remote sensing

MDPI

Article

Semi-Supervised Urban Change Detection Using Multi-Modal Sentinel-1 SAR and Sentinel-2 MSI Data

Sebastian Hafner, Yifang Ban * and Andrea Nascetti

Division of Geoinformatics, KTH Royal Institute of Technology, Teknikringen 10a, 114 28 Stockholm, Sweden; shafner@kth.se (S.H.); nascetti@kth.se (A.N.)
* Correspondence: yifang@kth.se

Abstract: Urbanization is progressing at an unprecedented rate in many places around the world. The Sentinel-1 synthetic aperture radar (SAR) and Sentinel-2 MultiSpectral Instrument (MSI) missions, combined with deep learning, offer new opportunities to accurately monitor urbanization at a global scale. Although the joint use of SAR and optical data has recently been investigated for urban change detection, existing data fusion methods rely heavily on the availability of sufficient training labels. Meanwhile, change detection methods addressing label scarcity are typically designed for single-sensor optical data. To overcome these limitations, we propose a semi-supervised urban change detection method that exploits unlabeled Sentinel-1 SAR and Sentinel-2 MSI data. Using bitemporal SAR and optical image pairs as inputs, the proposed multi-modal Siamese network predicts urban changes and performs built-up area segmentation for both timestamps. Additionally, we introduce a consistency loss, which penalizes inconsistent built-up area segmentation across sensor modalities on unlabeled data, leading to more robust features. To demonstrate the effectiveness of the proposed method, the SpaceNet 7 dataset, comprising multi-temporal building annotations from rapidly urbanizing areas across the globe, was enriched with Sentinel-1 SAR and Sentinel-2 MSI data. Subsequently, network performance was analyzed under label-scarce conditions by training the network on different fractions of the labeled training set. The proposed method achieved an F1 score of 0.555 when using all available training labels, and produced reasonable change detection results (F1 score of 0.491) even with as little as 10% of the labeled training data. In contrast, multi-modal supervised methods and semi-supervised methods using optical data failed to exceed an F1 score of 0.402 under this condition. Code and data are made publicly available.

Keywords: remote sensing; deep learning; data fusion; consistency regularization; urbanization monitoring

Citation: Hafner, S.; Ban, Y.; Nascetti, A. Semi-Supervised Urban Change Detection Using Multi-Modal Sentinel-1 SAR and Sentinel-2 MSI Data. *Remote Sens.* **2023**, *15*, 5135. https://doi.org/10.3390/rs15215135

Academic Editors: Jon Atli Benediktsson, Yuji Murayama, Zhiyong Lv, Zhou Zhang, Gang Yang, Nicola Falco and Weiwei Sun

Received: 10 August 2023
Revised: 6 October 2023
Accepted: 25 October 2023
Published: 27 October 2023

1. Introduction

While an increasing number of people are moving to cities, uncontrolled urban growth poses pressing threats, such as poverty and environmental degradation. In response to these threats, sustainable urban planning is essential. However, the lack of timely information on the sprawl of settlements is hampering current urban sustainability efforts. Earth observation (EO) is a crucial tool used to map land cover changes associated with urbanization [1]. Change detection is typically conducted by comparing images acquired at different times that cover the same geographical area in three consecutive steps: (1) image preprocessing, (2) derivation of change variables, and (3) classification of change variables. Image preprocessing includes making images acquired at different times radiometrically and spatially comparable. For example, radiometric correction can remove atmospheric effects in optical images [2], and speckle filtering can reduce noise in synthetic aperture radar (SAR) images [3]. For the derivation of change variables from optical images, various arithmetic methods have been developed, including image differencing, image ratioing, image regression, and change vector analysis [4]. A more recent example is a piecewise

distance to measure the change magnitude between bitemporal images [5]. In comparison, change detection in multi-temporal SAR images is commonly conducted using ratio-related operators, such as log-ratio, e.g., [6–9]. It should be noted that some research also focuses on change detection from heterogeneous images acquired by different sensors, e.g., [10]. Finally, the derived change variables are classified into changed/unchanged pixels or objects using either supervised or unsupervised algorithms [7,11–13].

