Article

Neural Network-Based Urban Change Monitoring with Deep-Temporal Multispectral and SAR Remote Sensing Data

Georg Zitzlsberger [1,*], Michal Podhorányi [1], Václav Svatoň [1], Milan Lazecký [1,2] and Jan Martinovič [1]

1 IT4Innovations, VŠB–Technical University of Ostrava, 708 00 Ostrava, Poruba, Czech Republic
2 School of Earth and Environment, University of Leeds, Leeds LS2 9JT, UK
* Correspondence: georg.zitzlsberger@vsb.cz; Tel.: +420-596-999-560

Abstract: Remote-sensing-driven urban change detection has been studied in many ways for decades for a wide field of applications, such as understanding socio-economic impacts, identifying new settlements, or analyzing trends of urban sprawl. Such kinds of analyses are usually carried out manually by selecting high-quality samples that binds them to small-scale scenarios, either temporarily limited or with low spatial or temporal resolution. We propose a fully automated method that uses a large amount of available remote sensing observations for a selected period without the need to manually select samples. This enables continuous urban monitoring in a fully automated process. Furthermore, we combine multispectral optical and synthetic aperture radar (SAR) data from two eras as two mission pairs with synthetic labeling to train a neural network for detecting urban changes and activities. As pairs, we consider European Remote Sensing (ERS-1/2) and Landsat 5 Thematic Mapper (TM) for 1991–2011 and Sentinel 1 and 2 for 2017–2021. For every era, we use three different urban sites—Limassol, Rotterdam, and Liège—with at least 500 km^2 each, and deep observation time series with hundreds and up to over a thousand of samples. These sites were selected to represent different challenges in training a common neural network due to atmospheric effects, different geographies, and observation coverage. We train one model for each of the two eras using synthetic but noisy labels, which are created automatically by combining state-of-the-art methods, without the availability of existing ground truth data. To combine the benefit of both remote sensing types, the network models are ensembles of optical- and SAR-specialized sub-networks. We study the sensitivity of urban and impervious changes and the contribution of optical and SAR data to the overall solution. Our implementation and trained models are available publicly to enable others to utilize fully automated continuous urban monitoring.

Keywords: urban change detection; continuous urban monitoring; neural network; SAR; optical multispectral; deep-temporal; ERS-1; ERS-2; Landsat 5 TM; Sentinel 1; Sentinel 2

1. Introduction

Digital remote-sensing-driven land cover and land use (LCLU) change detection with multi-temporal satellite data has been studied for decades, dating back to the early 1970s [1]. Due to the reliance of repeat cycles and sensor technologies of satellite-based remote sensing, it is sought after by urban change detection to understand socio-economic impacts and effects, identifying new settlements, monitoring urban growth, or analyzing trends of urban sprawl, just to name a few. For these fields, knowledge of how urban and human-made structures change over time is crucial and, depending on the exact application, can require high spatio-temporal resolution. In addition, automatic processing is desired if large-scale analysis is required to monitor larger areas over a longer time and to make available the methods and results to a wider range of interdisciplinary users to thrive on remote-sensing-based urban change detection and monitoring.

In our work, we propose an ensemble neural network architecture called the Ensemble of Recurrent Convolutional Neural Networks for Deep Remote Sensing (ERCNN-DRS).

Detecting changes is tuned for impervious and built-up land covers and is trained in two variants to detect urban-related changes and activities for two eras. The first era spans 1991–2011 with the use of ERS-1/2 and Landsat 5 TM missions as a pair. The second era is 2017–2021 with the Sentinel 1 and 2 pair. For the former, we consider time windows for analysis of one year (1y). Six months (6m) are used for the latter due to the higher temporal resolution between observations. We avoid the need for manual ground truth generation by automatically creating synthetic but noisy labels using a combination of state-of-the-art methods. These labels are used for training the ERCNN-DRS in a supervised setting. For every era, we combine SAR and optical missions as pairs to combine their remote sensing advantages, such as distinguishing impervious land covers with optical and higher temporal resolution with reduced atmospheric disturbance for SAR observations. Our method does not require the manual selection of observation samples. We use all available observations as long as they can be properly co-registered and do not contain corrupt information. This can usually be done with little effort, which paves the way for automatic processing. For optical data, we remove clouds using information already available directly at the optical data sources. Our solution aims at a large degree of automation to either train or use the neural network models for different Areas of Interest (AoI) without the need of manual ground truth generation or selecting high-quality observation samples. We use two AoIs (Rotterdam and Limassol) for training and validation and one (Liège) for testing. Each AoI is at least 500 km^2 with hundreds to thousands of combined SAR and optical observations (so-called deep-temporal). We consider 1y/6m windows for detecting urban changes that can be continuously carried out over longer periods, which we refer to as *continuous urban change and activity monitoring with deep-temporal remote sensing data*.

ERCNN-DRS is trained to detect urban changes with a varying number of optical or SAR observations of different qualities due to atmospheric effects, different solar irradiance and seasons, and partial updatesdue to out-of-swath conditions. The analysis windows required were intentionally kept short to allow flexibility in selecting time ranges of interest. Our goal is to offer a trained network that can be applied with minimal effort to an AoI for a short period of 1y for ERS-1/2 and Landsat 5 TM, or 6m for Sentinel 1 and 2 mission pairs.

Our work first summarizes related work in Section 2 and differentiates it to our study. In Section 3, we describe the data types used and their pre-processing to retrieve trainable data. Our proposed neural network architecture is discussed in Section 4, including the hyper-parameters for the two different eras and the training and validation methodology. Section 5 contains two ablation studies to document the performance of the trained models with respect to sensitivity to changes and the impact of SAR and optical data types. Futher, we show some examples from the Liège test site for both eras. A final discussion with current limitations and further improvements is carried out in Section 6. Section 7 concludes our work.

2. Related Work

The detection of urban changes has become one of the most important applications in remote sensing science. Urban objects include man-made structures such as airports, buildings, or roads. An urban structure type (UST) was proposed by Lehner et al. [2], and we alias urban as the defined built-up types. These artificial structures dynamically vary in time and space, and therefore, it is challenging to detect them accurately. Due to the public availability of various high-resolution remote sensing data sets, it has provided a large number of possibilities for extracting urban information.

As mentioned earlier, (urban) change detection and monitoring with remote sensing data has been of interest for over half a century. Table 1 gives a non-exhaustive overview of other work with different properties, which we differentiate our work from. A lot more work was carried out [3], and we only name some as reference. Our work is listed at the bottom of the table for direct comparison. We consider four important properties: change detection method, missions used, number of observations, and scaling of the method.

Throughout the last decade, change detection for remote sensing data has increasingly embraced neural networks, including deep neural networks with Multilayer Perceptron (MLP) [4,5] as one of their most simple representatives. Deep learning is able to fulfill the needs of remote sensing image processing, such as classification or segmentation [6]. Before the shift to deep neural networks, change detection was driven by Principal Component Analysis [7] (PCA), Decision Trees [8], Change Vector Analysis [9] (CVA), Difference or Ratio methods [10], Markov chains [11], and other methods [12]. Some neural network-based approaches existed before 2010. However, with the recent advent of deep neural networks, many architectures have been invented that have resulted in better performance and methods predesignated to processing large amounts data. This is fueled by the availability of dedicated and optimized hardware (e.g., GPU accelerators), the amount of large storage (big data) and easy-to-use software solutions, such as the training frameworks Tensorflow (https://www.tensorflow.org/; accessed on 29 July 2021) or Pytorch (https://pytorch.org/; accessed on 29 July 2021). Many studies in recent years have shown interest in deep learning technologies [13–15]. Deep neural networks also move feature extraction and mapping into the neural network architecture and its training process, which is easier to handle and can lead to unprecedented performance [16].

Another property is related to the mission types. Some works consider either optical multispectral or SAR solemnly to identify changes. The most commonly used satellite data for analyzing urban changes are Landsat space images [14,17]. The Landsat program has monitored the Earth's surface since 1972 and provided data from three different devices: Thematic Mapper (TM), Enhanced Thematic Mapper Plus (ETM+), and the Multispectral Scanner System (MSS). However, Landsat data, as well as other multispectral optical sensors, are extremely sensitive to weather changes and the time of day, which requires further pre-processing. Research on urban changes from Landsat data has attracted the interest of many scholars. For example, a novel approach for extracting high-frequency urban land cover changes by using clear-sky Landsat observations was developed by Jing et al. [18].

For optical multispectral observations, a large range of index methods exists, which are able to extract urban and impervious areas. To recognize urban objects, scientists have traditionally relied on built-up indices [19,20]. Built-up indices are often fast and accurate to extract urban information, and therefore, their usage is conventional. Several studies have used different indices to extract urban areas [21]. These are, for example, the Normalized Difference Built-up Index (NDBI) [22], the Normalized Difference Impervious Surface Index (NDISI) [23], the Enhanced Built-up and Bareness Index (EBBI) [24], Combinational Biophysical Composition Index (CBCI) [25], or the Enhanced Normalized Difference Impervious Surface Index (ENDISI) [26].

Even for SAR data, some works propose methods to identify urban changes [27,28], which demonstrate that information is immanent to designate urban changes. The advantages of microwave remote sensing with respect to optical features are in weather and atmospheric independence, the possibility to monitor during night time, and the high sensitivity to fine changes [29]. The potential of radar data was investigated much less due to the often difficult data pre-processing phase [30]. Other works [8,31,32] have identified positive synergies by combining optical and SAR observations. Nevertheless, different studies have shown that radar technology is an excellent approach for extracting urban information. For instance, Ansari et al. [33] used multi-resolution texture features from polarimetric SAR images to analyze urban changes. On the other side, a simple and fast method was developed to identify structural changes in urban environments using SAR imagery [34]. Urban mapping and delineation with the help of SAR data were presented in other studies as well [30,35,36].

Table 1. A non-exhaustive list of other work on (urban) change detection with remote sensing data. Our work is listed at the bottom for comparison.

Change Detection Method	Missions	Number of Observations	Resolution (m/pixel)	Area (km^2)	Authors/Reference
Fully Convolutional Siamese Networks (FCNN)	OSCD [37] (Sentinel 2), AC [39] (Aerial, RGB only)	24 pairs (2015–2018), 12 pairs (2000–2007)	down to 10, 1.5	864.0 (OSCD) 16.3 (AC)	Daudt et al. [38]
Multiple- Coherent Change Detection (M-CCD)	Sentinel 1	17 (2015–2016)	10	120.0	Manzoni et al. [34]
Pulse Coupled Neural Network	Sentinel 1 and 2	6 pairs (2015–2017)	10	1.75	Benedetti et al. [31]
Curvelet, Contourlet, and Wavelet Transforms	UAVSAR	2 pairs (2009 and 2015)	3.2	N/A (small)	Ansari et al. [33]
Maximum Likelihood	Landsat 2, 5, 7, and 8	5 (1978–2017)	30, 60 for Landsat 2	1129.4	Kundu et al. [40]
TSFLC: Temporal Segmentation and Trajectory Classification	Landsat TM/ETM+, OLI	68 (1986–2017)	N/A (ca. 30)	1997	Jing et al. [18]
Omnibus and Change Vector Analysis (CVA)	Sentinel 1 and Landsat 7,8	12+11 (2014–2016) 8+6 (2015)	both 30	490 and 3500	Muro et al. [32]
Decision Trees	ALOS PALSAR and Landsat TM/ETM+	174 (eff. 55) and 20 (2006–2011)	30	17,076	Qin et al. [8]
Regression Model	Landsat MSS/TM/ETM+ and QuickBird	10 (1977–2008) and 2 (calibration only)	30 and 0.6	3660	Lu et al. [41]
Modified Ratio Operator + Kittler-Illingworth	ENVISAT ASAR and ERS-2	2 pairs (1998/99–2008)	30	N/A (small)	Ban et al. [42]
Topologically Enabled Hierarchical Framework	Spot 2 and 5	3 (2004–2010)	10 and 5/2.5	940	Wania et al. [43]
Spectral and Textural Differences	IKONOS, GF-1, and Aerial	total of 6 (1, 2, and 3) pairs for 2000–2016	0.5–2	1.4, 0.4–0.45, and 0.24–1.1	Xiao et al. [10]
ERCNN-DRS	ERS-1/2 and Landsat 5 TM, Sentinel 1 and 2	1121 and 638 (1991–2011), 2067 and 640 (2017–2021)	30, 10	see Table 2	our study

The number of observations used for identifying urban change is the third property. Most of the existing work uses only a few high-quality observations (i.e., image pairs) to indicate changes. These samples are carefully identified and usually are from the same season to mitigate the effect of seasonal changes and with the best-case atmospheric conditions, if optical. Some studies attempt to use more observations, which are either averaged [14] or used for statistical analysis [8]. Both methods are less sensitive to changes as either the sampling interval is too coarse or averaging filters suppress changes. In addition, benchmark data sets, such as OSCD [37] or DeepGlobe 2018 [44], exist. Despite their high quality, these share the same limitations of including only a few multi-temporal samples from a single sensor.

Access to observations and high quality products is also made available through platforms such as the Thematic Exploitation Platform (TEP, https://eo4society.esa.int/platform-services/; accessed on 29 July 2021) or Data and Information Access Services (DIAS, https://earsc.org/dias-comparison/; accessed on 29 July 2021). The latter provides access to Copernicus (https://www.copernicus.eu/; accessed on 29 July 2021) and Sentinel data, information products from six operational services of Copernicus, and cloud-based tools for analysis. The former also provides the Urban TEP [45] (U-TEP), which focuses on the effective and efficient urban management and safeguarding of livable cities. It offers its users high-performance data access and processing, analysis and visualization services, and a customized workplace for communication and the means to share algorithms, data, and services. The platform offers a number of globally processed data products, such as the World Settlement Footprint 2015 [46] (WSF) derived from Landsat 8 and Sentinel 1 data, Global Urban Footprint [47] (GUF) derived from TerraSAR-X and TanDEM-X radar data, and TimeScan product derived from Landsat and Sentinel 1 and 2 imagery, providing a cloud-free representation of spectral and temporal characteristics of the land surface. The products on these platforms are of high quality but show similar temporal limitations as the aforementioned data and studies.

The last property is the applicability and demonstration of large-scale urban change detection. It is common that studies demonstrate their applicability over smaller areas, limited time frames, or lower temporal resolution. In addition, their requirement of high-quality data induces time-consuming and labor-intensive pre-processing that is difficult to automatize. Exceptions are works such as [8,14,18], which attempt larger scale demonstrations. Furthermore, the transfer to different remote sensing missions is underrepresented, which results in the use of either different missions of the same type (optical or SAR) or only one optical/SAR pair. The applicability to a broader range of sensors is not directly answered.

Compared to other studies, our work does not require high-quality samples, level 2 products, averaging over larger periods, or observations of matching seasons. Instead, we intend to use most of the data available as long as it can be properly co-registered, which gives access to higher frequency information. Furthermore, we extend the idea of combining SAR and multispectral optical observations in an ensemble neural network for overall better performance. We also demonstrate the flexibility of our solution to different sensor generations over two eras spanning almost 25 years combined. This is an important pre-requisite for long-term urban monitoring as it is not limited to a single remote sensing mission. Finally, due to no available ground truth data with the required temporal resolution, mission pairs, and temporal extent, we create synthetic noisy labels to train the neural network.

