automatically generated from a public registry which is known to be faulty at times. Another aerial change detection dataset is Hi-UCD [115] but its geographic focus is limited to Estonia with annual revisits. The MUDS [118] dataset provides monthly revisits with Planet imagery and building segmentations. LEVIR-CD [19] includes very-high resolution bi-temporal images from Google Earth (GE) from 20 different regions from Texas, USA. Similarly, DSIFN-CD [130] uses five large GE image pairs from cities in China for urban change detection. Many of these datasets are small, have a strong geographic limitation, cover only two or few steps in time and come with binary ground truth only which significantly limits change detection research.

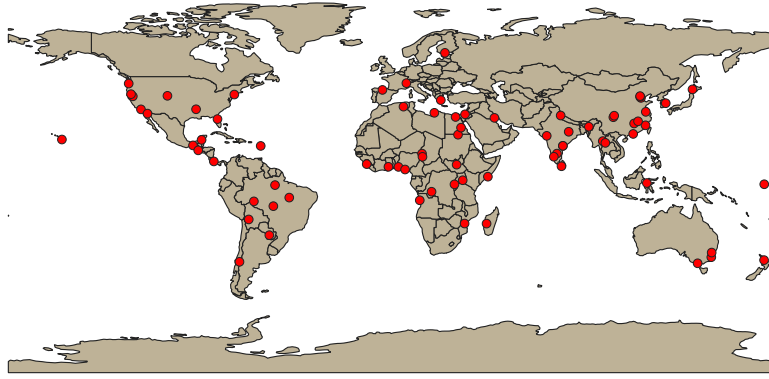


Figure 3: Locations of the DynamicEarthNet dataset.

### 3.4 DYNAMIC EARTHNET: MONTHLY SEMANTIC CHANGE DETECTION FROM DAILY DATA

#### 3.4.1 Motivation

To overcome dataset limitations in change detection we present the DynamicEarthNet dataset. It contains data from 75 areas of interest (AOIs) across the globe for the years 2018 and 2019 which are visible in Figure 3. AOIs are taken from all inhabited continents and cover a broad variety of changes such as deforestation, shoreline loss, or urbanization. The spatial dimensions of each AOI are  $1024 \times 1024$  pixels with daily Planet Fusion inputs. This adds up to an area of about  $10\text{km}^2$  per image with the 3m resolution. For every AOI, the dataset contains monthly, multi-class ground truth data which was manually annotated. Additionally, we provide monthly Sentinel-1 and 2 imagery for all locations for additional experiments.

Table 4 explores the specifications in more depth and compares them to existing datasets for change detection. The OSCD dataset is based on Sentinel-2 and is small with only 24 AOIs and two points in time. MUDS has monthly ground truth based on PlanetScope imagery but only for buildings. Technically, they do not provide segmentation masks but building polygons but the task can be converted to change segmentation as well. LEVIR-CD and DSIFN-CD are bi-temporal very-high resolution CD datasets obtained from Google Earth imagery but cover long intervals and are spatially limited to a small number of cities in Texas (LEVIR) and China (DSIFN). Change detection datasets based on satellites thus lack temporal coverage, multi-class annotations, scale and spatial diversity.

A second pillar of change detection datasets comes from aerial imagery. The imagery for these datasets comes from dedicated campaigns with limited spatial extent. The imagery for the WHU dataset comes from New Zealand and has binary ground truth on urban changes. The SECOND dataset in comparison is larger in scale and

Table 4: Existing Datasets for Change Detection.

	Inputs	GSD	AI	RT	# Images	Seg. Mask	Objects
OSCD [28]	S2	10m	2-3 years	2-3 years	48	Binary	Buildings
MUDS [118]	Planet	3m	Monthly	Monthly	2,389	Binary	Buildings
LEVIR-CD [19]	Google	0.5m	5-14 years	5-14 years	637	Binary	Buildings
DSIFN-CD [130]	Google	0.5m	Varies	Varies	788	Binary	Urban
WHU [55]	Aerial	0.3m	Yearly	Yearly	928	Binary	Urban
SECOND [127]	Aerial	<0.5m	Yearly	Yearly	9324	Semantic	Urban
HRSCD [29]	Aerial	0.5m	Yearly	Yearly	582	Semantic	Multiple
Hi-UCD [115]	Aerial	0.1m	Yearly	Yearly	2586	Semantic	Multiple
DynamicEarthNet	PF+S2+S1	3m	Monthly	Daily	54,750	Semantic	Multiple

contains semantic changes from several Chinese cities. Still, the observation times are long apart. HRSCD is based in France and the corresponding ground truth can at best serve a weak label [29]. Hi-UCD collects changes in Estonia with semantic ground truth but again the observation points are far apart and beyond the one in Estonia there is no AOI. Aerial data has the additional disadvantage that in the case of an event such as a natural disaster it is not collected for the whole globe in recurring circles but drones or planes have to be actively flown over the region of interest. Still, aerial imagery can provide valuable inputs to these situations but given the resolution difference, its use to train satellite based change detection methods is limited. Overall, Table 4 calls for a large benchmark dataset for CD with frequent observation times, multiple classes and global coverage. This is what we deliver with DynamicEarthNet. The monthly multi-class ground truth is a clear differentiator from other datasets. On top, the community can explore up to daily imagery for this purpose as supplementary input together with S2 and S1. Overall the dataset contains over 50,000 1024x1024 images for scientific analysis and we hope the dataset will allow CD research to transition from detection to continuous monitoring.

### 3.4.2 Multitemporal Semantic Segmentation Baselines

To provide a starting point for research on the DynamicEarthNet dataset, we benchmark a variety of methods for multi-temporal semantic segmentation. Change detection is conceptually related to multitemporal semantic segmentation as the change maps can be computed directly from the semantic predictions between two points in time. Our goal is to analyze if current time-series segmentation methods for satellite imagery are competitive for change detection on our dataset. At first, we provide a basic U-Net baseline [98] which uses monthly images to segment the respective land cover classes of each month. Beyond a one-to-one relationship of input and output, we also explore a number of time-series segmentation tasks where we use up to 31 images per month as input to map the land cover classes. Based

on a U-Net [98] feature extractor, U-ConvLSTM [75] is a U-Net segmentation model with an LSTM based temporal backbone. Similarly, 3D-UNet [75] makes use of 3D convolutions to process the temporal dimension. Finally, U-TAE uses self-attention for the temporal dimension following [107]. This model shares a similar temporal backbone to the PseTae [108] model used for crop type mapping above but is designed for segmentation tasks. Beyond these supervised baselines, we also benchmark a semi-supervised approach. In this baseline, additional imagery between the labeled scenes is exploited for consistency regularization during training following [64] with DeepLabv3+ [22] as segmentation backbone.

### 3.4.3 Evaluation Metric

We report the performance of supervised and semi-supervised baselines on DynamicEarthNet with mean intersection of union (MIOU) and a Semantic Change Segmentation (SCS) metric which we specifically design based on the requirements of the task at hand. SCS allows us to distinguish between detecting a change away from the current land cover class but incorrectly identifying the new class. In this case, a change would be recognized correctly in a binary setting which is arguably better than missing it completely. Formally, we define SCS therefore as the arithmetic mean of a binary Intersection over Union ('Binary Change': BC) and a per-class IoU of the semantic change (SC) scores.

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

where  $y$  is the actual semantic change map,  $\hat{y}$  the corresponding prediction and  $b$  is analogously defined but for change/no-change only. Details about the formal definition of BC and SC can be found in Appendix A.2.

### 3.4.4 Results

Table 5 presents the results of baseline experiments in three panels separated by input data. The first panel contains results for monthly inputs where CAC is the consistency regularization baseline with DeepLabv3+. The CAC model performs marginally better than the U-Net baseline with an edge of 0.4 p.p. in SCS. It segments both binary and semantic change slightly better than the U-Net. The insights of the SCS metric are also confirmed with mIoU. Going from monthly to weekly inputs does not improve the CAC baseline. A small gain in semantic change performance is offset by weaker binary change performance. This also points to a strength of our SCS metric: It allows for a more fine-grained understanding of change segmentation

performance. For the other weekly baselines, we combine the U-Net with a temporal backbone to process the time-series input. U-TAE reaches a high score in SCS with 19.1 because of a high semantic change score of 28.7. Similarly, the ConvLSTM temporal backbone reaches an SCS score just below U-TAE but is notably better in the binary category. The 3D-UNet performs slightly above yet in the range of the monthly U-Net. This implies that the additional temporal information is not effectively used in comparison to the TAE or ConvLSTM temporal backbone with weekly inputs. However, both CAC and 3D-UNet improve notably when going to daily data. CAC reaches a mIOU high score with a significant margin and 3D-UNet reaches the best binary change detection performance. On the other hand, performances of U-TAE and particularly U-ConvLSTM decline in comparison to weekly data. For the ConvLSTM temporal backbone, the performance even falls below the monthly U-Net baseline which uses 30x fewer inputs. These baseline experiments underline a number of things. First, current methods for multi-temporal semantic segmentation do not transform well off-the-shelf to change detection. Particularly, binary change accuracies are often only around 10%. This underlines the necessity of datasets like ours to offer opportunities to develop methods specifically tailored to the spatial and temporal depth of recent Earth observation advances. Second, much is yet to be understood about when and how the temporal depth is useful. In the cases of 3D-UNet and CAC going to daily data improved results notably but effects were the opposite for U-TAE and U-ConvLSTM. For research on these issues and many more, our dataset is well-tailored and we are excited about the kind of research advances that will be possible in the community with it. For additional details, please refer to [Appendix A.2](#).

Table 5: Monthly, weekly and daily baselines on DynamicEarthNet’s Semantic Change Segmentation test set

	SCS (↑)	BC (↑)	SC (↑)	MIOU (↑)
<i>Monthly Images</i>				
CAC [64]	17.7	10.7	24.7	37.9
U-Net [98]	17.3	10.1	24.4	37.6
<i>Weekly Images</i>				
CAC [64]	17.8	10.1	25.4	37.9
U-TAE [107]	<b>19.1</b>	9.5	<b>28.7</b>	39.7
U-ConvLSTM [75]	19.0	10.2	27.8	39.1
3D-Unet [75]	17.6	10.2	25.0	37.2
<i>Daily Images</i>				
CAC [64]	18.5	10.3	26.7	<b>43.6</b>
U-TAE [107]	17.8	10.4	25.3	36.1
U-ConvLSTM [75]	15.6	7.0	24.2	30.9
3D-Unet [75]	18.8	<b>11.5</b>	26.1	38.8



## METHODOLOGICAL CONTRIBUTIONS TO CHANGE DETECTION

---

### 4.1 STATE-OF-THE-ART

Beyond improving the availability of large-scale datasets for time-series analysis of optical satellite imagery, the second ambition of this thesis is advancing the methodological frontier in change detection. At first, this requires a thorough review of current change detection methods and their challenges. Significant methodological progress has been made in recent times in change detection fueled by better data availability [2] and progress in deep learning for image recognition [65]. Change detection methods can, among other things, be differentiated by their propensity to use labeled data about changes during training.

#### 4.1.1 *Unsupervised Methods*

Unsupervised change detection methods do not rely on labels during training. Since large-scale annotated data for change detection is still scarce [116], unsupervised methods can therefore be advantageous. One early example of such a method is change vector analysis (CVA) [76]. The image pair is subtracted from one another to obtain a difference image (DI). The absolute value of the difference image is thresholded to obtain a binary change segmentation. CVA has been refined and extended for several purposes. Robust CVA (RCVA) [114] aims to make CVA less sensitive to potential coregistration errors. Object size is modeled with histograms in object-based CVA (OCVA) [68] to transition from the image to the object level. A different approach to bridge the gap between pixels and objects is combining CVA with morphological operations [35]. Parcel CVA (PCVA) [11] is aimed at high-resolution imagery and operates at multiple scales. It combines hierarchical segmentations with the use of spatial context in comparison to standard CVA. Other methods such as local binary patterns [50] or graph structures of an image [113] were also explored to use the neighborhood information of a pixel for change detection. For example, local binary similarity patterns (LBSP) [9] compute a similarity between points of interest and if the similarity between two points in time is low, this can be indicative of a change [45].

More recently, however, the rise of deep learning has also ignited progress among unsupervised change detection methods. Within the CVA framework, Deep CVA (DCVA) [104] extracts features with a

neural network from the pre and post images. The difference image comparison is then calculated based on the deep feature difference rather than on the images themselves. As DCVA was originally designed for high-resolution imagery with lower spectral depth, an extension for Sentinel-2 which has more channels but lower resolution also exists [106]. I refer to this method as DCVAMR. In similar spirit, feature change analysis (FCA) [133] combines feature extraction from a deep belief network with a difference analysis on the extracted features. The difference image can also be predicted with a generative adversarial network (GAN) [42]. GANs can also refine the coregistration of images for change detection prior to the actual prediction which improves unsupervised change detection methods [94].

A number of approaches take the route of predicting an initial change map that is further refined with additional steps [37]. In [129], superpixels are segmented on the difference image as a first step. In the second step, high-confidence predictions are used as pseudo labels for training a different classifier. Similarly, Lv et al. [72] rely on clustering of features obtained with stacked contractive autoencoders that are used as pseudo labels in a superpixel set-up. Pseudo labels are obtained based on saliency guided deep neural networks and prioritized with hierarchical clustering in [40]. The final change prediction is obtained with an autoencoder based model.

#### 4.1.2 Supervised Methods

If labels are indeed available during the training stage, supervised change detection methods can exploit them. After the landmark success of AlexNet and other convolutional methods, CNNs have been widely applied to change detection problems. U-Net [98] is a popular fully convolutional neural network which has provided a starting point for a number of change detection algorithms. A Siamese U-Net is proposed in [15] where the siamese structure jointly takes pre and post images as an input. In two different versions of the method, the extracted features are either concatenated (FC-Siam-Conc) or subtracted (FC-Siam-Diff). U-Net++ is an extension of U-Net and was adapted to a change detection framework for very-high resolution imagery in [90]. Superpixel segmentation and CNNs are combined in ESCNet [132] ReCNN employs a recurrent model together with CNNs for multiple change detection [79].

In recent years, attention-based mechanisms have gained popularity over CNNs in general image recognition [119] which resulted in the creation of the vision transformer architecture [70]. Bitemporal Image Transformer [18] provides a siamese deep features extraction framework based on transformer encoder and image differencing. ChangeFormer [6] is based on related ideas. It contains a hierarchical transformer encoder and a lightweight multi-layer percep-

tron as a decoder. Feature difference modules are employed at several depths of the siamese structure to obtain the final change segmentation. In ChangeMask, a transformer network is inserted in between a semantic-aware encoder and a multi-task decoder for semantic change detection [136]. SwinTransformers [70] are a popular transformer architecture for segmentation and have been explored for change detection in [131]. Transformer models and convolutional models can also be combined for improved change detection performance [20, 32, 130].

#### 4.1.3 *Semi-Supervised Methods*

A variety of approaches try to merge supervised and unsupervised methods. Semi-supervised methods exploit large quantities of unlabeled data to support the supervised training process and are also popular in change detection. Bovolo et al. [12] design a support vector machine (S<sup>3</sup>VM) based on semi-supervision and Bayesian thresholding. Consistency regularization is exploited in [7] to constrain the output change probability map by adding an unsupervised element to a cross-entropy loss. This allows going beyond the necessity of available image pairs in semi-supervised change detection. A Self-Organizing Feature Map (SOFM) is presented in [41] where only a small number of labels is used initially and soft labels for unlabeled data are then generated with fuzzy set theory. One of the findings of the review of unsupervised methods was that some predicted an initial change map that is further refined by another method. In semi-supervised learning, a similar tendency exists. For example, [21] obtain an initial classification of change with Gaussian Processes (GP) which is then further processed with a Markov Random Field. This is conceptually related to student-teacher models [124]. The teacher model is trained first with available labels and obtains additional predictions, so-called pseudo labels, on unlabeled scenes. Typically, the student is then trained with the pseudo labels first which are of lower quality than the real labels. However, there is enough signal in the pseudo labels to enrich the training with the real labels. Pseudo labels in the semi-supervised context have, for example, been successfully used for hyperspectral image classification [123].

## 4.2 SIROC: SIBLING-REGRESSION FOR OPTICAL CHANGE DETECTION

### 4.2.1 *Motivation*

Detecting changes between two images naturally is not a task that is exclusive to remote sensing per se. This matters, for example, in autonomous driving [1] where up-to-date maps require continuous

change detection. One potentially surprising domain where similar problems pose a challenge is exoplanet detection in astrophysics. This is because a change in the light intensity of distant stars in a telescope can be indicative of a transient object between the sensor and the light source. A method that makes use of this logic is Half-Sibling Regression (HSR) or sometimes also called the Causal Pixel Model. It models a pixel as a function of its distant neighborhood of pixels. This is because changes are typically local in an image and distant pixels are likely unaffected by the same change. For an image later in time, the model predicts a pixel of interest based on its neighbors at a later point in time with the relationship learned from the previous image. If the deviation of the predicted pixel value from its actual value is large, this is indicative of a change.

Although a few methods include neighborhood information for change detection in Earth observation such as LBSP [9] the use of the distant neighborhood of a pixel is limited until now. We therefore apply and refine HSR to unsupervised change detection in Earth observation as Sibling Regression for Optical Change Detection (SiROC).

#### 4.2.2 Methodology

The first part of this section outlines Half-Sibling Regression Image Differencing whereas the second part explores our modifications for Earth observation data.

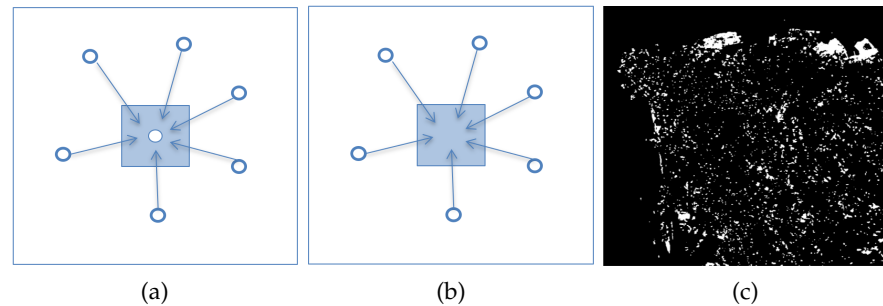


Figure 4: Half-Sibling Regression (HSR) for Change Detection

*Half-Sibling Regression Image Differencing:* HSR was originally designed for time-series data of the Kepler telescope and its core principle is outlined in Figure 4. At time  $t$  in Figure 4a, the pixel of interest is fitted as a linear combination based on the neighbors outside of the square. In practice, the number of pixels used is much higher but is limited here for visualization purposes. Figure 4b shows  $t+1$  where the neighboring pixels in  $t+1$  are used to obtain a prediction for the center pixel. Then, the predicted image in  $t+1$  is subtracted from the actual image and thresholded which results in the change segmentation of Figure 4c.

Formally, we estimate the residual of the predicted image  $\hat{I}_{t+1}$  and  $I_{t+1}$  as

$$\mathbf{e}_{t+1} = \hat{\mathbf{I}}_{t+1} - \mathbf{I}_{t+1} \quad (3)$$

where the prediction for the pixel with coordinates  $x$  and  $y$  is given as:

$$\hat{I}_{x,y,t+1} \equiv g_{t+1} I_{x,y,t} = \sum_{(i,j) \in N_{x,y}} \beta_{i,j,x,y} I_{i,j,t+1} \quad (4)$$

where  $g_{t+1}$  resembles a growth rate of pixels between  $t$  and  $t+1$  in the neighborhood of  $I_{x,y}$ . The model assumes that the growth rate of  $I_{x,y}$  should be similar if nothing changed. More explicitly,  $\hat{I}_{x,y,t+1}$  is defined as the sum over all selected neighborhood points from  $N$  with coordinates  $i$  and  $j$  times the respective linear coefficient  $\beta_{i,j,x,y}$ . Here,  $\beta$  depends on the pixel of interest  $I_{x,y}$ , the respective neighboring pixel for which  $\beta$  is estimated and the squared sum of all other neighbors. This is the closed-form solution of the least squares problem:

$$\beta_{i,j,x,y} = \frac{I_{i,j,t}}{\sum_{(i',j') \in N_{x,y}} I_{i',j',t}^2} I_{x,y,t} \quad (5)$$

This definition of  $\beta$  allows us to directly obtain  $\hat{I}_{x,y,t+1}$  without the explicit calculation of  $\beta$  as:

$$\hat{I}_{x,y,t+1} = \frac{\sum_{(i,j) \in N_{x,y}} I_{i,j,t+1} I_{i,j,t}}{\sum_{(i,j) \in N_{x,y}} I_{i,j,t}^2} I_{x,y,t} \quad (6)$$

Here, the resemblance of the fraction to a growth rate becomes more obvious: The numerator is a sum over the product of neighborhood pixels in  $t$  and  $t+1$ . On the other hand, the denominator represents only pixels in  $t$  in a similar term. While this is not an explicit growth rate, the changes in the neighborhood pixel values dictate whether the factor becomes larger or smaller than one. Depending on this factor, the model expects  $I_{i,j}$  in a similar intensity and predicts  $\hat{I}_{x,y,t+1}$  based on this. This can be extended to multi-channel images by summing over the absolute change signal per channel.

*Sibling Regression for Optical Change Detection:* In comparison to HSR, we make two major modifications for change detection in Earth observation.

1. We iterate over mutually exclusive neighborhoods and use the ensemble of resulting models to estimate an uncertainty of our predictions

2. We combine the pixel-level segmentations with morphological profiles (MP) to bridge the gap to the object level

---

**Algorithm 1** : SiROC
 

---

**Input:**  $I_t, I_{t+1}, s, n\_max, e\_start$

**Output:** Binary Change Segmentation

```

1:  $e = e\_start, n = e\_start + s$ 
2:  $Uncertainty\_CM = zeros\_like(I_t)$ 
3: while  $n < n\_max$  do
4:   for (channel in channels) do
5:     for (pixel in  $I_t$ ) do
6:       Apply HSR( $n, e$ ) to get  $\hat{I}_{t+1}$ 
7:     end for
8:      $Channel\_Difference\_Image = \hat{I}_{t+1} - I_{t+1}$ 
9:   end for
10:   $Diff\_Image = Sum(|Channel\_Difference\_Images|)$ 
11:   $Binary\_CM = Otsu\_Thresholding(Diff\_Image)$ 
12:   $Binary\_CM\_Object = Morph\_Profile(Binary\_CM)$ 
13:   $Uncertainty\_CM = Uncertainty\_CM + Binary\_CM\_Object$ 
14:   $n = n + s$ 
15:   $e = e + s$ 
16: end while
17:  $Final\_Segmentation = Threshold(Uncertainty\_CM)$ 

```

---

Algorithm 1 presents SiROC in Python style pseudocode. We aim to output a binary change segmentation of a pair of bi-temporal images  $I_t$  and  $I_{t+1}$ . As additional parameters,  $e\_start$  is the initial size of the exclusion window,  $s$  the step size, and  $n\_max$  the maximum size of the neighborhood. All these parameters are measured in the number of rows/columns from the pixel of interest. For example, if  $n\_max=20$  and  $s=e\_start=5$ , any pixel that is more than five but less than twenty-one rows *and* columns away will be included in a model. Note that this criterion has to be fulfilled for both, rows and columns. In this case, there are three models which consider the distance 5-10, 10-15, and 15-20 respectively.

We start with defining an uncertainty map filled with zeros in the shape of the pre or post image. As long as the current neighborhood window of interest has not reached its maximum (line 3), we obtain HSR change predictions given the current neighborhood window for every pixel in the image and repeat this for all channels. Then, we take the difference of the predicted and actual image in  $t+1$  based on equation 3. After this, we sum the absolute value of the change signal across channels and threshold this with Otsu thresholding [86]. The result is a binary change prediction for every pixel. To transition to the object level we apply a morphological profile which removes spurious predictions and closes gaps in larger objects. Finally, we add

the resulting binary predictions at the object level to the uncertainty change maps and update the neighborhood parameters. After the iteration over the neighborhoods has concluded, the uncertainty map is similar to a voting map where every pixel has an integer between 0 and the number of models. The higher the number of votes, the more models view this pixel as changed based on mutually exclusive neighborhoods. The final binary segmentation is obtained by applying a threshold to this number of votes. The idea behind using mutually exclusive neighborhoods is to exploit different trends at varying distances to the pixel.

### 4.2.3 Data

For our experiments, we use four different binary change detection datasets which cover a variety of different applications: Urban expansion, disaster response, agriculture, and alpine regions.

*OSCD*: At first, we rely on OSCD [28] with its 24 bi-temporal pairs of Sentinel-2 imagery and manual change annotations from across the globe. To be consistent with other evaluations on this dataset, we follow their standard train (14 cities) and test (10 cities) split. The dataset is focused on urban areas and mainly contains annotations of new roads or buildings.

*Beirut Harbor Explosion Dataset (BHED)*: Additionally, we construct a dataset around the Beirut Harbor explosion of August 2020. We base this on a pair of cloud-free PlanetScope images from before and after the explosion. This allows us to explore SiROC also with higher resolution data than Sentinel-2. The destruction reference data comes from the Center for Satellite-based Crisis Information (ZKI). ZKI bases this annotation on ground reports and manual inspection of very-high resolution satellite imagery.

*Agriculture Dataset*: Based on [104], this dataset explores manually annotated agricultural changes in Barrax, Spain. The imagery comes from Sentinel-2 with a size of  $600 \times 600$  pixels. Pictures are taken 10 days apart. Annotated events here are changes in the visual appearance of the field. Note that this may not necessarily align with a land cover/use change but can still be an event of interest (e.g. harvest).

*Alpine Dataset*: This dataset is also explored in [104] and concerns a fire close to Trento in the Italian Alps. The imagery is also based on Sentinel-2 with a pixel size of  $350 \times 350$  and heavily influenced by winter conditions in the respective region.

### 4.2.4 Experiments

First, I describe the selection of competing methods. The second part concerns evaluation criteria and selected hyperparameters.



*Competing Methods:* We select a number of competitive unsupervised change detection baselines from traditional as well as deep learning methods which are described in section 4.1.1. In terms of traditional methods, we benchmark RCVA [114] and PCVA [11] which are more recent versions of CVA. On the deep learning side, we include DCVA [104] and its extension for multispectral imagery, DCVAMR [106]. For the BHED dataset, we also use a high-resolution version that is based on self-supervised learning which we call SSD-CVA here [105].

*Evaluation Criteria:* To be consistent with previous evaluations on the used datasets, we select four criteria. At first, we use F1 score which was already described in section 3.2. However, previous contributions often only inspect the F1 score of the change class and we follow this convention here. Additionally, we include sensitivity and specificity. Sensitivity is defined as

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (7)$$

and specificity as

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (8)$$

Sensitivity is equivalent to recall which is also a commonly used expression. Both criteria have elements of a per-class accuracy since they consider all positive (sensitivity) or negative (specificity) samples and assess what fraction was correctly classified. The recall is often also combined with precision which is our final criterion:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (9)$$

Precision reports how many of the change predictions are actually changing. Together with sensitivity/recall, the F1 score is calculated as described earlier.

*Hyperparameters:* Based on tuning on the OSCD training set, we select the following hyperparameter configuration:

1. Maximum neighborhood size: n\_max=200
2. Initial exclusion window: e\_start=0
3. Step size of ensemble: s=8
4. Filter Size of Morphological Operations: p=5

In contrast to standard HSR, we find an initial exclusion window of zero to be optimal in our case. Our ensembling approach potentially incorporates the advantages of using distant neighborhoods in some of the models already. Therefore, the necessity to exclude close neighbors may be limited in our case here. For details on the hyperparameter tuning and a sensitivity analysis, please refer to Appendix A.3 section III.C



Table 6: Quantitative Results OSCD Test Set

	Specificity	Sensitivity	Precision	F1
SiROC	88.31%	70.71%	24.80%	36.72%
DCVAMR	78.38%	64.63%	14.01%	23.03%
DCVA	76.96%	69.02%	14.03%	23.33%
PCVA	75.61%	47.00%	9.50%	15.81%
RCVA	76.96%	64.08%	13.16%	21.84%
Ablation Scores				
No MP	80.64%	69.88%	16.44%	26.62%
HSR	79.45%	70.24%	15.70%	25.66%

#### 4.2.5 Results

Table 6 shows the results of SiROC and other baselines on the OSCD test set. Overall, SiROC is superior to the evaluated baselines from the deep learning side as well as CVA-based models with high scores in all categories for OSCD. The gap for specificity is at least 10 p.p. with DCVAMR coming in second place. This implies that SiROC captures the no-change class notably better than the other methods here. The edge for sensitivity is smaller where DCVA is a close second about 2 p.p. behind but the gap to other methods remains large. The gap is substantial for precision and F1 score as well.

The ablation scores evaluate the effectiveness of the morphological profiles and ensembling in SiROC on OSCD. Without morphological operations, the performance of SiROC drops mostly in specificity but also marginally in sensitivity. It seems that the morphological operations mostly help to reduce spurious false positives rather than adding in false negatives here. Still, even without the morphological operations, SiROC surpasses the competing models although by a small margin only. Not ensembling over mutually exclusive neighborhoods reduces performance marginally in precision and F1 but the effect is small here. HSR as a baseline method is also quite competitive here and also surpasses the scores of competing models which points to the effectiveness of the underlying logic to also use distant neighbors.

Qualitative results in Figure 5 on the Las Vegas scene from OSCD confirm the impressions of Table 6. Panel 5a shows the confidence outputs of SiROC which are the number of votes from the ensemble methods. The thresholded, final prediction is in Panel 5b, Panel 5c is equivalent to 5b without morphological operations, 5d - 5g are competing models and the reference data is in 5h.

Comparing the SiROC predictions to the ground truth, changes seem to generally be identified well. This is based not only on identi-

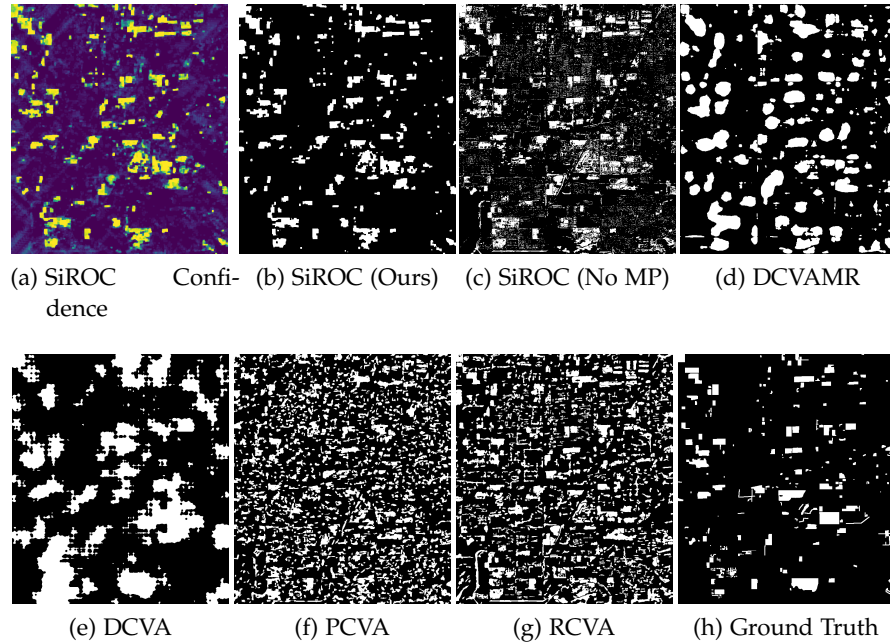
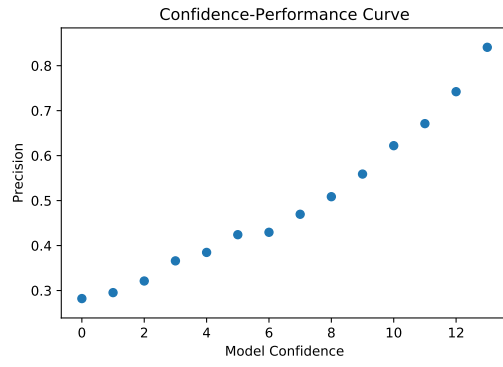


Figure 5: Qualitative Comparison OSCD - Las Vegas.

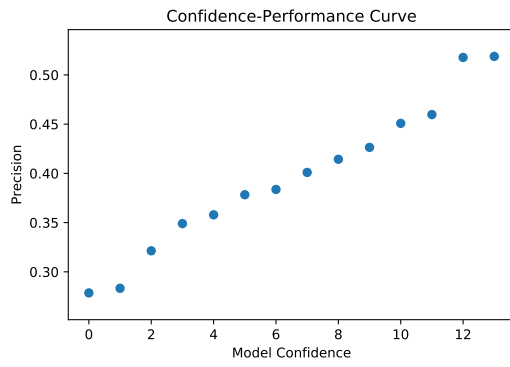
ifying the locations of changes but also fitting their shape comparably well. One could argue that this is merely thanks to the morphological operations but [5c](#) shows that the shapes are also well identified before that. Rather, the MP removes spurious false positives. The competing methods struggle more with these scenes for different reasons. DCVA and DCVAMR identify changing regions generally but fit blobs rather than refined shapes to these areas. RCVA and PCVA, on the other hand, seem capable of identifying object shapes on the ground here. However, they largely overestimate the number of changes on the ground. Therefore, SiROC seems best to extract changing buildings accurately here.

To analyze the correspondence of the confidence measure in [5a](#) within SiROC to classification performance, [Figure 6](#) plots a calibration curve for several AOIs in the OSCD test set. We separate our predictions by confidence levels and evaluate these buckets separately. Performance seems generally non-decreasing for all four cities which is most pronounced for Las Vegas. This implies that beyond accurate predictions, SiROC also returns a built-in confidence that corresponds well to its actual classification performance. This can help in practical applications to prioritize results of high confidence and verify predictions with low confidence.

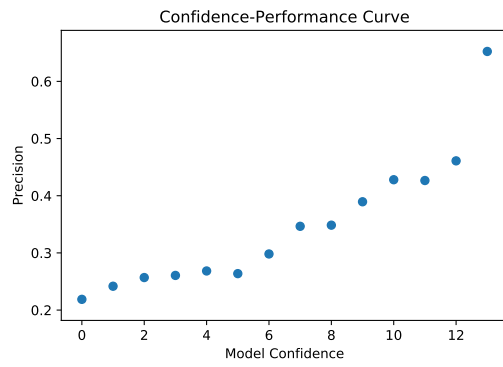
Beyond OSCD, [Table 7](#) outlines score for the Beirut Harbor Explosion dataset. This is based on 3m resolution imagery which is why we include SSDCVA instead of DCVAMR as the latter is tailored more towards medium-resolution imagery. We apply SiROC with its defaults calibrated on OSCD to this dataset without tuning it. Again SiROC is



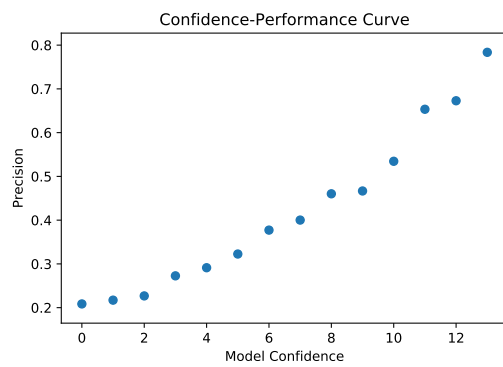
(a) Las Vegas



(b) Dubai



(c) Chongqing



(d) Montpellier

Figure 6: Confidence-Performance Plots on four cities of the OSCD Dataset.

Table 7: Quantitative Results Beirut Explosion

	Specificity	Sensitivity	Precision	F1
SiROC	<b>92.01%</b>	<b>83.38%</b>	<b>19.89%</b>	<b>32.12%</b>
DCVA	91.87%	79.85%	11.37%	19.93%
SSDCVA	88.25%	81.08%	8.80%	15.95%
PCVA	88.61%	58.56%	6.74%	12.10%
RCVA	86.56%	66.71%	6.52%	11.89%
Ablation Scores				
SiROC (p=10)	<b>92.34%</b>	<b>91.89%</b>	<b>22.20%</b>	<b>35.76%</b>
no MP	88.02%	79.67%	13.65%	23.30%
HSR	86.65%	71.63%	11.31%	19.54%

the strongest model on this dataset. Compared to DCVA, specificity is similar here but sensitivity is higher with about a 3 p.p. margin. Therefore, SiROC misses fewer changes here but handles false positives similarly well to DCVA. SSDCVA falls behind in both criteria and PCVA and RCVA are not really competitive here. The gap becomes more obvious when looking at precision and F1 where the margin is even larger.

As the image resolution is significantly higher in the input here, we test a version of SiROC with doubled size of morphological operations as an ablation study. This is because objects will be notably larger with increased input resolution which the SiROC defaults do not reflect automatically. The scores improve further which underlines that there is likely more potential in SiROC for this dataset. We restrain from tuning this, however, because only one scene is available for testing. Again, we disentangle the effect of SiROC compared to baseline HSR. In this case, both ensembling and MPs are effective and combine for optimized performance. MPs improve precision and F1 scores particularly while ensembling mostly helps with sensitivity. Therefore, both components play an important role in SiROC's effective use for Earth observation data.

Outside of urban applications, Table 8 outlines that SiROC can also be useful in agricultural applications. In line with [106], we restrict the input to vegetation and near-infrared channels of Sentinel-2. Again, SiROC is merely applied with its OSCD defaults to this dataset. SiROC with near-infrared input reaches the highest specificity and precision but falls short of DCVAMR in sensitivity and F1. While DCVAMR arguably has a slight edge here, SiROC still performs competitively just with its defaults.

SiROCs effectiveness for alpine applications is evaluated in Table 9. In accordance with [106], we stick to NIR and SWIR channels here as inputs for SiROC. SiROC outperforms RCVA and PCVA but falls

Table 8: Quantitative Results Agriculture Dataset

	Specificity	Sensitivity	Precision	F1
SiROC (VEG)	90.69%	86.38%	73.53%	79.44%
SiROC (NIR)	<b>90.81%</b>	88.70%	<b>74.28%</b>	80.85%
DCVAMR	88.88%	<b>94.26%</b>	71.73%	<b>81.47%</b>
PCVA (VEG)	88.83%	83.18%	69.04%	75.45%
PCVA (NIR)	86.60%	84.56%	65.38%	73.74%
RCVA (VEG)	88.91%	91.95%	71.28%	80.31%
RCVA (NIR)	87.39%	92.36%	68.67%	78.77%

Table 9: Quantitative Results Alpine Dataset

	Specificity	Sensitivity	Precision	F1
SiROC (NIR)	98.92%	75.71%	52.28%	61.85%
SiROC (SWIR)	<b>99.28%</b>	59.51%	56.10%	57.76%
DCVAMR	99.06%	<b>94.99%</b>	<b>61.23%</b>	<b>74.46%</b>
PCVA (NIR)	98.95%	46.99%	41.04%	43.82%
PCVA (SWIR)	95.48%	35.80%	10.98%	16.80%
RCVA (NIR)	99.22%	63.99%	56.20%	59.84%
RCVA (SWIR)	86.56%	66.71%	6.52%	11.89%

short of DCVAMR on this dataset. While it reaches the highest specificity, DCVAMR is superior in three out of four evaluation criteria. However, the overall results with SiROC defaults are still decent overall which underlines the versatility of our method even with its defaults.

#### 4.2.6 Discussion

Our experiments show that SiROC can be an effective method for change detection in medium as well as high-resolution optical imagery. Further, it scores competitively across applications in urban expansion, disaster response, agricultural monitoring and alpine change detection. Compared to other image differencing techniques, SiROC enables the inspection of local trends in an image. On the other hand, CVA and versions of it, often assume that a trend affects the whole image similarly. This may be intuitive for some aspects but for shadows or small clouds the necessity to analyze local trends is essential. This may explain a large part of the performance advantage of SiROC compared to other image differencing methods.

In comparison to unsupervised deep learning methods, SiROC still performs well. This may be different on larger datasets as OSCD is still comparably small. However, large-scale reference data is still scarce in change detection so the scenario of limited input data is not necessarily unlikely. Our intention is also not to compete with deep learning methods but rather to augment them. The calibrated confidence of SiROC combined with its performance makes it an interesting candidate for use in combination with deep learning methods. As the method is fast and performant, it can, for example, be used for pseudo labeling. In this scenario, SiROC is used to obtain pseudo labels which are technically predictions but can get larger methods already far in training. Then, the actual labels are used to finetune the models to maximize performance. This is explored in the following section below.

### 4.3 SEMISIROC: SEMI-SUPERVISED CHANGE DETECTION WITH OPTICAL IMAGERY

#### 4.3.1 Motivation

As the previous section outlined, SiROC can potentially be effectively combined with deep learning based methods for semi-supervised student-teacher learning. This is because SiROC's unsupervised predictions are not only accurate but come with an insightful confidence score.