In recent years, deep learning has become the state-of-the-art technology used to process and analyze EO data [14]. As a result, the proportion of deep learning-based change detection methods has significantly increased since 2016 [15]. Another driver of the gain in popularity of deep learning has been the availability of open high-resolution (10–30 m) data, provided by EO programs, such as the European Union's Copernicus program. Specifically, the Sentinel-1 (S1) C-band SAR mission with dual polarization capability and the Sentinel-2 (S2) MultiSpectral Instrument (MSI) mission (13 spectral bands) collect a large volume of EO data at spatial resolutions of 20 m and 10 m, respectively. Moreover, and particularly relevant for change detection, both missions provide frequent revisits of the same geographic area (i.e., sub-weekly).

To date, numerous urban change detection methods have combined deep learning techniques with S1 SAR and/or S2 MSI data. For example, Daudt et al. [16] proposed a Siamese network consisting of two encoders with shared weights to detect changes in urban environments from bitemporal S2 image pairs. Their so-called Siam-diff network was found to be better for change detection, compared to treating image pairs as a single input by concatenating them along the channel axis (i.e., early fusion) [16]. In follow-up work, Daudt et al. [17] incorporated this concept into fully convolutional neural networks (CNNs) using the U-Net architecture as the backbone [18]. Numerous improvements to Siamese networks have since been proposed, e.g., [19–24]. To improve change detection for very-high-resolution (VHR) imagery, some research has focused on incorporating more powerful CNN backbones into Siamese networks. For example, SNUNet employs a nested U-Net to maintain high-resolution fine-grained representations through dense skip connections [20], and HDA-Net employs a high-resolution net in combination with a difference attention module [22]. Other research explored methods to detect the edges of changed areas better. Basavaraju et al. [23], for example, incorporated a new spatial pyramid pooling block into a Siamese network to preserve the shape of change areas, which resulted in better change predictions from bitemporal S2 images. Another improvement to Siamese networks is multi-task learning, where urban change detection and building segmentation are learned simultaneously during training [25,26]. For example, Daudt et al. [25] proposed a dual-task Siamese network that employs an additional decoder for the semantic segmentation of buildings, and Liu et al. [26] demonstrated that the dual-task concept is effective in learning more discriminative features from the input images. Adding a semantic segmentation task to the change detection task was also explored in Papadomanolaki et al. [27] for a fully convolutional long short-term memory (LSTM) network using S2 time series data as input.

In recent years, many urban change detection methods have employed the self-attention mechanism to improve the modeling of long-range dependencies in VHR imagery [28–31]. Both Chen and Shi [28] and Chen et al. [29] extract image features with a CNN and employ self-attention modules to learn more discriminative features. Transformers were also employed in combination with spatial and channel attention modules for feature refinement in Liu et al. [30]. Bandara and Patel [31], on the other hand, proposed a fully transformer-based change detection method. Specifically, ChangeFormer combines two hierarchically structured transformer encoders with shared weights and a multi-layer perception decoder in a Siamese network architecture. While these transformer-based methods are considered state-of-the-art for urban change detection, it should, however, be noted that the effectiveness of these methods has been predominately demonstrated on VHR datasets.

The recent development of deep learning-based methods for the fusion of SAR and optical data, e.g., [32–34], is highly relevant for urban change detection from S1 SAR

and S2 MSI imagery. Importantly, it should be noted that SAR-optical data fusion has already been found useful for urban change detection using traditional machine learning algorithms [35]. Ebel et al. [32] proposed a multi-modal extension of the Siam-diff network by incorporating a separate encoder branch for each sensor modality. The extracted features from the branches are concatenated and forwarded via skip connections to a single decoder. Consequently, the fusion takes place at the different decoder levels. Following a similar concept, the authors in [33] introduced a dual-stream U-Net architecture to fuse SAR and optical data. Specifically, bitemporal image pairs from each sensor are initially concatenated along the channel axis in an early fusion fashion. Subsequently, these image pairs are fed separately to the respective U-Net stream to extract modality-specific change features. Finally, the extracted features are fused at the decision level.