3. Data Pre-Processing

For the two eras, we use two different remote sensing mission pairs to include both optical multispectral and SAR observations: ERS-1/2 and Landsat 5 TM for 1991–2011, and Sentinel 1 and 2 for 2017–2021. Three different AoIs are used, as shown in Table 2, with different resolutions (m/pixel) across every era due to sensor capabilities. The selection of the AoIs, the data acquisition and pre-processing methods, and time series preparation we applied are described below.

3.1. Selection of AoIs

The three AoIs—Rotterdam, Liège, and Limassol (see Table 2)—were selected due to different weather and seasonal conditions. Rotterdam is subject to high cloud coverage [48], snow/ice, whereas Limassol has a lower probability of these but also offers less observations over time. In addition, both sites are in different geographies and show different characteristics of impervious areas. As we discuss in Section 3.3.2, they require different parameters for creating the synthetic labels. Furthermore, the inclusion of two different eras gives access to a large time horizon and demonstrates the utility of our approach for different remote sensing missions. We intentionally use L1 products due to their wide temporal and spatial availability; our solution, however, is not bound to L1 products alone and can encompass other, higher quality levels. We used different reference systems due to our existing processing pipelines, which are not crucial for our study.

Table 2. AoIs of the three sites for both eras with bounding box coordinates and reference systems. See Appendix A for a map of AoIs.

	Site	AoI Bounding Box (Long, Lat)/(East, North)	Reference System	Area (km^2)
ERS-1/2 and Landsat 5 TM	Rotterdam	(3,923,101, 3,202,549), (3,959,241, 3,222,337)	EPSG:3035	712.9
	Liège	(3,988,121, 3,058,430), (4,024,261, 3,078,219)	EPSG:3035	713.3
	Limassol	(6,403,482, 1,601,833), (6,439,623, 1,621,621)	EPSG:3035	718.6
	Resolution	SAR: 12.5 m/pixel Optical: 30 m/pixel (interpolated to 12.5 m/pixel)		
Sentinel 1 and Sentinel 2	Rotterdam	(4.2033, 51.7913), (4.5566, 51.9854)	EPSG:4326	523.6
	Liège	(5.3827, 50.5474), (5.7280, 50.7350)	EPSG:4326	508.4
	Limassol	(32.8925, 34.6159), (33.1482, 34.8365)	EPSG:4326	576.2
	Resolution	SAR/Optical: 10 m/pixel		

Figure 1 shows the available observations over time, within each possible observation window of one year or six months, as we discuss later. Usually, remote sensing missions are configured with fixed repeat cycles so that sampling of observational data follows certain patterns. However, we consider larger time ranges where transition phases of missions and processing changes are common. In addition, further observations are discarded due to failed co-registration because of atmospheric effects, such as clouds, dust, and water vapor, and other effects caused by anomalies, such as banding, detector failure, and data loss. As a result, usable observations are irregularly spaced, and hence, every fixed observation window can include a different amount of observations. These variations add an additional level of complexity when it comes to detecting changes. In some cases, the number of available observations can even drop below a meaningful threshold. We combine both optical and SAR observations to provide enough information for detecting changes, even for shorter time frames of up to one year.

Liège was selected as the testing AoI with different observation frequencies and a higher likeliness of cloud coverage degrading optical information quality. The same window properties as for the other two AoIs are used (as described later).

3.2. Data Acquisition and Pre-Processing

The four different mission products used were obtained from three different data sources. Table 3 summarizes the sources, used products, and the number of observations available. The number of observations that were removed due to failed co-registration or corrupted data are given in parenthesis. In the following, we briefly describe the pre-processing applied to the products.
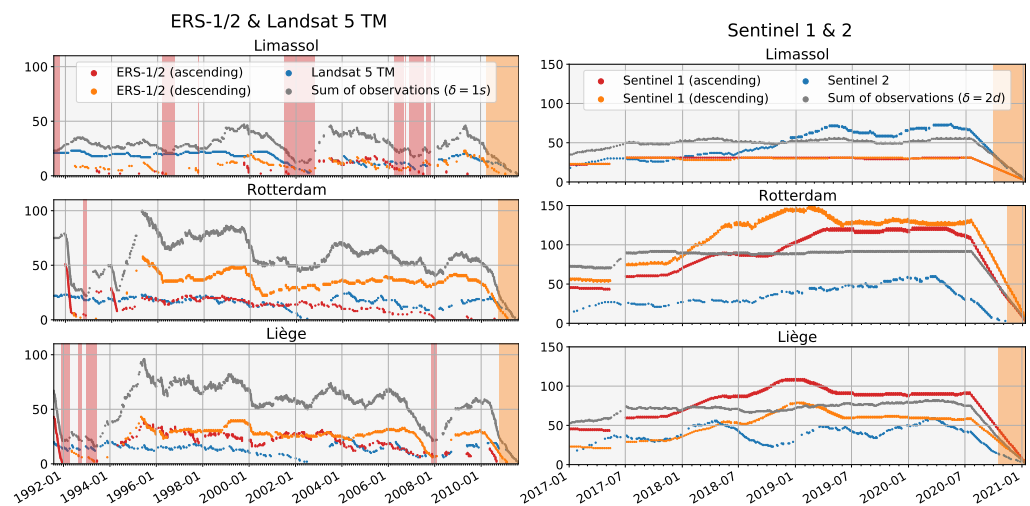
Figure 1. The number of remote sensing observations for the windows $w_{i,j}^t$ for ERS-1/2 and Landsat 5 TM (**left**) and the Sentinel 1 and 2 (**right**) era. The x-axis denotes the start times t of the according windows $w_{i,j}^t$. The accumulated observations ($\leq \Omega$) for the given step size δ are in gray. Areas in red/orange denote windows with observations $<\omega$; orange marks the end of mission/data.

Table 3. The number of total observations available with removed ones in parenthesis, and corresponding data sources with a short descriptor for the selected product.

	Site	SAR Observations (Ascending and Descending)	Optical Multispectral Observations
ERS-1/2 and Landsat 5 TM	Rotterdam	974 (-118)	753 (-434)
	Liège	934 (-89)	888 (-620)
	Limassol	291 (-27)	380 (-61)
	Source/Product	ESA/SAR_IMP_1P	USGS/L4-5 TM C1 L1
Sentinel 1 and Sentinel 2	Rotterdam	1603 (-4)	278 (-10)
	Liège	1040 (-0)	332 (-35)
	Limassol	468 (-0)	407 (-35)
	Source/Product	Sentinel Hub/ SENTINEL1_IW_[ASC | DSC]	Sentinel Hub/L1C

3.2.1. ERS-1/2

ESA's Earth Online *SAR_IMP_1P* (https://earth.esa.int/web/guest/data-access/browse-data-products/-/article/sar-precision-image-product-1477; accessed on 29 July 2021) product is used for ERS-1/2 observations. They are single-polarization (VV) backscatter intensity in Ground Range Detected (GRD) with a pixel size of 12.5×12.5 m and a number of looks of three. Orbits are separated into ascending and descending directions throughout the processing pipeline. Further processing is applied to geometrically terrain correct and co-register with *SRTM 1Sec HGT* Digital Elevation Model (DEM) using the ESA SNAP (https://step.esa.int/main/; accessed on 29 July 2021) toolset with a final clipping to the AoI and re-projection to EPSG:3035. All transformations use nearest neighbor interpolation to retain the values of the original data. Values are retained as a Digital Number (DN) and stored as a single-precision floating point data type.

3.2.2. Landsat 5 TM

Landsat 5 TM *L4-5 TM C1 L1* data are obtained from USGS (https://earthexplorer.usgs.gov/; accessed on 29 July 2021) with all seven bands and the 16-bit quality band (BQA). We apply top-of-atmosphere (TOA) reflection correction and co-register a final re-projection to EPSG:3035 with nearest neighbor interpolation. We use the 16-bit quality bands to filter out clouds, cloud shadows, snow/ice, and over-saturation (value: 0b10101011100

for masking clouds, cloud shadows, and snow/ice with medium confidence; masking saturation for five or more bands). All bands are used and super-sampled with nearest neighbor interpolation to 12.5×12.5 m to fit the resolution of ERS-1/2. Values are stored as single precision floating point numbers and normalized to a range of $(0.0, 1.0)$.

3.2.3. Sentinel 1 and Sentinel 2

Both Sentinel products are retrieved from Sentinel Hub (https://www.sentinel-hub.com/; accessed on 29 July 2021) using the eo-learn API (https://github.com/sentinel-hub/eo-learn; accessed on 29 July 2021). Sentinel 1 *SENTINEL1_IW_[ASC|DSC]* contains the orthorectified and terrain corrected (with *COPERNICUS* DEM) observations as GRD. They are dual-polarized (VV+VH) intensity backscatter in Interferometric Wide (IW) swath mode. The equivalent number of looks (ENL) is 4.4. Same as for ERS-1/2, we kept the ascending and descending orbit directions separate. Sentinel 2 *L1C* data are comprised of 13 bands with TOA reflection correction. Clouds are masked with the available L1C cloud mask (CLM). Observations with a maximum cloud coverage of over 80% were removed. Both Sentinel 1 and 2 are left in EPSG:4326 (WGS84) projection with a resolution of 10×10 m per pixel. The Sentinel Hub provides all bands interpolated to this resolution already. For Sentinel 1, values are retained as DN and stored as a single-precision floating point data type. For Sentinel 2, values are stored as single-precision floating point numbers and normalized to a range of $(0.0, 1.0)$.

3.3. Time Series Preparation

The observations are further processed by temporal stacking, assembling, tiling, and windowing into periods used for the change detection with the assignment of a label. Hence, each window is a data sample containing spatial and temporal constraint observations. These steps are carried out for every era separately. Figure 2 shows the basic steps, which we explain in detail in the following.
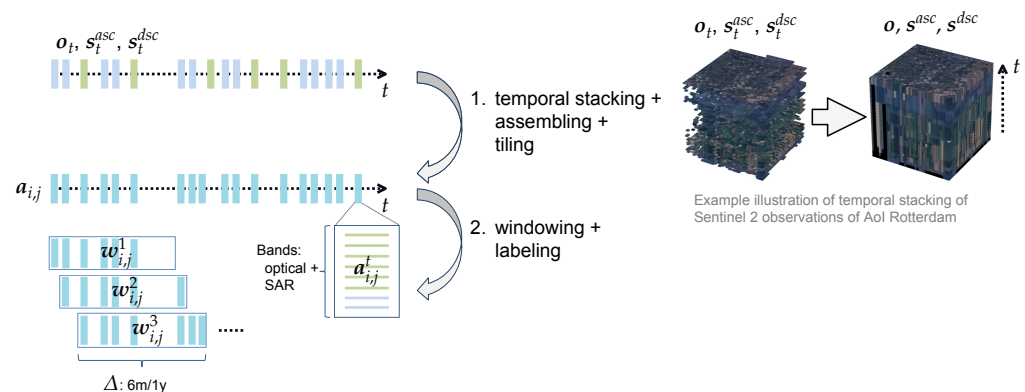

Figure 2. Two-step procedure for data preparation: From the series of available observations to the windows used for training/inference.

3.3.1. Temporal Stacking, Assembling, and Tiling

Optical observations $\boldsymbol{o}_t \in \mathbb{R}^{w \times h \times (b_{OPT}+1)}$ at time t over an AoI with a width w and height h contain the available optical multispectral bands b_{OPT} plus one mask band. The mask band indicates invalid pixels, e.g., due to cloud coverage, over-saturation, snow/ice, or being outside of swath. SAR observations are separated by orbit directions, such that $\boldsymbol{s}_t^{asc} \in \mathbb{R}^{w \times h \times (b_{SAR}+1)}$ are the intensity backscatter polarizations (b_{SAR}) plus a mask band of ascending orbit direction. SAR observations from descending orbit directions $\boldsymbol{s}_t^{dsc}$ are defined accordingly. For each era, the aforementioned data are temporally stacked independently to form a time series of observations. This results in $\boldsymbol{o} \in \mathbb{R}^{t_{OPT} \times w \times h \times b_{OPT}}$, $\boldsymbol{s}^{asc} \in \mathbb{R}^{t_{SAR}^{asc} \times w \times h \times b_{SAR}}$, and $\boldsymbol{s}^{dsc} \in \mathbb{R}^{t_{SAR}^{dsc} \times w \times h \times b_{SAR}}$, with a total of t_{OPT}, t_{SAR}^{asc}, and t_{SAR}^{dsc} observations, respectively. The masks are used to denote the pixels from each observation

that are to be ignored. Their values are substituted by the previous observation (or zero if not available). This guarantees that, at every observation, there is full coverage of the AoI with acceptable approximations. However, the more pixels are masked out over a longer period, the uncertainty increases (e.g., due to cloud conditions or over-saturation). The remote sensing kinds are combined into one time series with every observation aggregating optical and ascending and descending SAR data. This introduces two risks: at an observation time point, inconsistencies between the remote sensing kinds can occur; the amount of redundant data is increased (i.e., non-sparse representation). However, urban changes are expected to span across a longer time frame, which mitigates these effects. We also expect the neural network to become tolerant to smaller inconsistencies. Even though redundant data are introduced, it simplifies the ingestion during the training phase and further online processing is not required.

All three time series observations o, s^{asc}, and s^{dsc} are assembled into one time series $\boldsymbol{a} \in \mathbb{R}^{t_{all} \times w \times h \times b}$, with $t_{all} = t_{OPT} + t_{SAR}^{asc} + t_{SAR}^{dsc}$ as the accumulated observations of all three kinds and $b = b_{OPT} + 2 * b_{SAR}$ as the accumulated bands. We assume that no two observations happen simultaneously. Otherwise, the total number of accumulated observations would be less as some individual observations share the same time stamp.

The accumulated time series of each AoI is tiled into $x \times y$ pixel blocks with consecutive non-overlapping tiles. Only full tiles are considered with no padding applied. In the following, we continue with each tile so that we can use $\boldsymbol{a}_{i,j} \in \mathbb{R}^{t_{all} \times x \times y \times b}$ with x and y as spatial tile extent and $i \in \mathbb{N} : i \in [0, \lfloor w/x \rfloor]$, $j \in \mathbb{N} : j \in [0, \lfloor h/y \rfloor]$. One aggregated observation at time t for a tile (i, j) would then be $\boldsymbol{a}_{i,j}^t \in \mathbb{R}^{x \times y \times b}$, which contains all bands, optical and SAR. Our method does not depend on the tile size, which can be smaller or larger, as long as the applied filters of the neural network do not exceed these tiles. Based on memory considerations of our training platform, a size of $x = y = 32$ was selected and is used in the following to facilitate reading.