This is related to other teacher-student set-ups in earth observation [69]. However, we focus on an *unsupervised* teacher model here. While

this may sound counterintuitive at first, the motivation is to reduce the label dependency of change detection for generalization purposes. Transitioning from one dataset to the other can sometimes be an issue in change detection [103]. This is because different datasets consider different regions, applications and class balances, among other things. SiROC is a versatile method that is not tailored to a specific application or trained with a specific set of labels. Our hypothesis is that this may foster generalization abilities.

#### 4.3.2 Methodology

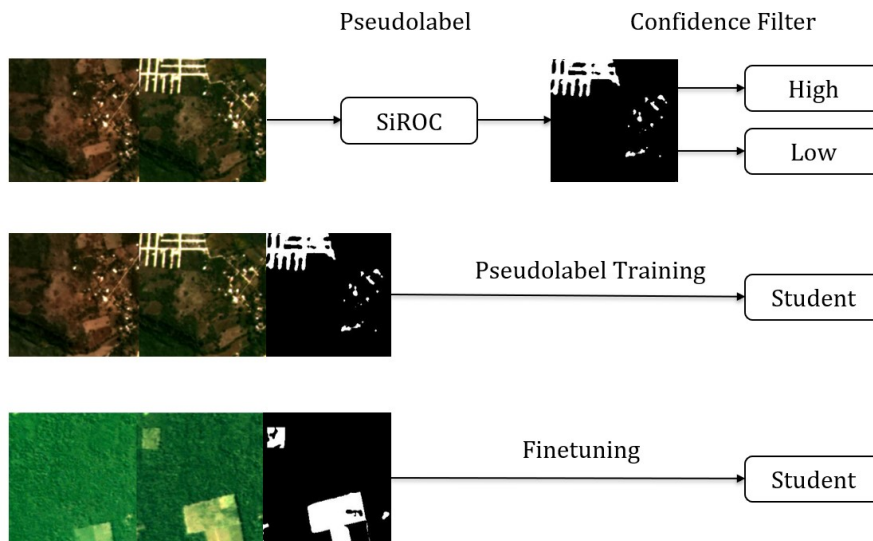


Figure 7: An overview of SemiSiROC.

Figure 7 visualizes the methodology behind SemiSiROC. SiROC is applied to obtain unsupervised predictions for unlabeled imagery. On the left, we see the pre and post pair where most of the change happens in the upper left corner. Then, the top quarter of AOIs in terms of confidence is used to train a deep learning model before using the actual ground truth. Therefore, the confidence score can be used to prioritize the pseudo labels.

In the second step, high-confidence pseudo labels are used for pre-training a student model. This allows deep learning models to learn from more reliable pseudo labels before going onto the real labels. Finally, the student model is finetuned with ground truth data in the third step. Note that only the student model uses the ground truth in our approach which is different from many other teacher-student approaches [125].

More formally, assume there are two collections of bi-temporal image pairs  $D$  and  $U$ .  $D$  contains change segmentation masks for every pair while  $U$  does not. Typically  $U$  would be much larger than  $D$  since labeled data is scarce. Therefore, the goal is to make effective

use of  $U$  and support the training of a change detection model with this. This could be done, for example, with consistency regularization during training with unlabeled scenes as explored in section 5 for DynamicEarthNet baselines. Here, we take a different approach within semi-supervision. We employ an unsupervised teacher model to label  $U$  with its change predictions. Therefore, a student model can exploit the label space  $D \cup U$  instead of just  $D$  although the pseudo labels for  $U$  will be of lower quality. Particularly in the early stages of training, however, pseudo labels can be beneficial and the higher quality of real labels only becomes more important in later stages of the training [125]. However, we do not label all of  $U$  since our ambition is to restrict the pseudo label training process to examples where the teacher model is fairly certain. Therefore, we only include the top quarter of AOIs by confidence.

As a teacher model, we use SiROC which is described at length in the previous section. As student models, we employ a number of competitive supervised change detection models.

- *FC-Siam-Diff* [15] is a UNet-based change detection method. It has a siamese structure with shared weights for post and pre images. Features extracted from the image pair are joined after the convolutional layers and then subtracted from each other. The approach to use temporal differences of deep features is common in change detection and also finds use, for example, in the unsupervised method DCVA [104] discussed in the previous section. For supervised approaches, such as FC-Siam-Diff, however, the decision layers after the feature difference can be explicitly trained with labeled data.
- *ChangeFormer* [6] is a transformer based siamese network. It processes image feature differences at multiple spatial scales with a hierarchical transformer encoder. This way, the model can scan for changes at several abstraction levels. The final decision decoder is a multi-layer perceptron which is fairly lightweight.
- *Bitemporal Image Transformer (BIT)* [18] also uses a siamese structure with a transformer encoder but the initial feature extraction is still based on a CNN backbone whereas ChangeFormer is purely transformer based. For BIT, the transformer encodes and decodes semantic tokens that are produced based on the CNN features. The siamese output of the transformer decoder is subtracted from one another which is fed into a shallow CNN for the final change segmentation. In a way, the transformer elements here are embedded by CNNs which is in contrast to [6] who remove CNNs completely.



## 4.3.3 Data and Evaluation

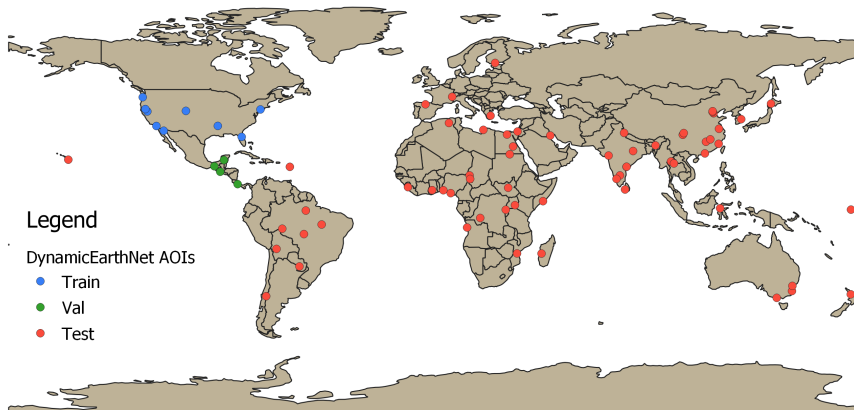


Figure 8: DynamicEarthNet spatial split

We rely on two datasets for our semi-supervised experiments. At first, we use the DynamicEarthNet dataset discussed in section 3.4. For this purpose, we take the first and last scene for each AOI and compute a binary change map from the ground truth data. Figure 8 shows our train/val/test split in more detail. We simulate a scenario where labeled data is scarce and primarily from specific regions which is the US in our case. From this, we try to generalize to unseen regions using validation data from Central America and the rest is withheld for global testing. If not mentioned otherwise, this is our baseline split although we evaluate a number of alternative scenarios below. We split the original  $1024 \times 1024$  scenes into 16  $256 \times 256$  chips to be consistent with [6]. The second dataset is OSCD [28] which is described in Table 4. From OSCD, only the test set is used for evaluation purposes here to evaluate if a model trained on DynamicEarthNet can generalize to this new dataset as well. However, OSCD images do not have consistent sizes and further are not square. Therefore, we pad the images to the next multiple of 256 and exclude the padded pixels during evaluation.

To test the effectiveness of SemiSiROC, we make two main comparisons. At first, we compare to finetuning only with the real labels where no pseudo label pretraining is involved. Second, as additional baselines, we include DCVA and CVA as competing pseudo label sources. In additional ablation studies, we further test the robustness of our results against a larger training set, alternative training loss choices, and splitting the locations of pseudo label pretraining and finetuning.

Each model is trained for 50 epochs with ADAM as an optimizer, a batch size of 32, an initial learning rate of 0.0001, and linear weight decay. Pseudo label pretraining is based on a focal loss. Models are scored based on accuracy, mean F1 score, and MIOU which have been introduced above. However, note that to be consistent with [6, 18] we

Table 10: Quantitative Results DynamicEarthNet grouped by pseudo label use.

Model	PL	Loss PL	Loss	Accuracy	MIOU	MF1
FC-Siam-diff [15]	✓	FL	MIOU	<b>0.7812</b> (+0.0104)	<b>0.4854</b> (+0.0037)	<b>0.6029</b> (+0.0018)
FC-Siam-diff [15]			MIOU	0.6359 (+0.0405)	0.419 (+0.0288)	0.5706 (+0.0244)
ChangeFormer [6]	✓	FL	MIOU	0.736 (+0.0448)	0.4586 (+0.0185)	0.584 (+0.0101)
ChangeFormer [6]			MIOU	0.4848 (+0.0923)	0.305 (+0.0627)	0.4545 (+0.0644)
BIT [18]	✓	FL	MIOU	0.7303 (+0.0158)	0.4598 (+0.0086)	0.5887 (+0.0058)
BIT [18]			MIOU	0.6242 (+0.0418)	0.4074 (+0.0227)	0.5587 (+0.0151)
SiROC [61]				0.6946	0.4408	0.5769

report the mean of the F1 score for the change and the no change class. This is in contrast to section 4.2 where only the change class is reported for consistency with [15, 104]. Each experiment is executed five times with different seeds and we report the mean as well as the standard deviation for each criterion.

#### 4.3.4 Results

Table 10 shows scores on DynamicEarthNet and compares the use of pseudo labels against just training with real labels. Pseudo label training is done with a focal loss, while the finetuning uses a MIOU loss for all specifications. For all three models, pseudo label pretraining notably improves performance. FC-Siam-Diff gains about 15 p.p. in accuracy while the gap for ChangeFormer is even larger. It seems that ChangeFormer without pseudo label pretraining is even not really applicable here as accuracy falls below 50%. For BIT, the gap is smaller but still significant at about 11 p.p. This result is confirmed by MIOU and MF1 as well. As an additional baseline, the performance of SiROC on this dataset is about 69% accuracy which is larger than the supervised models with real labels here. This may be the case because the amount of training data is comparably low here with 128 image pairs for training only. However, the training set of OSCD is even smaller where FC-Siam-Diff performs well so parts of this gap may also be due to distribution shift in the regions.

Still, our approach shows how the spatial robustness of the unsupervised approach and the ability to learn from labeled data combine nicely in SemiSiROC. The pseudo labels provide an effective way to push performance significantly compared to the supervised baselines.

Figure 9 adds a qualitative perspective to the results of Table 10 for 8 sample images from a rural region that is affected by deforestation. Changes are particularly present on the left and in the center of the rows as visible in the ground truth (9c). FC-Siam-Diff with pseudo label pretraining (9d) fits the scene better than without pseudo labels (9e). Particularly, without pseudo labels there are more false positives

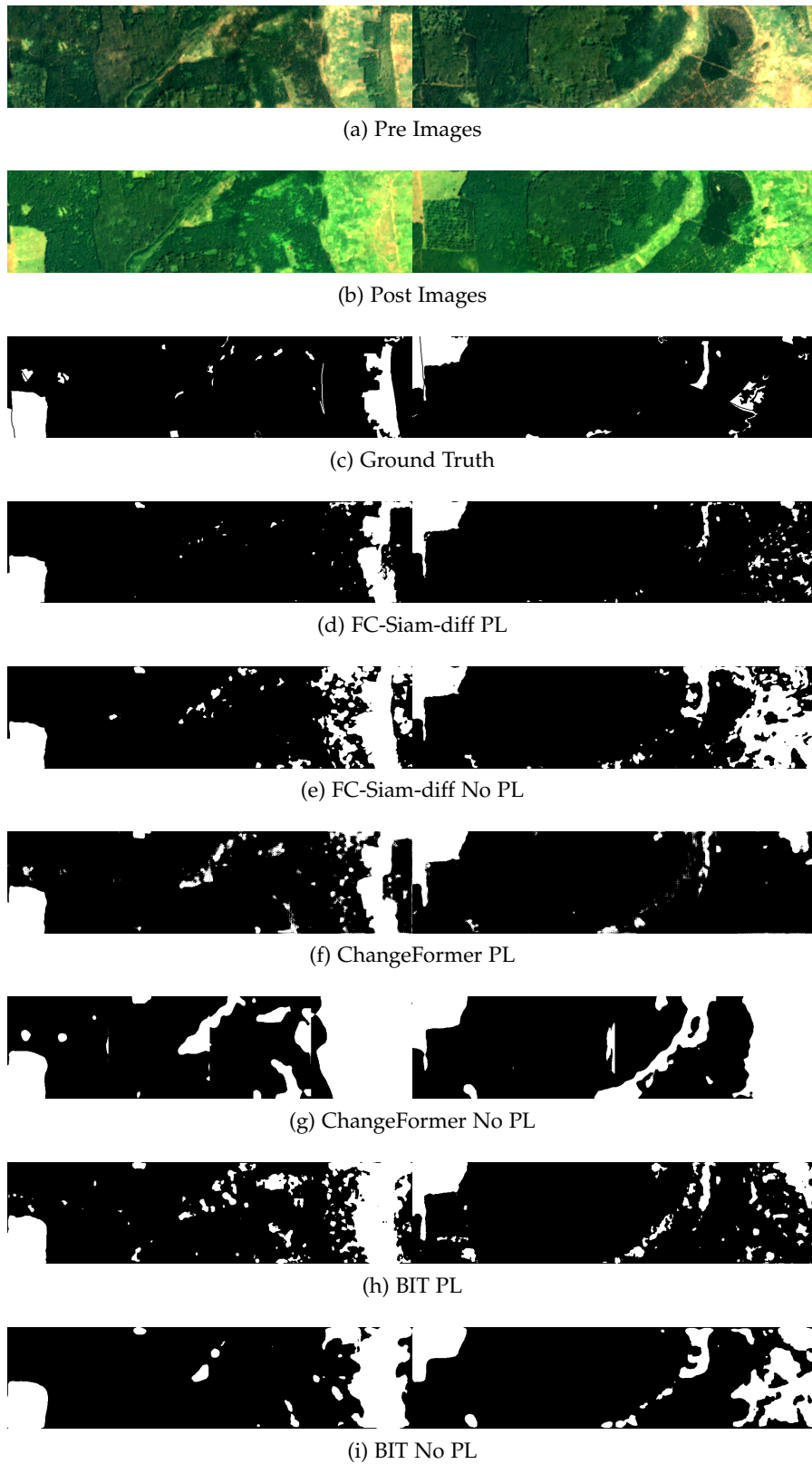


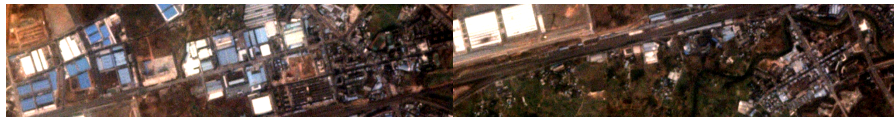
Figure 9: Qualitative results of 8 sample image pairs in a rural setting

which may stem from illumination differences between the pre and post images. This observation is even stronger for ChangeFormer (9f & 9g) and BIT (9h & 9i). Particularly, ChangeFormer seems much more sensitive to differences in acquisition conditions rather than actual changes on the ground. Pseudo label pretraining eliminates a large fraction of this oversensitivity for more robust change detection.

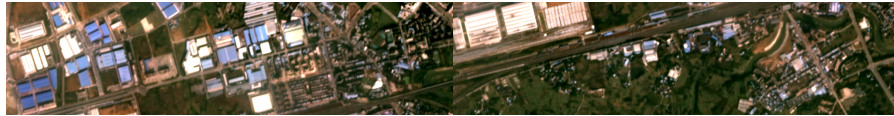
Figure 9 visualizes results for a rural scene with few but large and fairly distinct changes. In contrast, Figure 10 inspects more complex urban scenes with building and road changes as visible in the pre image (10a), the post image (10b), and the ground truth (10c). These scenes are more challenging overall. FC-Siam-Diff with pseudo labels is still arguably the best model here but misses some large changes for example on the bottom left. The no pseudo label version is less distinct in shapes but seems to identify changing regions generally. For ChangeFormer and BIT, the shapes of the changes get blurry and overestimated in size without pseudo label pretraining. This seems to suggest that pseudo label pretraining particularly helps supervised models to have a better understanding of the shapes of potential changes. This could be related to the morphological operations within SiROC since they provide prior understanding of change shapes. Overall, the qualitative inspection of scenes confirms the impressions of Table 10 that SemiSiROC is an effective strategy compared to its supervised baselines.

Table 11 compares SemiSiROC predictions to different pseudo label sources where DCVA and CVA are used to obtain alternative predictions. SiROC scores are identical to Table 10. For FC-Siam-Diff, SiROC pseudo labels reach the highest accuracy and MIOU but fall slightly behind on MF1. Further, the gap on MIOU is marginal at best. Therefore, it seems that there is no clear favorite for this model. SiROC pseudo labels nudge more towards higher accuracy whereas CVA and DCVA pseudo labels result in slightly more balanced decisions. On the other hand, for ChangeFormer former there is a notable difference in accuracy of up to 8 p.p. SiROC pseudo labels result in higher. MIOU is also higher with SiROC and the model reaches similar MF1 scores. The picture is similar for BIT. SiROC pseudo labels result in higher accuracy, slightly higher MIOU and similar MF1 scores. In total, it seems that SiROC pseudo labels give an advantage particularly for accuracy which is especially apparent with BIT and ChangeFormer.

One could argue, however, that the edge of our strategy is due to the limited availability of actual labels. Therefore, Table 12 inspects using more training data. Our initial split includes 8 but we test 16, 32, and 64 cubes out of 75 for training here. Note that the scores are not comparable to previous Tables here because they were obtained on different test sets. They are, however, consistent within the Table meaning that the rows can be compared against each other.



(a) Pre Images



(b) Post Images



(c) Ground Truth



(d) FC-Siam-diff PL



(e) FC-Siam-diff No PL



(f) ChangeFormer PL



(g) ChangeFormer No PL



(h) BIT PL



(i) BIT No PL

Figure 10: Qualitative results of 8 sample image pairs in an urban setting

Table 11: Quantitative Results DynamicEarthNet with different pseudo labels

Model	PL	Loss PL	Loss	Accuracy	MIOU	MF1
FC-Siam-diff [15]	SiROC	FL	MIOU	<b>0.7812</b> (+0.0104)	<b>0.4854</b> (+0.0037)	0.6029 (+0.0018)
FC-Siam-diff [15]	CVA	FL	MIOU	0.7599 (+0.0124)	0.4853 (+0.0072)	<b>0.6121</b> (+0.0048)
FC-Siam-diff [15]	DCVA	FL	MIOU	0.7553 (+0.0077)	0.4838 (+0.0015)	<b>0.6121</b> (+0.002)
ChangeFormer [6]	SiROC	FL	MIOU	0.736 (+0.0448)	0.4586 (+0.0185)	0.584 (+0.0101)
ChangeFormer [6]	CVA	FL	MIOU	0.6589 (+0.0414)	0.4232 (+0.0254)	0.5666 (+0.0196)
ChangeFormer [6]	DCVA	FL	MIOU	0.678 (+0.0264)	0.4423 (+0.0193)	0.5864 (+0.0166)
BIT [18]	SiROC	FL	MIOU	0.7303 (+0.0158)	0.4598 (+0.0086)	0.5887 (+0.0058)
BIT [18]	CVA	FL	MIOU	0.6886 (+0.0091)	0.4437 (+0.006)	0.5839 (+0.0049)
BIT [18]	DCVA	FL	MIOU	0.7004 (+0.0117)	0.4543 (+0.006)	0.594 (+0.0038)

Table 12: Ablation Study: Varying the Training Set Size

Model	PL	# Training Cubes	Loss	Accuracy	MIOU	MF1
FC-Siam-diff [15]	✓	64	MIOU	<b>0.9227</b> (+0.0038)	<b>0.5376</b> (+0.0012)	<b>0.6127</b> (+0.0028)
FC-Siam-diff [15]		64	MIOU	0.8538 (+0.0076)	0.4865 (+0.0055)	0.5685 (+0.0048)
ChangeFormer [6]	✓	64	MIOU	0.813 (+0.0113)	0.4613 (+0.0098)	0.5494 (+0.0102)
ChangeFormer [6]		64	MIOU	0.7792 (+0.0277)	0.4528 (+0.0239)	0.5516 (+0.0449)
FC-Siam-diff [15]	✓	32	MIOU	0.9159 (+0.0115)	0.5324 (+0.0094)	0.6082 (+0.0088)
FC-Siam-diff [15]		32	MIOU	0.8215 (+0.0715)	0.4764 (+0.031)	0.5681 (+0.0328)
FC-Siam-diff [15]	✓	16	MIOU	0.9162 (+0.0127)	0.5338 (+0.007)	0.6101 (+0.0047)
FC-Siam-diff [15]		16	MIOU	0.8488 (+0.0205)	0.4851 (+0.0107)	0.569 (+0.0074)

For all given numbers of training cubes here, the pseudo label variant remains more effective than pure supervised training. This even holds for using the sizeable maximum amount of training data with 64 cubes which amount to over 1000 image pairs. It seems that FC-Siam-Diff reaches much of its potential already with 16 cubes as performance increases only marginally - if at all - when we increase training data. One may argue that this could be an effect of the pseudo labels however similar trends can also be observed without them. Overall, Table 12 underlines that SemiSiROC still has potential to improve methods when more training data is available where the gap is still notable even with over 1000 training pairs.

Further, we validate our results against the choice of the finetuning loss and test for all three models not only a focal loss but also a MIOU and a cross-entropy (CE) loss in Table 13. For FC-Siam-Diff with pseudo labels, the finetuning loss seems to have limited effects only. As expected, a CE loss pushes the model more towards the majority class which results in marginally higher accuracy but substantially lower MIOU and MF1 scores. A focal loss changes little compared to the baseline of MIOU loss with and without pseudo labels. For ChangeFormer, the choice of the loss seems more relevant but the main finding of the effectiveness of SemiSiROC remains unchanged. For all three losses, the ChangeFormer model performs notably better with pseudo labels. A focal loss also paints a similar picture to the MIOU loss for BIT. Again, the supervised CE baseline gains ac-



Table 13: Ablation Study: Robustness to Finetuning Loss

Model	PL	Loss PL	Loss	Accuracy	MIOU	MF1
FC-Siam-diff [15]	✓	FL	FL	0.787 (+0.0088)	0.4858 (+0.0021)	0.6008 (+0.0051)
FC-Siam-diff [15]			FL	0.693 (+0.0657)	0.4426 (+0.0246)	0.5798 (+0.0163)
FC-Siam-diff [15]	✓	FL	MIOU	0.7812 (+0.0104)	0.4854 (+0.0037)	<b>0.6029</b> (+0.0018)
FC-Siam-diff [15]			MIOU	0.6359 (+0.0405)	0.419 (+0.0288)	0.5706 (+0.0244)
FC-Siam-diff [15]	✓	FL	CE	0.7945 (+0.0088)	<b>0.4868</b> (+0.0043)	0.5987 (+0.0096)
FC-Siam-diff [15]			CE	0.7988 (+0.0233)	0.4466 (+0.0219)	0.5304 (+0.0483)
ChangeFormer [6]	✓	FL	FL	0.6762 (+0.0538)	0.4355 (+0.034)	0.5769 (+0.027)
ChangeFormer [6]			FL	0.5644 (+0.0164)	0.3548 (+0.0085)	0.5036 (+0.0088)
ChangeFormer [6]	✓	FL	MIOU	0.736 (+0.0448)	0.4586 (+0.0185)	0.584 (+0.0101)
ChangeFormer [6]			MIOU	0.4848 (+0.0923)	0.305 (+0.0627)	0.4545 (+0.0644)
ChangeFormer [6]	✓	FL	CE	<b>0.8068</b> (+0.0122)	0.4399 (+0.0158)	0.5155 (+0.0321)
ChangeFormer [6]			CE	0.7735 (+0.0471)	0.4237 (+0.0088)	0.5067 (+0.0178)
BIT [18]	✓	FL	FL	0.7133 (+0.0203)	0.4531 (+0.0088)	0.5864 (+0.0066)
BIT [18]			FL	0.6673 (+0.0774)	0.412 (+0.0318)	0.5447 (+0.0222)
BIT [18]	✓	FL	MIOU	0.7303 (+0.0158)	0.4598 (+0.0086)	0.5887 (+0.0058)
BIT [18]			MIOU	0.6242 (+0.0418)	0.4074 (+0.0227)	0.5587 (+0.0151)
BIT [18]	✓	FL	CE	0.7593 (+0.0145)	0.4639 (+0.0027)	0.581 (+0.0098)
BIT [18]			CE	0.7876 (+0.0256)	0.4236 (+0.0055)	0.4984 (+0.0139)
SiROC				0.6946	0.4408	0.5769

Table 14: Ablation Study: PL Training not on Test Images with SiamUNet

Model	PL	Loss PL	Loss	Accuracy	MIOU	MF1
FC-Siam-diff [15]	✓	FL	MIOU	<b>0.7541</b> (+0.0115)	<b>0.4621</b> (+0.004)	<b>0.581</b> (+0.0017)
FC-Siam-diff [15]			MIOU	0.5965 (+0.0419)	0.3902 (+0.0286)	0.5448 (+0.0246)

accuracy on the pseudo label version but this is largely overfitting to the majority class and the model lacks far behind in other criteria. To summarize, our results are robust to a variety of loss choices.

One advantage for the pseudo labels may stem from the fact that test scenes are in some cases already visible during pseudo label pre-training. Table 14 explores this and restricts the images from the test set not to appear during pseudo label pretraining. For this purpose, we split the cubes in Figure 8 into the west for pseudo label training and the east for evaluation. Still, SemiSiROC remains substantially more effective than its supervised baselines with a gap of over 15 p.p. in accuracy.

Finally, we apply the models of Table 10 to the OSCD test set as an additional validation exercise. The respective scores are reported in Table 15. For all three models, the margins to the supervised baselines are substantial. The minimum accuracy gap is at 15% for FC-Siam-Diff and is even larger for the other models. The range of accuracy is even in the range of supervised models in [28] which were trained on the OSCD training set while our model was not. The ChangeFormer model without pseudo models seems to collapse at times for this application. Even when focusing on better runs, however, the maximum accuracy is below 75%. In terms of balance, BIT reaches the best model here with high scores in MIOU and MF1 when using pseudo

Table 15: Quantitative Results OSCD Test Set trained on DynamicEarthNet and grouped by pseudo label use.

Model	PL	Loss PL	Loss	Accuracy	MIOU	MF1
FC-Siam-diff [15]	✓	FL	MIOU	<b>0.9575</b> (+0.0096)	0.5547 (+0.0185)	0.6206 (+0.0252)
FC-Siam-diff [15]			MIOU	0.8083 (+0.1035)	0.4927 (+0.0892)	0.5966 (+0.082)
ChangeFormer [6]	✓	FL	MIOU	0.8592 (+0.0692)	0.5145 (+0.0457)	0.6085 (+0.0356)
ChangeFormer [6]			MIOU	0.384 (+0.2976)	0.2139 (+0.1703)	0.2984 (+0.1892)
BIT [18]	✓	FL	MIOU	0.9248 (+0.0154)	<b>0.5585</b> (+0.0115)	<b>0.6422</b> (+0.012)
BIT [18]			MIOU	0.7273 (+0.059)	0.4082 (+0.0321)	0.5066 (+0.0238)

labels with a large gap to the no pseudo label baseline. All in all, SemiSiROC scores on OSCD as an unseen dataset look convincing. Particularly, the gaps to the supervised baselines are substantial.

#### 4.3.5 Discussion

In our experiments, SemiSiROC appears to be an effective strategy for semi-supervised learning which is robust to different losses and training set choices. Its performance is competitive on two benchmark datasets and against other pseudo label sources. Its off-the-shelf performance for an unseen dataset such as OSCD is particularly promising and it seems more robust to unseen regions than supervised baselines.

The mechanism with which these improvements materialize seems largely related to less sensitivity to false positives. Particularly, the use of pseudo labels made the models less sensitive to different illumination conditions in Figure 10. Further, the pseudo labels seem to help in fitting the shape of changes better. This seems natural since this is one of the key strengths of SiROC which may be passed down from the teacher to the student.

A second noteworthy thing is the relative weakness of the two transformer models in our experiments. As these models are data-hungry, one reason could be the limited availability of labels. ChangeFormer reached competitive performance on the LEVIR-CD dataset which has about 10x as many labeled pixels as the version of DynamicEarthNet we use. However, ChangeFormer is also a top model on the DSIFN-CD dataset which only contains 25% more labeled pixels than our version of DynamicEarthNet. Still, however, they are all from the same region of the Earth and focused on the same application. So the amount of training data *per use case* is still substantially larger which is likely the reason for our observations. In either case, the SemiSiROC strategy helps the Transformer models to bridge large parts of this gap by providing synthetic, additional training labels. SemiSiROC can be used in many applications potentially since its teacher model is application and label agnostic by design.



Part III

CONCLUSION AND OUTLOOK



## CONCLUSION AND OUTLOOK

---

### 5.1 CONCLUSION

This thesis set out to achieve two main goals:

1. To improve the dataset landscape for time-series applications of optical satellite images for deep learning.
2. To advance label-efficient change detection approaches with optical satellite imagery.

Section 3 addresses the first goal of dataset availability. I introduce two large-scale datasets: Denethor and DynamicEarthNet. Denethor is the first dataset for crop type mapping with daily data. It also presents the first opportunity to use novel Planet Fusion data for research purposes. Its use cases are not necessarily restricted to agriculture as other potential applications of the dataset are in declouding or super-resolution. This is because Sentinel-1 and 2 data are also provided as part of Denethor.

Our baseline experiments on the dataset explore the applicability of current methods for spatial and temporal generalization in crop type mapping. We test three categories of temporal models for Planet Fusion data: Spatial average of pixels per field, padded or upsampled snapshot of the full field with lightweight spatial encoder, and randomly sampled pixels per field.

Interestingly, simple pixel averages outperform our lightweight spatial encoders such as MobileNet. This underlines that there is significant information in the temporal evolution of the field alone. This is because the spatial average discards any spatial variation within the field. The strongest results are, however, obtained with spatial sampling of a small number of pixels per field. Here, we explore PseTae and its lightweight version which are based on temporal self-attention and reached the highest scores in our baseline experiments.

Still, random forest baselines with Sentinel data were competitive and surpassed most but not all deep learning models. What makes Denethor challenging for all baseline models is the spatial and temporal shift from training to test set. This is because the test dataset is not only a different (but geographically close) tile but the data also comes from a different year.

Our results underline that current models do not transform well off-the-shelf to the novel Planet Fusion data on this task. Custom models will potentially have to be developed to deal with the high spatial and temporal cadence of next-generation EO products.

As the second dataset, DynamicEarthNet is presented. It contains Planet Fusion and Sentinel-1 and 2 imagery for 75 AOIs across the globe with monthly, manually annotated, semantic ground truth for 2 years. This presents a significant step forward for change detection datasets since no other dataset so far provides multi-temporal and multi-class annotations of change events from across the globe. Many of the current change detection datasets hardly fulfill several of these aspects while DynamicEarthNet offers them several sources of input data and daily Planet Fusion imagery.

For the benchmark experiments on DynamicEarthNet, we explore a variety of multi-temporal semantic segmentation methods. We evaluate baselines with monthly, weekly and daily input imagery for semantic segmentation every month. Our supervised methods take inspiration from UNet. Weekly and daily supervised methods are extensions of UNet with a temporal backbone such as a ConvLSTM, temporal self attention or 3D convolutions. Additionally, we add a semi-supervised baseline built on DeepLabv3+ with a consistency loss during training. Additionally, we propose the Semantic Change Score (SCS) which combines binary and semantic change detection performance as a new metric.

A semi-supervised consistency element during training seems to only pay off with daily imagery but does not have much effect in the weekly case compared to the monthly baseline. In the supervised case for weekly inputs, U-TAE and U-ConvLSTM have an edge over the standard UNet with monthly data. However, this edge seems to disappear with daily inputs. On the contrary, 3D convolutions are not particularly effective with additional inputs in the weekly case but this flips with daily data.

Our results outline that much is yet to be understood about these next-generation products in multitemporal Earth observation. It appears that current methods can not yet effectively extract potentially relevant information from dense temporal inputs for these tasks. This is somewhat consistent with the findings on Denethor. Current methods leave ample potential for methodological innovation on both presented datasets since existing methods do not seem tailored to the abundance of temporal input. Before Denethor and DynamicEarthNet, this was not necessarily clear and the possibility to tailor methods to these applications not given. We therefore encourage the community to make use of these datasets and develop custom approaches for next-generation time-series tasks in Earth observation.

The second ambition of this thesis is the advancement of label-efficient change detection methods. In this scope, section 4 presents two contributions. At first, I introduce Sibling Regression for Optical Change Detection (SiROC). The method is inspired by exoplanet search in astronomy and models pixels as a function of their distant neighborhood. In subsequent time periods, this relationship can be

used to predict a pixel based on its neighboring pixels at that time. Then, the predicted pixel value can be compared to its actual value at that time. The absolute value of the difference contains a change signal as large differences point towards a potential update in the relationship between the pixel and its neighborhood.

We ensemble over mutually exclusive neighborhoods and add morphological operations to transition from pixel to object level. With SiROC, we achieve competitive results on four different change detection datasets which include urban, alpine and agricultural applications. Even without morphological profiles or ensembling, the effectiveness of SiROC is still high compared to recent baselines.

SiROC also offers a confidence score which allows for an effective combination with data-hungry deep learning methods. SemiSiROC is the second contribution in section 4. It is a semi-supervised method that uses SiROC predictions and its confidence score for pseudo label generation and prioritization.

We test this framework with a binary version of DynamicEarthNet and OSCD. SemiSiROC improves change detection performance compared to the supervised baseline by a substantial margin. This is robust to training set size, loss choices or splitting pseudo labels and actual training labels geographically. Additionally, the performance is still competitive when compared to other pseudo label sources or on OSCD where the method was not trained. It seems that the channel through which SemiSiROC works is that it is less sensitive to false positives and learns change shapes more accurately. This is intuitive since SiROC, its teacher, relies on morphological operations for shape refining. It seems that this knowledge is passed down by the teacher to the student to some extent. SemiSiROC is an effective combination of the advantages of traditional methods such as efficiency and spatial robustness with deep learning. By combining their advantages, SemiSiROC builds a bridge between these areas and pushes the methodological frontier in optical change detection further.

## 5.2 OUTLOOK

This thesis presents four advancements in multitemporal Earth observation. Particularly in change detection, the contributions include a first-off-its-kind dataset and two methodological contributions. Still, as change detection is a complex problem much is yet to be explored in this area. This section suggests promising avenues for future research and discusses potential stepping stones.

At first, next-generation Earth observation products as explored in this thesis will likely play a larger role in multi-temporal Earth observation research. Effectively utilizing higher spatial and temporal resolutions remains a research challenge with many unknowns. Particularly, Planet Fusion for example is a commercial product and Sentinel-

2 data is openly available. This begs the question under which circumstances commercial data has an advantage over public data also in cost-effectiveness. This may depend on the specific application and the relevance of timely observations. Even if commercial products may result in better performance, it may not be worth the additional cost or scalability issues. Our datasets provide a starting point for researchers in this regard since they provide an opportunity to explore this kind of data without cost with baseline models and results. We hope that this can guide future inquiries in research and practice disentangling public versus commercial data in Earth observation.

A second opportunity for future work lies in semantic change detection. Available ground truth data provided a significant bottleneck for this application. While this remains an issue, DynamicEarthNet provides a significant step towards better availability of global, multi-class change detection data from satellites. The baseline experiments show that off-the-shelf approaches struggle for multi-temporal applications on this dataset which points to a clear need for designated methods for this. Given the ongoing advance of image recognition and a better understanding of tailoring these methods to Earth observation, there may be significant potential for semantic change detection in upcoming years.

Third, recent innovations in self-supervised learning may provide ample opportunities for change detection research. Earth observation imagery is practically abundant. With decreasing computational constraints and additional metadata stored with satellite imagery, a well-designed pretext task can achieve a lot without the necessity for ground truth data. Two developments could be particularly useful in this regard. First, ongoing efforts provide more insights into using metadata in satellite imagery for self-supervised learning more effectively. For example, images from the same location in different seasons are contrasted in [77]. This way, the model learns season-invariant representations of multi-temporal Earth observation data. Among other results, they manage to reach competitive performance on OSCD. Overlapping patches are identified as a pretext task in [67] which could be seen as a version of spatial positives. In similar spirit, contrastive methods are also used in combination with temporal positives and geography-awareness for remote sensing data in [3].

The second beneficial trend is the rise of masked autoencoders, a widely popular self-supervised technique in language processing, also for images [48]. These autoencoders mask patches of the input which natively fits to Vision Transformers where images are split into patches anyway. In contrast to language, however, masking large parts of the image is the key in image recognition. First attempts to customize these advances for Earth observation data are already appearing with likely many more in the making. SatMAE [24] includes temporal embeddings as well as coordinate and date infor-

mation in pretraining. AdaMAE [5] prioritizes beneficial patches for masking with spatio-temporal data. With their auxiliary sampling network, they manage to get masking ratios up to 95% compared to 75% in regular MAE. This further enhances the training efficiency of spatiotemporal MAEs without compromising performance. MAEs are by no means the only self-supervised technique with promise for change detection. Pretrained diffusion models, for example, are taking many applications in image processing by storm [97]. While they are a generative method in principle, they can still be utilized in a self-supervised way with pretraining on unlabeled data. For example, a denoising diffusion probabilistic model is pretrained on a large-scale unlabeled Earth observation data set for change detection in [4]. In summary, the abundance of Earth observation data, the richness of the available metadata and progress in self-supervised techniques such as MAEs for spatio-temporal data provide significant opportunities for future change detection research. In combination with the availability of next-generation EO data and large-scale semantic change detection datasets such as DynamicEarthNet, this will likely fuel a range of methodological developments in change detection.

Even if data availability and methods are increasing, however, some fundamental pitfalls for change detection research remain. At first, change detection is an umbrella term used to describe a variety of different applications. In each of these use cases, the definition of what constitutes a change may vary and even contradict each other. Recall the example of harvesting crops from section 3. What some may consider the change event of interest here could be disregarded by others because it represents no difference in the land cover class. This is a fundamental roadblock for any ‘generic’ change detection method. Therefore, a one-method-fits-all approach for change detection may be hard to realize in practice without finetuning at least somewhat to the designated application.

A second roadblock is differences in benchmark datasets and practical applications of change detection. Any change detection dataset will typically abstract from the real world as changes are extremely rare in practice. A Copernicus report<sup>1</sup> randomly samples pixels for change analysis and finds a change rate of about 0.4% between 2015 and 2018. This is hardly a meaningful ratio for a benchmark dataset where the fraction of changes is typically at least in the single digits. Often, the frequency of changes varies between the datasets as well which may lead to the fact that pretraining on one of them may be harmful for other evaluations let alone practical applications. Evaluations on multiple datasets are becoming the norm for new methods but transparency, data and code sharing could still be more frequent to overcome this. Common benchmark datasets are a first starting

<sup>1</sup> [https://land.copernicus.eu/global/sites/cgls.vito.be/files/products/CGLOPS1\\_VR\\_LC100m-V3.0\\_I1.10.pdf](https://land.copernicus.eu/global/sites/cgls.vito.be/files/products/CGLOPS1_VR_LC100m-V3.0_I1.10.pdf)

point to overcome this but the use of open-source best practices will be critical to achieving common goals in better change detection methods.

All in all, there are exciting developments likely ahead of us in change detection fueled by methodological innovation and dataset availability. This thesis contributes to both of these pillars. While there is still a significant way to go for change detection, the contributions towards dataset availability and label-efficient learning provide a step towards this vision.