However, a major limitation of supervised deep learning is that models require large amounts of labeled data which are costly and time-consuming to obtain, particularly for change detection tasks. Therefore, several papers investigated unsupervised learning for change detection. For example, Saha et al. [36] proposed a deep change vector analysis to model spatial relationships among neighboring pixels. Deep change vector analysis uses a pre-trained CNN to obtain deep change vectors from multi-temporal images. Since the vast majority of pre-trained networks can only deal with RGB images, generative adversarial networks were leveraged to learn robust feature representations in an unsupervised fashion. This pretraining technique proved to be effective for the detection of changes in bitemporal S2 images using the deep change vector analysis framework [37]. Others developed an unsupervised change detection method by leveraging the high temporal resolution of S1 using an LSTM network [37]. Specifically, change detection was treated as an anomaly detection problem where a shuffled time series was fed to the LSTM which was tasked to rearrange the input in the correct order. While the model can rearrange pixels representing no change in the correct order, the model was expected to fail for change pixels, which enables the unsupervised detection of changes [37]. Recently, Kondmann et al. [38] introduced an unsupervised bitemporal change detection method that first models pixels in an image as linear combinations of their distant neighbors and then uses these models for spatial context-based predictions for the subsequent image. Differences between the actual values and the predictions based on several mutually exclusive neighborhood models are used to derive changes via majority voting.

Although unsupervised change detection models require no labeled samples to learn from, they often fall short of performances achieved by their supervised counterparts. Therefore, it is desirable to investigate semi-supervised learning. The goal of semi-supervised learning is to incorporate unlabeled data—alongside labeled data—into network training to improve performance in supervised networks [39]. Considering the plethora of satellite data acquired by S1 and S2, this idea holds great potential for remote sensing applications. The state-of-the-art for semi-supervised learning can be broadly grouped into two techniques [40]. First, consistency regularization, following the underlying idea that perturbations of a sample should not significantly change the model output [41,42]; and, second, entropy minimization, which encourages more confident predictions on unlabeled data. Several recent papers applied the former technique, consistency regularization, to urban change detection problems using multi-task Siamese networks [43–45]. In particular, Bandara and Patel [43] first used an encoder with shared weights to extract features from unlabeled bitemporal VHR images. Then, consistency was enforced between the change prediction obtained from decoding the subtracted features, i.e., deep feature difference maps, and change predictions obtained from decoding the deep feature maps with small random perturbations using perturbation-specific decoders. Another study proposed a Siamese dual-task network to exploit unlabeled bitemporal Planet image pairs by encouraging consistency between change predictions and changes derived from the semantic segmentation of the images [44]. On the other hand, a more recent work aimed to improve change detection by incorporating additional building labels into network training since building labels are less costly to obtain than change labels. To leverage

additional building labels, Shu et al. [45] proposed a network that encourages consistency between the semantic segmentation of the pre-change image and a building prediction for the pre-change image derived from the change decoder features and the features of the post-change image produced by the semantic decoder. However, despite the fact that these works address the limited availability of labels for urban change detection, up to now, research on semi-supervised change detection has been limited to unimodal EO data from optical sensors.

In this study, we propose a semi-supervised urban change detection method using multi-modal S1 SAR and S2 MSI data. Specifically, a multi-modal Siamese network is modified to perform not only change detection between multi-modal image pairs but also semantic segmentation for both timestamps and sensor modalities. The network is trained in a semi-supervised fashion using consistency regularization to learn more robust features by penalizing inconsistent semantic outputs across sensor modalities. Therefore, we hypothesize that the capability of extracting more robust features for semantic segmentation also improves the change detection ability of the network. The effectiveness of this hypothesis is experimentally tested on the urban change detection problem posed by the SpaceNet 7 dataset [46] using satellite images from the S1 SAR and S2 MSI missions. The testing includes investigating model performance under varying limited labeled conditions, where only a fraction of the training data is used for supervised training and the remaining training data are used for unsupervised training via multi-modal consistency regularization.

2. Methods

This section introduces the methods of this paper in detail. First, a formal description of the problem is presented in Section 2.1. Second, the dataset preparation is described in Section 2.2. Thereafter, the proposed method is described in two parts: (1) the network architecture (Section 2.3.1) and (2) the training process (Section 2.3.2). Finally, a description of the experimental setup is presented in Section 2.4.

2.1. Problem Formulation

We consider a multi-modal image pair with timestamp t that consists of an S1 SAR image and an S2 MSI image referred to as x^t_{S1} and x^t_{S2}, respectively. We denote a multi-modal dataset by $\mathcal{D}$. This dataset consists of multiple bitemporal image pair instances $(x^{t1}_{\mathrm{S1}}, x^{t2}_{\mathrm{S1}}, x^{t1}_{\mathrm{S2}}, x^{t2}_{\mathrm{S2}})$, where $t1$ and $t2$ correspond to the pre-change and post-change timestamp, respectively. A limited fraction of these instances also contains building labels y_s for $t1$ (y^{t1}_s) and $t2$ (y^{t2}_s), as well as change labels y_c derived from the semantic labels using a basic arithmetic operator ($y^{t2}_s - y^{t1}_s$). We partition the labeled fraction of the dataset into training, validation, and test sets, denoted by $\mathcal{D}^{\mathrm{train}}_l$, $\mathcal{D}^{\mathrm{val}}_l$, and $\mathcal{D}^{\mathrm{test}}_l$, respectively. All unlabeled instances, on the other hand, constitute the unlabeled training set $\mathcal{D}^{\mathrm{train}}_u$. The goal is to incorporate unlabeled data into model training to train a model that predicts urban changes with higher accuracy than a model trained exclusively on labeled data. A model's ability to predict urban change is evaluated on the test set.