3.3.2. Windowing and Labeling

The tiled time series is further windowed into partially overlapping periods to provide all data for the change detection. Each window tile $\boldsymbol{w}_{i,j}^t$, starting at time t spans a fixed observation period Δ. We use observation periods for ERS-1/2 and Landsat 5 TM of one year and six months for Sentinel 1 and 2. The higher repeat cycles of Sentinel missions (ca. 6/5 days for Sentinel 1 and 2 versus ca. 35/16 days for ERS-1/2 and Landsat 5 TM) allows shorter periods while still covering the usual time frame to detect urban changes. We receive $\boldsymbol{w}_{i,j}^t \in \mathbb{R}^{l \times 32 \times 32 \times b}$ with l as the number of observations within the observation period. As window sizes are based on calendar months, the number of observations $l \in \mathbb{N} : l \in [\omega, \Omega]$ varies, with ω as the lower limit and Ω as the upper limit. Windows with less than ω observations are less likely contain enough information for change detection and are discarded. In Figure 1, these discarded windows are in the highlighted (red/orange) time frames. The upper bound Ω is added to limit the maximum memory footprint of each window and is defined to hold all observations of every window possible. This is controlled with a parameter δ to specify the minimum difference between two observations. For the ERS-1/2 and Landsat 5 TM era, it is set to $\delta = 1 \frac{\text{second}}{\text{observation}}$, which considers all observations. One second is the highest temporal resolution for observation data; we assume no two observations occur at the same second. For Sentinel 1 and 2, it is set to $\delta = 2 \frac{\text{days}}{\text{observation}}$. For the latter, this significantly reduces the potential maximum window size of over 300 (see Figure 1) down to a maximum of $\lceil \frac{6 \text{ months}}{2 \frac{\text{days}}{\text{observation}}} \rceil = 92$ observation(s). This is acceptable due to urban changes occurring at a much larger time scale and the original high number of observations per window is a result of nearby paths/rows where only a subset of the AoI is updated. These partial observations are mostly taken into account due to the temporal stacking after all, only discarding overlapping swaths.

In the following, we consider windows of the form $\boldsymbol{w}_{i,j}^t \in \mathbb{R}^{\Omega \times 32 \times 32 \times b}$. As the number l of observations varies for each window, we pad the remaining $\Omega - l$ observations.

As we show in Section 4.1, we apply masking in the neural network to discard these padded observations. Moreover, for each window, a synthetic label is created for the training sites Rotterdam and Limassol to provide a guideline for the supervised training. A synthetic label $\hat{y}_{i,j}^t \in \mathbb{R}^{32 \times 32}$ is created for each tile and window at time t as

$$\hat{y}_{i,j}^t := s_{cm}(w_{i,j}^t[b_{SAR}^{asc}], w_{i,j}^t[b_{SAR}^{dsc}]) \cdot o_{cm}(w_{i,j}^{t-\Delta}[b_{OPT}], w_{i,j}^{t+\Delta}[b_{OPT}]) \tag{1}$$

with

$$s_{cm}(\boldsymbol{p}^{asc}, \boldsymbol{p}^{dsc}) = \frac{OMNI(\boldsymbol{p}^{asc}, \sigma, \eta) + OMNI(\boldsymbol{p}^{dsc}, \sigma, \eta)}{2} \tag{2}$$

and

$$o_{cm}(\boldsymbol{p}^{-1}, \boldsymbol{p}^{+1}) = abs\left(ENDISI^c\left((\boldsymbol{p}^{-1})_{mean^t}, \alpha, \gamma\right) - ENDISI^c\left((\boldsymbol{p}^{+1})_{mean^t}, \alpha, \gamma\right) \right). \tag{3}$$

The operator $[\cdot]$ returns the subset of bands from the window. Hence, $w_{i,j}^t[b_{OPT}] \in \mathbb{R}^{\Omega \times 32 \times 32 \times b_{OPT}} \subsetneq \mathbb{R}^{\Omega \times 32 \times 32 \times b}$ and accordingly for the SAR bands in ascending and descending orbit directions. The variable Δ is the observation period for the windows and is fixed for each era. Hence, windows $w_{i,j}^{t-\Delta}[b_{OPT}]$ and $w_{i,j}^{t+\Delta}[b_{OPT}]$ denote the windows of the immediate preceding and subsequent observation periods relative to the window starting at time t.

The function s_{cm} creates a change map using the omnibus test statistic, as proposed by Conradsen et al. [12], and herein denoted as $OMNI(\cdot, \sigma, \eta)$. Its parameters σ and η are the significance and ENL for the multi-looking SAR data, respectively. ENL has been rounded to the closest integer. The significance is empirically identified (see Table 4), and the SAR related changes from the omnibus test statistic are averaged over the two orbit directions.

Table 4. Parameters of our proposed method used for creating the windows $w_{i,j}^t$ and synthetic labels $\hat{y}_{i,j}^t$ as needed for the training and validation processes.

Parameter	Mnemonic	ERS-1/2 and Landsat 5 TM		Sentinel 1 and 2	
		Rotterdam and Liège	Limassol	Rotterdam and Liège	Limassol
$b_{SAR}^{[asc \mid dsc]}$	SAR bands	1 (VV)	1 (VV)	2 (VV + VH)	2 (VV + VH)
b_{OPT}	optical bands	7	7	13	13
α	shift	0.25	0.5	0.25	0.5
γ	scale	30.0	30.0	10.0	10.0
σ	significance	0.1	0.1	0.001	0.001
η	ENL	3	3	4	4
δ	step $\left(\frac{\cdot}{observation}\right)$	1 s	1 s	2 days	2 days
ω	min. window size	25	25	35	35
Ω	max. window size	110	110	92	92

It is important to note that we did not clip the DN values of ERS-1/2 or Sentinel 1 and used single-precision floating point numbers, as recommended for the omnibus test statistic in [49]:

> Saturating the extreme pixel values [...] is unfortunate in our situation where the dominating changes detected are due precisely to those strongly reflecting human-made objects [...]. Pixels that are saturated at several timepoints may not be detected as change pixels, which is potentially wrong. The best way to handle this is to store the data as floats [...].

An optical change map is created with o_{cm}, which takes the preceding and subsequent observation period. The operation $(\cdot)_{mean^t} \subset \mathbb{R}^{32 \times 32 \times b_{OPT}}$ is the arithmetic mean over time applied individually to each band. A clipped index $ENDISI^c$ is computed thereof, which is an extended version of the Enhanced Normalized Difference Impervious Surfaces Index (ENDISI) that was originally proposed in [50]. According to a recent study [21]: "The ENDISI algorithm [...] effectively distinguished impervious surface from the background. It not only eliminated the impacts of water bodies, but also suppressed the impacts of most bare soil and arid land". $ENDISI^c \subset \mathbb{R}^{32 \times 32}$ is constructed as

$$ENDISI^c = \begin{cases} 0, & \text{if } ENDISI^e < 0 \\ 1 & \text{if } ENDISI^e > 1 \\ ENDISI^e, & \text{otherwise} \end{cases} \tag{4}$$

with

$$ENDISI^e = (ENDISI + \alpha - MNDBI^+ - 2 \cdot MNDWI^+)\gamma, \tag{5}$$

$$MNDBI^+ = \begin{cases} MNDBI, & \text{if } MNDBI > 0 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

and

$$MNDWI^+ = \begin{cases} MNDWI, & \text{if } MNDWI > 0 \\ 0, & \text{otherwise} \end{cases}. \tag{7}$$

For better readability, we omit the function arguments and consider the $ENDISI^c$ and subsequent functions to operate element-wise on each pixel. Function $ENDISI^c$ clips the values of $ENDISI^e$ between $(0, 1)$, denoting impervious areas that are urban due to their built-up/impervious reflectance characteristic with values closer to one. Values closer to zero indicate no urban pixels. The parameter α shifts the values of $ENDISI$ depending on the site to separate urban and non-urban pixels. The shifting values were empirically identified once for each site (see Table 4). Our values are in accordance with the study [21], where dark impervious surfaces (IS) have the lowest values of $ENDISI$ between ca. $(-0.5, 0)$. To increase the values of differences, $ENDISI^e$ is scaled by the factor γ. This ensures larger change detection values.

$ENDISI^e$ further incorporates $MNDWI$ [51] and $MNDBI$ [52]. The use of $MNDWI$ reduces the false positives of water bodies, while $MNDBI$ mitigates effects of farming where bare soil occurs frequently and can appear as an impervious area. The use of $MNDWI$ also removes false positives due to incomplete cloud removal with clouds detected erroneously as urban by $ENDISI$. Only their respective positive values are used by $MNDWI^+$ and $MNDBI^+$ as they denote water or farming areas and values below zero appear for other types of land cover, including impervious areas. The index $MNDWI^+$ is subtracted with a factor of two because $ENDISI$ tends to assign higher values to cloud fragments. The index methods $ENDISI$, $MNDBI$, and $MNDWI$ follow their original definitions [50–52], which are

$$ENDISI = \frac{\rho_{Blue} - \beta\left(\frac{\rho_{SWIR1}}{\rho_{SWIR2}} + (MNDWI)^2\right)}{\rho_{Blue} + \beta\left(\frac{\rho_{SWIR1}}{\rho_{SWIR2}} + (MNDWI)^2\right)} \tag{8}$$

with

$$\beta = \frac{2(\rho_{Blue})_{mean}}{\left(\frac{\rho_{SWIR1}}{\rho_{SWIR2}}\right)_{mean} + \left((MNDWI)^2\right)_{mean}}, \tag{9}$$

$$MNDBI = \frac{\rho_{SWIR1} - \rho_{Blue}}{\rho_{SWIR1} + \rho_{Blue}} \tag{10}$$

and

$$MNDWI = \frac{\rho_{Green} - \rho_{SWIR1}}{\rho_{Green} + \rho_{SWIR1}}. \tag{11}$$

The individual bands ρ are assigned, as shown in Table 5. The correction factor β, which is defined by $ENDISI$, is determined for the entire AoI for every observation instead of individually for every tile. This is to avoid biases due to limited spatial context, such as impervious surfaces covering most of the tile. The operation $(\cdot)_{mean} \subset \mathbb{R}$ hence computes the spatial mean of all pixels.

Table 5. The bands used for the synthetic label generation.

Symbol	Landsat 5 TM			Sentinel 2		
	Band	Spectral Range (nm)	Resolution (m/pixel)	Band	Spectral Range (nm)	Resolution (m/pixel)
ρ_{Blue}	TM1	450–520	30	B2	459–525	10
ρ_{Green}	TM2	520–600	30	B3	541–578	10
ρ_{Red}	TM3	630–690	30	B4	649–680	10
ρ_{NIR}	TM4	760–900	30	B8	780–886	10
ρ_{SWIR1}	TM5	1550–1750	30	B11	1565–1659	20
ρ_{SWIR2}	TM7	2080–2350	30	B12	2098–2290	20

3.3.3. Expected Noise of Synthetic Labels

The generation of synthetic labels on unfiltered data results in outliers. We clarify the expectations and risks of our proposed approach to later guide the discussion with the obtained results.

The multiplication of the SAR and optical change maps serves multiple purposes. First, the omnibus test statistic-derived changes do not distinguish between urban and non-urban pixels. Hence, the multiplication with optically detected urban changes reduces the false positives, e.g., by detecting changes due to farming or in water bodies. Second, due to atmospheric conditions, outliers are quite common in optical data (e.g., clouds, cloud shadows, changes of irradiation). This can lead to false positives where urbanized pixels are falsely detected. An advantage of the omnibus test statistic-based change detection is that urban areas receive a lower count of detected changes compared to non-urban pixels due to rigid structures. The multiplication further reduces false positives thereof.

On the other hand, our method relies on a certain amount of optical observations to distinguish between urban and non-urban pixels. False negatives occur when there are not enough observations, either by too little SAR observations, which leads to a low multiplication factor through s_{cm}, or by no optical data due to cloud coverage or incorrect cloud removal. We observed the latter for some parts of the Rotterdam AoI where the cloud mask removed some buildings with high reflective white roofs. False-positive scenarios are also possible where erroneously urban pixels are detected optically in non-urban areas. This can happen for bare and arid land that appears impervious for a longer time due to droughts. Depending on the SAR detected changes, they can result as falsely labeled urban changes. To mitigate these errors, we use optical means over the period before and after and ensure there are at least ω observations per each window available. Windows with a number of observations lower than ω are discarded for training. The use of means, however, causes a problem of changes from previous and subsequent observation periods attributed to some degree to the window of interest. As the network under training shall only see the window of interest, it will not be able to exactly match the label but find a general approximate solution.

Our framework is not tied to the described labeling method. Other indices, change detection, and even prediction methods could be used. We, however, focus on long-range changes as they happen in urban development. Shorter term urban changes will be more dependent on the quality of optical observations.

4. Proposed Neural Network

In the following, we describe our proposed neural network architecture to detect urban changes using the data sets discussed earlier for the two eras of interest. Furthermore, we elaborate on the training and validation methodology using the synthetic labels to train the neural network.

4.1. Architecture

Semantic labeling of remote sensing images with high spatial resolution is of great interest to a wide range of researchers. This task consists of assigning a category to every pixel in an image. The pixel classification in urban environments is still a challenging task because extracting spatially consistent information requires quality data processing and high model accuracy. Authors Diakogiannis et al. [53] named two main reasons why urban features are hard to extract/classify: (i) high variability within a class as objects with similar spectral reflectance can belong to completely different classes; (ii) interactions between urban features (e.g., 3D character of buildings, shadows).

For the last decade, Deep Neural Network (DNN) research has been rapidly developing, and therefore, DNN models became very popular as classifiers [54]. The main advantage of these models is their high performance and modular design, which allows their architectures to be easily adapted to specific needs. In our work, we use classification principles using a Convolutional Neural Network (CNN) [55] with the combination of Recurrent Neural Network (RNN) layers [56] to add spatio-temporal awareness. This awareness is needed to identify changes of a certain type, which are urban/impervious-related changes in our case.

We propose a novel neural network architecture that is a concatenation ensemble of two sub-models: the Ensemble of Recurrent Convolutional Neural Networks for Deep Remote Sensing (ERCNN-DRS). The sub-models are designed to ingest multispectral optical and SAR data. Figure 3 shows the ensemble architecture. We denote the network as a function f_{NN} returning a prediction $y_{i,j}^t := f_{NN}(w_{i,j}^t)$ of urban changes that happened in window $w_{i,j}^t$ or with separate data types for clarity: $y_{i,j}^t := f_{NN}(w_{i,j}^t[b_{OPT}], w_{i,j}^t[b_{SAR}^{asc}], w_{i,j}^t[b_{SAR}^{dsc}])$. As we discuss in Section 5, both sub-models can be used individually when only one kind of data type is available but with a reduced performance.