Part IV

APPENDIX





## PUBLICATIONS

---

### A.1 DENETHOR

---

# DENETHOR: The DynamicEarthNET dataset for Harmonized, inter-Operable, analysis-Ready, daily crop monitoring from space

---

Lukas Kondmann<sup>1,2</sup>, Aysim Toker<sup>3</sup>, Marc Rußwurm<sup>1,7</sup>, Andrés Camero<sup>2</sup>,  
Devis Peressutti<sup>4</sup>, Grega Milcinski<sup>4</sup>, Nicolas Longépé<sup>5</sup>, Pierre-Philippe Mathieu<sup>5</sup>,  
Timothy Davis<sup>6</sup>, Giovanni Marchisio<sup>6</sup>, Laura Leal-Taixé<sup>3</sup>, Xiao Xiang Zhu<sup>1,2\*</sup>

<sup>1</sup>Data Science in Earth Observation, Technical University of Munich (TUM)

<sup>2</sup>Earth Observation Center, German Aerospace Center (DLR)

<sup>3</sup>Dynamic Vision and Learning Group, Technical University of Munich (TUM)

<sup>4</sup>Sinergise

<sup>5</sup>Φ-Lab, European Space Agency (ESA)

<sup>6</sup>Planet Labs

<sup>7</sup>Environmental Computational Science and Earth Observation Laboratory (ECEO),  
École Polytechnique Fédérale de Lausanne (EPFL)

## Abstract

Recent advances in remote sensing products allow near-real time monitoring of the Earth’s surface. Despite the increasing availability of near-daily time series of satellite imagery, there has been little exploration of deep learning methods to utilize the unprecedented temporal density of observations. This is particularly interesting in crop monitoring where time series remote sensing data has been used frequently to exploit phenological differences of crops in the growing cycle over time. In this work, we present **DENETHOR: The DynamicEarthNET**<sup>2</sup> dataset for **H**armonized, **i**nter-**O**perable, **a**nalysis-**R**eady, **d**aily crop monitoring from space. Our dataset contains daily, analysis-ready Planet Fusion data together with Sentinel-1 radar and Sentinel-2 optical time series for crop type classification in Northern Germany. Our baseline experiments underline that incorporating the available spatial and temporal information fully may not be straightforward and could require the design of tailored architectures. The dataset presents two main challenges to the community: Exploit the temporal dimension for improved crop classification and ensure that models can handle a domain shift to a different year.<sup>3</sup>

## 1 Introduction

Remote sensing is entering a new era of time series analysis. A growing number of commercial and public satellites take the pulse of our planet in unprecedented frequency and resolution. Modern satellites reimage the Earth in ever-shorter time intervals generating petabytes of data every year [49]. Additionally, the open data policy of the Landsat program in the USA [46] and the Copernicus program by the European Space Agency (ESA) [1] have enabled the use of Earth observation (EO) data for many applications.

---

\*Corresponding author: xiaoxiang.zhu@dlr.de.

<sup>2</sup>DynamicEarthNET is the larger project under which our institutions collaborate to make multi-temporal Earth observation (EO) data more accessible.

<sup>3</sup>All model implementations and data are available at <https://github.com/lukaskondmann/DENETHOR>

One task at the heart of remote sensing efforts is vegetation monitoring. Particularly for the study of vegetation, access to frequent time series data is essential for accurate and timely monitoring of forests [47] and agricultural activities [26, 29]. The Sentinel-2 mission, which collects multispectral data at up to 10 m resolution at least every 5 days, has become particularly popular for crop type classification since its launch in 2015 [27, 33, 34, 38].

Recent advances in remote sensing have, however, made it possible to go beyond the spatial and temporal resolution of Sentinel-2. The Planet Fusion product, for instance, provides daily coverage of the Earth in 3m resolution and is part of a larger group of next-generation EO products that deliver analysis-ready data in dense time intervals. In the future, high temporal acquisition frequencies with near-daily intervals may become the norm in vegetation monitoring. This allows observing the growing cycle of crops in near real-time which provides significant potential for this field. However, the current methods in crop type classification are not designed to make use of daily temporal imagery, particularly in combination with high spatial resolution.

Therefore, in this paper, we present the dataset DENETHOR which provides the first opportunity to explore analysis-ready, daily data for crop type mapping. We provide a combination of harmonized, declouded, daily Planet Fusion data at 3m resolution together with Sentinel-1 and 2 time series for high-quality field boundaries and crop ids in Northern Germany. Train and test tiles are spatially separated and taken from different years to encourage out-of-year generalization.

We explore three types of benchmark methods on the dataset with the daily data: At first, we take the mean per pixel per field over time as input to a temporal encoder which discards spatial information but scales well [36]. Second, we include a spatial encoder in combination with the temporal encoder. Third, we follow [38] by randomly sampling pixels from a parcel as input to a temporal self-attention model. We compare these approaches to a random forest baseline with handcrafted spectral features from Sentinel-1 and 2.

Our experiments with the daily time series provide a starting point for future methodological approaches and underline that current methods may not yet be able to use the full potential of available information. We find that simply including a spatial encoder in addition to the temporal backbone does not improve performance compared to a simple mean of pixels per field. Second, many competitive deep learning models tested struggle to surpass the random forest baseline based on Sentinel-1 and 2 on our test set. One of the only approaches which manage this is based on pixel-set encoding and temporal self-attention [12, 38] with a high score of about 2/3 in accuracy. Finally, we underline that the performance drop in out-of-year evaluation from 2018 and 2019 can be substantial and amounts to 12 percentage points in accuracy on DENETHOR. To summarize, our contribution is threefold:

- We introduce the first publicly available benchmark dataset that includes daily, analysis-ready remote sensing data. With this, we aim to incentivize the community to study when and how this data source can be useful for crop type monitoring.
- Our experiments outline that mapping crops from daily imagery may require new methods since exploiting the full potential of the inputs seems to be a hurdle for many baseline models. Therefore, the dataset provides a challenging opportunity for the machine learning community with a novel type of input data.
- We emphasize the necessity of crop type models to be robust to domain shift not only along the spatial but also along the temporal dimension. This may be an underestimated problem in practice which we outline in our baseline results.

## 2 Related Work

Crop type classification is a special case of land cover classification in agricultural monitoring where field boundaries are typically assumed to be known. Because of its necessity for yield prediction and food security estimations, the task has received considerable attention in the past. Especially, multitemporal EO data has been a primary source for crop type classification for decades [26, 29]. Initially, however, the temporal scale of information provided a computational challenge. Therefore, early methods relied mostly on feature extraction from the time series [35]. Popular approaches include the computation of vegetation indices [5, 6, 11, 13, 28] that are often combined with Random Forest Classifiers [42] or Support Vector Machines [8, 20, 48]. Further, Dynamic Time Warping (DTW) [25] has found many applications in phenological studies with time series EO data [2, 7, 24].

More recently, however, the rise of deep learning in artificial intelligence [21] has also fueled improvements in crop type classification. Recurrent neural networks such as LSTMs [15] are well-suited to capture the temporal dynamics of crop types in a satellite time series [33, 40, 34]. Conversely, convolutional approaches can exploit the spatial dependency of the data. Further, [27] introduce the use of temporal convolutional neural networks (TempCNNs) in crop type mapping where convolutions are applied also along the temporal dimension. Convolutional and recurrent approaches have also been combined to leverage spatial as well as temporal information [18].

Attention mechanisms can further improve upon the capabilities of recurrent models [43]. Self-attention can be particularly effective when applied to raw optical time series as attention mechanisms can distinguish between informative and cloudy images [35]. A temporal attention encoder with pixel-set encoders (PSE) is successfully applied to randomly sampled pixels of crop parcels in [38].

The majority of recent methodological developments are based on publicly available medium-resolution satellite data from Sentinel-2 (S2). S2 has a spatial resolution of up to 10m per pixel and collects 13 spectral bands. Its revisit time is 5 days meaning that any region on Earth with a size larger than 100km<sup>2</sup> will be reimaged at least every 5 days [9]. The S2 mission has driven substantial methodological progress in remote sensing, especially because of its open data policy [49]. Still, on average 55% of the earth's surface is covered by clouds [19]. This can decrease the temporal resolution in practice notably and impede the aim of missions, such as Copernicus, to provide frequent and reliable observations [10].

Recent advances in remote sensing technology have induced the next-generation of optical EO products with superior temporal coverage compared to S2. The increased temporal resolution can be especially helpful against clouds since it improves the chance of a cloud-free observation substantially. Commercial missions such as the PlanetScope constellation achieve *near daily* revisit times. The spatial and temporal resolution of next-generation EO products holds great promise for a variety of applications in monitoring our planet. However, current methods in deep learning have not been designed to fully exploit the available temporal information at scale. This may be primarily an issue of available temporal resolution in current benchmark datasets. Benchmark datasets for crop type mapping are scarce in remote sensing and are mostly based on S2 data. The rarity of datasets may primarily be a result of the low availability of high-quality reference data at scale. Large-scale products of crop types such as the Cropland Data Layer (CDL) [4] in the US exist for some regions. However, the CDL is technically still a prediction and may not provide sufficient quality for benchmark purposes.

Table 1 presents an overview of available benchmark datasets for crop type classification from multi-temporal EO data. Two datasets based in Europe provide a large evaluation ground for newly developed crop type mapping models. At first, Breizhcrops [36] is based on S2 and reference data from almost 800,000 fields from the Brittany region in France. Similarly, the TimeSen2Crop [45] dataset covers a large fraction of fields in Austria with S2 input data at about 1,200,000 parcels. Satellite data is averaged at the field level which makes it possible to include a high number of fields. This averaging, however, discards a lot of spatial information per field, essentially reducing each field to one averaged pixel. Both of these datasets prioritize geographical size since there are natural limits to the temporal resolution through the S2 revisit rate and cloud obstruction.

Further, several datasets from Africa have been open-sourced as competitions through the work of the Radiant Earth Foundation together with Zindi Africa and local governments. The first dataset from Kenya was part of the challenge at the computer vision for agriculture workshop at the International Conference for Learning Representations (ICLR) in 2020. Similarly, datasets from Uganda, Rwanda and South Africa have been or are used in competitions to develop innovative methods for crop type mapping based on S2 (Uganda), aerial images (Rwanda) and a combination of Sentinel-1 and 2 (South Africa).

None of these datasets, however, include the opportunity to push the frontier of current models further by including next-generation EO data. Our dataset DENETHOR aims to change this by releasing daily, declouded and harmonized Planet Fusion data in combination with S1 and S2 inputs for 4,500 fields in Germany. With this, we aim to incentivize the remote sensing and the machine learning community to develop methods to improve current approaches based on rich data for a challenging and relevant problem. Naturally, the focus on temporal and spatial resolution of images comes with the necessity to restrict the spatial scale of the dataset to keep access to it democratic and feasible. Still, the daily data with 3m resolution is the main driver of dataset size which sums up to about

Table 1: Existing Datasets for Crop Type Classification. GSD = Ground Sampling Distance, RT = Revisit Time

	Inputs	GSD	RT	#Fields	Size[GB]
Breizhcrops (FR) [36]	S2	10m	5 days	768,000	17.4
TimeSen2Crop (AUT) [45]	S2	10m	5 days	1,200,000	2.1
CV4A Kenya [30]	S2	10m	5 days	4,700	3.5
Crop Type Uganda [3]	S2	10m	5 days	52	59.4
Crop Type Rwanda [32]	UAV	3cm	Monthly	2,611	26.9
Spot the Crop Challenge (SA) [31]	S1+S2	10m	5 days	35,300	52.1
DENETHOR (Ours)	PF+S2+S1	3m	Daily	4,500	254.5

255GB. The inclusion of S1 and S2 enables users of the dataset to further explore multi-modal combinations of input data.

Our dataset is the first of its kind to publish a daily product for scientific development at the intercept of machine learning and remote sensing. This does not only hold in the context of crop type mapping but in EO in general to the best of our knowledge. DENETHOR will be released under a CC-BY license to encourage widespread use and adoption.

### 3 DENETHOR: Daily Time Series for Crop Type Classification

**Crop Type Classes.** Our dataset includes field boundaries and crop type information from Northern Germany. This data is collected as part of the Common Agricultural Policy of the European Union. Farmers self-report the crops they grow in their fields to receive subsidies. The data is not only geographically precise but also of high quality since a variety of checks via in-situ measurements or EO data can potentially expose cheating.

Given the high spatial and temporal resolution of our dataset, we restrict our focus to two tiles. Both tiles are identical in size with  $24\text{km} \times 24\text{km}$ . One tile is used for training and validation, the other for testing. For the training tile, we include field masks and crop types from the year 2018 together with the respective satellite imagery. Test evaluations are based on the 2019 data. With this, we aim to encourage methodological development that incorporates not only a spatial but also a potential temporal shift in the input data. We will further release the 2018 test tile data and 2019 training tile data for future ablation studies.

The raw crop information provides the fields in vector format with a crop id coded between 1-999. Fields with areas below  $1000\text{m}^2$  are excluded since they are often broken in shape and can not easily be incorporated. This affects around 1% of all fields. We aggregate the crop type into a limited set of high-level classes which is common practice in crop type mapping [36]. The nine classes with the respective number of fields in the training set in brackets are: Wheat (305), Rye (276), Barley (137), Oats (45), Corn (251), Oil Seeds (201), Root Crops (23), Meadows (954) and Forage Crops (339). The class imbalance provides a challenge for machine learning algorithms but it is representative of the geographic region and an imbalance is generally common in real-world crop type mapping tasks [36]. Crops that do not fit into these categories are rare in the reference data but in these instances, we remove the respective fields instead of collecting them in a tenth ‘Other’ class.

**Planet Fusion Imagery.** The main source of imagery is the Fusion Monitoring product<sup>4</sup> by Planet Labs, a commercial provider of high-resolution satellite imagery. It is based on the Planetscope constellation of Cubesats which collect images of the Earth from over 180 small satellites. The product has a spatial resolution of 3m and collects 4 spectral bands (RGB + Near-infrared (NIR)). Although Fusion imagery is primarily used for early crop detection, plant health monitoring and the classification of phenological plant cycles, it has several features that may make it promising for crop type classification. At first, it provides imagery in a unique daily time interval which allows studying the evolution of crops in unprecedented temporal density. Especially in combination with the high spatial resolution, this could enable classification methods to pick up on small, crop-specific details of the growing cycle.

<sup>4</sup><https://www.planet.com/pulse/planet-announces-powerful-new-products-at-planet-explore-2020/>

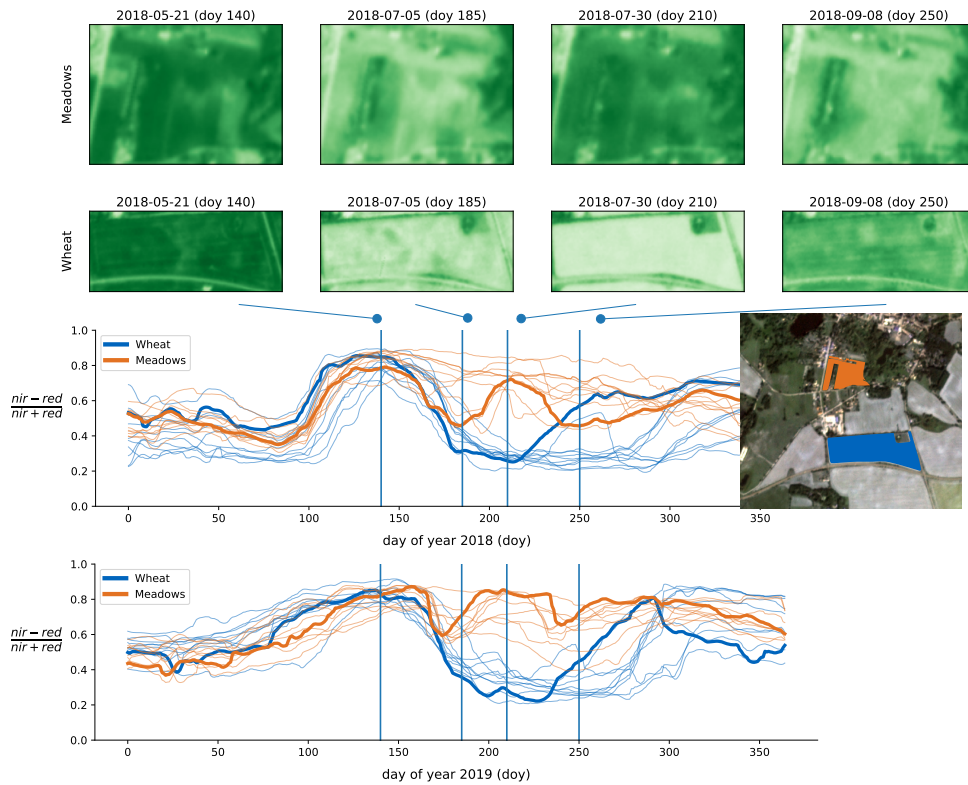


Figure 1: Examples of meadow and wheat parcels from the training dataset. The upper images show two parcels at four (of 365) acquisition times. The bottom plots shows the Normalized Difference Vegetation Index NDVI [41] averaged over all pixels of these parcels across the years 2018 (above) and 2019 (below). For reference, other fields meadow and wheat fields are plotted in thin lines which illustrates that wheat and meadow vary systematically between day 180 and 240 in both years. However, the vegetation activity varies notably between the years.

Second, it delivers a temporally consistent collection of images with removed clouds and shadows which in remote sensing is referred to as an analysis-ready (ARD) product. Potential gaps because of clouds are filled with different points in time. The data also includes quality assurance (QA) information that underlines from which source any pixel is taken from together with a confidence score if the observation is gap-filled. This may be particularly useful in combination with machine learning techniques since observation could be prioritized by their confidence measure.

Third, it is a Harmonized Landsat Sentinel-2 (HLS) time series.<sup>5</sup> To ensure interoperability with Landsat and Sentinel-2 data, the Fusion data is calibrated to the HLS spectrum which eliminates differences in the spectral signatures. These differences are subtle but important because they may, for example, lead to the fact that the red channel a Planetscope sensor collects is slightly different from the red band in Sentinel-2. The removal of these differences may also enable additional potential through data fusion. As Planet Fusion data is typically not freely available, our dataset provides an opportunity for the academic community to explore advantages and disadvantages of this data.

Figure 1 illustrates an example of what kind of signal multispectral time series data carry to map vegetation activity with two neighboring fields in our dataset. The two rows plot the normalized difference vegetation index (NDVI) around the meadow (top) and wheat (below) field for four points in time in a year. The 1D timelines at the bottom add the mean NDVI for the two selected parcels (thick lines) and other fields of the same crop (thin lines) in 2018 (above) and 2019 (below). After day 180 in the year, wheat has been harvested with a systematic decline in photosynthetic activity.

<sup>5</sup><https://earthdata.nasa.gov/esds/harmonized-landsat-sentinel-2>



In contrast, meadows are still active with a high NDVI which - among many other features - can be captured by algorithms. Comparing the NDVI across 2018 and 2019 underlines the difficulty of out-of-year generalization in crop type mapping. While patterns show some similarity across years, there are also significant differences, particularly for meadows in the middle of the year.

**Sentinel data.** To combine and compare Planet and the publicly available Sentinel data, we include imagery from Sentinel-1 and Sentinel-2 to the train and test tile. While the spatial and temporal depth of S2 is comparably low, the combination of spectral depth (S2) with spatial and temporal depth (Planet Fusion) may provide additional opportunities for crop type mapping. The S2 data is downloaded from Sentinel Hub <sup>6</sup> at processing level L2A. No observations are filtered because of cloud cover (maxcc=1) to maximize temporal coverage. We provide 12 bands - all resampled at 10m resolution - together with the valid data mask, scene classification (SCL) band and the s2cloudless probability map (SCP).

Sentinel-1 (S1) is a radar-based sensor with a revisit time of 6 days [49]. We provide S1 Ground Range Detected (GRD) data with included orthorectification from Sentinel Hub. Orthorectification is the removal of terrain distortions in raw satellite images which stem from the fact that satellites rarely take images directly above the area of interest ('off-nadir'). We include both, vertical-vertical (VV) and vertical-horizontal (VH) polarization. The distinction stems from the fact that radar-based sensors collect information from electromagnetic waves that may be repolarized when they reach the surface. HH measures the share of waves that were emitted in vertical polarization and return to the sensor in the same polarization. Conversely, HV measures the fraction of the waves which are repolarized. Radar-based sensors are not obstructed by clouds and provide information about vegetation from a different perspective. Hence, a multi-modal approach based on optical and radar data could be informative of phenological trends on the ground.

**Possible tasks.** While the main focus of this study is crop type classification, the uniquely high cadence of the data sources can be used for continuous monitoring of crop vigor and precise identification of crop growth stages and drive progress in sustainable agricultural practices. It may also be particularly insightful to study approaches for early crop detection in the season when the full time series is not yet available. Further, instead of taking field boundaries as given, the direct segmentation of crops and fields [37] provides a higher level of difficulty for models. This could also be seen in the context of instance segmentation and connected to approaches to the MS coco challenge [22] where the different modalities of inputs may provide an interesting challenge. Further, arable land classification could be an intermediary step towards the direct segmentation of crops.

Beyond applications in crop monitoring, DENETHOR could provide validation exercises in super-resolution and declouding since we provide long time series of multi-modal data at different resolutions. For declouding, downsampled Planet Fusion data could be treated as the desired output with raw and cloudy sentinel inputs. On the other hand, cloud-free sentinel images could be used as input for a superresolution network that tries to upsample to 3m resolution. Finally, declouding and superresolution could also be combined in a single task.

## 4 Model Descriptions

**Deep learning models.** The listed models are evaluated only with Planet data as inputs. When training solely on S2 inputs, models did not converge. This is likely because S2 models may need a higher number of fields to enable training because of missing temporal and spatial resolution. The Planet multi-temporal satellite image sequence provides data at high spatial (3m) and temporal resolution (1 day). We benchmark three different ways of operationalizing the crop type mapping task from field boundaries and daily satellite images as visible in Figure 2.

Following a common practice in crop type mapping [27, 34], we consider a simple pixel average encoder (Figure 2 left)  $f_{\text{spat}}(\mathbf{X}_t) = \frac{1}{hw} \sum_{r,c \in \text{mask}} \mathbf{x}_{r,c,t}$  over a field mask that averages each  $D$ -dimensional pixel  $\mathbf{x}_{r,c,t}$  of a field into a  $D$  dimensional vector. This discards spatial information for scalability.

Second, we include a rectangular image of each field (Figure 2 middle) at identical size (32x32). We sample down larger fields and zero pad smaller parcels. This preserves spatial and temporal information at the cost of increased input data of a factor of about  $10^3$ . We extract spatial and

<sup>6</sup><https://www.sentinel-hub.com>

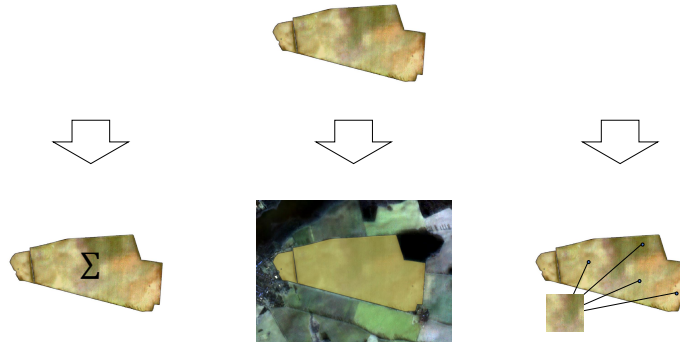


Figure 2: Three ways to operationalize the crop type mapping task with the field boundaries and satellite images as inputs (top): 1. Take the spatial mean of pixels within the field (left) 2. Take a rectangular image of identical size for all fields which can require padding or downsampling (middle) since fields vary in size. 3. Sample an identical number of random pixels from each field (right).

temporal features separately with two dedicated encoders. The spatial encoder  $f_{\text{spat}} : \mathbb{R}^{h \times w \times D} \mapsto \mathbb{R}^H$  maps a single  $D$ -dimensional image of certain height  $h$  and width  $w$  into a  $H$ -dimensional feature vector while the temporal encoder  $f_{\text{temp}} : \mathbb{R}^{T \times H} \mapsto [0, 1]^C$  maps a sequence of  $T$   $H$ -dimensional feature vectors directly into a probability for one of  $C$  classes. The complete model  $\{y_c\}_{c=1}^C = f_{\text{temp}}(\{f_{\text{spat}}(\mathbf{X}_t)\}_{t=1}^T)$  joins spatial encoder and temporal encoders such that the spatial encoder maps each image  $\mathbf{X}_t$  into a  $H$ -dimensional representation that is then mapped to a class probability  $y_c$  by the temporal encoder. In our case,  $T = 365$  because we use daily observations and  $C = 9$ .  $H$  depends on the model of the spatial encoder that is used.

For spatial encoders, we resort to standard, light-weight torchvision models [23], such as mobilenetv3 [16], squeezeNet [17], and resnet18 [14]. We use imagenet pre-trained weights but replace the first layer to accommodate  $D = 4$  input channels and use features before the classifier. The pixel average can be seen as a simple but scalable version of a spatial encoder.

For the temporal encoder, we utilize the provided implementations from the BreizhCrops [36] repository. We test the TempCNN [27] model that is based on three 1D-convolutions on the temporal dimension that are flattened and projected to class probabilities with a final dense layer. The 1D-Multi-Scale ResNet (MSResNet) model<sup>7</sup>, is a variant of the CSI Net [44] for pose estimation and uses three CNN streams each with a different kernel size of 3, 5, 7. Features in the CNN streams are joined by residual skip connections. A fixed-size feature vector is obtained by global max pooling before the decision layer. We also compare to a recurrent neural network with multiple stacked LSTM layers[15], as explored early for crop type mapping by [34] and a Transformer Encoder [43], as tested in [35].

A third variant of spatial and temporal encoding has been proposed by [12, 38] where  $f_{\text{spat}}$  is implemented as a Pixel Set Encoder (Pse). It transforms a set of random pixels (Figure 2 right) within a field parcel into a fixed representation by a pixel-wise MLP-based neural network with pooling. This strategy achieves good results when paired with a Temporal Attention Encoder (Tae) that is inspired by the self-attention-based Transformer architecture. This combination yields the PseTae [38] and a light-weight variant PseLTae [12].

**Data Fusion Baseline.** We compare the model accuracy with a random forest baseline on hand-designed features on openly available optical (Sentinel 2) and radar (Sentinel 1) data. For the 12-band Sentinel 2 and the 4-band RGB+NIR Planet data, we average all pixels of one field and consider only cloud-free observations which discards around 55% of S2 observations effectively cutting temporal resolution in half. Alongside the 12 or 4 spectral bands, we add the normalized difference vegetation index (NDVI) which is an index strongly related to vegetation. We calculate max, min, mean, median,

<sup>7</sup><https://github.com/geekfeiw/Multi-Scale-1D-ResNet>

Table 2: Accuracy of Benchmark Models with Planet Fusion data on the 2019 test set trained with 2018 data by spatial (rows) and temporal encoder (columns). Naively using standard spatial encoders such as ResNet18, SqueezeNet or MobileNet is not sufficient to lift performance over a simple pixel average. PselTae is the best deep learning model but we observe a notable drop in performance of more than 20% compared to the validation set. The majority of this drop can be attributed to out-of-year prediction rather than the spatial shift.

Spatial Encoder	Temporal Encoder			
	TempCNN [27]	MSResNet [44]	LSTM [34]	Transformer [35]
ResNet18 [14]	52.22%	49.53%	44.64%	43.61%
SqueezeNet [17]	53.94%	49.78%	35.89%	42.58%
MobileNetv3 [16]	53.20%	54.33%	43.46%	48.06%
Pixel Average [34]	64.46%	58.83%	48.40%	52.56%
Pixel-Set Encoding + Self-Attention				
PselTae [12]	<b>67.25%</b>			
PseTae [38]	64.95%			
Ablation Scores				
PselTae (2018)	78.77%			
PselTae (Val)	88.02%			

and standard deviation statistics on the resulting time series, as well as the date of max and min values for each of the 12 bands/4 bands plus NDVI index. For the Sentinel 1 radar data, we similarly average all pixels of one field at each time and calculate the same min, max, mean, median, std statistics on the vertical-vertical (VV) and vertical-horizontal (VH) polarisations. We additionally add the  $\frac{VV}{VH}$  ratio as the third band. We calculate these features for ascending and descending orbits separately and concatenate the features from each orbit type.

## 5 Benchmark Results

All temporal encoder models were trained with their BreizhCrops [36] defaults and all spatial encoders are initialized with pretrained imagenet weights from torchvision. We train with cross-entropy loss until convergence which typically occurs between epoch 50-100. Among the models, PselTae trained fastest with about 6 min per epoch with a batch size of 64 on a Nvidia GeForce GTX 1060.

Table 2 presents the accuracy test scores of our benchmarked approaches with the daily Planet data on the field level. The spatial encoder is given in the rows and the respective temporal encoder in the columns. Peak performances in this comparison group are reached by convolutional approaches with pixel average encoders with an accuracy of 64.46% for TempCNN and 58.83% for MSResNet. This stands in contrast to the results of [35] where self-attention outperforms convolutional approaches as temporal encoder. The notable difference is in the input data: When cloudy and raw Sentinel-2 data is used, convolutions may struggle to identify the relevant observations. In well-prepared, declouded images, temporal convolutions seem to excel and may hence be better suited for our dataset. All tested temporal encoders perform best with a simple pixel average as a spatial encoder. In alternatives to pixel averages, the choice of the spatial encoder seems to only make a marginal difference. This underlines that just including a deep-learning based spatial encoder does not work off-the-shelf and this kind of data may require more tailored approaches.

The highest score is reached by the light version of pixel-set encoding and self-attention (PselTae) with an accuracy of 67.25%. Given the validation accuracy of PselTae of 88.02%, this score is surprisingly low with a drop of over 20 percentage points (p.p.) between validation and test set. To split up this difference into a spatial and temporal shift, we evaluate PselTae as well on the test set in 2018 which results in 78.77% accuracy. Therefore, the spatial shift to a new tile accounts for about 9 p.p. and the temporal shift for 12 p.p. which is about 60% of the drop in total. Since accuracy scores may hide effects of class imbalance we also report macro-averaged F1 scores in the appendix in Table A1 which leaves the main impressions of Table 2 unchanged.

Table 3: Accuracy of different modalities with hand-designed features and a random forest classifier on the 2019 test set trained with 2018 data. Features are composed of 7 statistics (min, max, argmax, argmin, mean, median, std), for each band. Our Sentinel 2 data has 12 spectral bands plus a normalized red/near infrared ratio ( $7*(12+1)$ ), Planet has 4 bands plus normalized red/near infrared ratio. Sentinel 1 has two bands plus the band ratio.

Data Type	# Features	Accuracy	Macro F1-Score
Sentinel 1 (S1)	42	0.58	0.43
Sentinel 2 (S2)	91	0.59	0.42
Planet (PL)	35	0.37	0.12
S1 + S2	133	0.62	<b>0.46</b>
S1 + PL	77	0.60	0.42
S2 + PL	126	0.59	0.41
S1 + S2 + PL	168	<b>0.63</b>	<b>0.46</b>

Table 3 provides the scores of basic fusion experiments with random forests on the 2019 test set. S1 and S2 with hand-crafted features on their own reach accuracies of 58% and 59% respectively and 62% combined which surpasses all deep-learning models but PseLTae, PseTae and TempCNN with pixel average. The Random Forest approach does not work as well with the Planet data which is not surprising because it is tailored specifically for spectral rather than temporal or spatial depth. Beyond spectral features, textural features extracted from 3m data would likely contribute significantly to the differentiation of vegetation with Planet imagery. These textural features were not exploited in the current study but they could be highly complementary to the higher S2 spectral coverage.

The addition of Planet Fusion features to S1, S2 or to both adds some information that can be used by the random forest model but performance improves only marginally - if at all - in comparison to the models without PL. While the best deep learning models can surpass the performance of the Sentinel fusion baseline, this is not the case for most models implemented. Even if they do, the gap is fairly narrow at 2-5 p.p. accuracy. Extracting more advanced features from the Planet Fusion data with deep learning is therefore a promising route. Currently, however, it seems there may still be methodological potential in exploiting these novel kinds of inputs effectively at scale.

## 6 Discussion

The Planet Fusion data is a uniquely rich data source in the spatial and temporal dimension. However, our benchmark experiments suggest that our deep-learning baseline approaches may not be ideal to deal with the combination of high temporal and spatial resolution. The development of tailored architectures is opened as a challenge to the community to fully exploit the available information. One shortcoming of the tested models is that directly including spatial encoders before the temporal encoder makes performance worse compared to a simple spatial pixel average. PseLTae/PseTae with Planet Fusion data are the best performing models but only reach an accuracy of about two-thirds. Therefore, much potential may be in improving current deep learning methods for daily input data. One promising route might be to experiment with pixel-set encoders with temporal convolutions since they were superior to attention networks as temporal encoders in our dataset. Likely, performance from daily data could also be improved notably with the inclusion of a larger geographic area as the models encounter this type of data for the first time. Nevertheless, a large geographic scale may not always be available in practice which underlines the necessity to develop approaches that can also learn from smaller regions and adds to the items that make our dataset challenging.

Further, we find a significant performance drop of 21 percentage points between our validation data in 2018 taken from the training tile and the 2019 test tile. We show that in our case about 60% of this decrease can be attributed to the temporal shift of just one year which is about 12 percentage points in accuracy. This is large in magnitude and documents a challenge of crop type mapping in practice. Since the weather and growing cycles can vary notably from year to year, out-of-year generalization provides a challenge. The size of the performance drop shows that this could be underestimated in real-world applications of crop type mapping since the potential magnitude of this phenomenon seems not well documented yet. Therefore, it is a necessity that models incorporate this potential domain shift in the future to contribute to applications of crop type mapping in practice. To summarize, our

dataset presents two main challenges to the community: First, design new architectures which can effectively use spatial and temporal information for crop type mapping at scale. Second, ensure that models manage to generalize out of the year they have been trained on to make them applicable in real-world settings.

While we believe DENETHOR presents a significant step towards phenological monitoring near real-time, it has two main limitations. First, the area covered by our dataset is comparably small because we prioritize a high spatial and temporal resolution. Second, crop type datasets often lack geographic diversity because of label availability and class compatibility issues and our dataset is no exception. This limits the ability of developed algorithms to generalize to different geographies. Although resource-intensive initiatives begin to tackle this problem [39], this remains an obstacle with ample potential for future dataset work. One option for users interested in spatial generalization is to combine our dataset with a corpus of similar built for South Africa which is linked on our GitHub repository.

## 7 Conclusion

In this paper, we present **DENETHOR: The DynamicEarthNET dataset for Harmonized, interoperable, analysis-Ready, daily crop monitoring from space**. It is based on daily, analysis-ready Planet Fusion data in combination with Sentinel-1 and Sentinel-2 imagery. Ground truth of crop fields and types is taken from a public registry of farmer reports. We deliberately take the test data from a different year to ensure models that incorporate this temporal shift. Our experiments underline that the effects of this shift can be large and reduce performance by around 12 percentage points in accuracy.

Additionally, we point out that exploiting temporal, spatial and spectral information at scale is not a trivial task with current methods in crop type mapping. Tests with an off-the-shelf spatial encoder in combination with widely used temporal models fall short of simple pixel averages across a field with the same temporal model. Further, the best deep learning models tested barely outperform a random forest baseline of manually curated spectral features from S1 and S2 time series. Therefore, our dataset presents a challenging task to the machine learning community that may require the design of novel methods to push the frontiers of crop type mapping with next-generation EO data.

## Acknowledgments and Disclosure of Funding

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

## References

- [1] J. 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.
- [2] M. Belgiu, W. Bijker, O. Csillik, and A. Stein. Phenology-based sample generation for supervised crop type classification. *International Journal of Applied Earth Observation and Geoinformation*, 95:102264, 2021.
- [3] C. Bocquet. Dalberg data insights uganda crop classification. <https://doi.org/10.34911/RDNT.EH04X>, 2019.
- [4] C. Boryan, Z. Yang, R. Mueller, and M. Craig. Monitoring US agriculture: the US department of agriculture, national agricultural statistics service, cropland data layer program. *Geocarto International*, 26(5):341–358, 2011.
- [5] C. Conrad, S. Fritsch, J. Zeidler, G. Rücker, and S. Dech. Per-field irrigated crop classification in arid central asia using spot and aster data. *Remote Sensing*, 2(4):1035–1056, 2010.

- [6] C. Conrad, S. Dech, O. Dubovyk, S. Fritsch, D. Klein, F. Löw, G. Schorcht, and J. Zeidler. Derivation of temporal windows for accurate crop discrimination in heterogeneous croplands of uzbekistan using multitemporal rapideye images. *Computers and Electronics in Agriculture*, 103:63–74, 2014.
- [7] O. Csillik, M. Belgiu, G. P. Asner, and M. Kelly. Object-based time-constrained dynamic time warping classification of crops using sentinel-2. *Remote sensing*, 11(10):1257, 2019.
- [8] R. Devadas, R. Denham, and M. Pringle. Support vector machine classification of object-based data for crop mapping, using multi-temporal landsat imagery. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 39(1):185–190, 2012.
- [9] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, et al. Sentinel-2: Esa’s optical high-resolution mission for gmes operational services. *Remote Sensing of Environment*, 120:25–36, 2012.
- [10] P. Ebel, A. Meraner, M. Schmitt, and X. X. Zhu. Multisensor data fusion for cloud removal in global and all-season sentinel-2 imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 2020.
- [11] S. Foerster, K. Kaden, M. Foerster, and S. Itzerott. Crop type mapping using spectral–temporal profiles and phenological information. *Computers and Electronics in Agriculture*, 89:30–40, 2012.
- [12] V. S. F. Garnot and L. Landrieu. Lightweight temporal self-attention for classifying satellite image time series. URL <http://arxiv.org/abs/2007.00586>.
- [13] P. Hao, Y. Zhan, L. Wang, Z. Niu, and M. Shakir. Feature selection of time series modis data for early crop classification using random forest: A case study in Kansas, Usa. *Remote Sensing*, 7(5):5347–5369, 2015.
- [14] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer, 2016.
- [15] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- [16] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019.
- [17] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and < 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.
- [18] R. Interdonato, D. Ienco, R. Gaetano, and K. Ose. Duplo: A dual view point deep learning architecture for time series classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 149:91–104, 2019.
- [19] M. D. King, S. Platnick, W. P. Menzel, S. A. Ackerman, and P. A. Hubanks. Spatial and temporal distribution of clouds observed by modis onboard the terra and aqua satellites. *IEEE Transactions on Geoscience and Remote Sensing*, 51(7):3826–3852, 2013.
- [20] P. Kumar, D. K. Gupta, V. N. Mishra, and R. Prasad. Comparison of support vector machine, artificial neural network, and spectral angle mapper algorithms for crop classification using liss iv data. *International Journal of Remote Sensing*, 36(6):1604–1617, 2015.
- [21] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- [22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.
- [23] S. Marcel and Y. Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 1485–1488, 2010.

- [24] V. Maus, G. Câmara, R. Cartaxo, A. Sanchez, F. M. Ramos, and G. R. De Queiroz. A time-weighted dynamic time warping method for land-use and land-cover mapping. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(8):3729–3739, 2016.
- [25] C. S. Myers and L. R. Rabiner. A comparative study of several dynamic time-warping algorithms for connected-word recognition. *Bell System Technical Journal*, 60(7):1389–1409, 1981.
- [26] J. B. Odenweller and K. I. Johnson. Crop identification using landsat temporal-spectral profiles. *Remote Sensing of Environment*, 14(1-3):39–54, 1984.
- [27] C. Pelletier, G. I. Webb, and F. Petitjean. Temporal convolutional neural network for the classification of satellite image time series. *Remote Sensing*, 11(5):523, 2019.
- [28] J. M. Peña-Barragán, M. K. Ngugi, R. E. Plant, and J. Six. Object-based crop identification using multiple vegetation indices, textural features and crop phenology. *Remote Sensing of Environment*, 115(6):1301–1316, 2011.
- [29] B. C. Reed, J. F. Brown, D. VanderZee, T. R. Loveland, J. W. Merchant, and D. O. Ohlen. Measuring phenological variability from satellite imagery. *Journal of vegetation science*, 5(5):703–714, 1994.
- [30] REF. CV4A competition Kenya crop type dataset. <https://doi.org/10.34911/rdnt.dw605x>, 2020.
- [31] REF. Crop type classification dataset for Western Cape, South Africa. <https://doi.org/10.34911/rdnt.j0co8q>, 2021.
- [32] J. Rineer, R. Beach, D. Lapidus, M. O’Neil, D. Temple, N. Ujeneza, J. Cajka, and C. R. Drone imagery classification training dataset for crop types in Rwanda”. <https://doi.org/10.34911/rdnt.r4p1fr>, 2021.
- [33] M. Rußwurm and M. Körner. Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 11–19, 2017.
- [34] M. Rußwurm and M. Körner. Multi-temporal land cover classification with sequential recurrent encoders. *ISPRS International Journal of Geo-Information*, 7(4):129, 2018.
- [35] M. Rußwurm and M. Körner. Self-attention for raw optical satellite time series classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:421–435, 2020.
- [36] M. Rußwurm, C. Pelletier, M. Zollner, S. Lefèvre, and M. Körner. Breizhcrops: A time series dataset for crop type mapping. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences ISPRS (2020)*, 2020.
- [37] R. Rustowicz, R. Cheong, L. Wang, S. Ermon, M. Burke, and D. 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.
- [38] V. Sainte Fare Garnot, L. Landrieu, S. Giordano, and N. Chehata. Satellite image time series classification with pixel-set encoders and temporal self-attention. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12322–12331. IEEE. ISBN 978-1-72817-168-5. doi: 10.1109/CVPR42600.2020.01234. URL <https://ieeexplore.ieee.org/document/9157055/>.
- [39] M. Schneider, A. Broszeit, and M. Körner. Eurocrops: A pan-european dataset for time series crop type classification. *arXiv preprint arXiv:2106.08151*, 2021.
- [40] A. Sharma, X. Liu, and X. Yang. Land cover classification from multi-temporal, multi-spectral remotely sensed imagery using patch-based recurrent neural networks. *Neural Networks*, 105:346–355, 2018.
- [41] C. J. Tucker. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, 8(2):127–150, 1979.