2.2. Dataset Preparation

The proposed method requires building labels for the pre-change and the post-change image due to the dual-task nature of the underlying network. While popular change detection datasets such as LEVIR-CD [28] and WHU [47] fulfill this requirement, the pre-change images in these datasets were mainly acquired before the launch of the S1 and S2 missions (i.e., prior to 2014), rendering them unusable to test the proposed method. On the other hand, urban change detection datasets containing S2 images such as the Onera Satellite Change Detection (OSCD) dataset [16], provide change labels but lack building labels. Therefore, we chose to use the SpaceNet 7 multi-temporal urban development dataset as it not only provides multi-temporal building labels but also uses satellite imagery from 2017 onward. Specifically, the SpaceNet 7 dataset contains temporal stacks (approximately 24 images) of the VHR (i.e., ∼4 m) monthly planet composites, including corresponding

manually annotated building footprints (vector format) [46]. It should, however, be noted that not all planet composites are cloud-free and, consequently, not all building annotations are complete. The dataset covers 80 unique geographic sites split into 60 training sites and 20 test sites, where building labels are only available for the 60 training sites. We split the labeled SpaceNet 7 sites into training ($n = 30$), validation ($n = 15$), and test sites ($n = 15$), while the unlabeled sites are used for unsupervised learning (Table 1). Figure 1 shows the locations of the study sites, colored according to the set the sites belong to.

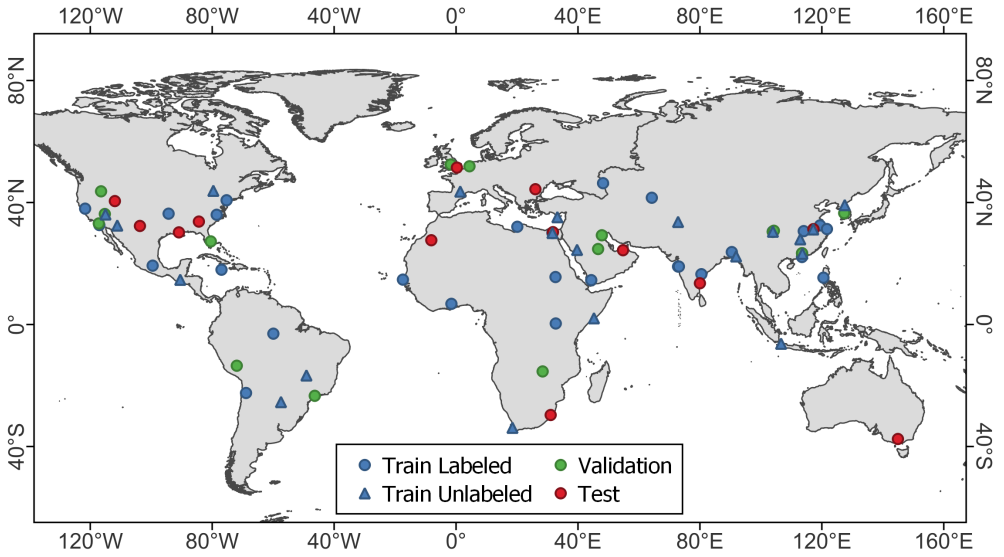

Figure 1. Locations of the study sites. The labeled SpaceNet 7 sites represent our labeled training, validation, and test sites, whereas the unlabeled SpaceNet 7 sites represent our unlabeled training sites.

Table 1. Number of sites per dataset split.