The sub-models are comprised of different convolutional layers [55] and one RNN layer (i.e., convolutional LSTM), which is the core of our architecture. The convolutional LSTM (ConvLSTM) layer learns the deep-temporal dependencies between the observations. This layer type has been introduced by Shi et al. [56] and solves the problem of applying Long Short-Term Memory (LSTM) layers [57] to spatial data. Instead of flattening the input, it retains the spatial information by accepting multi-dimensional inputs that are fed to the gates using further 3×3 convolutions. The ConvLSTM layer is fed from one (SAR) or two (optical) 3×3 convolutional layers. They apply the same set of filters to all observations of a window, which is referred to as the time-distributed convolution. As a result, these layers tend to learn filters to retrieve useful features for the following RNN layer from all or most of the observations in a window. In addition, they can imperfectly compensate co-registered observations caused by nearest neighbor interpolation, which can shift the pixels by one in each direction. After the RNN layers, the activations are concatenated to represent time-invariant features. The concatenation of sub-networks can lead to better performance, as demonstrated in [58,59], and natively facilitates the use of different data types. Two more 3×3 convolutions are applied to spatial filters after the time series is mapped to a set of representative samples. In the last layer, the activations are fed to a 1×1 convolutional layer to correct each pixel of the prediction independently without using any spatial context. All 3×3 convolutions apply batch normalization and subsequently a *ReLU* activation function [60] on their outputs. The ConvLSTM layers instead apply *tanh* activations and *hard sigmoid* for the recurrent step. In addition, they apply a dropout to the linear transformations of the recurrent states [61]. This is a regularization method used to minimize the impact of frequent window sizes biasing the predictions. The 1×1 layer does

not apply batch normalization and applies a *sigmoid* activation instead of *ReLU* to generate a heat map with values in the range of (0, 1) for every pixel in the input tile. Values close to one denote urban changes that are more likely, whereas values closer to zero indicate no urban-related change. All layers also train biases in addition to their kernels.
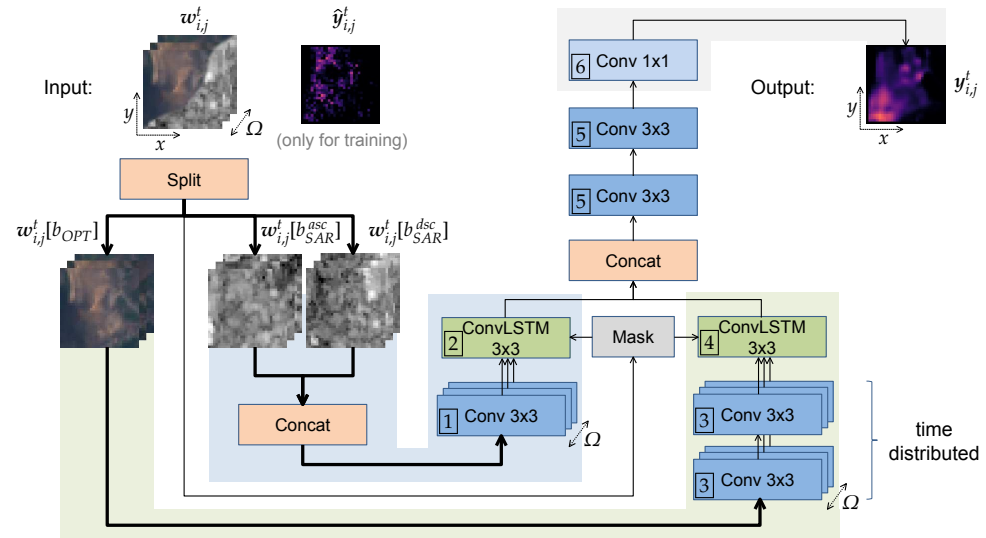

Figure 3. The architecture of ERCNN-DRS. Bold arrows indicate data flow of windows $w_{i,j}^t$ for the different data types (background blue: SAR, green: multispectral optical). The annotations $\boxed{1}\ldots\boxed{6}$ define the different hyper-parameters of the network listed in Table 6 for each era.

Throughout the network, each convolutional layer produces a different amount of filters, which are dependent on the complexity of the input data. The model for the ERS-1/2 and Landsat 5 TM era requires less filters due to less bands, whereas the Sentinel 1 and 2 era uses more bands containing more information. In addition, the higher spatial resolutions show finer spatial structures for which more filters can be useful. We intentionally left the convolutional kernels at a size of 3×3. Given the amount of data available per observation period, larger kernel sizes would result in an approximately quadratic increase of parameters of the neural network. We traded learning complex spatial structures to learning deeper temporal correlations. This is especially important as changes over time should have more weight compared to identifying urban micro-structures. As such, our trained network is aimed at an applicability for a wider range of AoIs as it learns the structure of urban (impervious) surfaces less, which can be different for different geographies. The limitation of spatial filters helps to avoid memorization of certain urban structures, such as reflective characteristics of rooftops, parking lots, or roads. Instead, general characteristics of impervious surfaces should be learned through their changes over time. Table 6 summarizes the important parameters used for the respective layers.

Due to the deep-temporal series with varying amounts of observations, we apply masking to the ConvLSTM layers to indicate which first l observations should be considered. This is different for each observation window, resulting in an individual mask. There is an upper bound Ω to the number of observations in the model architecture. To avoid losing observations, this bound needs to be set to the largest number of observations expected (see Table 4). As such, only the first l observations are used and the remaining $\Omega - l$ observations, which were padded and do not contain valid information, are masked out. Masking also helps to exploit the capacity of the network more efficiently as the network does not need to learn how to identity the valid observations and ignore the padded ones. As l is already known a priori, it does not need to be learned after all.

Table 6. The neural network model hyper-parameters used for the different eras.

Hyper-Parameters		Configurations					
		1	2	3	4	5	6
ERS-1/2 and Landsat 5 TM	Filters	4	4	20	20	8	1
	Kernel	3×3	3×3	3×3	3×3	3×3	1×1
	Stride	1×1	1×1	1×1	1×1	1×1	1×1
	Activation(s)	ReLU	tanh, hard sigmoid	ReLU	tanh, hard sigmoid	ReLU	sigmoid
	Dropout		0.4		0.4		
Sentinel 1 and Sentinel 2	Filters	10	10	26	26	8	1
	Kernel	3×3	3×3	3×3	3×3	3×3	1×1
	Stride	1×1	1×1	1×1	1×1	1×1	1×1
	Activation(s)	ReLU	tanh, hard sigmoid	ReLU	tanh, hard sigmoid	ReLU	sigmoid
	Dropout		0.4		0.4		

4.2. Training and Validation Methodology

The aforementioned data are separated into training and validation sets. For each era, AoIs Rotterdam and Limassol were used during the training process. For the training data set, tiles are drawn from $\{(i,j) \in \mathbb{N} \times \mathbb{N} \mid j \equiv \frac{i}{2} \pmod 2\}$, and for the validation data set, they are drawn from $\{(i,j) \in \mathbb{N} \times \mathbb{N} \mid j \equiv \frac{5-i}{2} \pmod 4\}$. This is approximately one-fourth of all tiles used for training and one-eighth for validation with both sets being disjunctive. The tile selections ensure non-adjacent tiles in the respective sets to avoid recurring and redundant patterns along the borders.

For each tile assigned to the training or validation data set, a subset of its windows $w_{i,j}^t$ is selected. Leaving aside corner cases where preceding $(w_{i,j}^{t-\Delta})$ or subsequent $(w_{i,j}^{t+\Delta})$ windows with more than ω observations are required for the synthetic labels, a window can be created for every observation that occurred at t. We select a random subset using a uniform distribution with $p = \frac{1}{10}$. Therefore, we expect approximately one-tenth of the overall windows to be selected. The random uniform distribution function uses different seeds for every tile for reproducibility and to avoid all tiles of the same AoI from having the same window starting times, which the neural network could learn. This is especially a concern as only two different AoIs are used for training, of which all the tiles would have the same sequence of observations within each AoI. If windows are identical for tiles of the same AoI, the sequence can be learned, but applying it to other AoIs with different sequences would result in bad prediction performance. This random selection also reduces redundancies as subsequent windows would almost fully overlap except for one or a few observations.

The selection of the loss function for the supervised learning has a significant effect on the results as it guides the network towards a solution during training. Several (region-based) loss functions for semantic segmentation and prediction problems have been developed previously. A few examples are the Dice, Focal Tversky, and Intersection over Union (IoU) loss [62]. In a recent study, a new loss as the extension of the Tanimoto coefficient—so-called Tanimoto with complement coefficient—was demonstrated by Diakogiannis et al. [53]. It shows improved gradients of first and second-order derivations. As our labels usually contain small values, a loss function is preferred that provides stronger and higher quality gradients. In addition, the Tanimoto coefficient is also an ideal candidate for segmentation tasks as it shares similarities with the Dice coefficient or IoU and ignores true negatives. Such true negatives would dominate the number of pixels as changes are less likely than no-changes. This unbalance would lead the network to bias the no-change background class.

The Tanimoto coefficient T is based on the Sorenson–Dice coefficient. T and it's complement $\widetilde{T}$ are defined as

$$\widetilde{T}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{T(\boldsymbol{y}, \hat{\boldsymbol{y}}) + T(1 - \boldsymbol{y}, 1 - \hat{\boldsymbol{y}})}{2} \tag{12}$$

with

$$T(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{\sum_i y_i \hat{y}_i}{\sum_i (y_i^2 + \hat{y}_i^2) - \sum_i (y_i \hat{y}_i)}. \tag{13}$$

For T and $\widetilde{T}$, $\hat{\boldsymbol{y}}$ and $\boldsymbol{y}$ are the ground truth label and prediction, respectively. We use $\mathscr{L}(\hat{\boldsymbol{y}}, \boldsymbol{y}) := 1 - \widetilde{T}(\hat{\boldsymbol{y}}, \boldsymbol{y})$ as the loss function.

Other studies [63–66] have found and discussed the adverse effects of tiling with respect to increased errors, especially towards the tile borders. As a result, we applied the Tanimoto with complement loss only to the center 30×30 pixels of every tile. The use of 3×3 convolutional filters in the neural network requires the information of its directly adjacent set of pixel neighbors for every pixel. This is only partially possible for pixels at the tile boundaries, such as the edges and more pronounced at the corners. As a result, the prediction quality at the boundaries would show a degraded quality and some capacity of the network would be used for such corner cases. Fully convolutional neural networks are not bound to any tile size, as already trained fully convolutional networks can use varying tile sizes during inference, irrespective of the tile size used during training. This enables an entire scene (AoI) used as one tile, as long as it fits into memory to speed up prediction and avoid the boundary effects. We only constrain the tile size to 32×32 pixel tiles to avoid excessive memory requirements during training.

5. Results

The neural network proposed in Section 4 was trained with different hyper-parameters and training data from Rotterdam and Limassol AoIs for each era, as previously described in Section 3. In the following, we describe our training environment and configurations. Due to the underlying methods and training objective, a detailed evaluation of prediction correctness is hard to accomplish. There are no second sources that provide enough data with similar sampling intervals and sufficient spatial resolutions to examine the performance of the neural network predictions. Instead, we carry out two ablation studies to evaluate the impact of urban changes over time and the contribution of optical and SAR observations. Finally, we show some example predictions for both eras, where we found historic Google Earth™ very high-resolution imagery, which explains (most of) the changes.

5.1. Training

We trained two models, one for every era, with the hyper-parameters shown in Table 6. The training was carried out on one compute node with four NVIDIA Tesla V100-SXM2 GPUs on the Barbora cluster at IT4Innovations (https://docs.it4i.cz/barbora/hardware -overview/; accessed on 29 July 2021). Tensorflow 2.4 with Keras and Horovod [67] was used for data parallel training, utilizing all four GPUs. We used a batch size of 32 per GPU (effective batch size of 128), a learning rate of 0.004 (fixed over all epochs), and the distributed synchronous minibatch stochastic gradient descent (Sync-SGD) solver [68,69] with a momentum of 0.8.

The development of the losses over the training epochs for the two models are shown in Figure 4. The ERS-1/2 and Landsat 5 TM model was trained for 126 epochs with a walltime of 40:23 h and the Sentinel 1 and 2 model for 115 epochs with a walltime of 23:30 h. For the former, we used the trained weights at epoch 86, and 98 for the latter (both zero-based), which showed the best validation loss until overfitting started to become visible. Figure 4 shows the start of overfitting marked with a green dashed vertical line. It also shows only one training session but with validation data sets split and assigned individually to each of the four GPUs (shuffling was only applied to training data). Hence,

the validation losses give insight into how well balanced the found solution is. In the ERS-1/2 and Landsat 5 TM model, it is visible that one loss (`val_1`) shows higher values, which likely stems from the Limassol AoI where there is a slight imbalance regarding the (validation) tiles containing water areas. Such tiles contain larger errors. We have rerun the training sessions multiple times and repeatedly observed the same behavior. This was also true during some limited hyper-parameter searches— no exhaustive search has yet been applied. The data sizes for the ERS-1/2 and Landsat 5 TM model are 58 GB (training) and 7 GB (validation) and 187 GB (training) and 24 GB (validation) for Sentinel 1 and 2. All are used as GZIP-compressed TFrecord files from the training sites Rotterdam and Limassol. To reduce the resource needs, only one-fourth of the validation data was used, and the majority of tiles covering sea for Limassol AoI have been removed.

In addition, we trained only a subset of the ensemble by using either of the two sub-networks $^{OPT}f_{NN}(\boldsymbol{w}^t_{i,j}[b_{OPT}])$ and $^{SAR}f_{NN}(\boldsymbol{w}^t_{i,j}[b^{asc}_{SAR}], \boldsymbol{w}^t_{i,j}[b^{dsc}_{SAR}])$. These correspond to the ensemble network with one input data type removed, e.g., $^{SAR}f_{NN}$ does not consider any optical input. All hyper-parameters were left unchanged. Figure 4 shows the results for $^{OPT}f_{NN}$ in the middle and $^{SAR}f_{NN}$ on the right of each era. It is visible that both do not reach the same low loss as with the ensemble configuration. In particular, $^{SAR}f_{NN}$ shows degraded performance. This likely stems from less information to distinguish urban-related changes and suggests that the overall solution is mostly driven by optical data. However, in the ensemble case, SAR can add additional information, leading to lower generalized losses. This is more visible for the Sentinel 1 and 2 era than for ERS-1/2 and Landsat 5 TM but is consistent across both eras. Furthermore, the training requires less epochs to reach the best general solution. We, however, have to stress that the hyper-parameters were not changed. A change of batch size or SGD momentum could result in less epochs, or more filters could result in an overall improved performance for the non-ensemble cases as well.
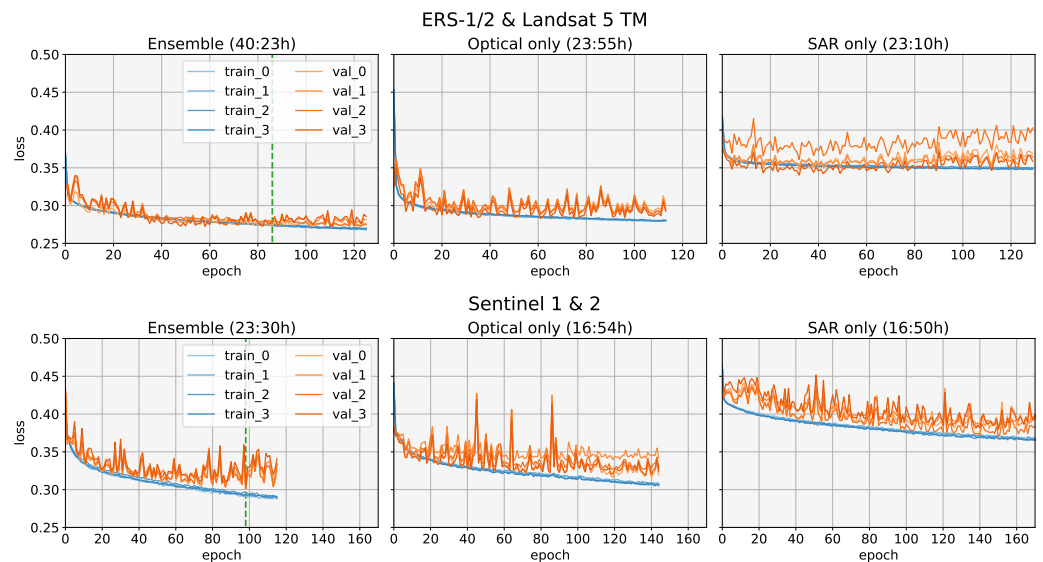

Figure 4. Loss values over epochs for both eras: on the left, the ensemble with both multispectral optical and SAR data; center and right, only optical and SAR data, respectively. The green vertical dashed line indicates the best ensemble parameters.