- [42] S. Valero, D. Morin, J. Inglada, G. Sepulcre, M. Arias, O. Hagolle, G. Dedieu, S. Bontemps, P. Defourny, and B. Koetz. Production of a dynamic cropland mask by processing remote sensing image series at high temporal and spatial resolutions. *Remote Sensing*, 8(1):55, 2016.
- [43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.
- [44] F. Wang, J. Han, S. Zhang, X. He, and D. Huang. Csi-net: Unified human body characterization and pose recognition. *arXiv preprint arXiv:1810.03064*, 2018.
- [45] G. Weikmann, C. Paris, and L. Bruzzone. TimeSen2crop: a million labeled samples dataset of sentinel 2 image time series for crop type classification. pages 1–1. ISSN 1939-1404, 2151-1535. doi: 10.1109/JSTARS.2021.3073965. URL <https://ieeexplore.ieee.org/document/9408357/>.
- [46] C. E. Woodcock, R. Allen, M. Anderson, A. Belward, R. Bindschadler, W. Cohen, F. Gao, S. N. Goward, D. Helder, E. Helmer, et al. Free access to landsat imagery. *Science VOL 320: 1011*, 2008.
- [47] M. A. Wulder, W. A. Kurz, M. Gillis, et al. National level forest monitoring and modeling in canada. *Progress in Planning*, 61(4):365–381, 2004.
- [48] B. Zheng, S. W. Myint, P. S. Thenkabail, and R. M. Aggarwal. A support vector machine to identify irrigated crop types using time-series landsat ndvi data. *International Journal of Applied Earth Observation and Geoinformation*, 34:103–112, 2015.
- [49] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017.



A.2 DYNAMIC EARTHNET

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

Aysim Toker<sup>1,\*</sup>, Lukas Kondmann<sup>1,2,\*</sup>, Mark Weber<sup>1</sup>, Marvin Eisenberger<sup>1</sup>, Andrés Camero<sup>2</sup>,  
 Jingliang Hu<sup>2</sup>, Ariadna Pregel Hoderlein<sup>1</sup>, Çağlar Şenaras<sup>3</sup>, Timothy Davis<sup>3</sup>,  
 Daniel Cremers<sup>1</sup>, Giovanni Marchisio<sup>3,†</sup>, Xiao Xiang Zhu<sup>1,2,†,‡</sup>, Laura Leal-Taixé<sup>1,†</sup>  
 Technical University of Munich<sup>1</sup>, German Aerospace Center<sup>2</sup>, Planet Labs<sup>3</sup>

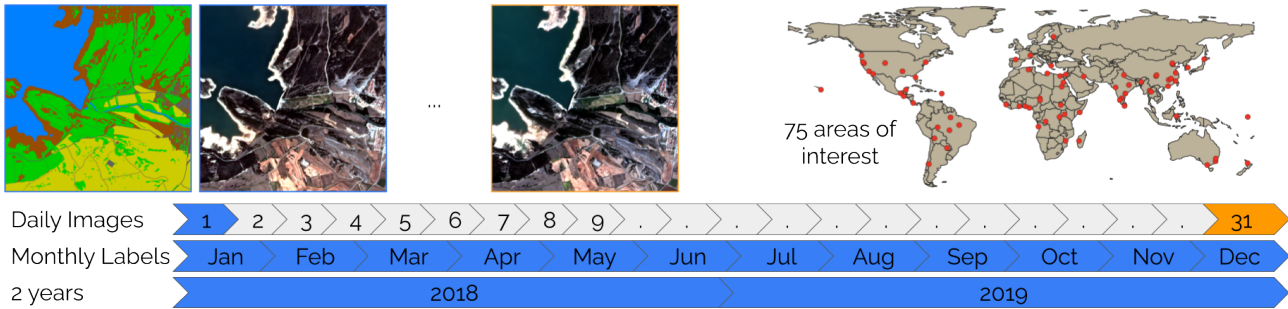


Figure 1. **Visualization of the *DynamicEarthNet* dataset.** For a specific area of interest, we show two satellite observations, 2019-08-01 and 2019-08-31, as well as the corresponding monthly ground-truth annotation (top left). The complete dataset consists of daily samples in the range from 2018-01-01 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

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

## 1. Introduction

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

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

millions of square kilometers of forest, automatic methods are required to detect these and other kinds of land changes. Current pixel-precise automatic methods are predominantly based on deep learning and thus require annotated data to learn.

In this work, we present *DynamicEarthNet*, a time-series satellite imagery dataset with daily revisits of 75 local regions across the globe. The dataset comprises consistent, occlusion-free 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 <sup>‡</sup>	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 <sup>‡</sup>	Oriented Bbox	Various
DeepGlobe [15]	✗	✗	1,146	Maxar	0.5	Seg. Mask	Various LULC
iSAID [43]	✗	✗	2,806	DOTA	0.15-12 <sup>‡</sup>	I. Seg Mask	Various
HRSCD [14]	✓	yearly	582	BD ORTHO	0.5	Seg. Mask	Various LULC
Hi-UCD [36]	✓	yearly	2,586	ELB <sup>†</sup>	0.1	Seg. Mask	Various LULC
<i>DynamicEarthNet</i>	✓	daily	54,750	PlanetFusion	3.0	Seg. Mask	Various LULC

<sup>†</sup> Estonian Land Board, <sup>‡</sup> 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 product<sup>1</sup> from Planet Labs, which provides multi-

<sup>1</sup><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)<sup>2</sup> 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.

<sup>2</sup><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 our the baseline [21] in combination with DeepLabv3+ [10] in Tab. 4. Similar to the spatio-temporal experiments, we consider different temporal densities. For the purely supervised setting, all unlabeled images are discarded (monthly). Additionally, we compare different semi-supervised settings with 6 (weekly), 28-31 (daily) unlabelled samples per month. Both, daily and weekly data help to improve over

	Sample Frequency	<i>per class IoU</i> ( $\uparrow$ )						Val	Test
		Imp. Surface	Agriculture	Forest	Wetlands	Soil	Water	mIoU ( $\uparrow$ )	mIoU ( $\uparrow$ )
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	<b>8.0</b>	77.3	0.0	<b>39.1</b>	58.1	<b>35.7</b>	<b>39.7</b>
	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	<b>77.7</b>	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	<b>32.4</b>	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	<b>66.0</b>	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	Test	
		Imp. Surface	Agriculture	Forest	Wetlands	Soil	Water	mIoU ( $\uparrow$ )	mIoU ( $\uparrow$ )	
CAC [21]	monthly	$\checkmark$	18.1	4.8	74.7	0.0	33.9	<b>55.9</b>	31.2	37.9
	weekly	$\times$	28.0	<b>7.2</b>	<b>75.7</b>	<b>8.3</b>	38.9	51.0	<b>34.9</b>	37.9
	daily	$\times$	<b>28.9</b>	4.0	75.5	0.5	<b>39.0</b>	55.6	33.9	<b>43.6</b>

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]	<b>19.1</b>	9.5	<b>28.7</b>	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	<b>43.6</b>
	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	<b>11.5</b>	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



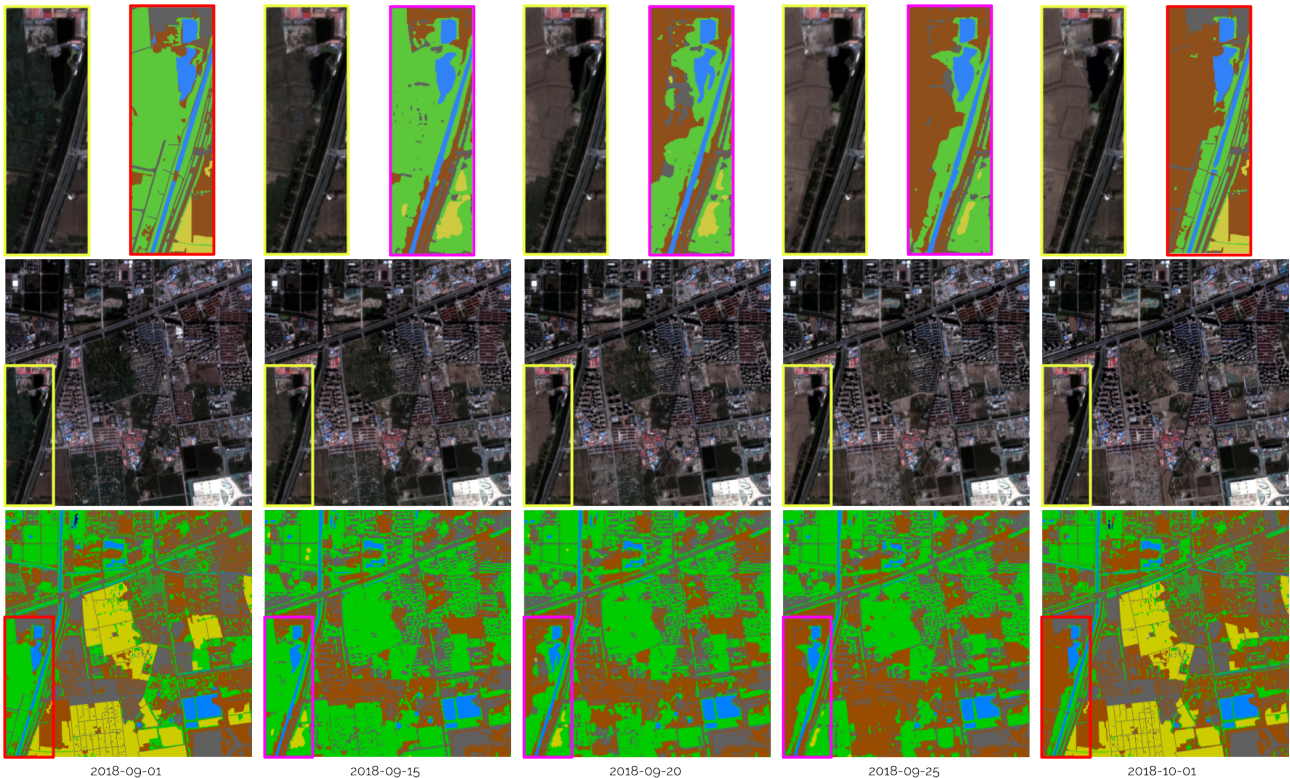


Figure 3. **Qualitative results on validation set.** Semantic maps (bottom row) of the semi-supervised baseline CAC [21] trained on daily images. The input sequence consists of 5 images (middle row) from September to October, spanning one month. For the first and last semantic map of the considered sequence,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

rics 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](#), [11](#)
- [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](#), [11](#)
- [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 Toket, 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 Toket, 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](#), [13](#), [14](#), [16](#), [17](#)
- [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, 13**
- [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, 13, 14, 15, 17**
- [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, 13**
- [29] A.H. Pickens, M.C. Hansen, B. Adusei, and Potapov P. Sentinel-2 forest loss alert. [www.globalforestwatch.org](http://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, 13, 15, 17**
- [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, 13**
- [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 Be-longie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liang-pei 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.3 SIROC



# Spatial Context Awareness for Unsupervised Change Detection in Optical Satellite Images

Lukas Kondmann<sup>1</sup>, Aysim Toker, Sudipan Saha<sup>2</sup>, *Member, IEEE*, Bernhard Schölkopf<sup>3</sup>, Laura Leal-Taixé<sup>4</sup>,  
and Xiao Xiang Zhu<sup>5</sup>, *Fellow, IEEE*

**Abstract**—Detecting changes on the ground in multitemporal Earth observation data is one of the key problems in remote sensing. In this article, we introduce Sibling Regression for Optical Change detection (SiROC), an unsupervised method for change detection (CD) in optical satellite images with medium and high resolutions. SiROC is a spatial context-based method that models a pixel as a linear combination of its distant neighbors. It uses this model to analyze differences in the pixel and its spatial context-based predictions in subsequent time periods for CD. We combine this spatial context-based CD with ensembling over mutually exclusive neighborhoods and transitioning from pixel to object-level changes with morphological operations. SiROC achieves competitive performance for CD with medium-resolution Sentinel-2 and high-resolution PlanetScope imagery on four datasets. Besides accurate predictions without the need for training, SiROC also provides a well-calibrated uncertainty of its predictions.

**Index Terms**—Change detection (CD), multitemporal, optical images, unsupervised, urban analysis.

## I. INTRODUCTION

CHANGE detection (CD) is at the heart of many impactful applications of remote sensing. Studying differences in land cover and land use over time with remote sensing imagery can shed light on urbanization trends [1], [2], ecosystem

Manuscript received July 14, 2021; revised October 14, 2021; accepted November 6, 2021. Date of publication November 25, 2021; date of current version February 21, 2022. This work was supported in part by the Helmholtz Association through the joint research school “Munich School for Data Science - MUDS,” in part by the Framework of Helmholtz AI under Grant ZT-I-PF-5-01—Local Unit “Munich Unit @Aeronautics, Space and Transport (MASTR),” in part by the Helmholtz Excellent Professorship “Data Science in Earth Observation-Big Data Fusion for Urban Research” under Grant W2-W3-100, in part by the European Research Council (ERC) through the European Union’s Horizon 2020 Research and Innovation Programme under Grant Agreement ERC-2016-StG-714087 (*So2Sat*), and in part by the German Federal Ministry of Education and Research (BMBF) in the framework of the International Future AI Lab “AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond” under Grant 01DD20001. (*Corresponding author: Xiao Xiang Zhu.*)

Lukas Kondmann and Xiao Xiang Zhu are with the Remote Sensing Technology Institute, German Aerospace Center (DLR), 82234 Weßling, Germany, and also with Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany (e-mail: lukas.kondmann@dlr.de; xiaoxiang.zhu@dlr.de).

Aysim Toker and Laura Leal-Taixé are with the Dynamic Vision and Learning Group, Technical University of Munich, 80333 Munich, Germany (e-mail: leal.taixe@tum.de; aysim.toker@tum.de).

Sudipan Saha is with Data Science in Earth Observation, Technical University of Munich (TUM), 80333 Munich, Germany (e-mail: sudipan.saha@tum.de).

Bernhard Schölkopf is with the Max Planck Institute for Intelligent Systems, 72076 Tübingen, Germany, and also with the ETH Zürich, 8092 Zürich, Switzerland.

Digital Object Identifier 10.1109/TGRS.2021.3130842

dynamics [3], surface water and sea ice trends [4], [5], and damages through natural disasters [6]–[8]. Because of rising spatial and temporal resolutions of Earth observation imagery, the possibilities of multitemporal analysis have increased significantly [9]. Combined with the open data policy of the Copernicus program, it is, for example, possible to acquire a Sentinel-2 image with 10-m resolution per pixel of any region of interest on any continent every five days [10] free of charge. Commercial providers of satellite imagery can even offer almost daily coverage with high-resolution imagery for large parts of the planet [11]. These trends emphasize the increasing opportunities in monitoring Earth from space and the relevance of CD as a field within remote sensing. Obtaining labeled data for CD, however, is costly in terms of time and effort, especially at scale. Therefore, a large focus of attention in the design of CD algorithms is unsupervised methods that do not require ground truth [12].

The applicability of unsupervised CD methods in multispectral satellite images varies depending on the spatial resolution of input images. For very-high-resolution (VHR) imagery with a spatial resolution up to 0.5 m, deep-learning-based methods tend to be in general preferable because of their elaborate capacity to model spatial context [12] although most of the work in this area focuses on supervised methods [13]–[18]. Since, for VHR imagery, an object, such as a building, consists of a number of pixels, modeling spatial context is essential to provide accurate unsupervised change segmentations. Saha *et al.* [12] introduce deep change vector analysis (DCVA), a VHR CD framework that combines ideas from image differencing with feature extraction based on pre-trained neural networks. DCVA has also been combined with self-supervised pretraining of the feature extractor specifically for remote sensing images [19]. MSDRL [20] is a scale-driven unsupervised method that uses deep feature extraction to obtain a pseudoclassification of change superpixels. Superpixels with high certainty pseudolabels are then taken as input to train a support vector machine, which eventually classifies the uncertain superpixels. Such preclassification schemes where pseudolabels are obtained based on another method have also been presented in conjunction with methods for unsupervised CD in synthetic aperture radar images [4], [21], [22], such as PCANet [23], [24]. Gong *et al.* [25] introduced modeling the difference image with a generative adversarial network (GAN). While the deep learning methods above were primarily designed for high-resolution imagery, some of them can be applied to medium-resolution imagery as well. In the case of

DCVA, there also exists a variant adjusted to the spatial and spectral scales of Sentinel-2 [26].

For medium-resolution CD, nondeep learning methods based and improved on change vector analysis (CVA) can still compete. CVA takes the difference of radiometric values or features derived from it over time [12] and applies a threshold to this difference image. Examples of features that have been derived from radiometric values as input for image differencing are vegetation indices [27] or tasseled cap transformation features [28]. Otsu thresholding [29] has been shown to be effective for thresholding the difference image [30] although a variety of approaches exist [31]–[33]. Beyond binary CD, the signal in the CVA difference image can also be used to uncover the type of change [34], [35].

CVA-based methods can still be insightful especially with medium resolution because the size of objects in these images is typically assumed to be similar to the spatial resolution of a pixel. However, extensions of CVA still fall short of the deep-learning-based DCVA for unsupervised CD on the OSCD benchmark. Still, the relative performance of traditional methods based on CVA improves with medium-resolution imagery compared to higher resolutions. More recent versions of CVA that can also be applied to higher resolution imagery tend to include close spatial context of pixels to some extent. Parcel CVA (PCVA) includes surrounding information of pixels by independent hierarchical segmentation at several scales [36]. Robust CVA (RCVA) improves on potential coregistration errors in the CVA framework by replacing a point in the difference image with the difference to a neighboring pixel if the difference to this neighbor is smaller [30]. Object CVA (OCVA) computes histograms of object sizes in an image and incorporates this information into a CVA framework [37]. Image differencing methods have also been successfully combined with morphological operations that allow transitioning from the pixel to the object level [38].

Although neighboring pixels are somewhat included in the change analysis of a pixel in these extensions of CVA, the spatial extent of incorporated information is small compared to the effective window of sequential convolutional operations in neural networks. Neighborhood in this context is defined not only as the immediate neighbors to a pixel of interest but also its larger spatial context up to a distance measure. The distant neighborhood of a pixel may help to identify changes because it is also affected by local trends in the image but unaffected by the change itself. For example, consider an explosion of a building between preimage and postimage, such as in the Beirut dataset used in the following. Analyzing the distant neighborhood allows to separate the actual change (destroyed building) from local trends, such as dust and dirt, which remains on surrounding buildings stirred by the explosion. However, the use of distant neighborhood context has only found limited application in CD thus far. This is particularly surprising since applications of image differencing in other domains, such as astronomy emphasize the importance of the relation of a pixel to its neighborhood [39].

Wang *et al.* [39] present the causal pixel model (CPM) for the study of multitemporal Kepler data that is used to spot transiting exoplanets in front of distant stars observed by the

space telescope. The method is also more generally known as half-sibling regression (HSR) [40]. Their task is conceptually related to a CD problem in remote sensing since it is also centered around spotting changes in multitemporal reflection intensities, which should be unrelated to the acquisition conditions. In their case, these deviations hint toward a transient object in front of a distant star rather than a change on the ground, but the fundamental principle is similar. They solve this task by modeling pixels as a function of their distant neighbors. With this model, it is possible to obtain a prediction for pixels in subsequent time steps based on their distant neighbors and compare the prediction to the actual value of the pixel. The size of the difference between predictions of pixels and their actual values is interpreted as the strength of the change signal.

HSR is related to the application of local binary patterns [41] for CD in more traditional image recognition problems. Bilodeau *et al.* [42] design a method based on local binary similarity patterns (LBSPs) to separate the image background from changes in multitemporal images. In their method, a binary similarity measure is computed between a pixel of interest and its closest neighbors within an image. If the binary similarity pattern updates notably between images, this is considered to be a change signal. A version of LBSPs has also been applied to CD in remote sensing where multitemporal images are split into overlapping blocks, and LBSPs of these blocks are compared across time [43]. Similarly, the graph structure of image patches across time has been used for homogenous and heterogenous CD [44]. The shared principle between LBSP and HSR is the approach to compare a pixel to its neighborhood and inspect how this relationship changes over time to discover potential changes. However, HSR relies mostly on distant neighborhood information rather than a small set of close neighbors and models this relationship explicitly to obtain a prediction for subsequent time periods.

One key property of HSR is the fact that it is by design comparably robust to registration errors and varying acquisition conditions for a given sensor [39]. This is because variations in the acquisition conditions can also affect distant neighbors in an image, whereas actual changes at the pixel level should be independent of distant context. Changing acquisition conditions and registration errors, however, are two of the primary sources of false positives (FPs) in CD [45]. Since HSR deals comparably well with these issues, it may work especially well for CD in remote sensing time-series data. In this article, we, therefore, apply HSR for CD in remote sensing. When we know from astronomy that distant spatial context can improve resilience against varying acquisition conditions, this might be particularly helpful in remote sensing CD.

We modify HSR in two major ways to apply it as SiROC for CD in remote sensing. First, we design an ensemble version of HSR based on mutually exclusive neighborhoods. Second, we make use of morphological operations to transition from pixel-level changes to object-level changes. SiROC is tested for urban CD in medium-resolution images on the Onera Satellite CD Dataset (OSCD) and high-resolution images from the Beirut Harbor Explosion of 2020 [Beirut Harbor Explosion

Dataset (BHED)]. Outside the urban context, we test SiROC on the Barrax Agriculture dataset and the Lamar Alpine dataset. Our main contributions are threefold.

- 1) We introduce SiROC, a robust method for unsupervised CD in optical remote sensing that combines ideas from HSR with ensembles over mutually exclusive neighborhoods and morphological operations.
- 2) SiROC achieves competitive performance for medium- and high-resolution unsupervised CDs with optical images.
- 3) SiROC also returns a built-in, well-calibrated uncertainty score with its change segmentation. The uncertainty measure allows distinguishing the predictions of the model by confidence, which is an important feature for pseudolabeling or detecting distribution shift.

## II. METHOD

### A. Half-Sibling Regression Image Differencing

The foundation of our method is HSR, which was originally developed for time-series analysis of the Kepler data in astronomy [39], [40], [46], [47]. Fig. 1 displays the intuition of HSR and how it is applied to obtain signals of changing pixels across time in three steps.

First, HSR models the pixel value of a star at time  $t$  as a linear combination of the pixel values of many other stars from the distant neighborhood of the pixel [see Fig. 1(a)]. The result of this first step is a linear coefficient for every included neighborhood pixel. In the second step [see Fig. 1(b)], predictions for the pixel at time  $t + 1$  are obtained with HSR based on the neighboring pixels at  $t + 1$  and the respective linear coefficients from step 1. If steps 1 and 2 are executed for the whole image, there is a prediction for any pixel at  $t + 1$  and its actual value. The predicted image at  $t + 1$  is then subtracted from the actual image value to obtain a change signal for every pixel [see Fig. 1(c)]. Intuitively, one expects a change in the pixels where the predictions based on the pixel's relation to its neighborhood in the past divert from the actual realization of the pixel. In the following, we elaborate on the formal definition of the three steps described. We restrict our description of HSR to the case of two time periods for simplicity since this is how we apply the method to remote sensing as well.

*Step 1:* Let  $I_{x,y,t}$  be a pixel in a single-channel, 2-D image time series ( $I$ ) at time  $t$  with coordinates  $(x, y)$ . HSR models the pixel  $I_{x,y,t}$  as a linear combination of a set of distant neighbors  $N$  from  $I$ . The neighborhood set has the points  $I_{i,j,t}$  as elements such that  $(i, j) \in N_{x,y}$

$$I_{x,y,t} = \sum_{(i,j) \in N_{x,y}} \beta_{i,j,x,y} I_{i,j,t} + \epsilon_{x,y,t} \quad (1)$$

where  $\beta_{i,j,x,y}$  is the coefficient of neighbor  $I_{i,j,t}$  to model the point of interest  $I_{x,y,t}$ .  $\epsilon_{x,y,t}$  is the residual of the model. Neighbors are chosen from the distant neighborhood because they might be subject to the same noise as the pixel of interest when they are selected to close to it. Wang *et al.* [39] require that an eligible neighbor has a distance of at least 20 pixels from the pixel of interest to be considered. This ensures that

the pixel of interest and the chosen neighbors have practically no overlap in stellar illumination. The number of neighboring pixels considered is generally large, and Wang *et al.* [39] select 4000 neighboring pixels in their original proposal of HSR to model one pixel of interest for Kepler data. Given the high temporal density of observations for each pixel (every 30 minutes) in Wang *et al.* [39], this is still solvable because the number of observed time periods exceeds the number of neighboring pixels used.

However, in the bitemporal case, where only one period is used for fitting, there are many potential combinations of  $\beta$ , which solves (1). We derive  $\beta$  as the closed form solution of the least-squares problem. It is a function of the pixel of interest  $I_{x,y}$ , the respective neighbor  $I_{i,j}$ , and the quadratic sum of all neighbors  $I_{i',j'}$

$$\beta_{i,j,x,y} = \frac{I_{i,j,t}}{\sum_{(i',j') \in N_{x,y}} I_{i',j',t}^2} I_{x,y,t}. \quad (2)$$

*Step 2:* With the coefficients obtained in step 1,  $I_{x,y,t+1}$  can be predicted as

$$\hat{I}_{x,y,t+1} = \sum_{(i,j) \in N_{x,y}} \beta_{i,j,x,y} I_{i,j,t+1}. \quad (3)$$

With the expression for  $\beta$  from (2), (3) can be rearranged to

$$\hat{I}_{x,y,t+1} = \frac{\sum_{(i,j) \in N_{x,y}} I_{i,j,t+1} I_{i,j,t}}{\sum_{(i,j) \in N_{x,y}} I_{i,j,t}^2} I_{x,y,t} \equiv g_{t+1} I_{x,y,t} \quad (4)$$

where  $g_{t+1}$  resembles a growth rate of the sum of pixel values in the selected neighbors from  $t$  to  $t + 1$ . In essence, the assumption is that, if the pixel values around  $I_{x,y,t}$  increase by a factor  $g_{t+1}$  and no changes occurred at this location,  $I_{x,y,t+1}$  should be close to  $I_{x,y,t} g_{t+1}$ . We can circumvent the explicit calculation of beta and directly obtain  $\hat{I}_{x,y,t+1}$  based on (4), which is computationally efficient.

*Step 3:* The difference between  $I_{x,y,t+1}$  and  $\hat{I}_{x,y,t+1}$  is taken as the change signal for pixel  $I_{x,y}$  between  $t$  and  $t + 1$

$$I_{x,y,t+1} = \hat{I}_{x,y,t+1} + \epsilon_{x,y,t+1}. \quad (5)$$

After obtaining  $I_{x,y,t+1}$  for all  $(x, y) \in I_{t+1}$ , the residual is given as the difference of the image matrices

$$\epsilon_{t+1} = \hat{\mathbf{I}}_{t+1} - \mathbf{I}_{t+1}. \quad (6)$$

Note that this is slightly different from the standard application of image differencing in CVA in multitemporal remote sensing. We do not directly take the difference of the image vectors at  $t$  and  $t + 1$ . Instead, we predict how the image would have looked like in  $t + 1$  if the local neighborhood relations persisted. Then, we use this predicted image as input for image differencing with the actual image in  $t + 1$ . The extension of HSR to images with several channels is straightforward as one can directly sum the absolute values of  $\epsilon$  for each channel to incorporate HSR information from all channels. Let  $\epsilon_{t+1,c}$  be the residual of channel  $c$  of a multispectral image with  $C$  channels. Then, the aggregated change signal can be computed as

$$\epsilon_{t+1} := \sum_{c=1}^C |\epsilon_{t+1,c}|. \quad (7)$$

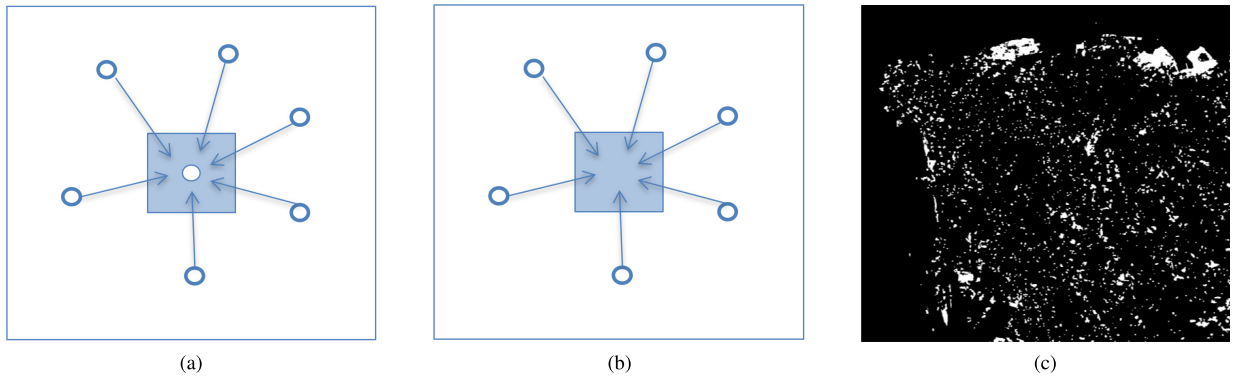


Fig. 1. HSR. (a) Set of neighbors is fit to a pixel of interest as a linear combination at time  $t$ . (b) At  $t + 1$ , the pixel values of the neighbors are used together with the coefficients obtained at time  $t$  to predict the pixel of interest in  $t + 1$ . (c) Predicted pixel values are compared with the actual pixel values at  $t + 1$  to obtain a change signal.

---

### Algorithm 1 SiROC

---

**Input:**  $I_t, I_{t+1}, s, n_{\max}, e_{\text{start}}$

**Output:** Binary Change Segmentation

```

1:  $e = e_{\text{start}}, n = e_{\text{start}} + s$ 
2:  $\text{Uncertainty\_CM} = \text{zeros\_like}(I_t)$ 
3: while  $n < n_{\max}$  do
4:   for (channel in channels) do
5:     for (pixel in  $I_t$ ) do
6:       Apply HSR( $n, e$ ) to get  $\hat{\mathbf{I}}_{t+1}$ 
7:     end for
8:      $\text{Channel\_Difference\_Image} = \hat{\mathbf{I}}_{t+1} - \mathbf{I}_{t+1}$ 
9:   end for
10:   $\text{Diff\_Image} = \text{Sum}(|\text{Channel\_Difference\_Images}|)$ 
11:   $\text{Binary\_CM} = \text{Otsu\_Thresholding}(\text{Diff\_Image})$ 
12:   $\text{Binary\_CM\_Object} = \text{Morph\_Profile}(\text{Binary\_CM})$ 
13:   $\text{Uncertainty\_CM} = \text{Uncertainty\_CM} +$ 
     $\text{Binary\_CM\_Object}$ 
14:   $n = n + s$ 
15:   $e = e + s$ 
16: end while
17:  $\text{Final\_Segmentation} = \text{Threshold}(\text{Uncertainty\_CM})$ 

```

---

### B. HSR for Earth Observation Data (SiROC)

We improve and adapt the standard HSR image differencing model to apply it effectively for CD in remote sensing as Sibling Regression for Optical CD (SiROC). Algorithm 1 outlines SiROC in pseudocode. In summary, there are two major differences between SiROC and HSR image differencing. First, we redesign the notion of included and excluded pixels in the neighborhood selection to create an ensemble version of HSR over mutually exclusive neighborhoods. This does not only improve performance but also allows us to obtain an uncertainty along with the prediction. The rationale of splitting neighborhoods by distance is to inspect trends at different distances separately instead of pooling the trends together. For example, two trends at different spatial scales might offset each other when pooling them although both may be a signal for change. Second, we combine HSR with morphological profiles to move from pixel- to object-level

changes since changes in remote sensing typically occur at the object level.

1) *Ensembling*: The starting point for SiROC is applying HSR to  $\mathbf{I}_t$  to obtain  $\hat{\mathbf{I}}_{t+1}$  based on a set of neighboring pixels. We use all pixels that have a distance of at least  $e$  but at most  $n$  rows or columns from the pixel of interest. Graphically, this corresponds to all points in a square with width  $2n$  and  $I_{x,y,t}$  in its center, which are not in the smaller square with width  $2e$  around  $I_{x,y,t}$ . Formally, a pixel  $I_{x',y',t}$  is included in the set of neighbors for  $I_{x,y,t}$  if

$$e < \max\{|x' - x|, |y' - y|\} \leq n. \quad (8)$$

With  $\hat{\mathbf{I}}_{t+1}$ , a channel-level difference image is obtained by taking the difference  $\hat{\mathbf{I}}_{t+1} - \mathbf{I}_{t+1}$ . The absolute value of the change signal is summed across the channels. We apply Otsu-thresholding [29] to the resulting difference image that has been successfully used for thresholding difference images in CD before [30]. Furthermore, the evaluation of competing methods is also based on this thresholding approach. This allows for comparing relevant methods in a fixed setting. Nevertheless, Otsu-thresholding is a design choice here with a variety of alternatives that can also be used in conjunction with SiROC, including the T-point method [48], the Rosin method [49], or the expectation-maximization (EM) algorithm [50], [51].

The result of the thresholding step is a binary segmentation of the difference image on the pixel level. However, in remote sensing applications, changes such as the construction of roads or buildings tend to occur at the object level. This is why object-based methods often tend to be superior for these applications [52]. We rely on morphological profiles that are an established tool to bridge the gap between pixel-level change segmentations and the object level [53].

2) *Morphological Profile*: A morphological profile is the sequential application of morphological opening and closing to an image [53]. We employ morphological opening and closing at one spatial filter size  $p$ . Intuitively, morphological closing helps to fill in missed pixels in detected change objects as changed. On the other hand, morphological opening removes spurious FPs when there are no other changes around them. After obtaining an object-level change segmentation for a given neighborhood size  $n$  and exclusion



window  $e$ , we repeat the procedure and use new neighbors that are further away than the current set. Both  $n$  and  $e$  increase by the same added factor  $s$ . In the next iteration, the previous neighborhood window becomes the exclusion window, and a new binary segmentation based on more distant neighbors is obtained. This procedure is repeated until a maximum neighborhood size is reached. The number of models  $F$  is given as  $F = ((n_{\max} - n_{\min})/s) + 1$ . Every model in the ensemble classifies each pixel either as change (1) or no change (0), which can be interpreted as a voting mechanism among models. Voting mechanisms across spatial scales [54] or different bands [26] are a common aggregation mechanism in CD. The number of votes per pixel ranges between 0 and  $F$ .

3) *Majority Voting*: The matrix of votes per pixel can be visualized as a heatmap of agreement between different sets of neighbors if a change occurred. This also directly transports a measure of uncertainty embedded in SiROC. If a pixel has no or the maximum number of votes, the agreement is high, and the method is confident in its prediction. If the number of votes is split, the model shows low confidence in its prediction for this point. We threshold these votes with a predefined voting share  $0 \leq v \leq 1$  that is required to classify a pixel as changed.  $v$  is the sensitivity of our model toward change. The choice of  $v$  contains a tradeoff between objectives. With a higher  $v$ , the number of false negatives rises but FPs decline (and vice versa). Since all models are equally weighted in the voting process, the importance of a single neighbor is decreasing in its distance to the pixel of interest. This is because the number of neighbors used per model is increasing in  $n$ . The underlying assumption is that pixels closer to the point of interest carry more information about its potential change. This assumption is domain-specific to Earth observation and stands in contrast to the idea of HSR in astronomy where there is no weighting based on distance. The application of the voting threshold is the last step of SiROC to obtain the final change segmentation. The voting matrix is normalized by the number of models before the percentage threshold is applied.

To summarize, SiROC has the following hyperparameters.

- 1) *Maximum Neighborhood Size*:  $n_{\max}$ .
- 2) *Initial Exclusion Window*:  $e_{\text{start}}$ .
- 3) *Step Size of Ensemble*:  $s$ .
- 4) *Filter Size of Morphological Operations*:  $p$ .
- 5) *Voting Threshold*:  $0 \leq v \leq 1$ .

The initial size of the neighborhood window  $n_{\text{start}}$  is given as  $e_{\text{start}} + s$ .

### III. EXPERIMENTS AND RESULTS

Section III-A describes the datasets used to assess the performance of SiROC and competing methods. The competing methods used as a benchmark and the evaluation criteria are described in more detail in Section III-B. The results on OSCD, BHED, the Agriculture Dataset, and the Alpine Dataset are presented in depth in Sections III-C–III-F, respectively.

#### A. Description of Datasets

1) *Onera Change Detection Dataset*: OSCD is a benchmark for bitemporal urban CD based on multispectral Sentinel-2

images [55]. It contains manual annotations of binary changes for 24 cities across the globe where 14 are used for training and 10 for testing. The labels focus on urban changes, such as newly constructed buildings, and natural changes, such as sea-level rise or differences in vegetation, are not annotated. The two images per city are selected to be cloud-free and are generally taken about one to three years apart. While there are 13 bands available in Sentinel-2 images, we restrict our focus to the RGB channels here. Although SiROC is able to handle channels outside of the visible spectrum as well, our experiments show that the inclusion of the NIR band does not add value in the urban applications considered here. This may be different in vegetation monitoring where NIR bands tend to be more insightful. Spatial bands beyond RGB and NIR do not have a spatial resolution of 10 m and are, therefore, excluded as well.

2) *Beirut Harbor Explosion Dataset*: On August 4, 2020, a devastating explosion of large amounts of ammonium nitrate occurred in the port of Beirut in Lebanon. It led to over 200 deaths and left more than 300000 people homeless because of heavy damages to buildings in the city.<sup>1</sup> We collect a pair of cloud-free Planetscope images with 3 m per pixel resolution on August 1 and 5 before and after the explosion. We combine these images with ground truth on destroyed buildings provided by the Center for Satellite Based Crisis Information (ZKI), German Aerospace Center.<sup>2</sup> The building destruction map is based on manual annotation of very-high resolution images and field reports on the ground. Note that the annotations contain building destruction rather than building damage. Therefore, partial damages to buildings that withstood the explosion are not included. With this dataset, we aim to test the applicability of SiROC not only in medium but also in higher resolution images in problems where fast and accurate annotations are essential.

3) *Agriculture Dataset*: To test SiROC also outside the urban domain, we include two other test datasets from Saha *et al.* [12] as reference points. The first one, the Agricultural dataset, is a scene with bitemporal Sentinel-2 images from July 2015 over Barrax, Spain, with  $600 \times 600$  pixels in size. Between the two images is a time period of 10 days between which agricultural field activity changed notably. The reference map was manually annotated by Saha *et al.* [12].

4) *Alpine Dataset*: The second dataset consists of pre and post Sentinel-2 images of a fire in an alpine region close to Trento, Italy, in spring 2019. A variety of other seasonal vegetation trends, such as ice and snow, complicate this dataset. The scene has a size of  $350 \times 350$  pixels with ground truth annotated manually by Saha *et al.* [12].

#### B. Competing Methods and Criteria

We compare our results to a variety of state-of-the-art unsupervised methods for optical CD in remote sensing. Since SiROC needs no training and does not rely on pretrained neural networks, its primary group of comparison consists of other image differencing-based methods. This makes SiROC

<sup>1</sup>[https://en.wikipedia.org/wiki/2020\\_Beirut\\_explosion](https://en.wikipedia.org/wiki/2020_Beirut_explosion)

<sup>2</sup><https://activations.zki.dlr.de/en/activations/items/ACT148.html>

fast compared to deep learning methods with comparable speed to traditional methods. We include several frameworks that improve on classical CVA. RCVA [30] incorporates close neighborhood information to make CVA more robust against misregistration. PCVA [36] uses CVA of multilevel parcels to improve on CVA. DCVA is based on deep feature extraction with a deep neural network pretrained on imagenet [12]. While DCVA was originally developed for high-resolution images, it is a resolution-agnostic framework relying on deep feature extraction from RGB channels. We also include a version of this method that we call DCVAMR specifically adjusted for medium-resolution, multispectral Sentinel-2 imagery [26] for the OSCD dataset. For BHED, we include DCVA, RCVA, and PCVA as baselines as DCVAMR is not capable of handling Planetscope input channels. The most recent advancement in unsupervised CD for high-resolution imagery is Saha *et al.* [19] who employ self-supervised pretraining on remote sensing images in combination with a DCVA framework. We call this refined version of DCVA “SSDCVA” and include it as the primary baseline besides general DCVA for BHED.