The ERS-1/2 and Landsat 5 TM model reaches lower loss values. This is a result of the lower spatial resolution of the remote sensing sensors. The observations hence contain less detailed structures that are easier to learn. The model for Sentinel 1 and 2, for example, is able to identify finer grained urban structures, such as roads and mid-sized buildings, whereas Landsat 5 is only able to detect very large urban structures, such as highways or logistics buildings, at best.

We would like to briefly elaborate on the aforementioned parameters Ω, the tile size, and their value selection. In theory, neither of them would need to be artificially

bound, leaving the maximum possible number of observations for a given window period to be only defined by the remote sensing mission parameters. Furthermore, any AoI could be considered as one single tile. In practice, however, we limit both parameters to accommodate the neural network model and data in the available (GPU) memory — in our case, $4 * 16$ GB. Similar to other works considered, the use of tiles (also called patches) "under a given GPU memory constraint" [66] as well as "the input to the network [is] fixed to a given size [...] mainly influenced by considerations of available memory size on the GPU" [70], and we also regard tiling as a compromise to enable deeper neural networks and data rather than a profound method. Similarly, we treat the limitation of observations per window with Ω as a compromise as well. Tiling can increase the prediction errors due to reduced spatial information, whereas the use of Ω and the related step size δ contribute to a reduction of temporal information.

5.2. Ablation Studies

We attempt two ablation studies to understand the quality of responses to urban changes and the contribution of each data type, optical and SAR, to the overall solution of the proposed ensemble network. Both document the performance of the neural network using deep time series and help to set the expectations of the predictions.

5.2.1. Urban Change Sensitivity

To analyze the sensitivity of the model predictions to urban changes over time, we study them using an artificial but well-controlled test. In this test, we select one 32×32 pixel tile for one window that does not cover an urban area and substitute the 16×16 pixel center with the a tile that contains an urban area. More specifically, we use one that is fully covering a forest as a reference tile and one from the airport of Liège for the center. The resulting perturbed tile is called hybrid ${}^{*}w_{i,j}^{t}$. The window covering the forest was selected so that it contained the maximum amount of observations among all other windows. Otherwise, windows with less observations would not allow to study the substitution effect towards Ω but only the first l observations. Figure 5 shows the substitutions for both eras. The substitution in a confined space in the center is deliberate as CNNs tend to show performance decay towards the edges of tiles as mentioned earlier. By that, we minimize such errors. Both locations are from the same time period, but the urban replacement is constant. As such, only changes occur at the beginning and end of the substitution length. The substitution is controlled via two parameters ζ and ψ for the start and length of the substitution, respectively. Both parameters are evaluated with single increments, and a prediction of the perturbed tile is applied. It is expected that the insertion of the urban sub-tile results in a positive prediction without dependence when the substitution happens (ζ). The longer the substitution persists, the stronger the predicted change should be, falling off after $\frac{l}{2}$. Of the predictions, we compute the mean over the 30×30 pixel center of the tile as a continuous representation of sensitivity.

Figure 6 shows the bivariate means and their intensity values, depending on ζ and ψ; for better readability, we only used every second value for ψ. The curves with $\psi = 0$ (i.e., no substitution) show constant values close to zero. Forests are neither urban nor impervious, and there is no such change in the reference tile for the window considered. As a result, every substitution with an impervious area should show values significantly larger than zero. What is more, we only consider sizes of ψ that fit into the window. Therefore, growing substitution lengths ψ delimit the start of substitution ζ towards the end.
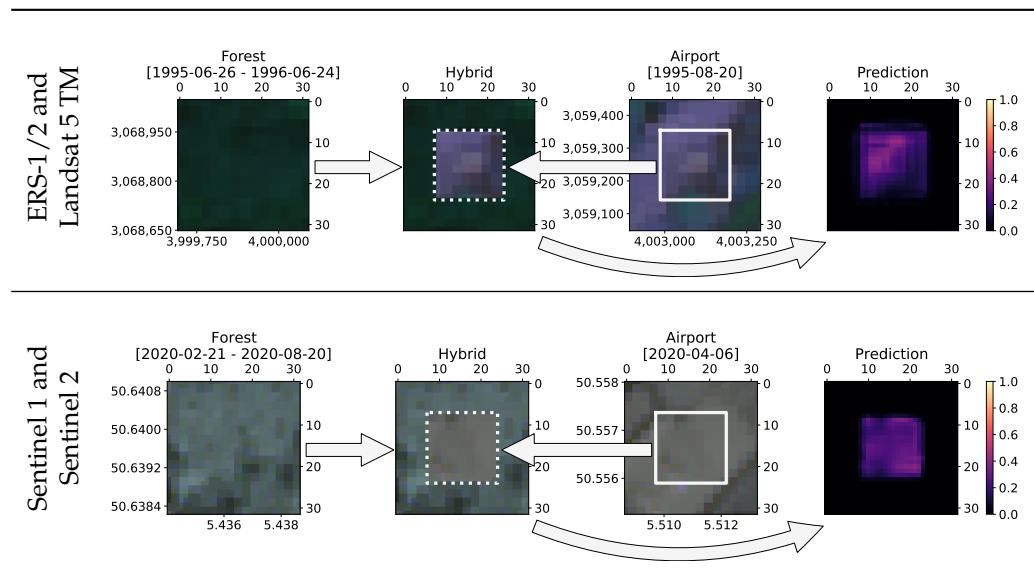
Figure 5. For each era, the substitution of the forest area in the center 16×16, taken from Liège airport. The forest tile is used with all observations $w_{i,j}^t$ whereas the airport sub-tile is from a fixed date. The predictions of the hybrid $^*w_{i,j}^t$ is on the right for $\psi = 30$ and $\zeta = 5$.
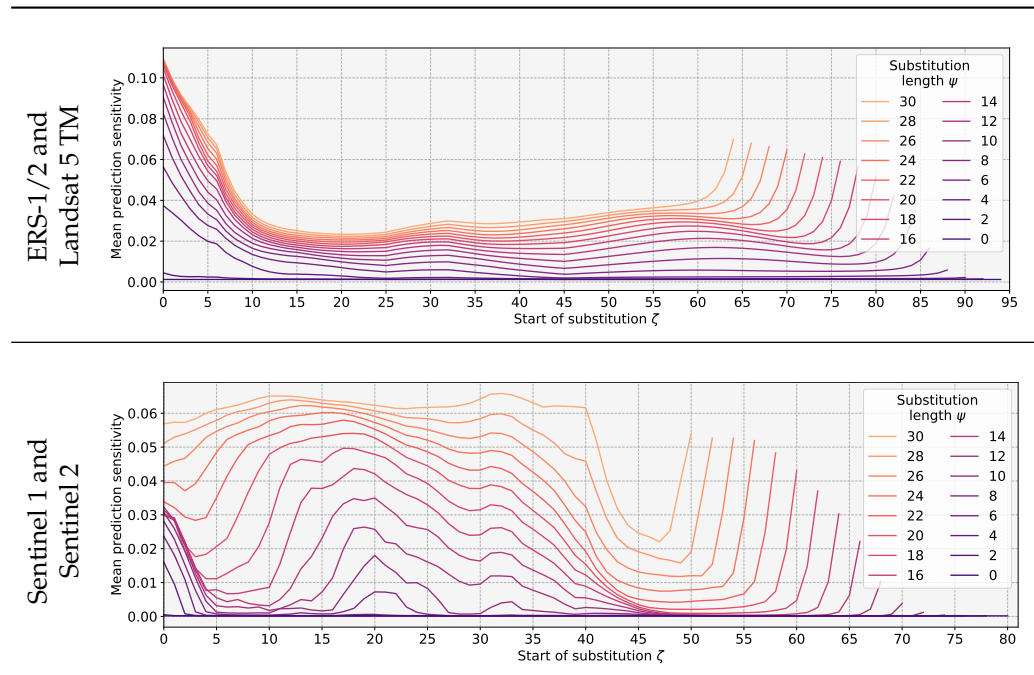

Figure 6. For each era, the sensitivity to the substitutions from Figure 5, depending on the substitution length ψ and the substitution start time ζ (zero based). For ERS-1/2 and Landsat 5 TM, the hybrid window $^*w_{i,j}^t$ contains 95 observations. For Sentinel 1 and 2, 81 observations are in the window. The curve with $\psi = 0$ shows the unmodified forest tile with a mean prediction close to zero.

With the substitutions carried out, different effects become visible. First, changes close to the beginning of the windows show a higher change intensity and tail off the further out as changes happen. These later scopes represent a dead spot of the predictions as any change there is attenuated. The reason for that behavior is the training of different periods with varying amounts of observations. With ω, a minimal amount of required observations was set. Periods with more observations are less likely and, hence, result in a lower weighting for the final predictions. At the end, the intensities rise again, leading to more weight put on the changes at the end of the windows. The ERS-1/2 and Landsat 5

TM model follows that pattern very clearly. For the Sentinel 1 and 2 model, the response is a bit more complex but with the same underlying pattern. Differences are mostly the strong attenuation towards the beginning of the window for smaller ψ with more pronounced plateauing up to the first ω observations.

Second, changes that are short are attenuated and increase the mean intensities with the growing ψ. This likely stems from errors in the observations where short outliers need more (temporal) filtering. For example, co-registration errors can lead to some pixels not detected as a change, unless there are enough observations (with the same error) increasing confidence in the change.

We hence can conclude that only changes that occur for a longer time and close to the beginning of the observation period (or right at the end) are well detected. The former is acceptable as urban changes typically occur over a longer time. The latter can be compensated by using repeated and overlapping incremental predictions to evenly cover a selected period.

5.2.2. Correlations of SAR and Optical Predictions

Since an ensemble of networks is trained, we analyze the contribution of these sub-networks to the overall solution. More specifically, we try to understand how well their ingested data types (SAR or optical) are influencing the result. We create two synthetic configurations. For each, we keep one data type constant, while the other remains unchanged. The constant data type is the mean over the entire period to avoid outliers biasing the predictions under analysis. As a result, one configuration contains a fixed mean of all the SAR observations (ascending and descending), while optical is unchanged, denoted as $^{OPT}y_{i,j}^t := f_{NN}(\boldsymbol{w}_{i,j}^t[b_{OPT}], (\boldsymbol{w}_{i,j}^t[b_{SAR}^{asc}])_{mean^t}, (\boldsymbol{w}_{i,j}^t[b_{SAR}^{dsc}])_{mean^t})$. The second variation is vice versa, using the averaged optical observations, while SAR observations are unchanged, denoted as $^{SAR}y_{i,j}^t := f_{NN}((\boldsymbol{w}_{i,j}^t[b_{OPT}])_{mean^t}, \boldsymbol{w}_{i,j}^t[b_{SAR}^{asc}], \boldsymbol{w}_{i,j}^t[b_{SAR}^{dsc}])$. As the network ensemble applies most of the spatial and all temporal operations individually for each data type, we expect to receive results close to each other if both data types contribute equally to the overall prediction. However, as we apply three more convolutions, two with *ReLU* and one with *sigmoid* activation functions after the concatenation layer, the two perturbed sub-network outputs are not directly comparable quantitatively. To correlate such non-linear continuous variables, we compute Kendall's τ coefficient [71] as a measure of closeness over all windows and tiles. Tiles are aggregated to one large prediction for each window, $^{OPT}y^t$ and $^{SAR}y^t$, respectively. We apply two correlations of $^{OPT}y^t$ and $^{SAR}y^t$ against y^t. The latter are the tile-aggregated predictions with both data types unchanged. As before, we ignore the border pixels and only consider the center 30×30 pixels of every tile.

As shown in Figure 7, both perturbations show a high positive correlation. The correlations of $^{OPT}y^t$ show up the highest. This indicates that it is harder for the network to identify changes as urban or impervious related from SAR data alone ($^{SAR}y^t$). Other work [27,28] demonstrated that urban areas can be detected to some extend from only SAR observations. Similarly, the SAR sub-network was also able to identify urban changes to a good degree, as indicated by the positive correlations. Seasonal changes are visible, resulting in recurring patterns of correlation values for $^{SAR}y^t$. The higher correlation to optical data suggests that the overall prediction y^t relies more on optical data, which are affected by seasonal changes. As SAR is less affected by seasonal changes, these patterns become visible but are still of good correlation. The outlier for the ERS-1/2 and Landsat 5 TM correlations between 2002 and 2003 stems from the lack of optical data in that period (see Figure 1). As expected, the SAR observations almost fully correlate (i.e., contribute) to the final solution due to the loss of optical data.
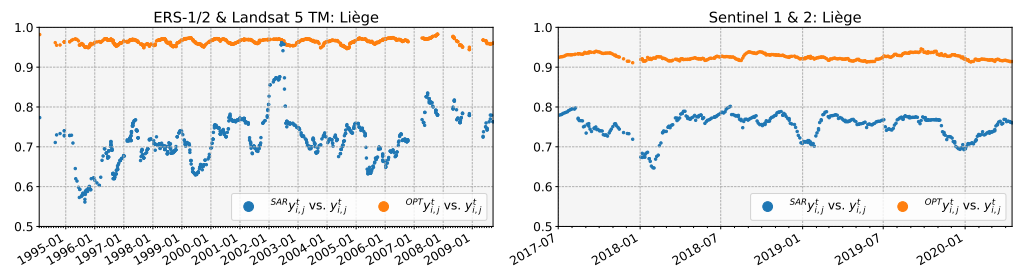
Figure 7. Kendall's τ coefficient for fixed SAR ($^{OPT}y^t$) or optical ($^{SAR}y^t$). Every data point is the start of a window w^t used for the correlation against its prediction y^t. (**Left**): ERS-1/2 and Landsat 5-TM; (**Right**): Sentinel 1 and 2.

On a related note, the positive correlation between the two data types and, in turn, the correlation of sub-network ensembles justifies the concatenation we applied. This is suitable for the task at hand instead of applying an averaging ensemble, which would be preferable for uncorrelated cases [59].

5.3. Qualitative Analysis

In the following, we present some examples of predictions for both trained models. We use as predictions $y_{i,j}^{max} := \left(f_{NN}(w_{i,j}^t)\right)_{max^t}$ with $(\cdot)_{max^t} \subset \mathbb{R}^{32 \times 32}$ as the maximum of every window's prediction throughout each era. We would like to note that this will promote outliers of the predictions. Other filtering methods in post-processing could be applied to mitigate these effects, such as computing the mean over time $(\cdot)_{mean^t}$ to only highlight changes over a longer period (i.e., constant background activities) or the difference $(\cdot)_{max^t} - (\cdot)_{mean^t}$ to suppress these background activities.

Figure 8 shows example predictions $y_{i,j}^{max}$ (*i.–iv.*), each for one tile, using the model trained for ERS-1/2 and Landsat 5 TM. As mentioned above, the prediction is the maximum over all the windows, with a total count of 843 with starting times ranging from 7 February 1994 to 19 September 2009. Due to our processing pipeline, we do not consider the first and last window period (1y/6m) of each era. The predictions are shown in the top middle. We used Google Earth's historic very-high resolution imagery (bottom row) to document the changes and show observations before and after the respective change. Due to the limitations of this resource, only a few samples are available, and not all changes can be investigated thoroughly. However, we selected samples that explain the detected changes concisely. In addition, we show the true color images from Landsat 5 TM as a comparison and as a scale for the resolution the network works with at an approximately similar time as the Google Earth imagery. The very high-resolution imagery at the bottom is overlaid with a red mask derived from $y_{i,j}^{max}$ to help comparing the images with the predictions. Examples *i.* and *iii.* from Figure 8 show changes over a larger area with (de-)constructions of buildings. In example *i.*, buildings were deconstructed. A series of predictions of this example is shown in *v.* Example *ii.* shows one hotspot (bottom right), which was used for longer ongoing construction purposes in the surrounding area. It also shows the limitation of sensor resolution as the new road on the top right is not detected. Example *iii.* shows exceptionally high prediction values due to a longer construction period. In *iv.*, a construction of a parking lot next to an existing building is shown. In the same example, the found changes on the top left in the prediction were due to SAR-detected changes in the time frame from approximate 2003-08 to 2003-10. We were not able to obtain other data to further investigate the root cause.
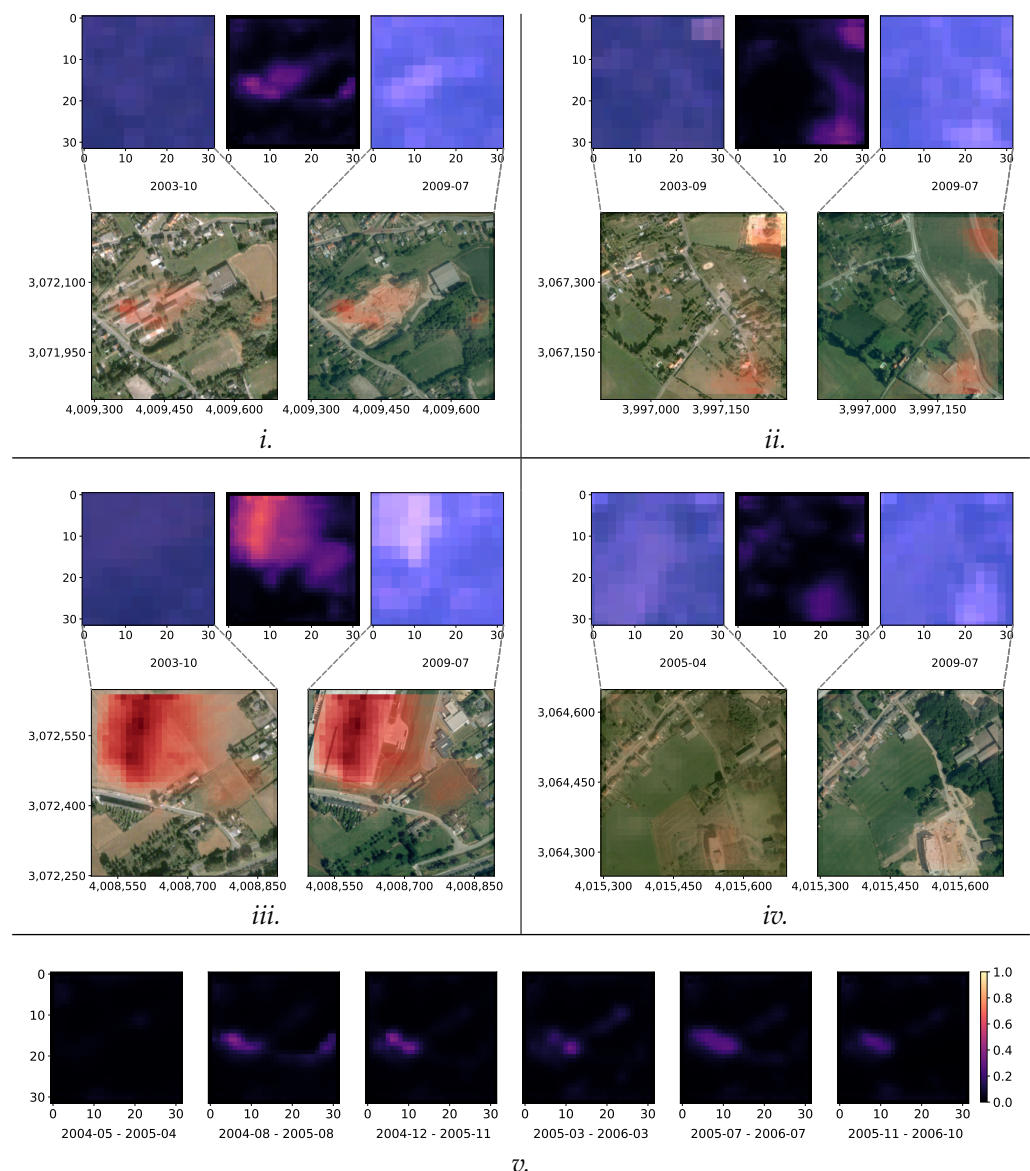
Figure 8. ERS-1/2 and Landsat 5 TM examples of Liège (*i.–iv.*). **Top** rows are Landsat 5 TM true color observations (**left**, **right**) with the change prediction (**middle**). **Bottom** rows are corresponding very-high resolution imagery from Google Earth, © 2021 Maxar Technologies, with predictions superimposed in red. Example *v.* shows a series of predictions from *i.*

Same as for the ERS-1/2 and Landsat 5 TM, we provide four example predictions $y_{i,j}^{max}$ (*i.–iv.*) in Figure 9 but with the model trained for Sentinel 1 and 2 data. For Liège, a total of 381 windows were used, ranging from 3 July 2017 to 15 April 2020. All intensities in $y_{i,j}^{max}$ are higher due to the temporal change characteristics, as discussed in Section 5.2.1. It is immediately visible that the higher sensor resolutions provide finer detailed changes, e.g., the construction of a single house in *i.* Larger constructions and deconstructions in *ii.* and *iii.* are also detected with hotspots where changes happened more often. Example *ii.* shows a hotspot on the bottom left that is caused by a (temporary) construction started in 2020-09. We were not able to obtain other data for further investigation. In *v.*, this is shown with a series of predictions. For example, in *iii.*, the building on the top right is not fully covered by the predictions due to the center of the building being unchanged, and co-registration variances lead to "smearing" of this unchanged area. Example *iv.* shows the beginning of open pit mining with hotspots where most activities happened.
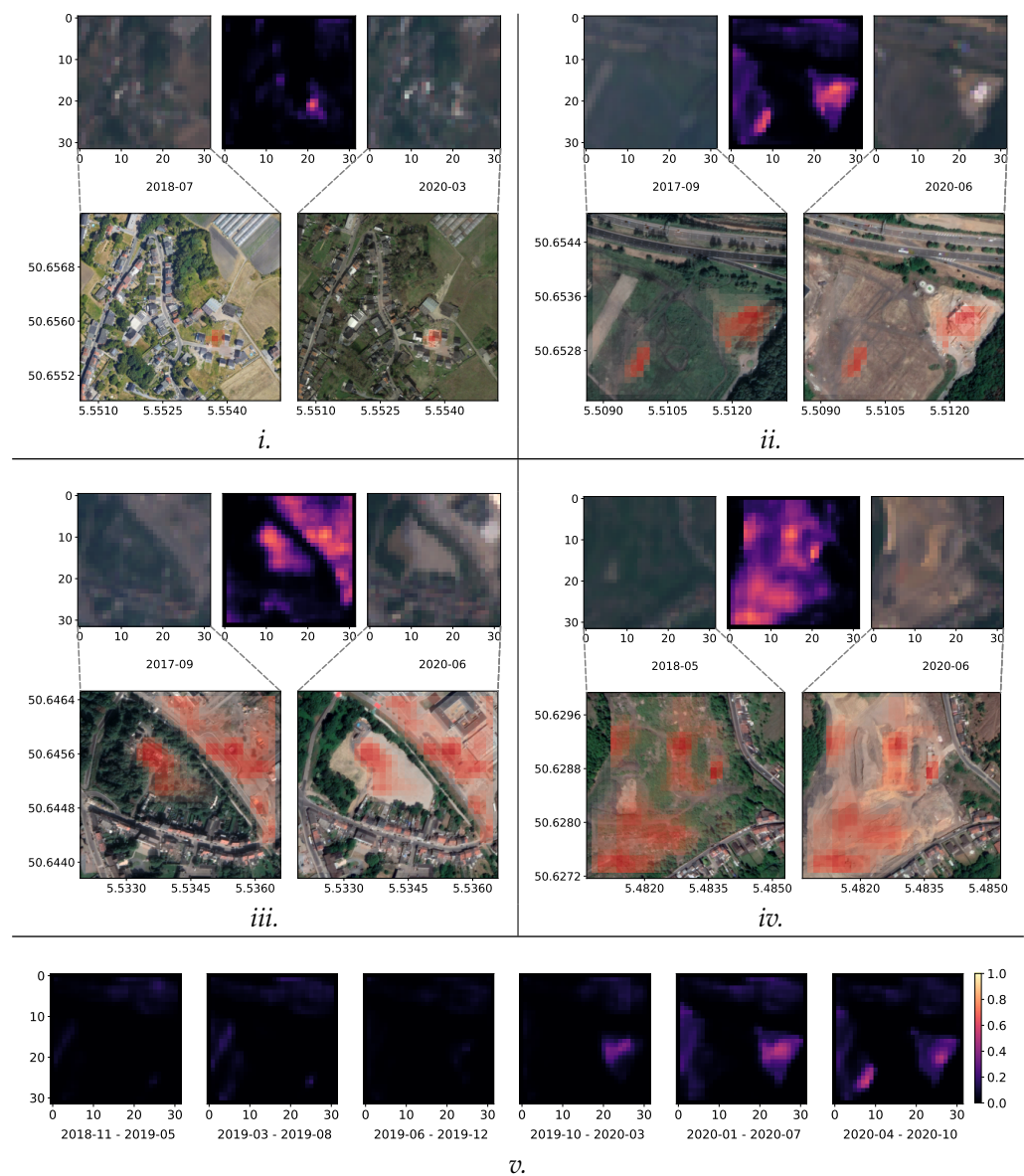
Figure 9. Sentinel 1 and 2 examples of Liège (*i.–iv.*). **Top** rows are Sentinel 2 true color observations (**left**, **right**) with the change prediction (**middle**). **Bottom** rows are corresponding very-high resolution imagery from Google Earth, © 2021 Maxar Technologies, with predictions superimposed in red. Example *v.* shows a series of predictions from *ii.*

6. Discussion

In the following, we discuss limitations of our proposed method and elaborate on improvements to mitigate or solve these limitations.

6.1. Current Limitations

Our proposed method uses the available observations as long as they can be co-registered. No further quality assessment was applied. On one hand, this enables a highly automatic process of training and inference, but it also introduces uncertainties and errors. The GUF [47] and WSF [46] clearly show the efforts needed for high-quality detection of urban areas. However, they only provide a single snapshot due to the high amount of resources needed. Our solution is designed for easy access to urban change detection for monitoring up to the immediate past but with higher uncertainties. We have inspected the predictions of the Liège AoI for both eras and discuss our observations with respect to limitations of our method in the following. Due to the higher available resolution of

Sentinel 1 and 2, we show examples only for this era. The limitations are the same for the ERS-1/2 and Landsat 5 TM era.

The remote sensing data are sampled with irregular intervals and varying quality due to clouds, shadows, irradiance changes, co-registration errors, atmospheric effects, snow/ice, or saturation. Hence, not all changes on the ground can be detected at the intensity they should be, and sampling gaps exist due to no or sparse observations. Furthermore, the imbalance of remote sensing types can cause degraded prediction quality, as shown in Figure 7 for Liège in the years 2002–2003. This all is embedded in the original data quality and could only be mitigated with down-sampling (e.g., moving window averaging) as done by other work [14]. This, however, yields temporal loss and dampens high frequency changes.

Artefacts of clouds, snow, and ice can become visible in the predictions, as shown in Figure 10*i*. This example shows incomplete cloud removal where the dense core of the cloud reflects a spectrum that appears as a sudden impervious/urban change. We have observed these cases very rarely, and light clouds do not seem to affect the predictions noticeably, as can also be seen in the same example.

Bright or reflective surfaces (optical) or metal objects (SAR) can cause outliers and saturation at the sensors. This leads to large gradients that are erroneously detected as changes. For SAR, Nielsen et al. [49] state that these "outliers are usually due to strong reflections from sharp angles on antennas and other human-made objects". The same is true for optical observations of high-reflective objects (e.g., with white roofs or solar panels), which can change their reflectance values due to weather, irradiance, and atmospheric effects quite significantly. Figure 10*ii,iii* show the optical saturation by the reflection of sunlight due to solar panels and the combination of optical and SAR saturation through metallic roofs, respectively. The mitigation of these effects would be possible by filtering over time to reduce outliers, which trades-off temporal resolution and results in a less sensitive change detection. Alternatively, saturations can be masked, which we applied to Landsat 5 TM data with the 16 bit BQA data. This is not directly available for the other data sources.

The distinction of uncultivated farmland from impervious and construction areas can at times be difficult, especially due to the phenological changes in combination with dry periods. We noticed that esp. during summer (e.g., the drought in central Europe in June–July 2018) some fields are more likely detected as urban changes due to their arid appearance. Figure 10*iv.* shows this effect and false detection where, only for a short time, arid land was present, which resulted in a significant response. We also experimented with other index methods that are more resilient to such effects, such as Combinational Biophysical Composition Index (CBCI) [25] or a different shift parameter α for the used ENDISI, which resulted in less sensitivity of early construction stages. We also applied the EBBI [24] to Landsat 5 TM with good results, but it cannot be applied to Sentinel 2, which lacks a needed thermal band. As a consequence thereof, open pit mining is recognized as urban changes (see Figure 9*iv*). In our proposed solution, we decided to accept these false positives in favor of a more rigorous detection. Farming and mining is usually a series of constant changes over time and can be filtered if longer time series observations exist. As an example, the mean prediction intensities of the year(s) of interest can be subtracted from the maximum of all predictions, as mentioned earlier in Section 5.3.