In line with previous evaluations on OSCD [26], [55], we analyze the performance of SiROC against the state-of-the-art binary change segmentation based on specificity and sensitivity. Specificity is defined as the number of true positives (TPs) over the sum of TPs and FPs:  $\text{specificity} = \text{TN}/(\text{TN} + \text{FP})$ . Sensitivity is the number of TP over the sum of TP and false negatives:  $\text{sensitivity} = \text{TP}/(\text{TP} + \text{FN})$ . This criterion is also known as recall. A method that is sensitive toward changes has a high sensitivity but a low specificity (and vice versa). A superior method should balance these objectives and evaluate better in both criteria. To further elaborate on the balance of change and no change class, we also report precision =  $\text{TP}/(\text{TP} + \text{FP})$  and F1-score =  $(2 * \text{Precision} * \text{Recall})/(\text{Precision} + \text{Recall})$ .

### C. Results on OSCD

1) *Parameters:* We tune the parameters of SiROC on the OSCD training set resulting in the following parameter specifications.

- 1) *Maximum Neighborhood Size:*  $n_{\text{max}} = 200$ .
- 2) *Initial Exclusion Window:*  $e_{\text{start}} = 0$ .
- 3) *Step Size of Ensemble:*  $s = 8$ .
- 4) *Filter Size of Morphological Operations:*  $p = 5$ .

The maximum neighborhood size is 200 with stepsize 8. Contrary to the original idea of HSR in astronomy, we do not find it to be optimal to exclude direct neighbors of the pixel of interest from the analysis resulting in an initial exclusion window of zero. While neighboring pixels may be subject to the same kind of object-level change on the ground, they still can contribute important information if their weight is moderate. We find the best results with morphological opening and closing with a filter size of 5. We do not tune the voting threshold because this parameter does not influence the change signal directly but rather how the method balances FPs and false negatives.

TABLE I  
QUANTITATIVE RESULTS’ OSCD TEST SET

	Specificity	Sensitivity	Precision	F1
SiROC	<b>88.31%</b>	<b>70.71%</b>	<b>24.80%</b>	<b>36.72%</b>
DCVAMR	78.38%	64.63%	14.01%	23.03%
DCVA	76.96%	69.02%	14.03%	23.33%
PCVA	75.61%	47.00%	9.50%	15.81%
RCVA	76.96%	64.08%	13.16%	21.84%
Ablation Scores				
No MP	80.64%	69.88%	16.44%	26.62%
HSR	79.45%	70.24%	15.70%	25.66%

2) *Quantitative Results:* Table I reports specificity and sensitivity scores of SiROC and competing methods on the OSCD test set. Scores are averaged on the city level. SiROC with a voting threshold of  $v = 1/2$  achieves a specificity of 88.31% with a sensitivity of 70.71% and 24.80% precision and 36.72% F1-score. This is a high score in all four categories by a significant margin. The difference to DCVAMR is about 6–13 percentage points (p.p.) depending on the category. DCVA achieves a sensitivity that is slightly below but close to SiROC but lacks behind in specificity, sensitivity, and F1-score by more than 10 p.p. Compared to the best results of methods without deep-learning-based feature extraction, SiROC gains about 12 p.p. in specificity, 7 p.p. in sensitivity, 12 p.p. in precision, and 15 p.p. in F1 on RCVA.

To understand the origin of the performance difference to the previous state of the art in more detail, we provide two ablation scores of SiROC. First, we remove the morphological operations in SiROC. While morphological profiles help to transition to an object-level change mask, SiROC still exceeds previous unsupervised performance without them. No MP performance improves by 2 p.p. in specificity and 5 p.p. in sensitivity versus DCVAMR and by 4 p.p. in specificity and 1 p.p. in sensitivity versus DCVA. The resulting F1-score is 3–4 p.p. higher than deep-learning-based methods and 5–10 p.p. higher than traditional methods here. To evaluate the effectiveness of ensembling, we also provide a score for a vanilla HSR with the same neighborhood size and no exclusion window. The Vanilla HSR performs slightly better but in the range of DCVA and DCVAMR with a specificity of 79.45% and a sensitivity of 70.24%. The F1-score is about 1 p.p. lower without ensemble voting.

Therefore, the majority voting mechanism is an effective tool to extract a more granular signal from the general HSR predictions. Furthermore, the use of wide spatial context pioneered in astronomy is advantageous for CD in remote sensing as well. In summary of Table I, SiROC sets a new state-of-the-art unsupervised CD in medium-resolution images on OSCD. Even without morphological filters, the method still notably outperforms previous scores, which points to a strong signal for change information in the original HSR method and the effectiveness of the majority voting mechanism. Combined with ensembling over different neighborhoods and morphological profiles, this exceeds previous quantitative results on the OSCD dataset.

3) *Qualitative Results*: The edge of SiROC compared to other unsupervised methods in the quality of change annotations for medium-resolution imagery is also visible when inspecting the predictions for specific scenes. Fig. 2 displays exemplary change masks for Las Vegas. For SiROC, a threshold of  $v = 1/2$  was used to obtain the images with a specificity of 95.28%, a sensitivity of 78.75%, and a precision of 58.14% for this scene. Fig. 2(a) visualizes the confidence of SiROC in the change propensity of a pixel as a heatmap from dark purple (0% votes) to yellow (100% votes). When comparing this to the ground truth on the bottom right, one can see that change confidence is strongly associated with the occurrence of a change. Fig. 2(b) shows the binary change map after applying the threshold to the uncertainty map. Not only does SiROC pick up on the changed areas in the image but it also fits the shapes of changing buildings fairly well. The visual similarities between Fig. 2(b) and (h) are striking, especially compared to the other segmentations of competing methods. Also, before applying the morphological operations, SiROC identifies the areas of interest in the image well although the predicted mask is naturally slightly more spurious. The morphological operations help to remove these spurious changes, but the change signal in the predictions is in line with the ground truth [see Fig. 2(c)]. DCVAMR is generally able to discover the changing regions of an image but struggles to identify the shapes of changing objects and rather fits round blobs [see Fig. 2(d)]. DCVA tends to discover large changes and overestimate their size, whereas smaller changes go undetected [see Fig. 2(e)]. This might be related to the fact that DCVA was originally designed for high-resolution optical imagery in which building changes are larger in terms of pixel size. This is in line with the fact that DCVAMR, which is explicitly adjusted for Sentinel 2, tends to fit the size of changes better even though it also struggles with change shapes. PCVA and RCVA seem to extract building footprints rather than building changes here, which leads to overcrowding of the segmentation mask.

A similar picture emerges when inspecting results for Dubai in Fig. 3, which is a slightly more complex scene since the shapes of changes differ widely. SiROC detects changing regions again well but seems to struggle with the shape of changes in the upper part of the image. The newly constructed road is identified well. Consequently, the quantitative scores on this scene are slightly lower compared to the Las Vegas Scene with 86.87 % specificity, 76.61% sensitivity, and 39.14% precision. The struggles of the competing methods are similar to the Las Vegas Scene: DCVAMR fits round shapes to any kind of change [see Fig. 3(d)], DCVA overestimates the size of large changes [see Fig. 3(e)], and PCVA extracts a spurious change map that rather looks, such as building footprints [see Fig. 3(f)]. Therefore, the quality inspection of visual results confirms that SiROC obtains superior results on OSCD.

4) *Uncertainty Estimation*: To properly analyze if the confidence of SiROC also corresponds to well-calibrated uncertainties, we test this with calibration curves. For this, we split pixels into subsets based on the SiROC confidence and analyze the respective performance for a level of confidence. If the performance of SiROC is in principle increasing with the

TABLE II  
SENSITIVITY TO HYPERPARAMETERS (OSCD TRAINING SET)

	Specificity	Sensitivity	Range	# Evaluations
Selected	89.48%	67.90%		
$N_{\max}$	87.96%	70.53%	[30,250]	20
$e_{\text{start}}$	89.37%	68.59%	[0,150]	20
$s$	88.89%	68.06%	[1,5]	5
$p$	91.81%	59.06%	[2,5]	4
Joint	90.34%	60.12%	All above	75

Average Specificity and Sensitivity on the OSCD training set when varying SiROC hyperparameters. The average scores are compared to the selected optimal selection of parameters and their training performance in the first line. For single parameters (rows 2-5), we vary only the mentioned parameter on the grid given in the column range and leave the others at the selected optimum. The column Evaluations gives the number of runs that were executed to obtain the average scores. Finally, in the last row "Joint" we vary all parameters on the given intervals in the rows above simultaneously.

TABLE III  
THRESHOLDING CHOICE (OSCD TRAINING SET)

	Specificity	Sensitivity
Otsu	89.48%	67.90%
EM	87.96%	74.22%
Triangle	88.98%	71.41%

Results on the OSCD training set with different binary thresholding techniques. We use Otsu thresholding to ensure comparability to previous results but alternatives such as expectation-maximization based thresholding give similar results.

confidence, the uncertainty levels, in fact, correspond to the certainty of the prediction that the model has. Fig. 4 plots these confidence–performance curves for four cities in the OSCD test set. For all four cities, we see that model precision is nondecreasing in the confidence of the SiROC. Most of the time, the prediction increases notably in the confidence, which means that SiROC not only performs well for this task but also returns well-calibrated uncertainties as part of its prediction.

5) *Sensitivity to Hyperparameters*: To allow effective use of SiROC in practice, we offer a sensitivity analysis of the hyperparameter choice on OSCD along with recommendations for this choice in other applications. This sensitivity analysis is executed on the training set to avoid multiple evaluations on the test set. The results are shown in Table II. While the performance of the method naturally varies with the choice of hyperparameters, SiROC looks fairly robust against its hyperparameter choices. The first row gives the training set performance based on the selected parameters described in this section as a comparison point. Varying only the maximum neighborhood  $N_{\max}$ , the number of rows excluded  $e_{\text{start}}$  and the stepsize  $s$  at the selected parameter specification influences the training performance marginally, at most. For all three parameters, the average specificity decreases, while average sensitivity increases slightly. These three parameters essentially navigate how to group and prioritize neighborhoods. Excluding close context ( $e_{\text{start}}$ ), including more distant context ( $N_{\max}$ ), and aggregating neighborhoods into larger groups ( $s$ ), therefore, do not seem to matter notably in practice to achieve good performance.



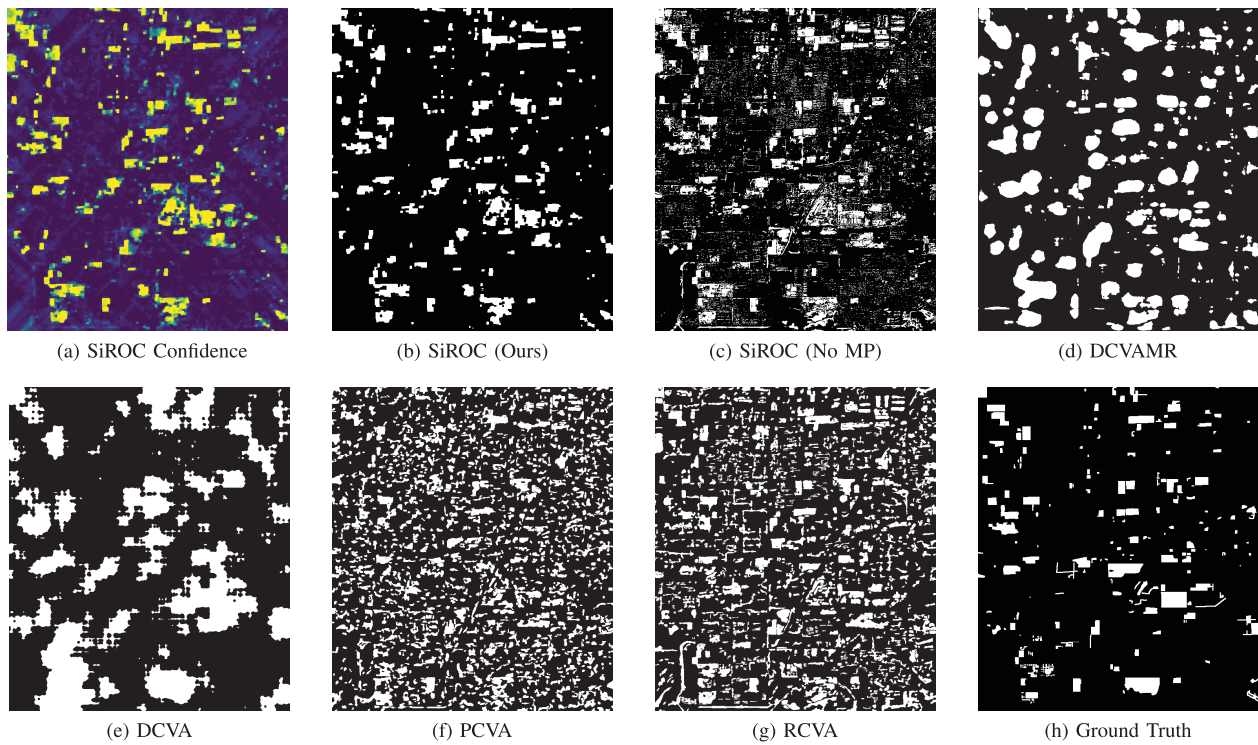


Fig. 2. Qualitative Comparison OSCD—Las Vegas. This figure visualizes the number of change votes per pixel in SiROC (a) and the corresponding binary predictions after (b) and before morphological operations (c). Competing models are visible in (d)–(g) and the ground truth in (h). SiROC predicts change regions and shapes of the ground truth well while competing methods struggle either with identifying the shapes visible in (d) and (e) or the areas of change in (f) and (g) for the Las Vegas Pair.

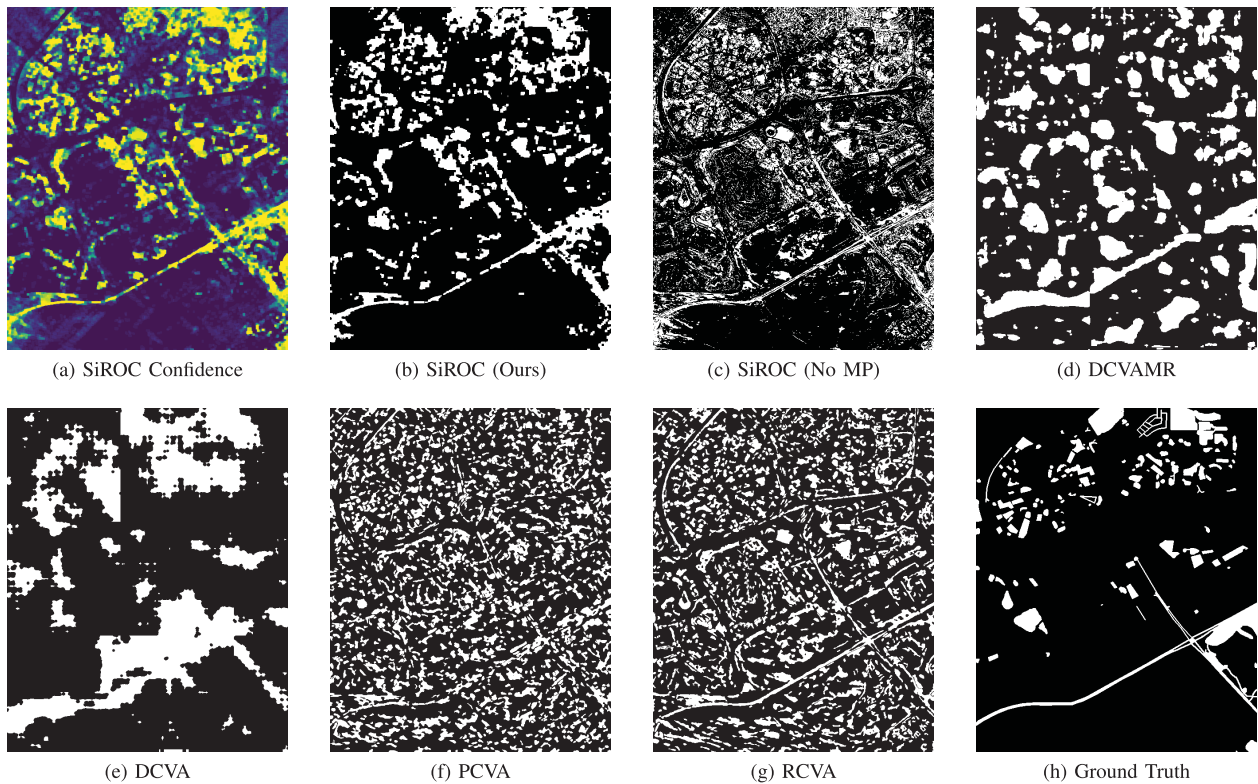


Fig. 3. Qualitative comparison OSCD—Dubai. The structure is identical to Fig. 2, but predictions and ground truth are presented for Dubai. Also, for this scene, SiROC predicts changing areas and their shape comparably well even. In contrast, competing methods miss the shapes of changing areas, such as the street in the lower part of the image or struggle to detect relevant regions.

The performance is slightly more sensitive toward the size of the morphological profile ( $p$ ) where average specificity increases marginally and sensitivity drops by 9 p.p. if this is

varied leaving other parameters untouched. Similarly, when varying all four parameters simultaneously in 75 random draws, performance drops with a difference of about 8 p.p.

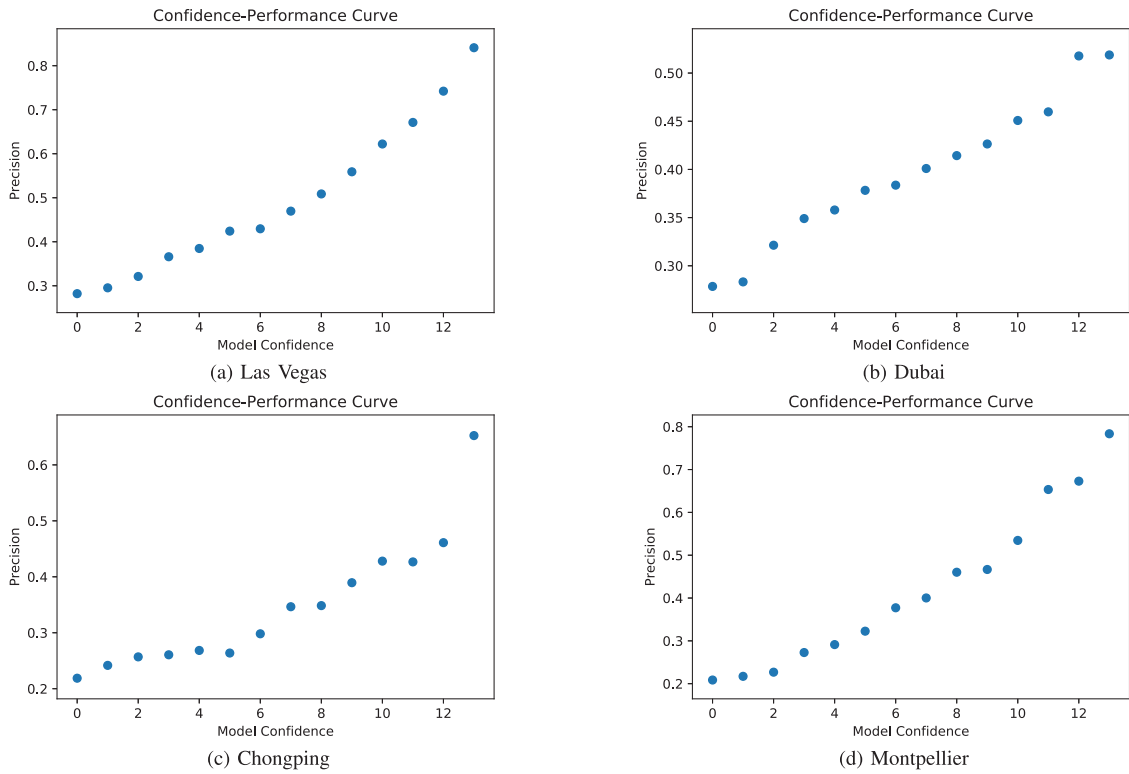


Fig. 4. Confidence–performance plots on four cities of the OSCD dataset. On the  $x$ -axis, points in the image are sorted into buckets by SiROC confidence. For each of these buckets, the performance is estimated separately. Since performance is generally nondecreasing in confidence, the uncertainty measure is considered well calibrated.

in sensitivity with similar specificity. In terms of magnitude, this performance drop is only a fraction of the difference between SiROC and its closest competitors on the OSCD test set. This implies that SiROC would likely outperform competing methods on this dataset for a variety of hyperparameter choices. To understand the sensitivity of the results to different thresholding techniques, we also benchmark SiROC based on the selected hyperparameters with EM-based thresholding [50], [51] and triangle thresholding following the OpenCV implementation.<sup>3</sup> Results are presented in Table III. While the results are similar, the different techniques balance the tradeoff between FPs and false negatives in a slightly different fashion. This may be relevant to consider for applications of SiROC in practice where this balance plays an important role.

For potential applications of SiROC in the future, we suggest using the obtained parameter combination initially. This provides a starting point for further analysis in different contexts. Since the performance seems to be comparably susceptible to the size of the morphological profile, this parameter may deserve special attention during tuning. In the following, results on the remaining three datasets are obtained with this parameter combination, which was the result of tuning on OSCD. Even though this may not necessarily give the best possible performance, we aim to validate that SiROC achieves convincing results in other applications without fine-tuning on single scenes.

<sup>3</sup>[https://docs.opencv.org/4.5.3/d7/d1b/group\\_\\_imgproc\\_\\_misc.html](https://docs.opencv.org/4.5.3/d7/d1b/group__imgproc__misc.html)

TABLE IV  
QUANTITATIVE RESULTS BEIRUT EXPLOSION

	Specificity	Sensitivity	Precision	F1
SiROC	<b>92.01%</b>	<b>83.38%</b>	<b>19.89%</b>	<b>32.12%</b>
DCVA	91.87%	79.85%	11.37%	19.93%
SSDCVA	88.25%	81.08%	8.80%	15.95%
PCVA	88.61%	58.56%	6.74%	12.10%
RCVA	86.56%	66.71%	6.52%	11.89%
Ablation Scores				
SiROC (p=10)	<b>92.34%</b>	<b>91.89%</b>	<b>22.20%</b>	<b>35.76%</b>
no MP	88.02%	79.67%	13.65%	23.30%
HSR	86.65%	71.63%	11.31%	19.54%

#### D. Results on BHED

1) *Quantitative Results:* Table IV displays specificity, sensitivity, precision, and F1-scores on the scene. Generally, scores on BHED are higher than on OSCD since the changes are centered around the same area and have similar shapes. SiROC with default parameters achieves a specificity of 92.01% and a sensitivity of 83.38%. DCVA achieves a similar specificity with 91.87% but falls short in terms of sensitivity by about 4 p.p. with a score of 79.85%. SSDCVA places slightly below DCVA with a specificity of 88.25% and a sensitivity of 81.08%. SiROC beats SSDCVA by about 3 p.p. in sensitivity and about 2 p.p. in specificity. PCVA and RCVA clearly fall behind SiROC and also DCVA-based methods. F1-score and precision results confirm the previous impressions with a gap of 12–20 p.p. in F1 and 9–13 p.p. in precision, respectively.

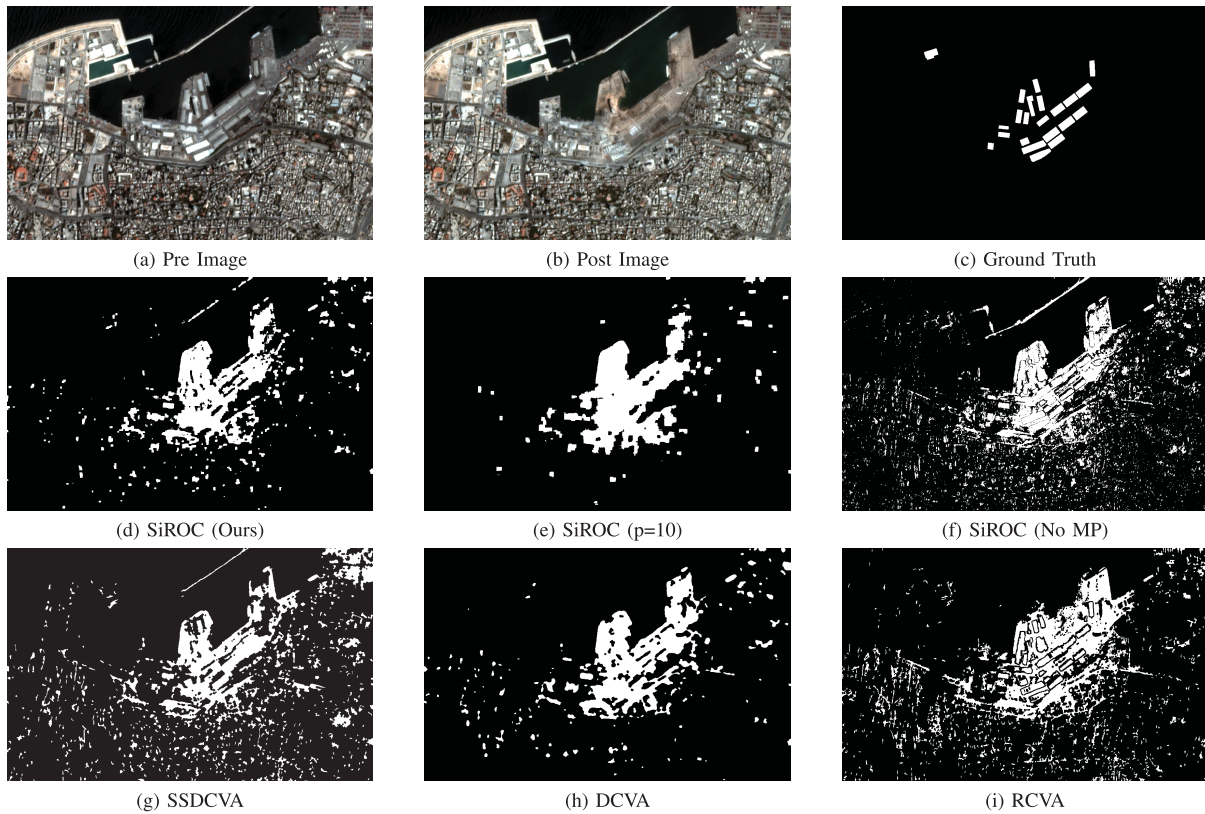


Fig. 5. Qualitative comparison Beirut explosion. This figure shows the PlanetScope image pairs (a) and (b), change ground truth (c), and model predictions (d)–(i) for the Beirut explosion scene. SiROC with main parameters (d) identifies the area of destroyed buildings around the epicenter correctly with few FPs although the shapes are lacking some granularity. Increasing the size of morphological operations improves accuracy but tends to fit one large blob with missing building shapes (e). Excluding morphological operations increases FPs although the main changes in the center are still identified well (f). Competing methods struggle not only with the shape of change but also detect a number of FPs far away from the explosion (g)–(i).

When we adjust the scale of morphological operations to 10, SiROC performs significantly better, which suggests that there may be notable tuning potential for higher resolution inputs. Still, the baseline parameters perform well on this scene. Therefore, SiROC demonstrates its usefulness beyond medium-resolution images and can also be used in conjunction with high-resolution images for CD.

The other ablation scores again point toward the most important steps within SiROC to achieve this performance. Without morphological profiles, the scores of SiROC drop about 4–9 p.p. in all four categories. Still, it achieves slightly superior precision and F1-scores but falls short to DCVA with a difference of about 3 p.p. in specificity and similar sensitivity. This is a notable difference to medium-resolution imagery on OSCD where the exclusion of morphological filters decreased the performance of SiROC, but it was still superior to DCVA-based methods. This is not necessarily surprising since deep-learning-based methods tend to relatively improve their CD performance compared to traditional methods with increasing spatial resolution. Without majority voting over different neighborhoods, the Vanilla HSR version performs better but in the range of RCVA and PCVA. Again, it is the combination of HSR, ensembling over different neighborhoods and transitioning to the object level with morphological operations that all contribute significantly to the overall performance of SiROC.

2) *Qualitative Results:* Fig. 5 shows visual comparisons of the discussed methods on BHED. The first row of images presents the preexplosion image [see Fig. 5(a)], the post image [see Fig. 5(b)], and the ground truth [see Fig. 5(c)]. The heart of the explosion in the port can be found in the middle of the image with almost the entirety of buildings completely destroyed around it. Fig. 5(d) presents the binary SiROC segmentation with baseline parameters obtained on OSCD. While SiROC is missing some granularity in its segmentation of destroyed building footprints, the changing areas are well identified with few FPs outside of the port. For a larger morphological filter size ( $p$ ), the main area is identified more densely with better quantitative results, but the shapes of buildings vanish [see Fig. 5(e)]. Without morphological operations, the core change is still well-segmented although the number of FPs in the outer regions of the image increases [see Fig. 5(f)]. SSDCVA shows similar tendencies to summarize the port area as one large change with a number of spurious FPs [see Fig. 5(g)]. DCVA shows fewer salt and pepper noise than SSDCVA here and generally segments the exploded buildings similar to SiROC, however, with a slightly more perforated shape [see Fig. 5(h)]. The segmentation by RCVA is not really competitive here since the maps are spurious and changes are not well identified [see Fig. 5(i)]. Results for PCVA are similar to RCVA and, hence, omitted.



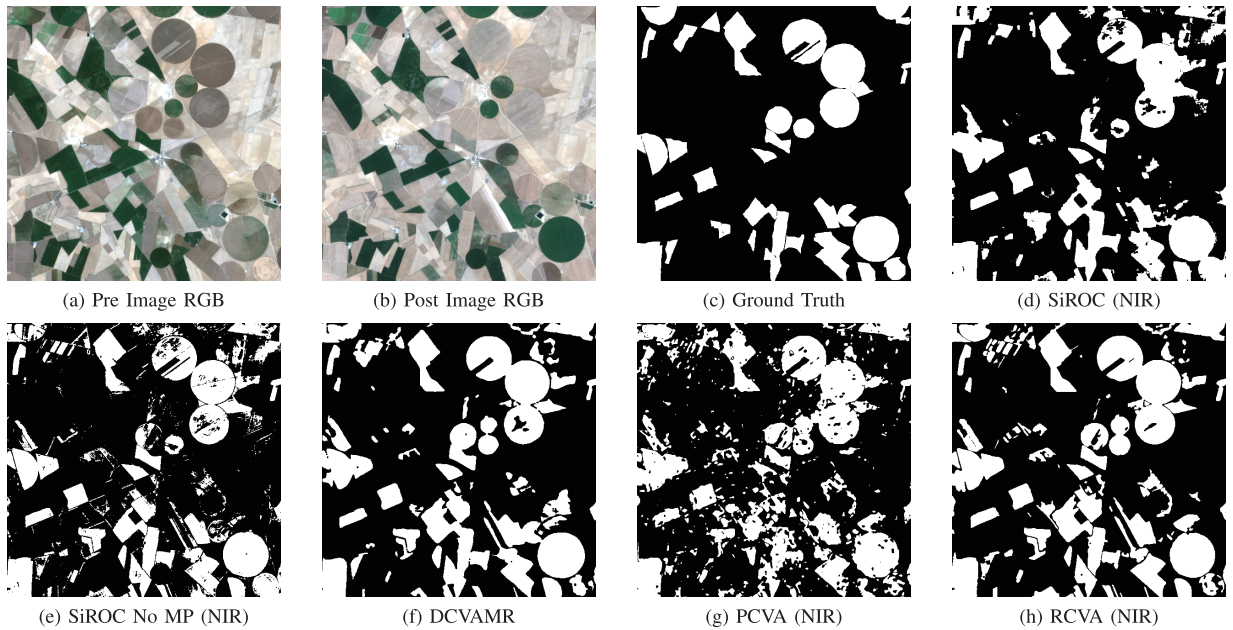


Fig. 6. Qualitative results agriculture dataset. Pre- (a) and post -GB (b) image with changing agricultural fields. The ground truth (c) shows high similarity to SiROC (d) but also to DCVA (f) and RCVA (h) while PCVA (g) has some FP areas and spurious change predictions. Change segmentations without morphological profiles (e) for SiROC still works well which is line with the quantitative results of Table V.

### E. Results on Agriculture Dataset

1) *Quantitative Results:* Table V displays the results of SiROC and competing methods on the agriculture scene. SiROC is applied to the dataset with the parameters obtained on Onera without further adjustment. Hence, the results that we provide are a validation exercise in the different context of nonvisible parts of the spectrum without parameter fine-tuning.

To be consistent with previous evaluations on this dataset [12], we compare SiROC with PCVA and RCVA based on vegetation (VEG) and near-infrared (NIR) channels of Sentinel-2 as inputs. The score for DCVAMR is based on the full Sentinel-2 input images as the method was deliberately designed to incorporate all channels.

While SiROC achieves the top score in terms of specificity and precision with 90.81% and 74.23%, respectively, it falls short of DCVAMR on sensitivity (88.70% versus 94.26%) and F1-score (80.85% versus 81.47%). DCVAMR seems to lean slightly more toward the change class, whereas SiROC rather classifies a pixel as no change in unclear cases. SiROC is superior to PCVA and comparable to RCVA in performance for both VEG and NIR channels as inputs.

The ablation scores underline that morphological profiles still help although the effects are smaller than in urban applications with an average difference in about 1–2 p.p. in all four criteria. Furthermore, excluding the majority voting mechanism does not hurt performance but actually improves it slightly here. The vanilla HSR performs slightly worse but in the range of RCVA and better than PCVA on its own. Smaller benefits of including majority voting and morphological profiles could be linked to the fact that parameters for these operations were tuned in an urban RGB context. Although already quite effective, the accuracy of SiROC could likely be further improved with parameter fine-tuning.

TABLE V  
QUANTITATIVE RESULTS AGRICULTURE DATASET

	Specificity	Sensitivity	Precision	F1
SiROC (VEG)	90.69%	86.38%	73.53%	79.44%
SiROC (NIR)	<b>90.81%</b>	<b>88.70%</b>	<b>74.28%</b>	<b>80.85%</b>
DCVAMR	88.88%	<b>94.26%</b>	71.73%	<b>81.47%</b>
PCVA (VEG)	88.83%	83.18%	69.04%	75.45%
PCVA (NIR)	86.60%	84.56%	65.38%	73.74%
RCVA (VEG)	88.91%	91.95%	71.28%	80.31%
RCVA (NIR)	87.39%	92.36%	68.67%	78.77%
Ablation Scores				
No MP (VEG)	89.87%	84.66%	71.44%	77.49%
No MP (NIR)	89.70%	87.15%	71.70%	78.67%
HSR (VEG)	89.96%	87.71%	72.35%	79.29%
HSR (NIR)	89.29%	90.64%	71.70%	80.06%

2) *Qualitative Results:* Fig. 6 presents pre and post RGB images [see Fig. 6(a) and (b)], the ground truth [see Fig. 6(c)], and change predictions [see Fig. 6(d)–(h)]. The visual impression of change predictions confirms the quantitative results. Predictions are fairly accurate on this scene, which suggests a comparably easy task relative to the more complex OSCD scenes. SiROC segments changing regions well and struggles with the varying field shapes only in rare instances. Similarly, the results of DCVAMR and RCVA are also fairly accurate with a slightly higher tendency to predict the change class. In comparison, the mask by PCVA produces some FP regions. Overall, SiROC shows similar performance to highly effective methods also in the agriculture domain.

### F. Results on Alpine Dataset

1) *Quantitative Results:* Results for the Alpine dataset can be found in Table VI. Even though SiROC reaches the highest specificity, it does not quite pass the overall performance of

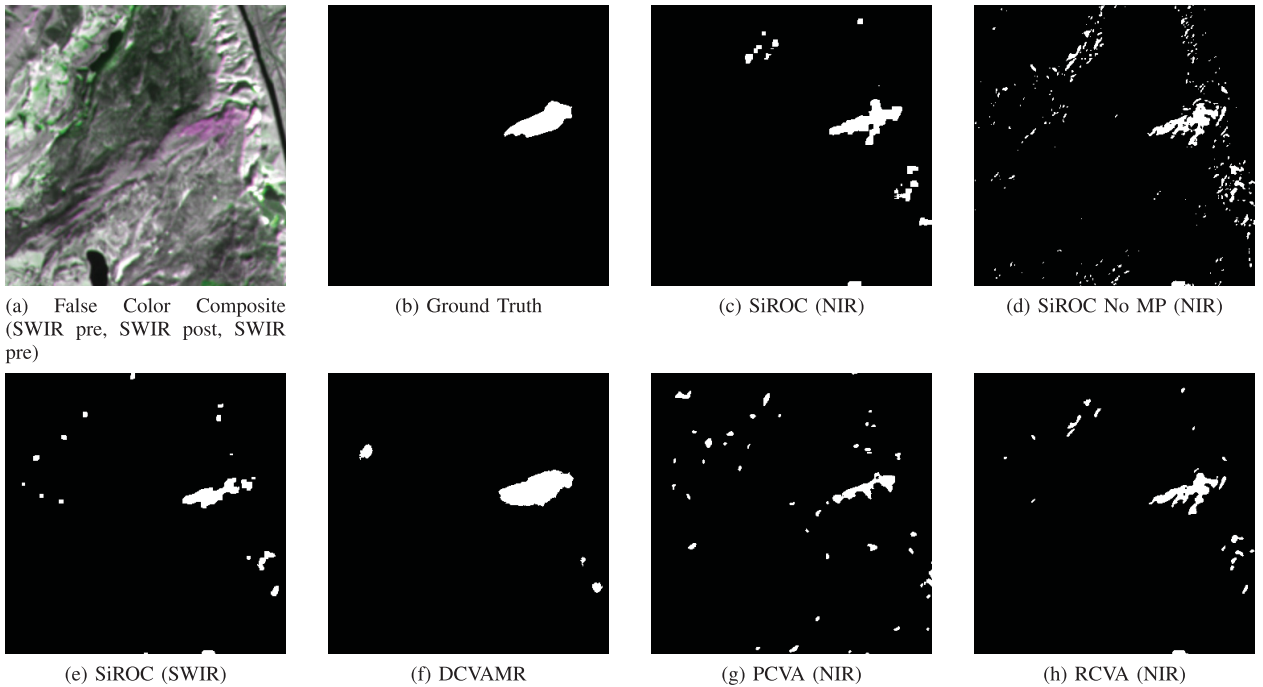


Fig. 7. Qualitative results Alpine dataset. The area of the fire can be seen in purple in the false color composite in (a) with the reference map in (b). SiROC (NIR) (c) and SiROC (SWIR) (e) both identify the changing area well although the shape is better approximated with NIR inputs. Without morphological profiles SiROC picks up more FPs (d). DCVAMR shows the most convincing results here (f). RCVA is roughly comparable (h) to SiROC while PCVA falls behind slightly (g).

DCVAMR from [12] on this dataset. Nevertheless, SiROC ranks highly also in sensitivity, precision, and F1-score, particularly based on NIR inputs with total scores of 98.92%, 75.71%, 52.28%, and 61.85%. RCVA with NIR inputs is comparable in performance, but SiROC is the only method that makes effective use of SWIR inputs compared to PCVA and RCVA.

The ablation scores underline the effectiveness of morphological transformations with about a 20 pp. drop in F1-score compared to SiROC for both NIR and SWIR. Removing the ensembling leads to a notable drop in F1-scores, particularly with NIR inputs.

2) *Qualitative Results*: Fig. 7 plots prediction masks for selected models for the Alpine dataset. In Fig. 7(a), the false color composite shows the annotated area of change affected [see Fig. 7(b)] by a fire in purple on the right. SiROC identifies this well although it is tempted to also classify a small number of FPs as change. While it is hard to control for seasonality in a bitemporal setting, SiROC (NIR) [see Fig. 7(c)] still excludes most other vegetation updates that are not the result of actual change here. The morphological profiles help on this scene to exclude spurious predictions [see Fig. 7(d)]. Compared to SiROC (SWIR) [see Fig. 7(e)], SiROC (NIR) segments the changing area slightly better although the shape is identified more clearly by DCVAMR [see Fig. 7(f)]. PCVA (NIR) [see Fig. 7(g)] seems to struggle slightly more with the shape of the burned area, whereas the results of RCVA (NIR) [see Fig. 7(h)] look similar to the results of SiROC (NIR), which is in line with the quantitative scores of Table VI.

TABLE VI  
QUANTITATIVE RESULTS ALPINE DATASET

	Specificity	Sensitivity	Precision	F1
SiROC (NIR)	98.92%	75.71%	52.28%	61.85%
SiROC (SWIR)	<b>99.28%</b>	59.51%	56.10%	57.76%
DCVAMR	99.06%	<b>94.99%</b>	<b>61.23%</b>	<b>74.46%</b>
PCVA (NIR)	98.95%	46.99%	41.04%	43.82%
PCVA (SWIR)	95.48%	35.80%	10.98%	16.80%
RCVA (NIR)	99.22%	63.99%	56.20%	59.84%
RCVA (SWIR)	86.56%	66.71%	6.52%	11.89%
Ablation Scores				
No MP (NIR)	97.81%	65.58%	31.74%	42.78%
No MP (SWIR)	97.39%	55.51%	24.87%	34.36%
HSR (NIR)	95.32%	82.10%	21.45%	34.01%
HSR (SWIR)	96.32%	66.12%	21.87%	32.87%

#### IV. DISCUSSION

SiROC is an effective method for CD in medium- and high-resolution optical imageries, which achieves competitive performance on four datasets. In the following, we elaborate on the intuition of SiROC's performance. When contrasting SiROC to image differencing methods, SiROC can be interpreted as an improvement over standard image differencing techniques because it does not assume the same changes in the acquisition conditions across time for the whole image. Rather, it allows for local changes in acquisition conditions. In standard CVA or RCVA, for example, an implicit assumption is that changes in the acquisition conditions across time affect each pixel similarly. SiROC releases this restriction and, instead, allows for local trends in regions of the image. If a



pixel deviates from the local trend around it, it is likely to undergo a change in SiROC. In RCVA or CVA, one would compare this pixel against trends in the whole image and not against its surrounding only. This might be unrealistic in complex scenes where pixels values highly depend on local trends in the surroundings. This is, for example, the case when a new building casts a shadow on a previously illuminated pixel. Similarly, a cloudy pixel in  $t + 1$  that was unobstructed in  $t$  might not necessarily be changing and is rather influenced by the local trend of a cloud rather than general image trends if large parts of the image are not obstructed by clouds. Hence, SiROC allows for a more granular analysis of deviations from trends in an image time series because, compared to previous methods, it makes full use of multitemporal information in close and distant neighbors. Although we compare our results to deep-learning-based methods, our intention is rather to augment these models than replace them, especially with high-resolution images. SiROC provides an efficient and accurate way to obtain change labels that could also be infused into deep learning models. One application of SiROC could be in self-supervised learning where pseudolabels are often obtained based on traditional image differencing techniques, such as CVA [56]. SiROC is not only superior in performance compared to image differencing. It also comes with a built-in, well-calibrated uncertainty of predictions. This could be especially beneficial in self-supervised settings since it automatically allows discriminating pseudolabels by confidence. For example, one could train only based on pseudolabels with high certainty and discard uncertain data points. Similarly, in some unsupervised methods, such as MSDRL for VHR imagery, an initial pseudoclassification is separated by confidence where high confidence examples are used for training a classifier that, subsequently, obtains predictions for leftover uncertain pixels [20]. In these methods, SiROC could also be used to obtain initial predictions and uncertainties to potentially improve not only the initial classification but maybe also the uncertainty categorization. The combination of deep-learning-based methods and SiROC may hence open up new potential for CD methods. While we restrict our focus to CD with optical images here, the framework of SiROC may be extended for applications on other multitemporal CD problems in remote sensing as well.

## V. CONCLUSION

We present SiROC, an efficient and accurate unsupervised method for CD in medium- and high-resolution optical images. SiROC is inspired by HSR that is used for exoplanet search in astronomy. It models a pixel of interest in  $t$  as a linear combination of its neighbors and applies this model to  $t + 1$  to obtain a prediction for the pixel based on its neighbors. The difference of the prediction for  $t + 1$  and the actual pixel value in  $t + 1$  is interpreted as the change signal. If the prediction is far from the actual value, trends in the neighboring pixels divert from the difference in the pixel of interest over time, which is seen as an indicator for change on the ground.

We refine and extend HSR in two major ways to apply it to optical satellite images as SiROC. First, we iterate over

several, mutually exclusive neighborhoods and apply HSR with all of these neighborhoods as input to obtain a distribution of change predictions. We combine these predictions with majority voting, which improves performance significantly and also returns a heatmap of votes per pixel, which can be interpreted as a well-calibrated uncertainty. Second, we use morphological opening and closing at one spatial filter scale to transition from pixel- to object-level predictions.

The results of SiROC are validated on four datasets. For urban CD with medium-resolution images, we verify the effectiveness of our method on OSCD, which contains binary change annotations for 24 cities across the globe. SiROC sets a new state-of-the-art unsupervised CD on OSCD, which surpasses previous methods by 10 p.p. in terms of specificity, 2 p.p. in sensitivity, 11 p.p. in precision, and 13 p.p. in F1-score. We further validate the performance of SiROC on high-resolution images with a dataset on the Beirut Harbor Explosion (BHED). Also, in this dataset, SiROC surpasses the performance of competing methods and underlines its abilities to segment urban change accurately at several scales. Furthermore, we provide two validation exercises on nonurban data with Sentinel-2 inputs. SiROC segments the effects of a fire in the Italian Alps accurately and in the range of competing methods. On the Agriculture dataset, SiROC falls short of DCVAMR in overall scores but still identifies the changing crop activity correctly.

While SiROC compares well against current deep-learning-based unsupervised methods in CD, SiROC should rather be seen as a complement than a substitute to these methods. Since it provides an accurate way to predict change signals with a built-in, well-calibrated uncertainty, it may be especially useful in conjunction with deep-learning-based methods to generate pseudolabels. Although we apply SiROC primarily to changes with multispectral data, the model may be applicable to other CD problems as well, which we plan to explore in future research.

## ACKNOWLEDGMENT

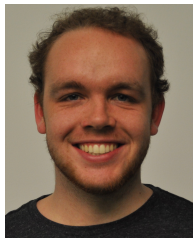
The authors are grateful to the Center for Satellite Based Crisis Information (ZKI), German Aerospace Center, for providing the ground truth of the Beirut Explosion scene.

## REFERENCES

- [1] D. Lu, E. Moran, and S. Hetrick, "Detection of impervious surface change with multitemporal Landsat images in an urban-rural frontier," *ISPRS J. Photogram. Remote Sens.*, vol. 66, no. 3, pp. 298–306, May 2011.
- [2] S. Ji, Y. Shen, M. Lu, and Y. Zhang, "Building instance change detection from large-scale aerial images using convolutional neural networks and simulated samples," *Remote Sens.*, vol. 11, no. 11, p. 1343, Jun. 2019.
- [3] G. Chen and G. J. Hay, "An airborne lidar sampling strategy to model forest canopy height from quickbird imagery and GEOBIA," *Remote Sens. Environ.*, vol. 115, no. 6, pp. 1532–1542, Jun. 2011.
- [4] Y. Gao, F. Gao, J. Dong, and S. Wang, "Transferred deep learning for sea ice change detection from synthetic-aperture radar images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 10, pp. 1655–1659, Oct. 2019.
- [5] K. Rokni, A. Ahmad, K. Solaimani, and S. Hazini, "A new approach for surface water change detection: Integration of pixel level image fusion and image classification techniques," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 34, pp. 226–234, Feb. 2015. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0303243414001780>
- [6] R. Gupta *et al.*, "Creating xBD: A dataset for assessing building damage from satellite imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2019, pp. 10–17.

- [7] L. Moya *et al.*, "Detecting urban changes using phase correlation and  $\ell_1$ -based sparse model for early disaster response: A case study of the 2018 Sulawesi Indonesia earthquake-tsunami," *Remote Sens. Environ.*, vol. 242, Jun. 2020, Art. no. 111743.
- [8] M. Zanetti *et al.*, "A system for burned area detection on multi-spectral imagery," *IEEE Trans. Geosci. Remote Sens.*, early access, Sep. 17, 2021, doi: [10.1109/TGRS.2021.3110280](https://doi.org/10.1109/TGRS.2021.3110280).
- [9] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.
- [10] M. Drusch *et al.*, "Sentinel-2: ESA's optical high-resolution mission for GMES operational services," *Remote Sens. Environ.*, vol. 120, pp. 25–36, May 2012.
- [11] C. Kwan *et al.*, "Assessment of spatiotemporal fusion algorithms for planet and worldview images," *Sensors*, vol. 18, no. 4, p. 1051, Mar. 2018.
- [12] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised deep change vector analysis for multiple-change detection in VHR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3677–3693, Jun. 2019.
- [13] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep Siamese convolutional network for optical aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1845–1849, Oct. 2017.
- [14] H. Lyu, H. Lu, and L. Mou, "Learning a transferable change rule from a recurrent neural network for land cover change detection," *Remote Sens.*, vol. 8, no. 6, p. 506, 2016.
- [15] L. Mou, L. Bruzzone, and X. X. Zhu, "Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 924–935, Feb. 2019.
- [16] P. Zhang, M. Gong, L. Su, J. Liu, and Z. Li, "Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images," *Photogram. Remote Sens.*, vol. 116, pp. 24–41, Sep. 2016.
- [17] S. Saha, L. Mou, X. X. Zhu, F. Bovolo, and L. Bruzzone, "Semisupervised change detection using graph convolutional network," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 4, pp. 607–611, Apr. 2021.
- [18] M. Gong, Y. Yang, T. Zhan, X. Niu, and S. Li, "A generative discriminatory classified network for change detection in multispectral imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 1, pp. 321–333, Jan. 2019.
- [19] S. Saha, L. Mou, C. Qiu, X. X. Zhu, F. Bovolo, and L. Bruzzone, "Unsupervised deep joint segmentation of multitemporal high-resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8780–8792, Dec. 2020.
- [20] T. Zhan, M. Gong, X. Jiang, and M. Zhang, "Unsupervised scale-driven change detection with deep spatial-spectral features for VHR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5653–5665, Aug. 2020.
- [21] M. Gong, H. Yang, and P. Zhang, "Feature learning and change feature classification based on deep learning for ternary change detection in SAR images," *J. Photogramm. Remote Sens.*, vol. 129, pp. 212–225, Jul. 2017.
- [22] F. Gao, X. Wang, Y. Gao, J. Dong, and S. Wang, "Sea ice change detection in SAR images based on convolutional-wavelet neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1240–1244, Aug. 2019.
- [23] F. Gao, J. Dong, B. Li, and Q. Xu, "Automatic change detection in synthetic aperture radar images based on PCANet," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1792–1796, Dec. 2016.
- [24] M. Li, M. Li, P. Zhang, Y. Wu, W. Song, and L. An, "SAR image change detection using PCANet guided by saliency detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 3, pp. 402–406, Mar. 2018.
- [25] M. Gong, X. Niu, P. Zhang, and Z. Li, "Generative adversarial networks for change detection in multispectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2310–2314, Nov. 2017.
- [26] S. Saha, Y. T. Solano-Correa, F. Bovolo, and L. Bruzzone, "Unsupervised deep transfer learning-based change detection for HR multispectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 856–860, May 2021.
- [27] A. Singh, "Review article digital change detection techniques using remotely-sensed data," *Int. J. Remote Sens.*, vol. 10, no. 6, pp. 989–1003, 1989.
- [28] Y. T. S. Correa, F. Bovolo, and L. Bruzzone, "Change detection in very high resolution multisensor images," in *Proc. 20th Image Signal Process. Remote Sens.*, vol. 9244, 2014, Art. no. 924410.
- [29] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.
- [30] F. Thonfeld, H. Feilhauer, M. Braun, and G. Menz, "Robust change vector analysis (RCVA) for multi-sensor very high resolution optical satellite data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 50, pp. 131–140, Aug. 2016.
- [31] L. Bruzzone and D. F. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 3, pp. 1171–1182, May 2000.
- [32] L. Bruzzone and D. F. Prieto, "A minimum-cost thresholding technique for unsupervised change detection," *Int. J. Remote Sens.*, vol. 21, no. 18, pp. 3539–3544, 2000.
- [33] T. Celik, "Unsupervised change detection in satellite images using principal component analysis and  $k$ -means clustering," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 772–776, Aug. 2009.
- [34] F. Bovolo and L. Bruzzone, "A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 1, pp. 218–236, Jan. 2006.
- [35] F. Bovolo, S. Marchesi, and L. Bruzzone, "A framework for automatic and unsupervised detection of multiple changes in multitemporal images," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 6, pp. 2196–2212, May 2011.
- [36] F. Bovolo, "A multilevel parcel-based approach to change detection in very high resolution multitemporal images," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 1, pp. 33–37, Jan. 2008.
- [37] L. Li, X. Li, Y. Zhang, L. Wang, and G. Ying, "Change detection for high-resolution remote sensing imagery using object-oriented change vector analysis method," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 2873–2876.
- [38] N. Falco, G. Cavallaro, P. R. Marpu, and J. A. Benediktsson, "Unsupervised change detection analysis to multi-channel scenario based on morphological contextual analysis," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 3374–3377.
- [39] D. Wang, D. W. Hogg, D. Foreman-Mackey, and B. Schölkopf, "A causal, data-driven approach to modeling the Kepler data," *Publications Astronomical Soc. Pacific*, vol. 128, no. 967, 2016, Art. no. 094503.
- [40] B. Schölkopf *et al.*, "Modeling confounding by half-sibling regression," *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 27, pp. 7391–7398, Jul. 2016.
- [41] M. Heikkilä, M. Pietikäinen, and C. Schmid, "Description of interest regions with local binary patterns," *Pattern Recognit.*, vol. 42, no. 3, pp. 425–436, 2009.
- [42] G.-A. Bilodeau, J.-P. Jodoin, and N. Saunier, "Change detection in feature space using local binary similarity patterns," in *Proc. Int. Conf. Comput. Robot. Vis.*, May 2013, pp. 106–112.
- [43] N. Gupta, G. V. Pillai, and S. Ari, "Change detection in optical satellite images based on local binary similarity pattern technique," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 3, pp. 389–393, Mar. 2018.
- [44] Y. Sun, L. Lei, X. Li, X. Tan, and G. Kuang, "Patch similarity graph matrix-based unsupervised remote sensing change detection with homogeneous and heterogeneous sensors," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 4841–4861, Jun. 2021.
- [45] L. Bruzzone and F. Bovolo, "A conceptual framework for change detection in very high resolution remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2010, pp. 2555–2558.
- [46] D. Wang, D. W. Hogg, D. Foreman-Mackey, and B. Schölkopf, "A pixel-level model for event discovery in time-domain imaging," 2017, *arXiv:1710.02428*.
- [47] T. D. Gebhard, M. J. Bonse, S. P. Quanz, and B. Schölkopf, "Physically constrained causal noise models for high-contrast imaging of exoplanets," Dec. 2020, *arXiv:2010.05591*. [Online]. Available: <https://arxiv.org/abs/2010.05591>
- [48] N. Coudray, J.-L. Buessler, and J.-P. Urban, "Robust threshold estimation for images with unimodal histograms," *Pattern Recognit. Lett.*, vol. 31, no. 9, pp. 1010–1019, Jul. 2010.
- [49] P. L. Rosin, "Unimodal thresholding," *Pattern Recognit.*, vol. 34, no. 11, pp. 2083–2096, 2001.
- [50] Y. Bazi, L. Bruzzone, and F. Melgani, "Image thresholding based on the EM algorithm and the generalized Gaussian distribution," *Pattern Recognit.*, vol. 40, no. 2, pp. 619–634, Feb. 2007.
- [51] M. Zanetti, F. Bovolo, and L. Bruzzone, "Rayleigh-Rice mixture parameter estimation via EM algorithm for change detection in multispectral images," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5004–5016, Dec. 2015.

- [52] A. Song, Y. Kim, and Y. Han, "Uncertainty analysis for object-based change detection in very high-resolution satellite images using deep learning network," *Remote Sens.*, vol. 12, no. 15, p. 2345, Jul. 2020.
- [53] M. D. Mura, J. A. Benediktsson, F. Bovolo, and L. Bruzzone, "An unsupervised technique based on morphological filters for change detection in very high resolution images," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 3, pp. 433–437, Jul. 2008.
- [54] S. Liu, Q. Du, X. Tong, A. Samat, L. Bruzzone, and F. Bovolo, "Multi-scale morphological compressed change vector analysis for unsupervised multiple change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 9, pp. 4124–4137, Sep. 2017.
- [55] R. C. Daudt, B. L. Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral Earth observation using convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 2115–2118.
- [56] H. Dong, W. Ma, Y. Wu, J. Zhang, and L. Jiao, "Self-supervised representation learning for remote sensing image change detection based on temporal prediction," *Remote Sens.*, vol. 12, no. 11, p. 1868, Jun. 2020.



**Lukas Kondmann** received the bachelor's degree in economics from the Ludwig Maximilian University of Munich, Munich, Germany, in 2016, the Honors degree in technology management from the Center for Digital Technology and Management, Munich, in 2017, and the master's degree in social data science from the University of Oxford, Oxford, U.K., in 2019. He is currently pursuing the Ph.D. degree in engineering with the Technical University of Munich, Munich, and the German Aerospace Center, Munich.

He was a Visiting Researcher working on big data for social good with the School of Information, University of California at Berkeley (UC Berkeley), Berkeley, CA, USA, in spring 2017. His research is centered around time-series analysis of multispectral remote sensing imagery with a focus on monitoring the sustainable development goals (SDGs).



**Aysim Toker** received the B.Sc. degree in computer engineering and the M.Sc. degree in computer science from the Technical University of Munich, Munich, Germany, in 2016 and 2018, respectively, where she is currently pursuing the Ph.D. degree with the Dynamic Vision and Learning Group, under the supervision of Prof. Dr. Laura Leal-Taixé for a joint project with Prof. Dr. Xiaoxiang Zhu.

She has been a programmer for many years, with broad experience in many languages. Her interest in computer vision and machine learning started with her M.Sc. degree. During her master's thesis, she concentrated on video object segmentation. Currently, she is doing research in deep learning, sequence analysis, and remote sensing with a focus on the intersection of these three domains.



**Sudipan Saha** (Member, IEEE) received the M.Tech. degree in electrical engineering from IIT Bombay, Mumbai, India, in 2014, and the Ph.D. degree in information and communication technologies from the University of Trento, Trento, Italy, and the Fondazione Bruno Kessler, Trento, in 2020.

He was an Engineer with TSMC Ltd., Hsinchu, Taiwan, from 2015 to 2016. In 2019, he was a Guest Researcher with the Technical University of Munich (TUM), Munich, Germany, where he is currently a Post-Doctoral Researcher. His research interests are related to multitemporal remote sensing image analysis, domain adaptation, time-series analysis, image segmentation, deep learning, image processing, and pattern recognition.

Dr. Saha was a recipient of the Fondazione Bruno Kessler Best Student Award in 2020. He is a reviewer for several international journals. He has served as a Guest Editor for *Remote Sensing* (MDPI) special issue on "Advanced Artificial Intelligence for Remote Sensing: Methodology and Application."



**Bernhard Schölkopf** has researched at AT&T Bell Labs, Holmdel, NJ, USA, GMD FIRST, Berlin, Germany, and Microsoft Research Cambridge, Cambridge, U.K., becoming a Max Planck Director in 2001. He is currently a Professor with ETH Zürich, Zürich, Switzerland. He co-initiated the MLSS series of Machine Learning Summer Schools, the Cyber Valley Initiative, and the ELLIS grassroots initiative. He has applied his methods to a number of different fields, ranging from biomedical problems to computational photography and astronomy. His scientific interests are in machine learning and causal inference.

Dr. Schölkopf is also a fellow of the Association for Computing Machinery (ACM) and the CIFAR Program Learning in Machines and Brains and a member of the German Academy of Sciences. He (co)received the Academy Prize of the Berlin-Brandenburg Academy of Sciences and Humanities, the Royal Society Milner Award, the Leibniz Award, the Koerber European Science Prize, and the BBVA Foundation Frontiers of Knowledge Award. He is the Co-Editor-in-Chief of the *Journal of Machine Learning Research*, an early development in open access and today the field's flagship journal.



**Laura Leal-Taixé** received the B.Sc. and M.Sc. degrees in telecommunications engineering from the Technical University of Catalonia (UPC), Barcelona, Spain, in 2005 and 2008, respectively, and the Ph.D. degree from the Leibniz University of Hannover, Hanover, Germany, in 2013.

She was a Visiting Scholar with the University of Michigan, Ann Arbor, MI, USA. She spent two years as a Post-Doctoral Researcher at ETH Zürich, Zürich, Switzerland, and a year as a Senior Post-Doctoral Researcher at the Computer Vision Group, Technical University of Munich, Munich, Germany. She went to Boston, MA, USA, to do her master's thesis at Northeastern University, Boston, with a fellowship from the Vodafone foundation. She is currently a tenure-track Professor (W2) with the Technical University of Munich, leading the Dynamic Vision and Learning Group.

Dr. Leal-Taixé was a recipient of the Sofja Kovalevskaja Award of 1.65 million euros for her project socialMaps and the Google Faculty Award.



**Xiao Xiang Zhu** (Fellow, IEEE) received the M.Sc., Dr.Ing., and "Habilitation" degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.




She was a Guest Scientist or a Visiting Professor with the Italian National Research Council (CNR-IREA), Naples, Italy, Fudan University, Shanghai, China, The University of Tokyo, Tokyo, Japan, and the University of California at Los Angeles, Los Angeles, CA, USA, in 2009, 2014, 2015, and 2016, respectively. Since 2019, she has been a Co-Coordinator of the Munich Data Science Research School, Munich. Since 2019, she has been heading the Helmholtz Artificial Intelligence—Research Field "Aeronautics, Space, and Transport." Since May 2020, she has been the Director of the international future AI lab "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond," Munich. Since October 2020, she has been the Co-Director of the Munich Data Science Institute (MDSI), TUM. She is currently a Professor with Data Science in Earth Observation (former: Signal Processing in Earth Observation), TUM, and the Head of the Department "EO Data Science," Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany. She is currently a Visiting AI Professor with ESA's Phi-Lab, Frascati, Italy. Her main research interests are remote sensing and Earth observation, signal processing, machine learning, and data science, with a special application focus on global urban mapping.

Dr. Zhu is also a member of the young academy, Junge Akademie/Junges Kolleg, at the Berlin-Brandenburg Academy of Sciences and Humanities, the German National Academy of Sciences Leopoldina, and the Bavarian Academy of Sciences and Humanities. She also serves on the scientific advisory board in several research organizations, including the German Research Center for Geosciences (GFZ) and the Potsdam Institute for Climate Impact Research (PIK). She is also an Associate Editor of *IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING*. She also serves as the Area Editor responsible for special issues of *IEEE Signal Processing Magazine*.

A.4 SEMISIROC



# SemiSiROC: Semi-Supervised Change Detection With Optical Imagery and an Unsupervised Teacher Model

Lukas Kondmann , Sudipan Saha , and Xiao Xiang Zhu 

**Abstract**—Change detection is an important yet challenging task in remote sensing. In this paper, we underline that the combination of unsupervised and supervised methods in a semi-supervised framework improves change detection performance. We rely on Half-Sibling Regression for Optical Change Detection (SiROC) as an unsupervised teacher model to generate pseudo labels and select only the most confident pseudo labels for pretraining different student models. Our results are robust to three different competitive student models, two semi-supervised pseudo label baselines, two benchmark datasets and a variety of loss functions. While the performance gains are highest with a limited number of labels, a notable effect of pseudo label pretraining persists when more labeled data is used. Further, we outline that the confidence selection of SiROC is indeed effective and that the performance gains generalize to scenes that were not used for pseudo label training. Through the pseudo label pretraining, SemiSiROC allows student models to learn more refined shapes of changes and makes them less sensitive to differences in acquisition conditions.

**Index Terms**—Change Detection, semi-supervised, unsupervised, optical images, multitemporal

## I. INTRODUCTION

CHANGE DETECTION (CD) is the task of segmenting changing pixels over time in multitemporal Earth observation data. In the face of a changing planet, CD is at the core of many relevant monitoring tasks. It allows us to study the temporal evolution of forests [1]–[3], urban areas [4], [5], coastal and maritime regions [6], [7] and the effects of natural disasters [8]–[12]. Change detection methods face a number of hurdles related to the acquisition conditions between the different times the images are collected. This includes but is not limited to illumination conditions, clouds

The work is jointly supported by the Helmholtz Association through the joint research school “Munich School for Data Science - MUDS” and Helmholtz Excellent Professorship “Data Science in Earth Observation - Big Data Fusion for Urban Research” (grant number: W2-W3-100), and by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab “AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond” (grant number: 01DD20001).

*Corresponding author: Xiao Xiang Zhu.*

Lukas Kondmann is with the Remote Sensing Technology Institute, German Aerospace Center (DLR), Wessling 82234, Germany, and also with Data Science in Earth Observation, Technical University of Munich, Ottobrunn 85521, Germany (e-mail: lukas.kondmann@dlr.de).

Sudipan Saha is with the Yardi School of Artificial Intelligence, Indian Institute of Technology Delhi, New Delhi 110016, India (e-mail: sudipan.saha@scai.iitd.ac.in)

Xiao Xiang Zhu is with Data Science in Earth Observation, Technical University of Munich (TUM), Ottobrunn 85521, Germany (e-mail: xiaoxiang.zhu@tum.de)

and shadows, acquisition angles and the definition of what constitutes a change [13]. Despite these challenges, several trends have been beneficial for the methodological progress in change detection in recent years. First, open data policies for example in the Copernicus program [14] increase accessibility and availability of multitemporal Earth observation data [15]. Second, technological progress results in increasing spatial and temporal resolution of satellite data with up to daily imagery [16]. Third, methodological progress in image recognition, particularly deep learning [17], has also fueled a variety of improvements in Artificial Intelligence (AI) for Earth observation including change detection [18]–[20].

Many recent advances are in supervised learning for binary change detection from optical imagery [21]–[29]. Following the success of convolutional neural networks (CNNs) in a variety of computer vision problems [17], CNNs have been used frequently for change detection problems as well. Daudt et al. [23] introduce a siamese change detection architecture inspired by UNet [30]. ESCNET is a combination of superpixel enhancement and a deep CNN [31]. For change detection in aerial images, Xu et al. [32] design a pseudo-siamese capsule network.

More recently, the success of vision transformers [33], [34] has induced increasing attention also from the remote sensing community. For example, Bandara and Patel [21] design ChangeFormer, a siamese transformer network for building change detection. In a similar spirit, Chen et al. [25] employ a self-attention based transformer method. Further, many approaches also combine convolutional and attention-based approaches with promising results [22], [35], [36].

However, obtaining large-scale labeled data for change detection remains a challenge. Unsupervised CD methods [37]–[41] therefore learn without labeled data to circumvent this issue. Many methods also utilize the advances in deep learning for unsupervised CD. For example, Saha et al. introduce Deep Change Vector Analysis (DCVA) for high-resolution imagery which combines ideas from classical image differencing with a deep convolutional feature extractor [37]. DCVA has also been further extended in combination with self-supervised pretraining [42] and refined further for medium-resolution images [38]. A generative approach is used in [43] to model the difference image in an unsupervised fashion. Zhan et al. [44] rely on an initial classification of changing superpixels with a fully convolutional neural network. These superpixels are then categorized by uncertainty and used to train a classifier in a second step.

Still, in unsupervised change detection, particularly with lower resolution, many methods reach high performance also without the use of deep features. SiROC [39] is inspired by exoplanet search and compares pixels against their distant neighborhood to identify changes in optical imagery. Further, image differencing also called change vector analysis [45] and its extensions [46]–[48] still play a role in practice.

Semi-supervised approaches bridge the gap between unsupervised and supervised approaches. These methods try to combine labeled data with larger amounts of unlabeled data to support the training process. Among the first to apply semi-supervised learning in change detection were Bovolo et al [49]. They use a Bayesian thresholding mechanism to set up an adequately defined binary semi-supervised support vector machine ( $S^3VM$ ). Modified Self-Organizing Feature App (SOFM) uses only a limited set of initial labels to compute soft labels for unlabeled additional input [50]. Chen et al [51] rely on probabilistic Gaussian Processes (GP) as a first step with labeled and unlabeled data. The outputs of the GP classifier are then refined with a Markov Random Field regularizer. A Laplacian Regularized Metric Learning mechanism is used in [52] to exploit unlabeled training data at scale for hyperspectral image change detection. For very high spatial resolutions, graph convolutional networks (GCNs) are also effective for semi-supervised learning by encoding multitemporal images as a graph [53].

One particularly effective direction in semi-supervised learning in general image recognition is student-teacher models [54]. Typically, there is a teacher model that is trained on labeled data and predicts additional labels for images where ground truth is not available. Then, a student model uses these additional labels, referred to as pseudo labels (PL), during the training. With Earth observation data, pseudo labels have also been shown to be effective for hyperspectral image classification [55]. Pseudo labels are also related to unsupervised CD approaches for small scenes which rely on an initial difference image or change classification and finetune this further with another unsupervised method [43], [56], [57]. This is similar to using pseudo labels although these approaches are purely unsupervised and are applied only to single scenes instead of large-scale training. Li et al. [58] use pseudo labels explicitly for change detection in SAR images but stay in the unsupervised domain. Similarly, Gao et al [59] train convolutional wavelet neural networks with automatically generated labels for sea ice change detection with SAR images.

In many student-teacher settings, the actual labels are used at least in some capacity in the pseudo labeling. However, this can be somewhat challenging in scenarios with limited labels as in change detection. Additionally, applications of methods in regions outside their training data often require some robustness to unseen regions [60]. In this paper, we therefore propose SemiSiROC where we use an unsupervised method with well-calibrated uncertainties for pseudo label training. The uncertainty score for each prediction allows us to filter only high-quality pseudo labels for pretraining. In the second step of the semi-supervised method, we finetune student models with the actual labels to improve optical change

detection performance. We evaluate our results on a binary version of the DynamicEarthNet benchmark [61] as well as the OSCD dataset [24] and compare the effectiveness of our strategy with five competitive change detection models as students: ChangeFormer [21], BIT [25], DTCDCN [29], FC-Siam-Diff [23] and FC-Siam-Conc [23]. Although SemiSiROC is most effective in limited label scenarios we also find that even with a sizeable amount of 1000 labeled image pairs, SemiSiROC boosts performance for all tested models notably. While student-teacher models themselves are not new in remote sensing, our ingenuity lies in the components specifically designed for change detection on large-scale datasets and further validation on a global dataset of such scale. We have three main contributions:

- 1) We present SemiSiROC, a semi-supervised change detection method in optical remote sensing that combines advanced supervised models with unsupervised pseudo labeling.
- 2) Building on the confidence filtering of SiROC, we devise a mechanism to prioritize relevant scenes during pseudo label filtering.
- 3) We propose a detailed experimental setup for change detection subject to geographic disparity, based on the recently launched publicly available DynamicEarthNet dataset [61]. This experimental setup will be helpful for other researchers to pursue research in this direction. Our experiments on this setup and the OSCD [23] benchmark show that semisupervised learning is indeed helpful.

## II. METHOD

### A. SemiSiROC

Let us assume, we have two different collections of images,  $D$  and  $U$ .  $D$  is a collection of  $N_D$  bi-temporal pairs with associated pixelwise change/unchanged label. On the other hand,  $U$  is a collection of  $N_U$  unlabeled bi-temporal pairs. Generally  $N_U > N_D$ , however this is not a strict assumption. The  $U$  and  $D$  can be acquired over different geographic areas/continents, thus they need not be representing the same geographic distribution. Our goal is to exploit both  $D$  and  $U$  to learn a change detection model. Towards this, we design a semisupervised pipeline that allows exploiting  $U$  for model training even if labels for it are not available. We exploit a teacher-student model where the teacher model labels the images and selects relevant samples from  $U$ . This allows its student to exploit the label space  $D \cup U$  instead of  $D$ . Therefore, we train with pseudo labels first before we go on to real labels. This is consistent with semi-supervised literature [62] and has the underlying assumption that the model can immensely benefit from pseudo labels as a first step of training, which can be subsequently refined with actual labels.

The pseudo labels for pretraining are based on SiROC [39], an unsupervised method for optical change detection. We average the confidence on the cube level and as a default choice use the top 25%. Then, we train a student model with the preselected locations and pseudo labels first before finetuning with the actual labels. Since the teacher model

exploits SiROC in a semi-supervised setting, we call our approach SemiSiROC.

Algorithm 1 outlines SemiSiROC in pseudocode in more depth. Given the unlabeled collection  $U$ , the labeled collection  $D$ , the corresponding labels  $L$  and a supervised change detection model, the desired output is a binary change segmentation. At first, we define a collection of confidence scores  $C$  and pseudo labels  $P$ . Then, we loop over the elements of  $U$  and obtain pseudo labels and confidence scores with SiROC for each image pair. Before semi-supervised pretraining we filter  $P$  and  $U$  to only use the scenes with the highest confidence which is defined as  $U_P$ . These scenes are used as input for the pretraining of the CD model before training with actual labels in the final step.

While the proposed SemiSiROC approach is similar to many semi-supervised learning strategies [62], note that our approach is distinct in three ways: (i) how we generate the pseudo labels with an unsupervised CD method, (ii) how we select the samples for student training based on a well-calibrated uncertainty, (iii) how we exploit them for global change detection.

---

**Algorithm 1** : SemiSiROC
 

---

**Input:**  $U, D, L$ , model

**Output:** Binary Change Segmentation

```

1:  $C = [], P = []$ 
2: for ( $u$  in  $U$ ) do
3:    $P_u, C_u = \text{SiROC}(u)$ 
4:    $C.append(C_u)$ 
5:    $P.append(P_u)$ 
6: end for
7:  $U_P = C_P.top\_quarter(C)$ 
8:  $P_P = P.top\_quarter(C)$ 
9:  $\text{model.train}(U_P, P_P)$  {Pseudo label training}
10:  $\text{model.train}(D, L)$  {Finetuning}

```

---

### B. Unsupervised teacher model

The goal of the teacher model is to assign pseudo labels to some samples from  $U$  with reasonable confidence that they can be used later for training the change detection (student) model. Since  $U$  and  $D$  may not necessarily be from the same distribution, the teacher model may use its learning from  $D$  and bias the distribution of pseudo labels for  $U$  by overfitting to  $D$ . This is particularly relevant in the geo context where different locations and points in time can quickly change the data-generating distribution [63]. We argue that the teacher label should refrain from using the actual labels in any form to obtain the pseudo labels. If the pseudo label extraction process uses the actual labels, this would make them interdependent and hamper generalization. Thus, the teacher model should be based on unsupervised learning in this case. Additionally, semi-supervised pretraining is more flexible with unsupervised pseudo labels and our pretrained model can serve as a starting point for other CD applications without the need to retrain the teacher model on new datasets with new labels to obtain other pseudo labels. Therefore, we propose to use an

unsupervised teacher model to incentivize more robustness to spatial generalization in the pseudo labels. This is in contrast to many other semi-supervised approaches with pseudo labels which rely on teacher models which have seen at least some of the actual labels [62].

As unsupervised teacher model, we employ SiROC [39]. While the method is highly performant, we pick it as pseudo label source or so-called teacher model mainly because it comes with a built-in well-calibrated confidence score ranging from 0 (low) to 1 (high) with its prediction for each pixel. This allows us to filter pseudo labels based on their confidence and only train on high confidence labels. As this confidence score is closely connected to the quality of the pseudo label, we hypothesize that algorithms should learn better with selected pseudo labels only. Out of  $N_U$  total samples in  $U$ ,  $N'_U$  are chosen after confidence filtering for pretraining. In the following, we explore SiROC in more depth.

*SiROC.* SiROC models a pixel as a linear combination of a set of neighboring pixels  $n$  at a certain time  $t$  in a time series. At time  $t + 1$ , the value of the respective pixel is predicted based on the neighbors  $n$  at  $t + 1$ . The deviation between the actual and the predicted pixel value is interpreted as a change signal. If the difference is high, this is seen as an indication of change as the pixel seems to have undergone a change compared to its neighborhood. The comparison against the neighborhood serves to eliminate local or image-wide trends as sources of false positives for changes.

More formally, given a channel of a multispectral image  $I$  at time  $t$  and  $t + 1$ , the core of the predicted change segmentation  $\hat{P}$  is based on the following equation:

$$\hat{P} = \begin{cases} 1, & \text{if } \hat{\mathbf{I}}_{t+1} - \mathbf{I}_{t+1} > o \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where  $o$  is the Otsu-Threshold [64].  $\hat{\mathbf{I}}_{t+1}$  is the predicted image at time  $t + 1$  based on half-sibling regression. To extend this to multiple channels  $C$ , the absolute sum of the difference between the predicted and the actual image is taken:

$$\hat{P} = \begin{cases} 1, & \text{if } \sum_{c=1}^C |\hat{\mathbf{I}}_{t+1,c} - \mathbf{I}_{t+1,c}| > o \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

For formal details on how  $\hat{\mathbf{I}}_{t+1,c}$  is obtained given a set of neighbors we refer to [39]. SiROC ensembles over many mutually exclusive neighborhoods and relies on majority voting between the models for its final prediction. This iterative process uses mutually exclusive sets of neighboring pixels that are increasingly more distant from the pixel of interest itself. Relevant parameters for this process are the maximum neighborhood size and the step size of the ensemble. The number of ensembles is given as the maximum neighborhood size divided by the step size. We use SiROC with its presented defaults in [39]. The respective parameter values are:

- 1) Maximum neighborhood size:  $n\_max=200$
- 2) Initial exclusion window:  $e\_start=0$
- 3) Step size of ensemble:  $s=2$
- 4) Filter Size of Morphological Operations:  $p=5$

One deviation is to reduce the step size of the ensemble from 8 to 2. This results in 100 models with a maximum neighborhood size of 200 and allows for more variation in the uncertainty estimates.

The number of votes, as shown in [39], can be interpreted as a well-calibrated uncertainty and is used in this work as a confidence score. This is because the performance of SiROC is increasing in its confidence. Therefore, we use SiROC in combination with three supervised student models for change detection.

### C. Student model

Once the teacher model is used to select the pseudo samples from  $U$ , ideally any machine learning based classifier model can be used to train the student model. The training involves two steps: 1) Training with pseudo labeled  $N'_U$  samples from  $U$ , obtained in Section II-B and 2) Fine-tuning with the labeled dataset  $D$ .

To illustrate that our SemiSiROC can work with a diverse set of classifiers, we chose several competitive supervised change detection architectures. They are outlined in more detail as follows:

*FC-Siam-diff* [23] is a fully convolutional Siamese neural network inspired by the UNet architecture [30]. Pre and post images are processed in two separate, parallel streams with shared weights which are only merged after the convolutional layers of the network. In contrast to a classic concatenation of features, this network takes the absolute difference of the encoding streams. This allows the model to focus on temporal differences in the image pair which is well suited for change detection tasks. These differences are infused as inputs to the upsampling steps. Allowing feature differences to be passed without further processing far into the network allows the network to treat simple decisions without unnecessary complexity.

*FC-Siam-conc* [23] is similar to *FC-Siam-diff* with one major distinction. Instead of taking feature differences of the encoding streams, the features are concatenated. This gives the model more flexibility but nudges it less directly towards a temporal comparison of features.

*DTCDCSN* [29] stands for Dual Task Constrained Deep Siamese Convolutional Network. It is a convolutional model which performs semantic segmentation and change detection simultaneously. This is helpful for change detection since a prior understanding of objects and their size from semantic segmentation can be utilized for the change detection task.

*ChangeFormer* [21] is also a Siamese network with a transformer-based encoder that reaches competitive performance on the LEVIR-CD [65] and DSIFN-CD [22] benchmarks. The hierarchical transformer encoder uses four transformer blocks in with shared weights in each branch. After every transformer block, a difference module is taken to compare differences at different abstraction levels. These differences are then passed to a lightweight multi-layer perceptron decoder which samples the features up and computes the final predicted change map.

*Bitemporal Image Transformer (BIT)* [25] also relies on self-attention rather than only deep convolutional features in a

transformer framework. It has three main elements: A siamese semantic tokenizer, a transformer encoder and a transformer decoder. The siamese backbone extracts convolutional features and inputs them into the semantic tokenizer. Inspired by advances in language processing, the tokenizer pools the image features into a compact set of vocabulary. The compact tokens are converted back to the pixel space and fed into a CNN prediction head. As a CNN backbone for the feature extraction, ResNet18 is used following the main paper.

## III. EXPERIMENTAL VALIDATION

### A. Data

*DynamicEarthNet*: We base our analysis on a modified version of the DynamicEarthNet dataset [61]. This is because it allows benchmarking change detection algorithms with areas of interest (AOIs) across the globe and covers a variety of different changes that are not specific to a certain use case such as buildings or urban regions only. Both of these properties make the dataset well-tailored to binary change detection in an application-agnostic way. It contains monthly, manual land cover annotations for two years with Planet imagery for 75 AOIs across the globe. The locations were selected to include a wide spectrum of land cover changes across seven classes.

We pick the labels of the first and last month of each AOI and compute a binary mask of changing land cover. This maximizes change and also ensures a certain difference in the scenes. The corresponding Planet Fusion images are highly preprocessed as an analysis-ready product which includes a variety of steps including temporal gap-filling of clouds and shadow removal. Each scene is  $1024 \times 1024$  pixels with 3m resolution per pixel in size which results in an area per scene of about  $10\text{km}^2$ . To be consistent with the image size in [21], we split each scene into  $16 \times 256 \times 256$  pixels RGB images. This results in a total of 1200 pairs of pre and post images taken 2 years apart. The class balance in the resulting dataset is about 80% no change and 20% change.