The networks were trained to highlight urban related activities (i.e., changes involving impervious areas). Sensors have a coarse resolution, which does not allow identifying mobile objects (such as cars, trucks, planes, containers) nor distinguishing them from construction changes due to the mixed pixel problem [72]. However, they cause changes of reflectance to individual pixels. As a result, the network detects changes of urban activity in general. Urban centers with high activities show higher intensities than rural areas. The same is true for highly utilized roads and harbors with more likely identified activities, as shown in Figure 10*v.,vi.*, respectively. Since these changes are typically also constant background activities, the same mitigation strategy as for farmland can be applied.
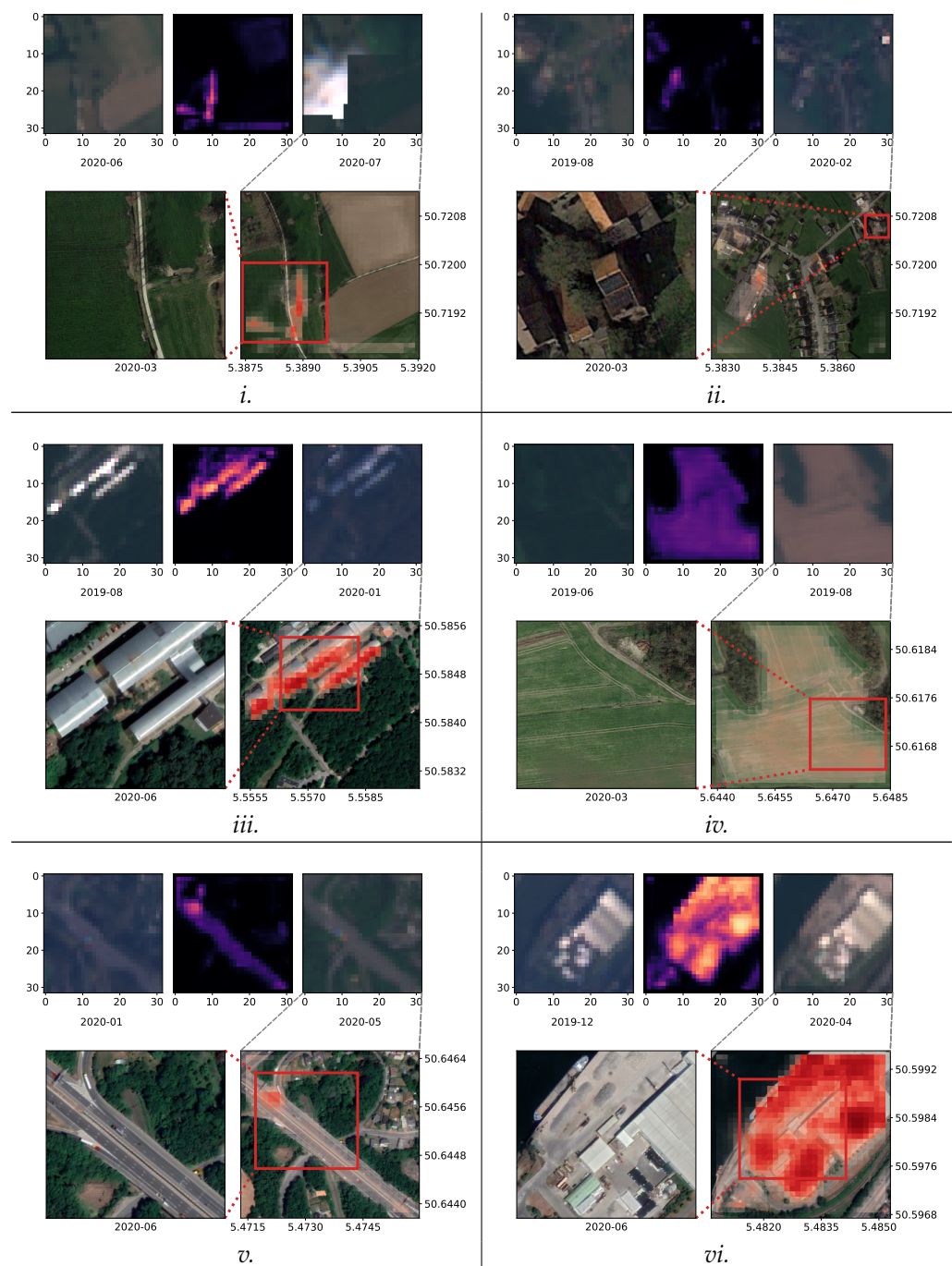
Figure 10. Limitations of the current method. We demonstrate with Sentinel 1 and 2 examples of Liège. **Top** rows are Sentinel 2 true color observations (**left**, **right**) with the change prediction (**middle**). **Bottom** rows show the corresponding very high-resolution imagery from Google Earth, © 2021 Maxar Technologies, with predictions superimposed in red.

The mixed pixel problem and the limitation of the sensor resolutions limit the detection capabilities of our method. For example, in Figure 9*ii*, a new road on the top right was not detected. It is shared by multiple pixels due to which the surrounding vegetation reduces impervious reflectance characteristics. As investigated earlier, our solution has a bias towards optical-based detection of changes. Even though ERS-1/2 has a higher resolution than Landsat 5 TM, it did not contribute to finding such small changes.

6.2. Further Improvements

For inference, we have retained the same tile sizes of 32 × 32 and only used the center 30 × 30 pixels for the predictions. As our proposed neural network architecture is fully convolutional, we can allow any tile size without retraining. In the ideal case, the entire AoI can be predicted at once. However, the deep time series induces significant memory requirements and depending on the used inference platform, practical limitations of tile sizes exist. In the work from Isensee et al. [66], a sliding window approach is proposed to enable the use of tiles (patches) with the mitigation of stitching errors:

> Images are predicted with a sliding window approach, where the window size equals the patch size used during training. Adjacent predictions overlap by half the size of a patch. The accuracy of segmentation decreases towards the borders of the window. To suppress stitching artifacts and reduce the influence of positions close to the borders, a Gaussian importance weighting is applied, increasing the weight of the center voxels in the softmax aggregation. Test time augmentation by mirroring along all axes is applied.

This solution can be applied to our two-dimensional predictions, using pixels instead of voxels, applied directly after the final *sigmoid* activation.

Furthermore the training process can be improved by considering the real number of observations per pixel. In our current work, we have temporally stacked the observations to fill the gaps caused by unavailable data due to anomalies, cloud coverage, orout-of-swath conditions. As a result, every pixel in a window can have a different amount of real (effective) observations as it is influenced by replications of the same value to cover the gaps. This can be taken into account with the regularization of the loss function. Pixels that have the least amount of effective observations should only minimally influence the loss. This can further be extended to inference to combine the prediction with a confidence map. Change predictions of individual pixels indicate lower confidence if there are not enough effective observations backing the prediction. This can help in identifying real changes by lowering the error noise.

Furthermore, the training with noisy labels can be further improved. Many approaches [73] exist to learn from noisy labels with deep neural networks. We would consider the *SELF* method from Nguyen et al. [74] as a viable approach. This could help to push out the start of overfitting to lower the training losses further, leading to more accurate predictions. A downside of *SELF* is its conception for classification and not prediction tasks. Further work would be required to make it available for our solution.

7. Conclusions

In our study, we trained two ensemble neural networks of the same architecture for the eras of ERS-1/2 and Landsat 5 TM, as well as Sentinel 1 and 2, using deep time series of multispectral optical and SAR data types. The supervised training with synthetic but noisy labels has shown good utility, which avoids the need to manually create ground truth labels to train the neural networks. The ensemble of mixing two data types showed synergies but with a tendency towards relying more on optical observations. We also analyzed the sensitivity to changes over time, which indicates that changes closer to the beginning of the observation window show a stronger response. This suggests to sample multiple window predictions preferably over a longer period in order to balance change responses.

Our fully automated and parametric environment allows training and inference with little manual work. In combination with services, such as the Sentinel Hub, the necessary data are directly accessible and can be used for training or inference, omitting further processing. We make available our work on Github (https://github.com/It4innovations/ERCNN-DRS_urban_change_monitoring; accessed on 29 July 2021), including the training environments, trained models and data samples.

Author Contributions: Conceptualization, methodology, software, validation, formal analysis, resources, data curation, visualization, and writing—original draft preparation: G.Z. (min. 90% of

overall work); investigation and writing—review and editing: G.Z., M.P., V.S. and M.L.; supervision, project administration, and funding acquisition: J.M. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: Source code, (trained) models and samples used in this study are available on https://github.com/It4innovations/ERCNN-DRS_urban_change_monitoring (accessed on 29 July 2021).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.
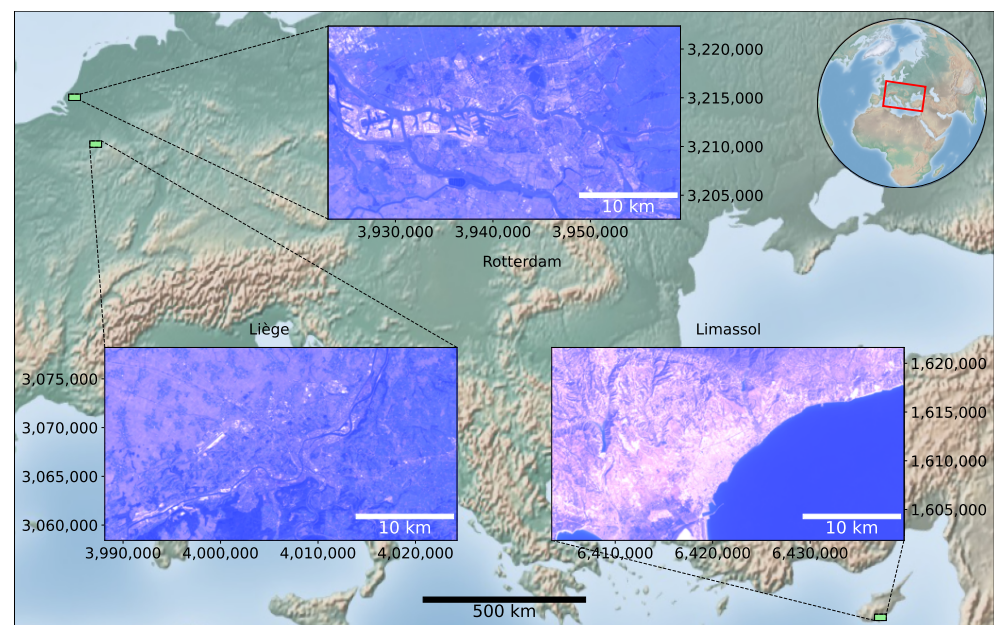
Appendix A. Areas of Interest

Figure A1. Bounding boxes of AoIs used in our study for ERS-1/2 and Landsat 5 TM era. Coordinates are in EPSG:3035. Zoomed in sections show the average of all optical observations in true color. Scales are approximations for comparison only.
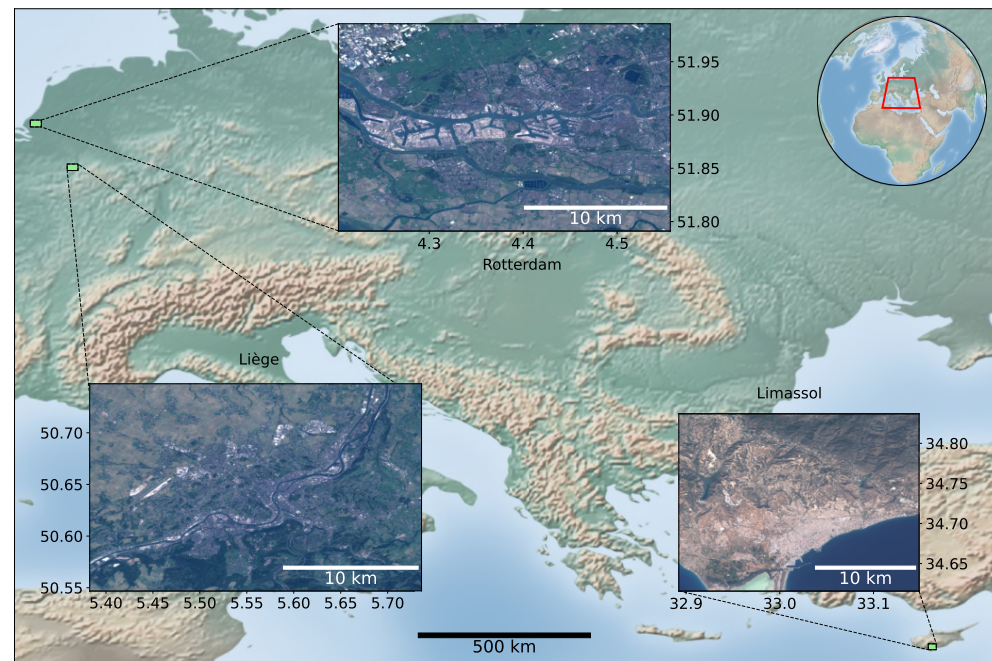
Figure A2. Bounding boxes of AoIs used in our study for Sentinel 1 and 2 era. Coordinates are in EPSG:4326. Zoomed in sections show the average of all optical observations in true color. Scales are approximations for comparison only.

References

1. Singh, A. Review Article Digital change detection techniques using remotely-sensed data. *Int. J. Remote Sens.* **1989**, *10*, 989–1003. [CrossRef]
2. Lehner, A.; Blaschke, T. A Generic Classification Scheme for Urban Structure Types. *Remote Sens.* **2019**, *11*, 173. [CrossRef]
3. You, Y.; Cao, J.; Zhou, W. A Survey of Change Detection Methods Based on Remote Sensing Images for Multi-Source and Multi-Objective Scenarios. *Remote Sens.* **2020**, *12*, 2460. [CrossRef]
4. Patra, S.; Ghosh, S.; Ghosh, A. Semi-supervised Learning with Multilayer Perceptron for Detecting Changes of Remote Sensing Images. In *Pattern Recognition and Machine Intelligence*; Ghosh, A., De, R.K., Pal, S.K., Eds.; Springer: Berlin/Heidelberg, Germany, 2007; pp. 161–168.
5. Roy, M.; Routaray, D.; Ghosh, S.; Ghosh, A. Ensemble of Multilayer Perceptrons for Change Detection in Remotely Sensed Images. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 49–53. [CrossRef]
6. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
7. Deng, J.S.; Wang, K.; Deng, Y.H.; Qi, G.J. PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data. *Int. J. Remote Sens.* **2008**, *29*, 4823–4838. [CrossRef]
8. Qin, Y.; Xiao, X.; Dong, J.; Chen, B.; Liu, F.; Zhang, G.; Zhang, Y.; Wang, J.; Wu, X. Quantifying annual changes in built-up area in complex urban-rural landscapes from analyses of PALSAR and Landsat images. *ISPRS J. Photogramm. Remote Sens.* **2017**, *124*, 89–105. [CrossRef]
9. Canty, M.J.; Nielsen, A.A. Automatic radiometric normalization of multitemporal satellite imagery with the iteratively re-weighted MAD transformation. *Remote Sens. Environ.* **2008**, *112*, 1025–1036. [CrossRef]
10. Xiao, P.; Zhang, X.; Wang, D.; Yuan, M.; Feng, X.; Kelly, M. Change detection of built-up land: A framework of combining pixel-based detection and object-based recognition. *ISPRS J. Photogramm. Remote Sens.* **2016**, *119*, 402–414. [CrossRef]
11. Jafari, M.; Majedi, H.; Monavari, S.M.; Alesheikh, A.A.; Zarkesh, M.K. Dynamic simulation of urban expansion through a CA-Markov model Case study: Hyrcanian region, Gilan, Iran. *Eur. J. Remote Sens.* **2016**, *49*, 513–529. [CrossRef]
12. Conradsen, K.; Nielsen, A.A.; Skriver, H. Determining the Points of Change in Time Series of Polarimetric SAR Data. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 3007–3024. [CrossRef]
13. Zhang, M.; Shi, W. A Feature Difference Convolutional Neural Network-Based Change Detection Method. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7232–7246. [CrossRef]
14. Chen, T.H.K.; Qiu, C.; Schmitt, M.; Zhu, X.X.; Sabel, C.E.; Prishchepov, A.V. Mapping horizontal and vertical urban densification in Denmark with Landsat time-series from 1985 to 2018: A semantic segmentation solution. *Remote Sens. Environ.* **2020**, *251*, 112096. [CrossRef]
15. Yoo, C.; Han, D.; Im, J.; Bechtel, B. Comparison between convolutional neural networks and random forest for local climate zone classification in mega urban areas using Landsat images. *ISPRS J. Photogramm. Remote Sens.* **2019**, *157*, 155–170. [CrossRef]

16. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*; Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2012; Volume 25.