Our baseline train, validation, and test split is visible in Figure 1. Locations are available across the globe which is relevant to test generalizability to unseen regions where all continents except Antarctica are covered. Following the DynamicEarthNet terminology, we refer to the locations also as cubes given that the 2d images also vary in time. The cubes do not only differ by their geography but also by the type of change. The dataset covers locations from coastal areas, islands, urban regions, agricultural areas, and forests. This shows the diversity of change in practical applications which makes this dataset challenging.

The cubes based in the continental US are used as training (blue), the validation data is taken from central America (green) and we test with the remaining cubes from across the globe. This simulates label scarcity in global change detection tasks where generalizability to unseen regions is a key requirement. Particularly, annotated data in low and middle-income countries are often relatively rare. However, to validate our results against this choice we use other splits with more training data (16, 32, 64 cubes) as an ablation study below.



*Onera Satellite Change Detection (OSCD)* [23]: As a secondary dataset, we rely on OSCD which in total contains 24 before and after pairs of Sentinel-2 images in urban areas across the globe but we only use the 10 pairs in the test set. To be consistent with our training efforts on DynamicEarthNet, we only include the RGB channels and crop  $256 \times 256$  images from the original scenes. As OSCD image pairs are not square and vary in size, we pad the images to the next multiple of 256 and mask the added points during evaluation of the change prediction.

### B. Training and Evaluation

Our goal is to evaluate the effectiveness of a pseudo label pretraining step. Therefore, we compare SiROC confidence pretraining for a variety of specifications including the above-mentioned models but also different choices of training sets, pseudo label sets and training losses. We train each model until convergence with and without a pretraining step. For this study, experiments were conducted with a single NVIDIA Quadro P4000. We acknowledge that semi-supervised pretraining requires an additional computational effort compared to finetuning. Pseudo label training for 50 epochs with the top quarter of scenes by confidence takes about 15 min with the P4000 for the FC-Siam-diff model. However, pseudo label training has to be done only once and allows for all kinds of change detection applications.

The following specifications are used for all experiments to ensure comparability. We train with Adam as an optimizer with a batch size of 32 and a starting learning rate of 0.0001 and linear weight decay. We evaluate our results based on three popular criteria: Accuracy, mean IOU (MIOU) and mean F1 Score. Formally, in terms of false positives (FP), true positives (TP), false negatives (FN), and true negatives (TN) these criteria have the following definitions:

$$Accuracy = (TP + TN)/(TP + TN + FN + FP) \quad (3)$$

Accuracy is simply asking how often is our prediction correct relative to the total number of predictions.

$$MIOU = (IOU_1 + IOU_0)/2 \quad (4)$$

with  $IOU = TP/(TP + FP + FN)$ . In comparison to accuracy, the IOU criterion eliminates TN from the picture per class. Similarly,

$$MF1 = (F1_1 + F1_0)/2 \quad (5)$$

with F1 balancing precision and recall.  $F1 = (2 * precision * recall)/(precision + recall)$ . Precision is defined as  $TP/(TP + FP)$  and recall as  $TP/(TP + FN)$ . Every model is run for five different seeds and reported scores are therefore a mean with the respective standard deviation in brackets.

### C. DynamicEarthNet Results

Table I outlines the main results of our paper. Overall, we test pseudo label pretraining with SiROC with four different competitive models. Each pair of rows for one model compares

the scores with and without pretraining on the confident pseudo labels (PL). All specifications are run five times with different seeds to increase the robustness of the result against an unrepresentative seed. Pseudo label training is done with a focal loss (FL) and training with the real labels with the split of Figure 1 and a MIOU loss with only the top 25% of cubes based on average SiROC confidence per cube.

At first, FC-Siam-diff with pseudo label pretraining reaches an overall accuracy of 0.7812 with a MIOU score of 0.4854 and a Mean F1 Score of 0.6029. This makes it the best model in the Table overall according to all three criteria and notably better than its counterpart without pretraining. FC-Siam-diff without SiROC pretraining is about 15 percentage points (p.p.) lower in accuracy, 7 p.p. lower in MIOU and about 3 p.p. lower in terms of mean F1 score. Further, standard deviations of performance are visibly lower with confidence-filtered pseudo label pretraining for FC-Siam-diff. FC-Siam-Conc does not seem competitive here in comparison with a fairly low accuracy of around 62% with pseudolabels and 56% without them. It seems that without the explicit feature difference the model is not incentivized to pay enough attention to temporal differences for the final change segmentation. Therefore, it has trouble to distinguish changes from non-changes. This is improved by the use of pseudo labels but the issue remains large in comparison to FC-Siam-diff.

Similarly, the scores of ChangeFormer improve and stabilize notably by an even larger margin although the baseline performance is comparably bad. The general effectiveness is also confirmed when looking at BIT and DTCDCSN although the margins seem slightly lower. Given that DTCDCSN and particularly FC-Siam-conc seem weaker convolutional baselines than FC-Siam-diff, we focus on the latter, ChangeFormer and BIT for the remainder of the paper for the sake of brevity. As an additional baseline, the performance of SiROC on the test set is given as a reference point.

Generally, SiROC places decently on the dataset given that it is an unsupervised method and often even outscores the supervised baselines with few labels. The information contained in the pseudo labels and the capacity of the methods combine effectively in our semi-supervised strategy. The respective scores are consistently substantially higher than in the SiROC baseline with the pseudo labels.

Figure 2 visualizes model predictions for eight image pairs of the models in Table I. On top are the pre (2a) and post image (2b) samples together with the ground truth (2c) from left to right. Large forest changes are for example visible in the image on the left or in middle. Notably, the illumination conditions between the pre and post images differ slightly which is often a challenge in change detection problems [37]. The first comparison is for FC-Siam-Diff with training on pseudo labels in 2d and the corresponding version without it in 2e. 2d was the best performing model quantitatively in Table I which is confirmed by the visual inspection of the predictions.

The location and the shape of large changes are segmented well with limited mistakes. While the model does miss some smaller changes on the right, regions in the middle are segmented well. In comparison to 2e without pseudo labels,

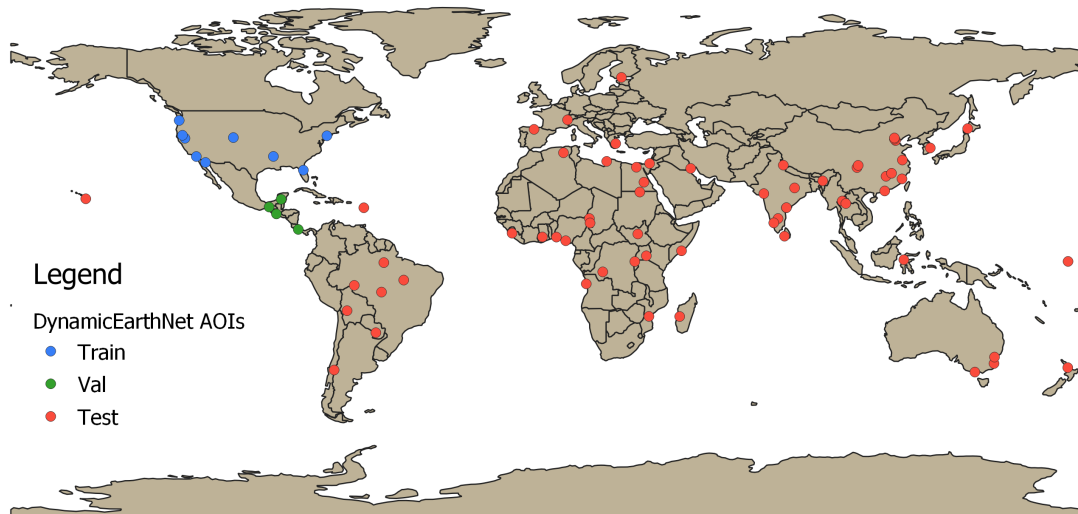


Fig. 1. Spatial Train/Validation/Test split used as a default. The limited amount of training data simulates real-world scenarios where training data is scarce and mainly from specific regions.

TABLE I  
QUANTITATIVE RESULTS DYNAMICEARTHNET GROUPED BY PSEUDO LABEL USE

Model	PL	Loss PL	Loss	Accuracy	MIOU	MF1
FC-Siam-diff [23]	✓	FL	MIOU	<b>0.7812</b> (+-0.0104)	<b>0.4854</b> (+-0.0037)	<b>0.6029</b> (+-0.0018)
FC-Siam-diff [23]			MIOU	0.6359 (+-0.0405)	0.419 (+-0.0288)	0.5706 (+-0.0244)
FC-Siam-conc [23]	✓	FL	MIOU	0.6174 (+-0.0306)	0.3862 (+-0.0146)	0.5288 (+-0.0109)
FC-Siam-conc [23]			MIOU	0.5592 (+-0.0892)	0.3481 (+-0.0366)	0.4942 (+-0.0156)
ChangeFormer [21]	✓	FL	MIOU	0.736 (+-0.0448)	0.4586 (+-0.0185)	0.584 (+-0.0101)
ChangeFormer [21]			MIOU	0.4848 (+-0.0923)	0.305 (+-0.0627)	0.4545 (+-0.0644)
DTCDSCN [29]	✓	FL	MIOU	0.7208 (+-0.0286)	0.4602 (+-0.0099)	0.5935 (+-0.002)
DTCDSCN [29]			MIOU	0.6844 (+-0.0393)	0.441 (+-0.0236)	0.5815 (+-0.0206)
BIT [25]	✓	FL	MIOU	0.7303 (+-0.0158)	0.4598 (+-0.0086)	0.5887 (+-0.0058)
BIT [25]			MIOU	0.6242 (+-0.0418)	0.4074 (+-0.0227)	0.5587 (+-0.0151)
SiROC [39]				0.6946	0.4408	0.5769

the results are visibly better in 2d. The plain FC-Siam-Diff is thrown off by different shades of green which results in false positives in the middle and on the right. The pseudo label version helps to reduce these false positives due to acquisition conditions and further seems to improve not only the location but also the shapes of segmented changes.

As also visible in Table I, the segmentation performance of ChangeFormer and BIT is generally worse in comparison to FC-Siam-Diff. SiROC pseudo labels brought the biggest improvement for ChangeFormer in Table I which is also visible in 2f and 2g. The no pseudo label version predicts change for virtually all grassland regions since it interprets the change in illumination as change. It is therefore too sensitive to the change class and struggles to extract meaningful change. This improves visibly with the pseudo label training. For example, the shapes in the middle are fit notably better.

Similarly, the pseudo labels bring improvement with BIT as shapes get more refined and there are fewer false positives on the right.

The impressions of Figure 2 are generally confirmed when inspecting predictions for a more complex urban scene in Figure 3. Again, the upper panels for each method show pre

and post images as well as the ground truth. For all three models, the upper prediction with pseudo label pretraining shows more refined shapes. This becomes particularly visible for ChangeFormer (3f, 3g) and BIT (3h, 3i) where the predictions without pseudo labels are visibly more blurry and overall worse. The difference is smaller for FC-Siam-diff but the no pseudo label version 3i predicts a number of false positives that are predicted correctly with pseudo labels 3d particularly on the left and center-right. On the other hand, both models miss key changes in this complex scene where the no pseudo label variant seems keener on classifying something as a change. Overall, the qualitative inspection of scenes confirms our finding that confidence-filtered pseudo labels help increase change detection performance.

Table I shows that pseudo label training is effective in addition to supervised use of labels. Table II outlines what happens when other pseudo labels based on CVA or DCVA are used as semi-supervised baselines. The training set-up is identical to Table I and the scores for SiROC PL are the same. What varies is the source of the pseudo labels in the pretraining step listed in the second column. FC-Siam-Diff with SiROC pseudo labels reaches high scores in accuracy and MIOU.

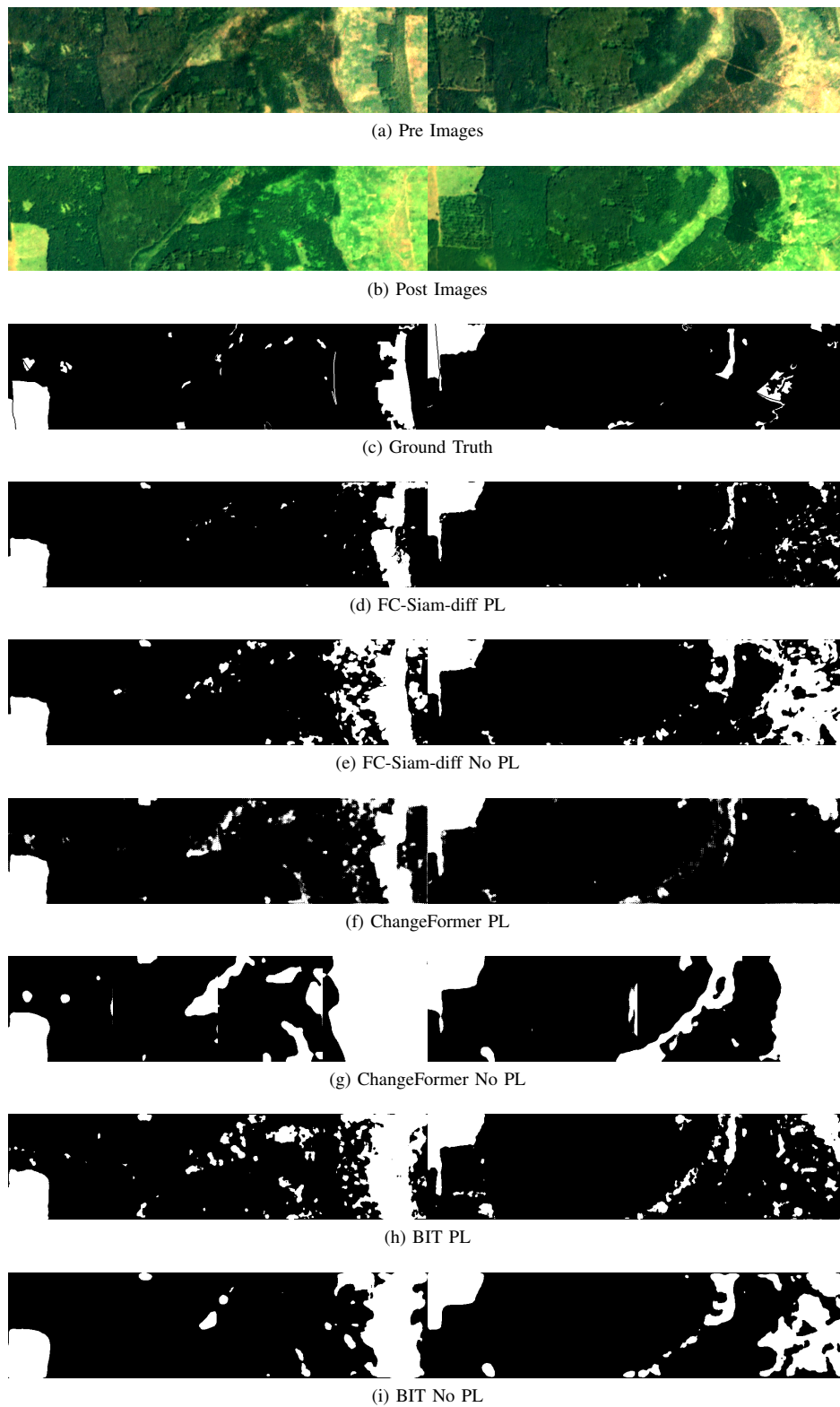


Fig. 2. Qualitative results of 8 sample image pairs with ground truth and respective model predictions with and without pseudo labels. In general, the pseudo labels seem to help the models reduce false positives based on illumination differences. Examples of this are deforestation in the middle and on the right.

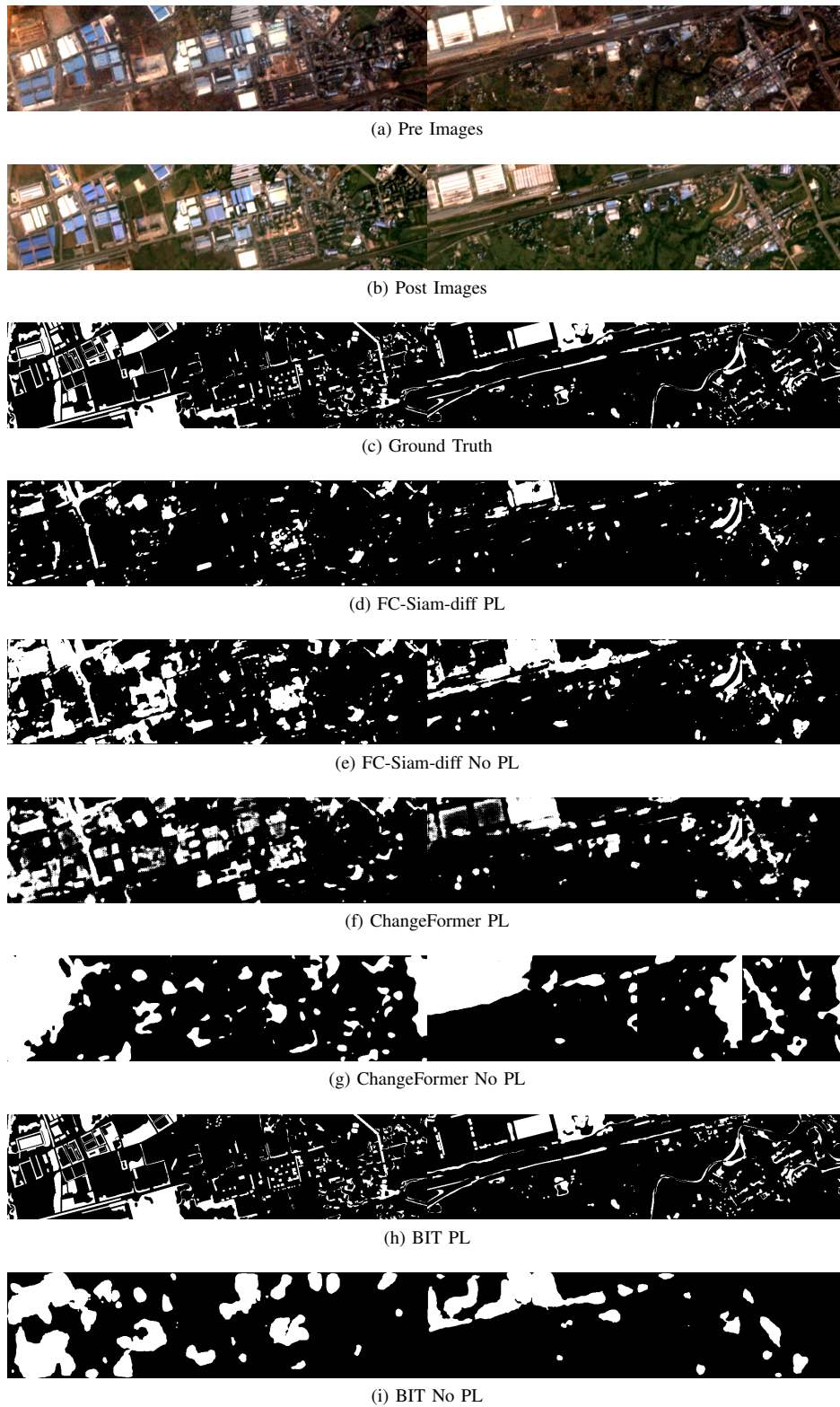


Fig. 3. Qualitative results of 8 sample image pairs with ground truth and respective model predictions with and without pseudo labels here for a complex urban scene.

TABLE II  
QUANTITATIVE RESULTS DYNAMIC EARTHNET WITH DIFFERENT PSEUDO LABELS

Model	PL	Loss PL	Loss	Accuracy	MIOU	MF1
FC-Siam-diff [23]	SiROC	FL	MIOU	<b>0.7812</b> (+-0.0104)	<b>0.4854</b> (+-0.0037)	0.6029 (+-0.0018)
FC-Siam-diff [23]	CVA	FL	MIOU	0.7599 (+-0.0124)	0.4853 (+-0.0072)	<b>0.6121</b> (+-0.0048)
FC-Siam-diff [23]	DCVA	FL	MIOU	0.7553 (+-0.0077)	0.4838 (+-0.0015)	<b>0.6121</b> (+-0.002)
ChangeFormer [21]	SiROC	FL	MIOU	0.736 (+-0.0448)	0.4586 (+-0.0185)	0.584 (+-0.0101)
ChangeFormer [21]	CVA	FL	MIOU	0.6589 (+-0.0414)	0.4232 (+-0.0254)	0.5666 (+-0.0196)
ChangeFormer [21]	DCVA	FL	MIOU	0.678 (+-0.0264)	0.4423 (+-0.0193)	0.5864 (+-0.0166)
BIT [25]	SiROC	FL	MIOU	0.7303 (+-0.0158)	0.4598 (+-0.0086)	0.5887 (+-0.0058)
BIT [25]	CVA	FL	MIOU	0.6886 (+-0.0091)	0.4437 (+-0.006)	0.5839 (+-0.0049)
BIT [25]	DCVA	FL	MIOU	0.7004 (+-0.0117)	0.4543 (+-0.006)	0.594 (+-0.0038)

Accuracy is 2-3 p.p. higher compared to other pseudo labels which is significant but the MIOU edge is rather small. For MF1 it seems that CVA and DCVA pseudo labels, although lacking behind in accuracy, reach a slightly more balanced classification with 61.21% MF1 each. For ChangeFormer and BIT the scores are again lower on average. Compared to CVA, the Change Former SiROC combination scores visibly better across all three categories (+ 8 p.p. accuracy, + 3 p.p. MIOU, + 2 p.p. MF1). ChangeFormer with SiROC pseudo labels notably exceeds accuracy and MIOU compared to its DCVA baseline and obtains a similar MF1 score. The picture for BIT is similar with higher accuracy and MIOU and slightly better (CVA) or marginally worse (DCVA) F1 scores. Overall, SiROC pseudo labels perform visibly better in accuracy and MIOU where the edge is particularly apparent for ChangeFormer and BIT.

#### D. DynamicEarthNet Ablation Studies

*Amount of training data.* One may be concerned that the edge of our approach is limited by the small number of training cubes with real labels. Therefore, we iteratively add more training cubes to explore differences in the edge depending on this parameter. Table III presents these scores on a harmonized test set for this Table. As we use up to 64 cubes for training and aim to keep the scores comparable, we use the respective test set for all specifications in this Table. All pseudo label specifications are again pretrained with the top 25% of cubes in confidence. We use all available training cubes with FC-Siam-diff and Change Former in the upper panel. Despite the increasing amount of training data, FC-Siam-diff remains better than ChangeFormer by a significant margin. In both specifications, SemiSiROC exceeds the no pseudo label baseline again visibly.

In the lower panel, we compare FC-Siam-diff against versions with fewer training data (25% and 50% of the above training set). Interestingly, the performance of SemiSiROC increases only marginally with additional real training data. This may indicate that a large part of potential gains through additional training data could already have been exploited by the pseudo labels. Conversely, the gap between PL and no PL gets smaller with 16 training cubes. Then, performance from 16 to 32 cubes drops slightly which is unexpected. One reason could be that the additional training cubes are somewhat more unrepresentative of the remaining cubes on the other side of the globe compared to the previous cubes. The highest scores with

and without pseudo labels are achieved with the maximum number of training cubes of 64 which is about 85% of our dataset with over 1000 image pairs where the rest is used for testing and validation. Still, the pseudo label specification remains better than its baseline with a sizeable gap. Overall, the main takeaway remains unaffected. With both a few and a larger amount of labels, SemiSiROC is an effective strategy for change detection on this dataset.

*Varying the finetuning loss.* However, the edge of our strategy may be specific to the loss combination used. Therefore, we test the robustness of our results with other losses at the finetuning step in Table IV for ChangeFormer, BIT and FC-Siam-diff. We do not vary the pseudo label loss here as this would leave the baselines without SiROC pretraining unaffected. In total, there are six specifications per model given three loss combinations each. The MIOU scores are identical to Table I.

The choice of the finetuning loss leaves SemiSiROC largely unaffected with minor differences in scores. It is marginally better in accuracy and MIOU compared to the MIOU loss and slightly lower in terms of Mean F1. The focal loss baseline with FC-Siam-diff is slightly stronger than with MIOU but still lacks behind the comparable SemiSiROC specification by about 9 p.p. in accuracy, 4 p.p. in MIOU and 2 p.p. in Mean F1.

Expectedly, training with a cross-entropy (CE) loss pushes the FC-Siam-diff baseline to almost exclusively predict the majority no change class. This results in an accuracy high score of almost 0.80 which even marginally surpasses the respective SemiSiROC score although with a higher standard deviation. However, the corresponding Mean F1 score which is comparably sensitive to large discrepancies in predictive performance across the classes falls behind by almost 7 p.p. to the SemiSiROC CE score.

For the ChangeFormer model, the observations of the MIOU finetuning seem to be confirmed. Similar to FC-Siam-diff, CE training leads to the prediction of mostly no change. The FL results are somewhat better than the MIOU results but still comparably bad. Overall, Table IV confirms the impression of the effectiveness of our semi-supervised strategy.

At last, the results for the BIT model mirror the above results. Pseudo labeling is highly effective across all categories with an FL or MIOU loss. With CE the model again tends to overfit largely to the no-change class which is why the

TABLE III  
ABLATION STUDY: VARYING THE TRAINING SET SIZE

Model	PL	# Training Cubes	Loss	Accuracy	MIOU	MF1
FC-Siam-diff [23]	✓	64	MIOU	<b>0.9227</b> (+-0.0038)	<b>0.5376</b> (+-0.0012)	<b>0.6127</b> (+-0.0028)
FC-Siam-diff [23]		64	MIOU	0.8538 (+-0.0076)	0.4865 (+-0.0055)	0.5685 (+-0.0048)
ChangeFormer [21]	✓	64	MIOU	0.813 (+-0.0113)	0.4613 (+-0.0098)	0.5494 (+-0.0102)
ChangeFormer [21]		64	MIOU	0.7792 (+-0.0277)	0.4528 (+-0.0239)	0.5516 (+-0.0449)
FC-Siam-diff [23]	✓	32	MIOU	0.9159 (+-0.0115)	0.5324 (+-0.0094)	0.6082 (+-0.0088)
FC-Siam-diff [23]		32	MIOU	0.8215 (+-0.0715)	0.4764 (+-0.031)	0.5681 (+-0.0328)
FC-Siam-diff [23]	✓	16	MIOU	0.9162 (+-0.0127)	0.5338 (+-0.007)	0.6101 (+-0.0047)
FC-Siam-diff [23]		16	MIOU	0.8488 (+-0.0205)	0.4851 (+-0.0107)	0.569 (+-0.0074)

TABLE IV  
ABLATION STUDY: ROBUSTNESS TO FINETUNING LOSS

Model	PL	Loss PL	Loss	Accuracy	MIOU	MF1
FC-Siam-diff [23]	✓	FL	FL	0.787 (+-0.0088)	0.4858 (+-0.0021)	0.6008 (+-0.0051)
FC-Siam-diff [23]			FL	0.693 (+-0.0657)	0.4426 (+-0.0246)	0.5798 (+-0.0163)
FC-Siam-diff [23]	✓	FL	MIOU	0.7812 (+-0.0104)	0.4854 (+-0.0037)	<b>0.6029</b> (+-0.0018)
FC-Siam-diff [23]			MIOU	0.6359 (+-0.0405)	0.419 (+-0.0288)	0.5706 (+-0.0244)
FC-Siam-diff [23]	✓	FL	CE	0.7945 (+-0.0088)	<b>0.4868</b> (+-0.0043)	0.5987 (+-0.0096)
FC-Siam-diff [23]			CE	0.7988 (+-0.0233)	0.4466 (+-0.0219)	0.5304 (+-0.0483)
ChangeFormer [21]	✓	FL	FL	0.6762 (+-0.0538)	0.4355 (+-0.034)	0.5769 (+-0.027)
ChangeFormer [21]			FL	0.5644 (+-0.0164)	0.3548 (+-0.0085)	0.5036 (+-0.0088)
ChangeFormer [21]	✓	FL	MIOU	0.736 (+-0.0448)	0.4586 (+-0.0185)	0.584 (+-0.0101)
ChangeFormer [21]			MIOU	0.4848 (+-0.0923)	0.305 (+-0.0627)	0.4545 (+-0.0644)
ChangeFormer [21]	✓	FL	CE	<b>0.8068</b> (+-0.0122)	0.4399 (+-0.0158)	0.5155 (+-0.0321)
ChangeFormer [21]			CE	0.7735 (+-0.0471)	0.4237 (+-0.0088)	0.5067 (+-0.0178)
BIT [25]	✓	FL	FL	0.7133 (+-0.0203)	0.4531 (+-0.0088)	0.5864 (+-0.0066)
BIT [25]			FL	0.6673 (+-0.0774)	0.412 (+-0.0318)	0.5447 (+-0.0222)
BIT [25]	✓	FL	MIOU	0.7303 (+-0.0158)	0.4598 (+-0.0086)	0.5887 (+-0.0058)
BIT [25]			MIOU	0.6242 (+-0.0418)	0.4074 (+-0.0227)	0.5587 (+-0.0151)
BIT [25]	✓	FL	CE	0.7593 (+-0.0145)	0.4639 (+-0.0027)	0.581 (+-0.0098)
BIT [25]			CE	0.7876 (+-0.0256)	0.4236 (+-0.0055)	0.4984 (+-0.0139)
SiROC				0.6946	0.4408	0.5769

accuracies are higher. Even though the no PL version with CE loss reaches the highest accuracy among BIT models, the results are visibly unbalanced. While the PL version lacks behind 3 p.p. in accuracy, it makes more balanced choices with more than 8 p.p. more MF1.

*Results on unseen geographic areas.* Note that for the two previous Tables, we did not restrict the pseudo labels to be outside of the test set. While during training no model sees any actual labels from the test set, one could argue that the images of the test set may be advantageous for our strategy.

To ensure that our strategy is effective also on cubes that were also not part of the pseudo label training, we split the former test set in two where we use the western half from the perspective of Figure 1 for pseudo label training and the eastern half for testing with the FC-Siam-diff as the most effective model overall. The respective scores are reported in Table V and can not be directly compared to the scores of previous Tables anymore because of the difference in the test cubes. Still, the pseudo label step remains better in comparison by a wide margin that seems even bigger than in previous comparisons. The gap is substantial at 15 p.p. in accuracy and 7 p.p. in MIOU.

*Pseudo label filtering.* Another ablation study concerns the effectiveness of the pseudo label filtering. Since labels are

limited, the preselection discards additional information which may be useful in training. Therefore, we mix up the cube selection with a random selection and the lowest 25% in confidence. The respective results are reported in Table VI. The top 25% cubes score best in terms of accuracy and MIOU and fall just short of the random selection in terms of MF1. Still, with a difference of almost 3 p.p. with similar MIOU and F1 values, it seems that the confidence prefiltering indeed extracts meaningful pseudo labels which result in more effective learning. Additionally, we notice decreasing marginal returns of adding a higher fraction of pseudo labels in our case. Using the top half or even all cubes with their respective pseudo labels results in similar performance than only using the top quarter. Therefore, we choose the threshold of 25% for more efficient training. Even though SiROC pseudo labels improve performance already without filtering, the confidence selection further pushes change detection performance.

#### E. OSCD results

To further investigate the transferability and generizability of the proposed approach, we evaluate SemiSiROC also on OSCD [23] which is a widely-used binary change detection benchmark based on Sentinel-2 with a focus on urban regions. The results of our experiments are presented in Table VII.

TABLE V  
ABLATION STUDY: PL TRAINING NOT ON TEST IMAGES WITH SIAMUNET

Model	PL	Loss PL	Loss	Accuracy	MIOU	MF1
FC-Siam-diff [23]	✓	FL	MIOU	<b>0.7541</b> (+-0.0115)	<b>0.4621</b> (+-0.004)	<b>0.581</b> (+-0.0017)
FC-Siam-diff [23]			MIOU	0.5965 (+-0.0419)	0.3902 (+-0.0286)	0.5448 (+-0.0246)

TABLE VI  
ABLATION STUDY: DIFFERENT CONFIDENCE SPLITS

Model	PL Split	Loss PL	Loss	Accuracy	MIOU	MF1
FC-Siam-diff [23]	Top 25%	FL	MIOU	0.7812 (+-0.0104)	<b>0.4854</b> (+-0.0037)	0.6029 (+-0.0018)
FC-Siam-diff [23]	Random 25%	FL	MIOU	0.7541 (+-0.0113)	0.4801 (+-0.002)	<b>0.6074</b> (+-0.003)
FC-Siam-diff [23]	Bottom 25%	FL	MIOU	0.7576 (+-0.0093)	0.4767 (+-0.005)	0.6009 (+-0.0075)
FC-Siam-diff [23]	Top 50%	FL	MIOU	0.7794 (+-0.0067)	0.4839 (+-0.004)	0.6016 (+-0.0031)
FC-Siam-diff [23]	All	FL	MIOU	<b>0.787</b> (+-0.0055)	0.4824 (+-0.0043)	0.5958 (+-0.0054)

TABLE VII  
QUANTITATIVE RESULTS OSCD TEST SET TRAINED ON DYNAMICEARTHNET AND GROUPED BY PSEUDO LABEL USE. THESE ARE THE MODELS OF TABLE I APPLIED TO THE OSCD TEST SET WITHOUT RETRAINING.

Model	PL	Loss PL	Loss	Accuracy	MIOU	MF1
FC-Siam-diff [23]	✓	FL	MIOU	<b>0.9575</b> (+-0.0096)	0.5547 (+-0.0185)	0.6206 (+-0.0252)
FC-Siam-diff [23]			MIOU	0.8083 (+-0.1035)	0.4927 (+-0.0892)	0.5966 (+-0.082)
ChangeFormer [21]	✓	FL	MIOU	0.8592 (+-0.0692)	0.5145 (+-0.0457)	0.6085 (+-0.0356)
ChangeFormer [21]			MIOU	0.384 (+-0.2976)	0.2139 (+-0.1703)	0.2984 (+-0.1892)
BIT [25]	✓	FL	MIOU	0.9248 (+-0.0154)	<b>0.5585</b> (+-0.0115)	<b>0.6422</b> (+-0.012)
BIT [25]			MIOU	0.7273 (+-0.059)	0.4082 (+-0.0321)	0.5066 (+-0.0238)

The models used are identical to the ones in Table I. We merely apply them to the OSCD test set instead of the DynamicEarthNet test set directly to analyze the transferability of models. Similar to Table I, we test a variation with additional pseudo label pretraining and without it for each model. At first, FC-Siam-diff [23] remains a strong model and achieves an average accuracy of above 95% with a MIOU of 55.47% and a MF1 score of 62.06% across the five runs. There is a notable difference across all three scoring criteria between the pseudo label and the no pseudo label version. Most significantly, accuracy drops about 15 p.p. without DynamicEarthNet based pseudo label pretraining. This is the case even though both models were trained with real DynamicEarthNet labels. Interestingly, the accuracies are in the range (94-96%) of FC-Siam models in [23] based on supervised training on OSCD whereas our approach does not use OSCD labels at all. The contrast to no pseudo labels gets even larger for ChangeFormer although some of the ChangerFormer models seem to tilt towards predicting mostly change on this dataset which results in unstable average performance. Even when excluding these runs, however, the maximum performance of ChangeFormer on the OSCD test set is 74.13% accuracy, 41.86% MIOU and 51.71% which is substantially below the average with pseudo labels. Third, BIT model pseudo labels is arguably the best model here since it is only slightly inferior to FC-Siam-diff in accuracy but achieves high scores in MIOU and MF1 with 55.85% and 64.22% respectively. Again, the difference to no pseudo labels is large across all categories. Overall, the

OSCD results confirm the previous impression that pseudo label pretraining with SemiSiROC can be highly effective in optical CD applications.

#### IV. DISCUSSION

*Comparing teacher and students* The previous section outlines the effectiveness of SiROC as an unsupervised teacher model for change detection with limited labels. This is because it is an effective method and can prioritize pseudo labels based on a well-calibrated confidence. The mechanism for these improvements seems to be higher robustness to false positives because of acquisition conditions and more refined shapes of changes.

Since SiROC models analyze how much a pixel changes in comparison to its neighborhood, it seems intuitive that it would guide a student model towards higher robustness to false positives. Consider the example of Figure 2. Grassland seems much greener in the post images but since this affects virtually all pixels in the grassland neighborhood of a pixel, SiROC would not necessarily view this as change. This is something the student models seem to pick up on without modeling this explicitly. Another property of SemiSiROC seems to be more refined change shapes which is also a strength of the initial SiROC model [39]. This may incentivize the student model to learn more about likely shapes and spatial dependencies of changes.

*Relative weakness of transformer models* Second, we notice that throughout our results the two transformer models seem



to perform worse compared to the siamese UNet. This results in large gains through pseudo label pretraining and underlines the effectiveness of our strategy. There are several possible explanations for this relative weakness. A likely candidate is model size and label availability. ChangeFormer, in particular, is a large model which makes it data hungry and its success on other datasets such as Levir-CD in [21] may be related to the fact that more labels are available there. This seems plausible for Levir-CD which was about 10x more labeled pixels than the binary DynamicEarthNet we use here.

However, DSIFN only has 25% more labeled pixels than our dataset. Therefore, another reason could be that both of these methods have been tested in the context of urban change detection only with a focus on buildings. Maybe the different kinds of change applications across the globe within DynamicEarthNet pose a challenge to these models and the smaller siamese model adjusts to this more quickly. Nevertheless, the SemiSiROC framework shows effectiveness for all the methods we tested here and shows promise for change detection applications with optical data in practice. Our model pre-trained with pseudo labels converges faster during fine-tuning (i.e., training with actual labels). Thus, our proposed method reduces the time requirement of the training phase with actual samples.

## V. CONCLUSION

Monitoring changes of the Earth's surface over time with satellite imagery is an integral part of remote sensing. In this paper, we combine unsupervised and supervised techniques in a semi-supervised framework. This framework, called SemiSiROC, relies on pretraining a student model with pseudo labels that we filter by confidence. This enables the student model to learn from additional, meaningful high-confidence examples in a pretraining step before finetuning with actual labels. We evaluate SemiSiROC with three different supervised backbones: FC-Siam-Diff, ChangeFormer, and BIT. We evaluate the models with and without filtered pseudo label pretraining on a binary version of the DynamicEarthNet benchmark that is based on Planet Fusion imagery with 3m resolution. We pick only the cubes with the 25% highest confidence scores during pretraining. For all three models, we find a notable boost in performance for our baseline specification in Table I with 8 cubes which corresponds to 124 training scene pairs with real labels. Additionally, we outline that SemiSiROC remains competitive in the eye of semi-supervised student-teacher baselines based on DCVA and CVA pseudo labels.

Further, we evaluate the SemiSiROC models on scenes not seen during pseudo label training which results in similar performance gains. This ensures that the learned features are not specific to scenes close to the pseudo labels. Even with 64 training cubes with over 1000 labeled pairs, SemiSiROC is effective compared to its non-pseudo label baseline where gains are still large. Additional evaluations on the OSCD benchmark confirm the effectiveness of our SemiSiROC strategy also on an urban CD dataset based on Sentinel-2. Qualitative inspections of the predictions shed light on what the teacher model

seems to teach its students: Compared to its no pseudo label counterparts, the SemiSiROC models predict more refined shapes and seem to be less sensitive to false positives.

Our results point towards several potentially promising future research directions. At first, our work could be applied to related tasks such as multi-class change detection or different input sensors. Second, more experiments are necessary to understand the role of teacher models in spatial generalization generally and particularly in change detection.