17. Deng, C.; Zhu, Z. Continuous subpixel monitoring of urban impervious surface using Landsat time series. *Remote Sens. Environ.* **2020**, *238*, 110929. [CrossRef]

18. Jing, C.; Zhou, W.; Qian, Y.; Yu, W.; Zheng, Z. A novel approach for quantifying high-frequency urban land cover changes at the block level with scarce clear-sky Landsat observations. *Remote Sens. Environ.* **2021**, *255*, 112293. [CrossRef]

19. Pandey, D.; Tiwari, K. Extraction of urban built-up surfaces and its subclasses using existing built-up indices with separability analysis of spectrally mixed classes in AVIRIS-NG imagery. *Adv. Space Res.* **2020**, *66*, 1829–1845. [CrossRef]

20. Estoque, R.C.; Murayama, Y. Classification and change detection of built-up lands from Landsat-7 ETM+ and Landsat-8 OLI/TIRS imageries: A comparative assessment of various spectral indices. *Ecol. Indic.* **2015**, *56*, 205–217. [CrossRef]

21. Chen, J.; Chen, S.; Yang, C.; He, L.; Hou, M.; Shi, T. A comparative study of impervious surface extraction using Sentinel-2 imagery. *Eur. J. Remote Sens.* **2020**, *53*, 274–292. [CrossRef]

22. Guha, S.; Govil, H.; Gill, N.; Dey, A. A long-term seasonal analysis on the relationship between LST and NDBI using Landsat data. *Quat. Int.* **2021**, *575–576*, 249–258. [CrossRef]

23. Xu, H. Analysis of Impervious Surface and its Impact on Urban Heat Environment using the Normalized Difference Impervious Surface Index (NDISI). *Photogramm. Eng. Remote Sens.* **2010**, *76*, 557–565. [CrossRef]

24. Mitra, D.; Banerji, S. Urbanisation and changing waterscapes: A case study of New Town, Kolkata, West Bengal, India. *Appl. Geogr.* **2018**, *97*, 109–118. [CrossRef]

25. Zhang, S.; Yang, K.; Li, M.; Ma, Y.; Sun, M. Combinational Biophysical Composition Index (CBCI) for Effective Mapping Biophysical Composition in Urban Areas. *IEEE Access* **2018**, *6*, 41224–41237. [CrossRef]

26. Zhong, J.; Li, Z.; Sun, Z.; Tian, Y.; Yang, F. The spatial equilibrium analysis of urban green space and human activity in Chengdu, China. *J. Clean. Prod.* **2020**, *259*, 120754. [CrossRef]

27. Gomez-Chova, L.; Fernández-Prieto, D.; Calpe, J.; Soria, E.; Vila, J.; Camps-Valls, G. Urban monitoring using multi-temporal SAR and multi-spectral data. *Pattern Recognit. Lett.* **2006**, *27*, 234–243. [CrossRef]

28. Sinha, S.; Santra, A.; Mitra, S.S. Semi-automated impervious feature extraction using built-up indices developed from space-borne optical and SAR remotely sensed sensors. *Adv. Space Res.* **2020**, *66*, 1372–1385. [CrossRef]

29. Moreira, A.; Prats-Iraola, P.; Younis, M.; Krieger, G.; Hajnsek, I.; Papathanassiou, K.P. A tutorial on synthetic aperture radar. *IEEE Geosci. Remote Sens. Mag.* **2013**, *1*, 6–43. [CrossRef]

30. Zhang, H.; Lin, H.; Li, Y.; Zhang, Y.; Fang, C. Mapping urban impervious surface with dual-polarimetric SAR data: An improved method. *Landsc. Urban Plan.* **2016**, *151*, 55–63. [CrossRef]

31. Benedetti, A.; Picchiani, M.; Del Frate, F. Sentinel-1 and Sentinel-2 Data Fusion for Urban Change Detection. In Proceedings of the IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 1962–1965. [CrossRef]

32. Muro, J.; Canty, M.; Conradsen, K.; Hüttich, C.; Nielsen, A.A.; Skriver, H.; Remy, F.; Strauch, A.; Thonfeld, F.; Menz, G. Short-Term Change Detection in Wetlands Using Sentinel-1 Time Series. *Remote Sens.* **2016**, *8*, 795. [CrossRef]

33. Ansari, R.A.; Buddhiraju, K.M.; Malhotra, R. Urban change detection analysis utilizing multiresolution texture features from polarimetric SAR images. *Remote Sens. Appl. Soc. Environ.* **2020**, *20*, 100418. [CrossRef]

34. Manzoni, M.; Monti-Guarnieri, A.; Molinari, M.E. Joint exploitation of spaceborne SAR images and GIS techniques for urban coherent change detection. *Remote Sens. Environ.* **2021**, *253*, 112152. [CrossRef]

35. Susaki, J.; Kajimoto, M.; Kishimoto, M. Urban density mapping of global megacities from polarimetric SAR images. *Remote Sens. Environ.* **2014**, *155*, 334–348. [CrossRef]

36. Jiang, M.; Hooper, A.; Tian, X.; Xu, J.; Chen, S.N.; Ma, Z.F.; Cheng, X. Delineation of built-up land change from SAR stack by analysing the coefficient of variation. *ISPRS J. Photogramm. Remote Sens.* **2020**, *169*, 93–108. [CrossRef]

37. Caye Daudt, R.; Le Saux, B.; Boulch, A.; Gousseau, Y. Urban Change Detection for Multispectral Earth Observation Using Convolutional Neural Networks. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Valencia, Spain, 22–27 July 2018.

38. Caye Daudt, R.; Le Saux, B.; Boulch, A. Fully Convolutional Siamese Networks for Change Detection. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 4063–4067. [CrossRef]

39. Benedek, C.; Sziranyi, T. Change Detection in Optical Aerial Images by a Multilayer Conditional Mixed Markov Model. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 3416–3430. [CrossRef]

40. Kundu, K.; Halder, P.; Mandal, J.K. Urban Change Detection Analysis during 1978–2017 at Kolkata, India, using Multi-temporal Satellite Data. *J. Indian Soc. Remote Sens.* **2020**, *48*, 1535–1554. [CrossRef]

41. Lu, D.; Moran, E.; Hetrick, S. Detection of impervious surface change with multitemporal Landsat images in an urban–rural frontier. *ISPRS J. Photogramm. Remote Sens.* **2011**, *66*, 298–306. [CrossRef] [PubMed]

42. Ban, Y.; Yousif, O.A. Multitemporal Spaceborne SAR Data for Urban Change Detection in China. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2012**, *5*, 1087–1094. [CrossRef]

43. Wania, A.; Kemper, T.; Tiede, D.; Zeil, P. Mapping recent built-up area changes in the city of Harare with high resolution satellite imagery. *Appl. Geogr.* **2014**, *46*, 35–44. [CrossRef]

44. Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; Raskar, R. DeepGlobe 2018: A Challenge to Parse the Earth through Satellite Images. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 172–181. [CrossRef]

45. Esch, T.; Asamer, H.; Bachofer, F.; Balhar, J.; Boettcher, M.; Boissier, E.; d' Angelo, P.; Gevaert, C.M.; Hirner, A.; Jupova, K.; et al. Digital world meets urban planet–new prospects for evidence-based urban studies arising from joint exploitation of big earth data, information technology and shared knowledge. *Int. J. Digit. Earth* **2020**, *13*, 136–157. [CrossRef]

46. Marconcini, M.; Metz-Marconcini, A.; Üreyen, S.; Palacios-Lopez, D.; Hanke, W.; Bachofer, F.; Zeidler, J.; Esch, T.; Gorelick, N.; Kakarla, A.; et al. Outlining where humans live, the World Settlement Footprint 2015. *Sci. Data* **2020**, *7*, 242. [CrossRef] [PubMed]

47. Esch, T.; Heldens, W.; Hirner, A.; Keil, M.; Marconcini, M.; Roth, A.; Zeidler, J.; Dech, S.; Strano, E. Breaking new ground in mapping human settlements from space–The Global Urban Footprint. *ISPRS J. Photogramm. Remote Sens.* **2017**, *134*, 30–42. [CrossRef]

48. King, M.D.; Platnick, S.; Menzel, W.P.; Ackerman, S.A.; Hubanks, P.A. Spatial and Temporal Distribution of Clouds Observed by MODIS Onboard the Terra and Aqua Satellites. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 3826–3852. [CrossRef]

49. Nielsen, A.A.; Canty, M.J.; Skriver, H.; Conradsen, K. Change detection in multi-temporal dual polarization Sentinel-1 data. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 3901–3904. [CrossRef]

50. Chen, J.; Yang, K.; Chen, S.; Yang, C.; Zhang, S.; He, L. Enhanced normalized difference index for impervious surface area estimation at the plateau basin scale. *J. Appl. Remote Sens.* **2019**, *13*, 016502. [CrossRef]

51. Xu, H. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *Int. J. Remote Sens.* **2006**, *27*, 3025–3033. [CrossRef]

52. Faridatul, M.I.; Wu, B. Automatic Classification of Major Urban Land Covers Based on Novel Spectral Indices. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 453. [CrossRef]

53. Diakogiannis, F.; Waldner, F.; Caccetta, P.; Wu, C. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 94–114. [CrossRef]

54. Shahroudnejad, A. A Survey on Understanding, Visualizations, and Explanation of Deep Neural Networks. *arXiv* **2021**, arXiv:2102.01792.

55. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput.* **1989**, *1*, 541–551. [CrossRef]

56. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.k.; Woo, W.c. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1, Montreal, QC, Canada, 7–12 December 2015; MIT Press: Cambridge, MA, USA, 2015; pp. 802–810.

57. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]

58. Nguyen, L.D.; Gao, R.; Lin, D.; Lin, Z. Biomedical image classification based on a feature concatenation and ensemble of deep CNNs. *J. Ambient. Intell. Humaniz. Comput.* **2019**, 1–13. [CrossRef]

59. Ju, C.; Bibaut, A.; van der Laan, M. The relative performance of ensemble methods with deep convolutional neural networks for image classification. *J. Appl. Stat.* **2018**, *45*, 2800–2818. [CrossRef]

60. Glorot, X.; Bordes, A.; Bengio, Y. Deep Sparse Rectifier Neural Networks. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 11–13 April 2011; Gordon, G., Dunson, D., Dudík, M., Eds.; PMLR: Cambridge, MA, USA, 2011; Volume 15, pp. 315–323.

61. Gal, Y.; Ghahramani, Z. A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5 December 2016; Curran Associates Inc.: Red Hook, NY, USA, 2016; pp. 1027–1035.

62. Ma, J.; Chen, J.; Ng, M.; Huang, R.; Li, Y.; Li, C.; Yang, X.; Martel, A.L. Loss Odyssey in Medical Image Segmentation. *Med Image Anal.* **2021**, *71*, 102035. [CrossRef] [PubMed]

63. Reina, G.A.; Panchumarthy, R.; Thakur, S.P.; Bastidas, A.; Bakas, S. Systematic Evaluation of Image Tiling Adverse Effects on Deep Learning Semantic Segmentation. *Front. Neurosci.* **2020**, *14*, 65. [CrossRef]

64. Huang, B.; Reichman, D.; Collins, L.M.; Bradbury, K.; Malof, J.M. Dense labeling of large remote sensing imagery with convolutional neural networks: A simple and faster alternative to stitching output label maps. *arXiv* **2018**, arXiv:1805.12219.

65. Huang, B.; Reichman, D.; Collins, L.M.; Bradbury, K.; Malof, J.M. Tiling and Stitching Segmentation Output for Remote Sensing: Basic Challenges and Recommendations. *arXiv* **2019**, arXiv:1805.12219.

66. Isensee, F.; Jaeger, P.F.; Kohl, S.A.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **2021**, *18*, 203–211. [CrossRef]

67. Sergeev, A.; Balso, M.D. Horovod: Fast and easy distributed deep learning in TensorFlow. *arXiv* **2018**, arXiv:1802.05799.

68. Zinkevich, M.; Weimer, M.; Li, L.; Smola, A. Parallelized Stochastic Gradient Descent. In *Advances in Neural Information Processing Systems*; Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., Culotta, A., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2010; Volume 23.

69. Chen, J.; Monga, R.; Bengio, S.; Jozefowicz, R. Revisiting Distributed Synchronous SGD. In Proceedings of the Workshop Track of the International Conference on Learning Representations, ICLR, San Juan, Puerto Rico, 2–4 May 2016.

70. Roth, H.R.; Oda, H.; Zhou, X.; Shimizu, N.; Yang, Y.; Hayashi, Y.; Oda, M.; Fujiwara, M.; Misawa, K.; Mori, K. An application of cascaded 3D fully convolutional networks for medical image segmentation. *Comput. Med Imaging Graph.* **2018**, *66*, 90–99. [CrossRef] [PubMed]

71. Kendall, M.G. A new measure of rank correlation. *Biometrika* **1938**, *30*, 81–93. [CrossRef]

72. Fisher, P. The pixel: A snare and a delusion. *Int. J. Remote Sens.* **1997**, *18*, 679–685. [CrossRef]

73. Song, H.; Kim, M.; Park, D.; Lee, J. Learning from Noisy Labels with Deep Neural Networks: A Survey. *arXiv* **2020**, arXiv:2007.08199.

74. Nguyen, T.; Mummadi, C.K.; Ngo, T.P.N.; Nguyen, T.H.P.; Beggel, L.; Brox, T. SELF: Learning to filter noisy labels with self-ensembling. In Proceedings of the International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 26–30 April 2020.

75. Valentin, B.; Gale, L.; Boulahya, H.; Charalampopoulou, B.; Christos, K.C.; Poursanidis, D.; Chrysoulakis, N.; Svatoň, V.; Zitzlsberger, G.; Podhoranyi, M.; et al. Blended-using blockchain and deep learning for space data processing. In Proceedings of the 2021 Conference on Big Data from Space, Virtual Event, 18–20 May 2021; Soille, P., Loekken, S., Albani, S., Eds.; Publications Office of the European Union: Luxembourg, 2021; pp. 97–100. [CrossRef]