## REFERENCES

- [1] J. A. Cardille, E. Perez, M. A. Crowley, M. A. Wulder, J. C. White, and T. Hermosilla, "Multi-sensor change detection for within-year capture and labelling of forest disturbance," *Remote Sensing of Environment*, vol. 268, p. 112741, 2022.
- [2] G. Chen and G. J. Hay, "An airborne lidar sampling strategy to model forest canopy height from quickbird imagery and geobias," *Remote Sensing of Environment*, vol. 115, no. 6, pp. 1532–1542, 2011.
- [3] C. Senf and R. Seidl, "Mapping the forest disturbance regimes of europe," *Nature Sustainability*, vol. 4, no. 1, pp. 63–70, 2021.
- [4] D. Lu, E. Moran, and S. Hetrick, "Detection of impervious surface change with multitemporal landsat images in an urban–rural frontier," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 66, no. 3, pp. 298–306, 2011.
- [5] S. Ji, Y. Shen, M. Lu, and Y. Zhang, "Building instance change detection from large-scale aerial images using convolutional neural networks and simulated samples," *Remote Sensing*, vol. 11, no. 11, p. 1343, 2019.
- [6] Y. Gao, F. Gao, J. Dong, and S. Wang, "Transferred deep learning for sea ice change detection from synthetic-aperture radar images," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 10, pp. 1655–1659, 2019.
- [7] K. Rokni, A. Ahmad, K. Solaimani, and S. Hazini, "A new approach for surface water change detection: Integration of pixel level image fusion and image classification techniques," *International Journal of Applied Earth Observation and Geoinformation*, vol. 34, pp. 226 – 234, 2015. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0303243414001780>
- [8] R. Gupta, B. Goodman, N. Patel, R. Hosfelt, S. Sajeew, E. Heim, J. Doshi, K. Lucas, H. Choset, and M. Gaston, "Creating xbd: A dataset for assessing building damage from satellite imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 10–17.
- [9] Z. Lv, H. Huang, L. Gao, J. A. Benediktsson, M. Zhao, and C. Shi, "Simple multiscale unet for change detection with heterogeneous remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, 2022.
- [10] Z. Lv, F. Wang, G. Cui, J. A. Benediktsson, T. Lei, and W. Sun, "Spatial–spectral attention network guided with change magnitude image for land cover change detection using remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.
- [11] L. Moya, A. Muhari, B. Adriano, S. Koshimura, E. Mas, L. R. Marval-Perez, and N. Yokoya, "Detecting urban changes using phase correlation and l1-based sparse model for early disaster response: A case study of the 2018 sulawesi indonesia earthquake-tsunami," *Remote Sensing of Environment*, vol. 242, p. 111743, 2020.
- [12] M. Zanetti, S. Saha, D. Marinelli, M. L. Magliozzi, M. Zavagli, M. Costantini, F. Bovolo, and L. Bruzzone, "A system for burned area detection on multispectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.
- [13] Z. Lv, T. Liu, J. A. Benediktsson, and N. Falco, "Land cover change detection techniques: Very-high-resolution optical images: A review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 1, pp. 44–63, 2022.
- [14] J. Aschbacher, "Esa's earth observation strategy and copernicus," in *Satellite earth observations and their impact on society and policy*. Springer, Singapore, 2017, pp. 81–86.
- [15] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort *et al.*, "Sentinel-2: Esa's optical high-resolution mission for gmes operational services," *Remote sensing of Environment*, vol. 120, pp. 25–36, 2012.
- [16] C. Kwan, X. Zhu, F. Gao, B. Chou, D. Perez, J. Li, Y. Shen, K. Koperski, and G. Marchisio, "Assessment of spatiotemporal fusion algorithms for planet and worldview images," *Sensors*, vol. 18, no. 4, p. 1051, 2018.

- [17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [18] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 5966–5978, 2021.
- [19] M. Wang, Q. Wang, D. Hong, S. K. Roy, and J. Chanussot, "Learning tensor low-rank representation for hyperspectral anomaly detection," *IEEE Transactions on Cybernetics*, 2022.
- [20] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.
- [21] W. G. C. Bandara and V. M. Patel, "A transformer-based siamese network for change detection," 2022.
- [22] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangquan, L. Huang, and G. Liu, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 183–200, 2020.
- [23] R. Caye Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *IEEE International Conference on Image Processing (ICIP)*, October 2018.
- [24] R. Caye Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral earth observation using convolutional neural networks," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, July 2018.
- [25] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.
- [26] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep siamese convolutional network for optical aerial images," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1845–1849, 2017.
- [27] H. Lyu, H. Lu, and L. Mou, "Learning a transferable change rule from a recurrent neural network for land cover change detection," *Remote Sensing*, vol. 8, no. 6, p. 506, 2016.
- [28] L. Mou, L. Bruzzone, and X. X. 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*, vol. 57, no. 2, pp. 924–935, 2018.
- [29] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 5, pp. 811–815, 2020.
- [30] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [31] H. Zhang, M. Lin, G. Yang, and L. Zhang, "Escnet: An end-to-end superpixel-enhanced change detection network for very-high-resolution remote sensing images," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
- [32] Q. Xu, K. Chen, X. Sun, Y. Zhang, H. Li, and G. Xu, "Pseudo-siamese capsule network for aerial remote sensing images change detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2020.
- [33] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10012–10022.
- [34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [35] J. Chen, Z. Yuan, J. Peng, L. Chen, H. Huang, J. Zhu, Y. Liu, and H. Li, "Dasnet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 1194–1206, 2020.
- [36] F. I. Diakogiannis, F. Waldner, and P. Caccetta, "Looking for change? roll the dice and demand attention," *Remote Sensing*, vol. 13, no. 18, p. 3707, 2021.
- [37] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised deep change vector analysis for multiple-change detection in vhr images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 6, pp. 3677–3693, 2019.
- [38] S. Saha, Y. T. Solano-Correa, F. Bovolo, and L. Bruzzone, "Unsupervised deep transfer learning-based change detection for hr multispectral images," *IEEE Geoscience and Remote Sensing Letters*, 2020.
- [39] L. Kondmann, A. Toker, S. Saha, B. Schölkopf, L. Leal-Taixé, and X. X. Zhu, "Spatial context awareness for unsupervised change detection in optical satellite images," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.
- [40] C. Ren, X. Wang, J. Gao, and H. Chen, "Unsupervised change detection in satellite images with generative adversarial network," *arXiv preprint arXiv:2009.03630*, 2020.
- [41] N. Falco, G. Cavallaro, P. R. Marpu, and J. A. Benediktsson, "Unsupervised change detection analysis to multi-channel scenario based on morphological contextual analysis," in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2016, pp. 3374–3377.
- [42] S. Saha, L. Mou, C. Qiu, X. X. Zhu, F. Bovolo, and L. Bruzzone, "Unsupervised deep joint segmentation of multitemporal high-resolution images," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.
- [43] M. Gong, X. Niu, P. Zhang, and Z. Li, "Generative adversarial networks for change detection in multispectral imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 12, pp. 2310–2314, 2017.
- [44] T. Zhan, M. Gong, X. Jiang, and M. Zhang, "Unsupervised scale-driven change detection with deep spatial-spectral features for vhr images," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.
- [45] W. A. Malila, "Change vector analysis: an approach for detecting forest changes with landsat," in *LARS symposia*, 1980, p. 385.
- [46] F. Bovolo, "A multilevel parcel-based approach to change detection in very high resolution multitemporal images," *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 1, pp. 33–37, 2008.
- [47] F. Thonfeld, H. Feilhauer, M. Braun, and G. Menz, "Robust change vector analysis (rcva) for multi-sensor very high resolution optical satellite data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 50, pp. 131–140, 2016.
- [48] L. Li, X. Li, Y. Zhang, L. Wang, and G. Ying, "Change detection for high-resolution remote sensing imagery using object-oriented change vector analysis method," in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2016, pp. 2873–2876.
- [49] F. Bovolo, L. Bruzzone, and M. Marconcini, "A novel approach to unsupervised change detection based on a semisupervised svm and a similarity measure," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 7, pp. 2070–2082, 2008.
- [50] S. Ghosh, M. Roy, and A. Ghosh, "Semi-supervised change detection using modified self-organizing feature map neural network," *Applied Soft Computing*, vol. 15, pp. 1–20, 2014.
- [51] K. Chen, Z. Zhou, C. Huo, X. Sun, and K. Fu, "A semisupervised context-sensitive change detection technique via gaussian process," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 2, pp. 236–240, 2012.
- [52] Y. Yuan, H. Lv, and X. Lu, "Semi-supervised change detection method for multi-temporal hyperspectral images," *Neurocomputing*, vol. 148, pp. 363–375, 2015.
- [53] S. Saha, L. Mou, X. X. Zhu, F. Bovolo, and L. Bruzzone, "Semisupervised change detection using graph convolutional network," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 4, pp. 607–611, 2021.
- [54] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10687–10698.
- [55] H. Wu and S. Prasad, "Semi-supervised deep learning using pseudo labels for hyperspectral image classification," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1259–1270, 2017.
- [56] M. Gong, H. Yang, and P. Zhang, "Feature learning and change feature classification based on deep learning for ternary change detection in sar images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 129, pp. 212–225, 2017.
- [57] M. Gong, Y. Yang, T. Zhan, X. Niu, and S. Li, "A generative discriminatory classified network for change detection in multispectral imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 1, pp. 321–333, 2019.
- [58] Y. Li, C. Peng, Y. Chen, L. Jiao, L. Zhou, and R. Shang, "A deep learning method for change detection in synthetic aperture radar images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 5751–5763, 2019.
- [59] F. Gao, X. Wang, Y. Gao, J. Dong, and S. Wang, "Sea ice change detection in sar images based on convolutional-wavelet neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 8, pp. 1240–1244, 2019.

- [60] S. Saha, B. Banerjee, and X. X. Zhu, "Trusting small training dataset for supervised change detection," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE, 2021, pp. 2031–2034.
- [61] A. Toker, L. Kondmann, M. Weber, M. Eisenberger, A. Camero, J. Hu, A. P. Hoderlein, C. Senaras, T. Davis, D. Cremers, G. Marchisio, X. X. Zhu, and L. Leal-Taié, "Dynamicearthnet: Daily multi-spectral satellite dataset for semantic change segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
- [62] I. Z. Yalniz, H. Jégou, K. Chen, M. Paluri, and D. Mahajan, "Billion-scale semi-supervised learning for image classification," *arXiv preprint arXiv:1905.00546*, 2019.
- [63] L. Kondmann, A. Toker, M. Rußwurm, A. Camero Unzueta, D. Peresuti, G. Milcinski, N. Longépé, P.-P. Mathieu, T. Davis, G. Marchisio *et al.*, "Denethor: The dynamicearthnet dataset for harmonized, interoperable, analysis-ready, daily crop monitoring from space," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.
- [64] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.
- [65] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sensing*, vol. 12, no. 10, 2020.



**Lukas Kondmann** received the bachelor degree in Economics from the Ludwig-Maximilians-University Munich in 2016 and the master degree in Social Data Science from the University of Oxford in 2019. He holds an honors degree in Technology Management (2015-2017) from the Center for Digital Technology Management in Munich and was a visiting researcher working on big data for social good at the UC Berkeley School of Information in spring 2017.

He is currently pursuing the Ph.D. degree in engineering at the Technical University of Munich and the German Aerospace Center in Munich. His research is centered around time-series analysis of multispectral remote sensing imagery with a focus on monitoring the Sustainable Development Goals (SDGs).



**Sudipan Saha** (S'16–M'20) received the PhD degree in information and communication technologies from the University of Trento, Trento, Italy, and Fondazione Bruno Kessler, Trento, Italy in 2020. Previously, he obtained the M.Tech. degree in electrical engineering from IIT Bombay, Mumbai, India, in 2014.

He is currently an assistant professor at the Yardi School of Artificial Intelligence, Indian Institute of Technology Delhi. Previously, he was a postdoctoral researcher at Technical University of Munich (TUM), Munich, Germany and he worked as an Engineer with TSMC Limited, Hsinchu, Taiwan, from 2015 to 2016. In 2019, he was a Guest Researcher with the Technical University of Munich (TUM), Munich, Germany. He is the recipient of Fondazione Bruno Kessler Best Student Award 2020. His research interests are related to multitemporal remote sensing image analysis, domain adaptation, time-series analysis, image segmentation, deep learning, image processing, and pattern recognition. Dr. Saha is a Reviewer for several international journals and served as a guest editor at Remote Sensing (MDPI) special issue on "Advanced Artificial Intelligence for Remote Sensing: Methodology and Application".



**Xiao Xiang Zhu** (S'10–M'12–SM'14–F'21) received the Master (M.Sc.) degree, her doctor of engineering (Dr.-Ing.) degree and her "Habilitation" in the field of signal processing from Technical University of Munich (TUM), Munich, Germany, in 2008, 2011 and 2013, respectively.

She is the Chair Professor for Data Science in Earth Observation at Technical University of Munich (TUM) and was the founding Head of the Department "EO Data Science" at the Remote Sensing Technology Institute, German Aerospace Center (DLR). Since 2019, Zhu is a co-coordinator of the Munich Data Science Research School ([www.mu-ds.de](http://www.mu-ds.de)). Since 2019 She also heads the Helmholtz Artificial Intelligence – Research Field "Aeronautics, Space and Transport". Since May 2020, she is the PI and director of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond", Munich, Germany. Since October 2020, she also serves as a co-director of the Munich Data Science Institute (MDSI), TUM. Prof. Zhu was a guest scientist or visiting professor at the Italian National Research Council (CNR-IREA), Naples, Italy, Fudan University, Shanghai, China, the University of Tokyo, Tokyo, Japan and University of California, Los Angeles, United States in 2009, 2014, 2015 and 2016, respectively. She is currently a visiting AI professor at ESA's Phi-lab. Her main research interests are remote sensing and Earth observation, signal processing, machine learning and data science, with their applications in tackling societal grand challenges, e.g. Global Urbanization, UN's SDGs and Climate Change.

Dr. Zhu is a member of young academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina and the Bavarian Academy of Sciences and Humanities. She serves in the scientific advisory board in several research organizations, among others the German Research Center for Geosciences (GFZ) and Potsdam Institute for Climate Impact Research (PIK). She is an associate Editor of IEEE Transactions on Geoscience and Remote Sensing and serves as the area editor responsible for special issues of IEEE Signal Processing Magazine. She is a Fellow of IEEE.

## BIBLIOGRAPHY

---

- [1] Pablo F Alcantarilla, Simon Stent, German Ros, Roberto Arroyo, and Riccardo Gherardi. "Street-view change detection with deconvolutional networks." In: *Autonomous Robots* 42 (2018), pp. 1301–1322.
- [2] Josef Aschbacher. "ESA's earth observation strategy and Copernicus." In: *Satellite earth observations and their impact on society and policy*. Springer, Singapore, 2017, pp. 81–86.
- [3] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. "Geography-aware self-supervised learning." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 10181–10190.
- [4] W Gedara Chaminda Bandara, N Gopalakrishnan Nair, and VM Patel. "Remote sensing change detection (segmentation) using denoising diffusion probabilistic models." In: *arXiv e-prints*, pp. arXiv-2206 (2022).
- [5] Wele Gedara Chaminda Bandara, Naman Patel, Ali Gholami, Mehdi Nikkhah, Motilal Agrawal, and Vishal M Patel. "AdaMAE: Adaptive Masking for Efficient Spatiotemporal Learning with Masked Autoencoders." In: *arXiv preprint arXiv:2211.09120* (2022).
- [6] Wele Gedara Chaminda Bandara and Vishal M. Patel. *A Transformer Based Siamese Network for Change Detection*. 2022.
- [7] Wele Gedara Chaminda Bandara and Vishal M Patel. "Revisiting consistency regularization for semi-supervised change detection in remote sensing images." In: (2022).
- [8] Mariana Belgiu, Wietske Bijker, Ovidiu Csillik, and Alfred Stein. "Phenology-based sample generation for supervised crop type classification." In: *International Journal of Applied Earth Observation and Geoinformation* 95 (2021), p. 102264.
- [9] Guillaume-Alexandre Bilodeau, Jean-Philippe Jodoin, and Nicolas Saunier. "Change detection in feature space using local binary similarity patterns." In: *2013 International Conference on Computer and Robot Vision*. IEEE. 2013, pp. 106–112.
- [10] Christopher Bocquet. "Dalberg Data Insights Uganda Crop Classification." In: <https://doi.org/10.34911/RDNT.EII04X> (2019).
- [11] Francesca Bovolo. "A multilevel parcel-based approach to change detection in very high resolution multitemporal images." In: *IEEE Geoscience and Remote Sensing Letters* 6.1 (2008), pp. 33–37.

- [12] Francesca Bovolo, Lorenzo Bruzzone, and Mattia Marconcini. "A novel approach to unsupervised change detection based on a semisupervised SVM and a similarity measure." In: *IEEE Transactions on Geoscience and Remote Sensing* 46.7 (2008), pp. 2070–2082.
- [13] György Büttner. "CORINE land cover and land cover change products." In: *Land use and land cover mapping in Europe: practices & trends* (2014), pp. 55–74.
- [14] Jeffrey A Cardille, Elijah Perez, Morgan A Crowley, Michael A Wulder, Joanne C White, and Txomin Hermosilla. "Multi-sensor change detection for within-year capture and labelling of forest disturbance." In: *Remote Sensing of Environment* 268 (2022), p. 112741.
- [15] R. Caye Daudt, B. Le Saux, and A. Boulch. "Fully Convolutional Siamese Networks for Change Detection." In: *IEEE International Conference on Image Processing (ICIP)* (Athens, Greece). 2018.
- [16] R. Caye Daudt, B. Le Saux, A. Boulch, and Y. Gousseau. "Urban Change Detection for Multispectral Earth Observation Using Convolutional Neural Networks." In: *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* (Valencia, Spain). 2018.
- [17] Gang Chen and Geoffrey J Hay. "An airborne lidar sampling strategy to model forest canopy height from Quickbird imagery and GEOBIA." In: *Remote Sensing of Environment* 115.6 (2011), pp. 1532–1542.
- [18] Hao Chen, Zipeng Qi, and Zhenwei Shi. "Remote sensing image change detection with transformers." In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021), pp. 1–14.
- [19] Hao Chen and Zhenwei Shi. "A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection." In: *Remote Sensing* 12.10 (2020).
- [20] Jie Chen, Ziyang Yuan, Jian Peng, Li Chen, Haozhe Huang, Jiawei Zhu, Yu Liu, and Haifeng Li. "DASNet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images." In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14 (2020), pp. 1194–1206.
- [21] Keming Chen, Zhixin Zhou, Chunlei Huo, Xian Sun, and Kun Fu. "A semisupervised context-sensitive change detection technique via Gaussian process." In: *IEEE Geoscience and Remote Sensing Letters* 10.2 (2012), pp. 236–240.

- [22] 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." In: *IEEE transactions on pattern analysis and machine intelligence* 40.4 (2017), pp. 834–848.
- [23] Zhaohua Chen and Jinfei Wang. "Land use and land cover change detection using satellite remote sensing techniques in the mountainous Three Gorges Area, China." In: *International Journal of Remote Sensing* 31.6 (2010), pp. 1519–1542.
- [24] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. "SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery." In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho. 2022.
- [25] Christopher Conrad, Stefan Dech, Olena Dubovyk, Sebastian Fritsch, Doris Klein, Fabian Löw, Gunther Schorcht, and Julian Zeidler. "Derivation of temporal windows for accurate crop discrimination in heterogeneous croplands of Uzbekistan using multitemporal RapidEye images." In: *Computers and Electronics in Agriculture* 103 (2014), pp. 63–74.
- [26] Christopher Conrad, Sebastian Fritsch, Julian Zeidler, Gerd Rucker, and Stefan Dech. "Per-field irrigated crop classification in arid Central Asia using SPOT and ASTER data." In: *Remote Sensing* 2.4 (2010), pp. 1035–1056.
- [27] Ovidiu Csillik, Mariana Belgiu, Gregory P Asner, and Maggi Kelly. "Object-based time-constrained dynamic time warping classification of crops using Sentinel-2." In: *Remote sensing* 11.10 (2019), p. 1257.
- [28] 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*. IEEE. 2018, pp. 2115–2118.
- [29] Rodrigo Caye Daudt, Bertrand Le Saux, Alexandre Boulch, and Yann Gousseau. "Multitask learning for large-scale semantic change detection." In: *Computer Vision and Image Understanding* 187 (2019), p. 102783.
- [30] Baudouin Desclée, Patrick Bogaert, and Pierre Defourny. "Forest change detection by statistical object-based method." In: *Remote sensing of environment* 102.1-2 (2006), pp. 1–11.

- [31] R Devadas, RJ Denham, and M Pringle. "Support vector machine classification of object-based data for crop mapping, using multi-temporal Landsat imagery." In: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 39.1 (2012), pp. 185–190.
- [32] Foivos I Diakogiannis, François Waldner, and Peter Caccetta. "Looking for change? Roll the dice and demand attention." In: *Remote Sensing* 13.18 (2021), p. 3707.
- [33] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." In: *International Conference on Learning Representations*. 2021.
- [34] 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." In: *Remote Sensing of Environment* 120 (2012), pp. 25–36.
- [35] Nicola Falco, Gabriele Cavallaro, Prashanth R Marpu, and Jon Atli Benediktsson. "Unsupervised change detection analysis to multi-channel scenario based on morphological contextual analysis." In: *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE. 2016, pp. 3374–3377.
- [36] Saskia Foerster, Klaus Kaden, Michael Foerster, and Sibylle Itzerott. "Crop type mapping using spectral–temporal profiles and phenological information." In: *Computers and Electronics in Agriculture* 89 (2012), pp. 30–40.
- [37] Feng Gao, Xiaopeng Liu, Junyu Dong, Guoqiang Zhong, and Muwei Jian. "Change detection in SAR images based on deep semi-NMF and SVD networks." In: *Remote Sensing* 9.5 (2017), p. 435.
- [38] Yunhao Gao, Feng Gao, Junyu Dong, and Shengke Wang. "Transferred deep learning for sea ice change detection from synthetic-aperture radar images." In: *IEEE Geoscience and Remote Sensing Letters* 16.10 (2019), pp. 1655–1659.
- [39] Vivien Sainte Fare Garnot and Loic Landrieu. "Lightweight Temporal Self-Attention for Classifying Satellite Image Time Series." In: *arXiv:2007.00586 [cs]* (July 1, 2020). (Visited on 07/06/2020).
- [40] Jie Geng, Xiaorui Ma, Xiaojun Zhou, and Hongyu Wang. "Saliency-guided deep neural networks for SAR image change detection." In: *IEEE Transactions on Geoscience and Remote Sensing* 57.10 (2019), pp. 7365–7377.



- [41] Susmita Ghosh, Moumita Roy, and Ashish Ghosh. "Semi-supervised change detection using modified self-organizing feature map neural network." In: *Applied Soft Computing* 15 (2014), pp. 1–20.
- [42] Maoguo Gong, Xudong Niu, Puzhao Zhang, and Zhetao Li. "Generative adversarial networks for change detection in multispectral imagery." In: *IEEE Geoscience and Remote Sensing Letters* 14.12 (2017), pp. 2310–2314.
- [43] Wyn Grant. *The common agricultural policy*. Bloomsbury Publishing, 1997.
- [44] Sabine Grunwald, Gustavo M Vasques, and Rosanna G Rivero. "Fusion of soil and remote sensing data to model soil properties." In: *Advances in Agronomy* 131 (2015), pp. 1–109.
- [45] Neha Gupta, Gargi V Pillai, and Samit Ari. "Change detection in optical satellite images based on local binary similarity pattern technique." In: *IEEE Geoscience and Remote Sensing Letters* 15.3 (2018), pp. 389–393.
- [46] Ritwik Gupta, Bryce Goodman, Nirav Patel, Ricky Hosfelt, Sandra Sajeev, Eric Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew Gaston. "Creating xBD: A dataset for assessing building damage from satellite imagery." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2019, pp. 10–17.
- [47] Pengyu Hao, Yulin Zhan, Li Wang, Zheng Niu, and Muhammad Shakir. "Feature selection of time series MODIS data for early crop classification using random forest: A case study in Kansas, USA." In: *Remote Sensing* 7.5 (2015), pp. 5347–5369.
- [48] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. "Masked autoencoders are scalable vision learners." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 16000–16009.
- [49] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Identity mappings in deep residual networks." In: *European Conference on Computer Vision*. Springer. 2016, pp. 630–645.
- [50] Marko Heikkilä, Matti Pietikäinen, and Cordelia Schmid. "Description of interest regions with local binary patterns." In: *Pattern recognition* 42.3 (2009), pp. 425–436.
- [51] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory." In: *Neural computation* 9.8 (1997), pp. 1735–1780.
- [52] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. "Searching for mobilenetv3." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 1314–1324.

- [53] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5 MB model size." In: *arXiv preprint arXiv:1602.07360* (2016).
- [54] Shunping Ji, Yanyun Shen, Meng Lu, and Yongjun Zhang. "Building instance change detection from large-scale aerial images using convolutional neural networks and simulated samples." In: *Remote Sensing* 11.11 (2019), p. 1343.
- [55] Shunping Ji, Shiqing Wei, and Meng Lu. "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set." In: *IEEE Transactions on Geoscience and Remote Sensing* 57.1 (2018), pp. 574–586.
- [56] Kasper Johansen, Matteo G Ziliani, Rasmus Houborg, Trenton E Franz, and Matthew F McCabe. "CubeSat constellations provide enhanced crop phenology and digital agricultural insights using daily leaf area index retrievals." In: *Scientific reports* 12.1 (2022), p. 5244.
- [57] A Kääh, C Huggel, L Fischer, S Guex, F Paul, I Roer, N Salzmann, S Schläefli, K Schmutz, D Schneider, et al. "Remote sensing of glacier-and permafrost-related hazards in high mountains: an overview." In: *Natural Hazards and Earth System Sciences* 5.4 (2005), pp. 527–554.
- [58] Salman H Khan, Xuming He, Fatih Porikli, and Mohammed Bennamoun. "Forest change detection in incomplete satellite images with deep neural networks." In: *IEEE Transactions on Geoscience and Remote Sensing* 55.9 (2017), pp. 5407–5423.
- [59] Lukas Kondmann, Sebastian Boeck, Rogerio Bonifacio, and Xiao Xiang Zhu. "Early Crop Type Classification With Satellite Imagery-An Empirical Analysis." In: (2022).
- [60] Lukas Kondmann, Aysim Toker, Marc Rußwurm, Andres Camero Unzueta, Devis Peressuti, Grega Milcinski, Nicolas Longépé, Pierre-Philippe Mathieu, 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*. 2021.
- [61] 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." In: *IEEE Transactions on Geoscience and Remote Sensing* (2021).

- [62] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger. Vol. 25. Curran Associates, Inc., 2012.
- [63] Pradeep Kumar, Dileep Kumar Gupta, Varun Narayan Mishra, and Rajendra Prasad. "Comparison of support vector machine, artificial neural network, and spectral angle mapper algorithms for crop classification using LISS IV data." In: *International Journal of Remote Sensing* 36.6 (2015), pp. 1604–1617.
- [64] 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*. 2021, pp. 1205–1214.
- [65] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." In: *Nature* 521.7553 (2015), pp. 436–444.
- [66] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. "Gradient-based learning applied to document recognition." In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.
- [67] Marrit Leenstra, Diego Marcos, Francesca Bovolo, and Devis Tuia. "Self-supervised pre-training enhances change detection in Sentinel-2 imagery." In: *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VII*. Springer. 2021, pp. 578–590.
- [68] Liang Li, Xue Li, Yun Zhang, Lei Wang, and Guowei Ying. "Change detection for high-resolution remote sensing imagery using object-oriented change vector analysis method." In: *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE. 2016, pp. 2873–2876.
- [69] Yangyang Li, Cheng Peng, Yanqiao Chen, Licheng Jiao, Linhao Zhou, and Ronghua Shang. "A deep learning method for change detection in synthetic aperture radar images." In: *IEEE Transactions on Geoscience and Remote Sensing* 57.8 (2019), pp. 5751–5763.
- [70] 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*. 2021, pp. 10012–10022.
- [71] Dengsheng Lu, Emilio Moran, and Scott Hetrick. "Detection of impervious surface change with multitemporal Landsat images in an urban–rural frontier." In: *ISPRS Journal of Photogrammetry and Remote Sensing* 66.3 (2011), pp. 298–306.

- [72] Ning Lv, Chen Chen, Tie Qiu, and Arun Kumar Sangaiah. "Deep learning and superpixel feature extraction based on contractive autoencoder for change detection in SAR images." In: *IEEE transactions on industrial informatics* 14.12 (2018), pp. 5530–5538.
- [73] Zhiyong Lv, HaiTao Huang, LiPeng Gao, Jón Atli Benediktsson, MingHua Zhao, and Cheng Shi. "Simple Multiscale UNet for Change Detection with Heterogeneous Remote Sensing Images." In: *IEEE Geoscience and Remote Sensing Letters* (2022).
- [74] Zhiyong Lv, Fengjun Wang, Guoqing Cui, Jón Atli Benediktsson, Tao Lei, and Weiwei Sun. "Spatial-Spectral Attention Network Guided With Change Magnitude Image for Land Cover Change Detection Using Remote Sensing Images." In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), pp. 1–12.
- [75] 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*. 2019, pp. 75–82.
- [76] William A Malila. "Change vector analysis: an approach for detecting forest changes with Landsat." In: *LARS symposia*. 1980, p. 385.
- [77] Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 9414–9423.
- [78] Victor Maus, Gilberto Câmara, Ricardo Cartaxo, Alber Sanchez, Fernando M Ramos, and Gilberto R De Queiroz. "A time-weighted dynamic time warping method for land-use and land-cover mapping." In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9.8 (2016), pp. 3729–3739.
- [79] 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." In: *IEEE Transactions on Geoscience and Remote Sensing* 57.2 (2018), pp. 924–935.
- [80] Luis Moya, Abdul Muhari, Bruno Adriano, Shunichi Koshimura, Erick Mas, Luis R Marval-Perez, and Naoto Yokoya. "Detecting urban changes using phase correlation and l1-based sparse model for early disaster response: A case study of the 2018 Su-

- lawesi Indonesia earthquake-tsunami." In: *Remote Sensing of Environment* 242 (2020), p. 111743.
- [81] Cory S Myers and Lawrence R Rabiner. "A comparative study of several dynamic time-warping algorithms for connected-word recognition." In: *Bell System Technical Journal* 60.7 (1981), pp. 1389–1409.
- [82] J Nichol and Man Sing Wong. "Satellite remote sensing for detailed landslide inventories using change detection and image fusion." In: *International journal of remote sensing* 26.9 (2005), pp. 1913–1926.
- [83] Joachim Nyborg, Charlotte Pelletier, Sébastien Lefèvre, and Ira Assent. "TimeMatch: Unsupervised cross-region adaptation by temporal shift estimation." In: *ISPRS Journal of Photogrammetry and Remote Sensing* 188 (2022), pp. 301–313.
- [84] Julie B Odenweller and Karen I Johnson. "Crop identification using Landsat temporal-spectral profiles." In: *Remote Sensing of Environment* 14.1-3 (1984), pp. 39–54.
- [85] Aiym Orynbaikyzy, Ursula Gessner, and Christopher Conrad. "Crop type classification using a combination of optical and radar remote sensing data: a review." In: *International Journal of Remote Sensing* 40.17 (2019), pp. 6553–6595.
- [86] Nobuyuki Otsu. "A threshold selection method from gray-level histograms." In: *IEEE Transactions on Systems, Man, and Cybernetics* 9.1 (1979), pp. 62–66.
- [87] Mahesh Pal. "Random forest classifier for remote sensing classification." In: *International journal of remote sensing* 26.1 (2005), pp. 217–222.
- [88] Charlotte Pelletier, Geoffrey I Webb, and François Petitjean. "Temporal convolutional neural network for the classification of satellite image time series." In: *Remote Sensing* 11.5 (2019), p. 523.
- [89] José M Peña-Barragán, Moffatt K Ngugi, Richard E Plant, and Johan Six. "Object-based crop identification using multiple vegetation indices, textural features and crop phenology." In: *Remote Sensing of Environment* 115.6 (2011), pp. 1301–1316.
- [90] Daifeng Peng, Yongjun Zhang, and Haiyan Guan. "End-to-end change detection for high resolution satellite images using improved UNet++." In: *Remote Sensing* 11.11 (2019), p. 1382.
- [91] REF. "CV4A Competition Kenya Crop Type Dataset." In: (2020).
- [92] REF. "Crop Type Classification Dataset for Western Cape, South Africa." In: (2021).

- [93] “Red and photographic infrared linear combinations for monitoring vegetation.” In: *Remote Sensing of Environment* 8.2 (1979), pp. 127–150.
- [94] Caijun Ren, Xiangyu Wang, Jian Gao, Xiren Zhou, and Huanhuan Chen. “Unsupervised Change Detection in Satellite Images With Generative Adversarial Network.” In: *IEEE Transactions on Geoscience and Remote Sensing* 59.12 (2021), pp. 10047–10061.
- [95] William J Ripple. “Asymptotic reflectance characteristics of grass vegetation.” In: *Photogrammetric Engineering and Remote Sensing* 51.2 (1985), pp. 1915–1921.
- [96] Komeil Rokni, Anuar Ahmad, Karim Solaimani, and Sharifeh Hazini. “A new approach for surface water change detection: Integration of pixel level image fusion and image classification techniques.” In: *International Journal of Applied Earth Observation and Geoinformation* 34 (2015), pp. 226–234.
- [97] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. “High-resolution image synthesis with latent diffusion models.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10684–10695.
- [98] 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*. Springer. 2015, pp. 234–241.
- [99] Marc Rußwurm and Marco Körner. “Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images.” In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2017, pp. 11–19.
- [100] Marc Rußwurm and Marco Körner. “Multi-temporal land cover classification with sequential recurrent encoders.” In: *ISPRS International Journal of Geo-Information* 7.4 (2018), p. 129.
- [101] Marc Rußwurm and Marco Körner. “Self-attention for raw optical satellite time series classification.” In: *ISPRS Journal of Photogrammetry and Remote Sensing* 169 (2020), pp. 421–435.
- [102] Marc Rußwurm, Charlotte Pelletier, Maximilian Zollner, Sébastien Lefèvre, and Marco Körner. “BreizhCrops: A Time Series Dataset for Crop Type Mapping.” In: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences ISPRS (2020)* (2020).

- [103] Sudipan Saha, Biplab Banerjee, and Xiao Xiang Zhu. "Trusting small training dataset for supervised change detection." In: *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE. 2021, pp. 2031–2034.
- [104] Sudipan Saha, Francesca Bovolo, and Lorenzo Bruzzone. "Unsupervised deep change vector analysis for multiple-change detection in VHR images." In: *IEEE Transactions on Geoscience and Remote Sensing* 57.6 (2019), pp. 3677–3693.
- [105] Sudipan Saha, Lichao Mou, Chunping Qiu, Xiao Xiang Zhu, Francesca Bovolo, and Lorenzo Bruzzone. "Unsupervised Deep Joint Segmentation of Multitemporal High-Resolution Images." In: *IEEE Transactions on Geoscience and Remote Sensing* (2020).
- [106] Sudipan Saha, Yady Tatiana Solano-Correa, Francesca Bovolo, and Lorenzo Bruzzone. "Unsupervised Deep Transfer Learning-Based Change Detection for HR Multispectral Images." In: *IEEE Geoscience and Remote Sensing Letters* (2020).
- [107] Vivien Sainte Fare Garnot and Loic Landrieu. "Panoptic Segmentation of Satellite Image Time Series with Convolutional Temporal Attention Networks." In: *ICCV* (2021).
- [108] 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: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, June 2020, pp. 12322–12331. (Visited on 07/22/2021).
- [109] Maja Schneider, Amelie Broszeit, and Marco Körner. "EuroCrops: A Pan-European Dataset for Time Series Crop Type Classification." In: *arXiv preprint arXiv:2106.08151* (2021).
- [110] Cornelius Senf and Rupert Seidl. "Mapping the forest disturbance regimes of Europe." In: *Nature Sustainability* 4.1 (2021), pp. 63–70.
- [111] Atharva Sharma, Xiuwen Liu, and Xiaojun Yang. "Land cover classification from multi-temporal, multi-spectral remotely sensed imagery using patch-based recurrent neural networks." In: *Neural Networks* 105 (2018), pp. 346–355.
- [112] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." In: *arXiv preprint arXiv:1409.1556* (2014).
- [113] Yuli Sun, Lin Lei, Xiao Li, Xiang Tan, and Gangyao Kuang. "Patch Similarity Graph Matrix-Based Unsupervised Remote Sensing Change Detection With Homogeneous and Heterogeneous Sensors." In: *IEEE Transactions on Geoscience and Remote Sensing* (2020).



- [114] Frank Thonfeld, Hannes Feilhauer, Matthias Braun, and Gunter Menz. "Robust Change Vector Analysis (RCVA) for multi-sensor very high resolution optical satellite data." In: *International Journal of Applied Earth Observation and Geoinformation* 50 (2016), pp. 131–140.
- [115] Shiqi Tian, Ailong Ma, Zhuo Zheng, and Yanfei Zhong. "Hi-UCD: A large-scale dataset for urban semantic change detection in remote sensing imagery." In: *arXiv preprint arXiv:2011.03247* (2020).
- [116] Aysim Toker et al. "DynamicEarthNet: Daily Multi-Spectral Satellite Dataset for Semantic Change Segmentation." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
- [117] Silvia Valero, David Morin, Jordi Inglada, Guadalupe Sepulcre, Marcela Arias, Olivier Hagolle, Gérard Dedieu, Sophie Bontemps, Pierre Defourny, and Benjamin Koetz. "Production of a dynamic cropland mask by processing remote sensing image series at high temporal and spatial resolutions." In: *Remote Sensing* 8.1 (2016), p. 55.
- [118] 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*. 2021, pp. 6398–6407.
- [119] 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* 30 (2017).
- [120] Fei Wang, Jinsong Han, Shiyuan Zhang, Xu He, and Dong Huang. "CSI-Net: Unified human body characterization and pose recognition." In: *arXiv preprint arXiv:1810.03064* (2018).
- [121] Sherrie Wang, François Waldner, and David B Lobell. "Unlocking large-scale crop field delineation in smallholder farming systems with transfer learning and weak supervision." In: *Remote Sensing* 14.22 (2022), p. 5738.
- [122] Giulio Weikmann, Claudia Paris, and Lorenzo Bruzzone. "Time-Sen2Crop: a Million Labeled Samples Dataset of Sentinel 2 Image Time Series for Crop Type Classification." In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2021), pp. 1–1. (Visited on 04/29/2021).
- [123] Hao Wu and Saurabh Prasad. "Semi-supervised deep learning using pseudo labels for hyperspectral image classification." In: *IEEE Transactions on Image Processing* 27.3 (2017), pp. 1259–1270.

- [124] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. "Self-training with noisy student improves imagenet classification." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10687–10698.
- [125] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. "Billion-scale semi-supervised learning for image classification." In: *arXiv preprint arXiv:1905.00546* (2019).
- [126] Chenghai Yang, James H Everitt, Qian Du, Bin Luo, and Jocelyn Chanussot. "Using high-resolution airborne and satellite imagery to assess crop growth and yield variability for precision agriculture." In: *Proceedings of the IEEE* 101.3 (2012), pp. 582–592.
- [127] Kunping Yang, Gui-Song Xia, Zicheng Liu, Bo Du, Wen Yang, Marcello Pelillo, and Liangpei Zhang. "Semantic change detection with asymmetric Siamese networks." In: *arXiv preprint arXiv:2010.05687* (2020).
- [128] Massimo Zanetti, Sudipan Saha, Daniele Marinelli, Maria Lucia Magliozzi, Massimo Zavagli, Mario Costantini, Francesca Bovolo, and Lorenzo Bruzzone. "A System for Burned Area Detection on Multispectral Imagery." In: *IEEE Transactions on Geoscience and Remote Sensing* (2021).
- [129] Tao Zhan, Maoguo Gong, Xiangming Jiang, and Mingyang Zhang. "Unsupervised Scale-Driven Change Detection With Deep Spatial-Spectral Features for VHR Images." In: *IEEE Transactions on Geoscience and Remote Sensing* (2020).
- [130] Chenxiao Zhang, Peng Yue, Deodato Tapete, Liangcun Jiang, Boyi Shangguan, Li Huang, and Guangchao Liu. "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images." In: *ISPRS Journal of Photogrammetry and Remote Sensing* 166 (2020), pp. 183–200.
- [131] Cui Zhang, Liejun Wang, Shuli Cheng, and Yongming Li. "SwinSUNet: Pure transformer network for remote sensing image change detection." In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), pp. 1–13.
- [132] Hongyan Zhang, Manhui Lin, Guangyi Yang, and Liangpei Zhang. "Escnet: An end-to-end superpixel-enhanced change detection network for very-high-resolution remote sensing images." In: *IEEE Transactions on Neural Networks and Learning Systems* (2021).
- [133] Hui Zhang, Maoguo Gong, Puzhao Zhang, Linzhi Su, and Jiao Shi. "Feature-level change detection using deep representation and feature change analysis for multispectral imagery." In:

- IEEE Geoscience and Remote Sensing Letters* 13.11 (2016), pp. 1666–1670.
- [134] Baojuan Zheng, James B Campbell, Guy Serbin, and John M Galbraith. “Remote sensing of crop residue and tillage practices: Present capabilities and future prospects.” In: *Soil and Tillage Research* 138 (2014), pp. 26–34.
- [135] Baojuan Zheng, Soe W Myint, Prasad S Thenkabail, and Rimjhim M Aggarwal. “A support vector machine to identify irrigated crop types using time-series Landsat NDVI data.” In: *International Journal of Applied Earth Observation and Geoinformation* 34 (2015), pp. 103–112.
- [136] Zhuo Zheng, Yanfei Zhong, Shiqi Tian, Ailong Ma, and Liangpei Zhang. “ChangeMask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection.” In: *ISPRS Journal of Photogrammetry and Remote Sensing* 183 (2022), pp. 228–239.
- [137] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. “Deep learning in remote sensing: A comprehensive review and list of resources.” In: *IEEE Geoscience and Remote Sensing Magazine* 5.4 (2017), pp. 8–36.