	Train		Validation	Test
	Labeled	**Unlabeled**		
Number of sites	30	20	15	15

In order to replace the monthly planet composites, S1 SAR and S2 MSI images are generated using the cloud-based platform Google Earth Engine (GEE) [48]. An overview of the data preparation workflow is illustrated in Figure 2. One of the advantages of GEE is that S1 SAR data and S2 optical data are directly available as analysis-ready data cubes. Specifically, S1 interferometric wide swath SAR scenes with dual polarization (VV+VH band) are available as ground range detected (GRD) products, processed using the S1 Toolbox. Processing includes thermal noise removal, radiometric calibration, terrain correction, and the conversion of backscatter coefficients (σ) to decibels via log scaling ($10 \log_{10} x$). Furthermore, S1 SAR scenes were resampled to a spatial resolution of 10 m from their native resolution of 20 m. On the other hand, S2 MSI scenes are available in GEE as ortho-corrected top-of-atmosphere reflectance (Level-1A) scenes scaled by a factor of 10,000. Although S2 scenes contain 13 spectral bands with various spatial resolutions, only the bands acquired at a 10 m spatial resolution, i.e., B2 (blue), B3 (green), B4 (red), and B8 (near-infrared), are considered. To produce an S1 and S2 image for a given timestamp of a site, all acquisitions within that month are obtained. For S1, ascending and descending scenes are separated due to the strong influence of the incidence angle on the backscatter coefficients of buildings. Consequently, scenes from the pass with better data availability in terms of absolute image count are selected. After masking backscatter coefficients lower than -25 dB in each scene, the per-pixel temporal mean is computed for both polarization bands to remove speckle noise without reducing the spatial resolution [49]. This workflow is consistent with the one in [50] used to prepare S1 images for the OSCD

dataset. For S2 scenes, on the other hand, temporal aggregation is not applied to preserve the information that was actually measured by S2, as recommended in [51]. Instead, the least cloudy scene among all scenes acquired within a month is selected based on the cloud probabilities layer, retrieved via the Sentinel Hub's cloud detector (https://github.com/sentinel-hub/sentinel2-cloud-detector) and available in GEE as a precomputed dataset. Specifically, the goodness of a scene is defined as the sum of per-pixel cloud probability values. Finally, pixel values are normalized to the range $[0, 1]$ from the range $[-25, 0]$ and $[0, 10,000]$ for S1 and S2, respectively. The EO data are publicly available on Zenodo (https://doi.org/10.5281/zenodo.7794693), including corresponding, rasterized building footprints.

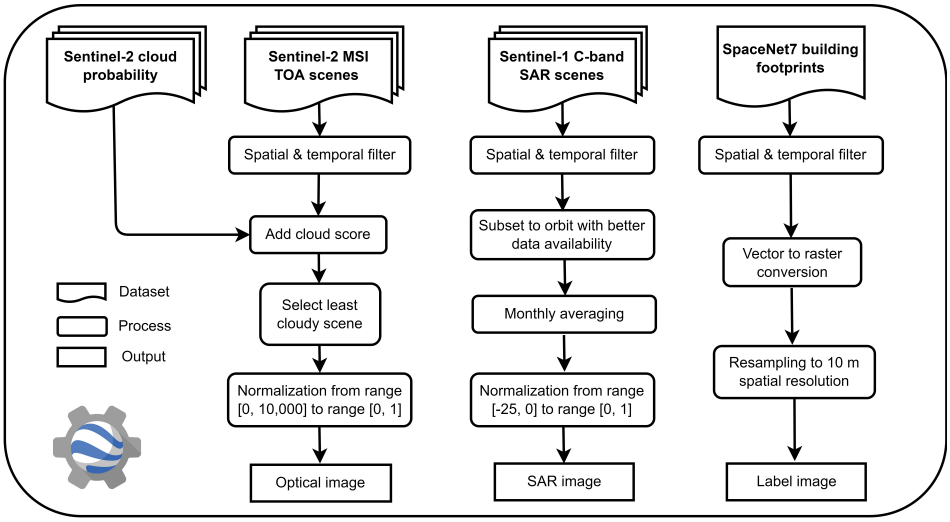

Figure 2. Overview of the data preparation workflow, implemented in GEE [48], to generate S1 SAR and S2 MSI images for all sites of the SpaceNet 7 dataset [46]. Additionally, building labels are derived from the SpaceNet 7 training sites for which manually annotated building footprints are available.

2.3. Proposed Method

We propose a novel semi-supervised change detection method that combines multi-modal data fusion, multi-task learning, and consistency regularization. To that end, we design a multi-modal network architecture that performs two tasks, namely urban change detection and semantic segmentation of buildings. Furthermore, a loss function consisting of a supervised term and an unsupervised term to train the model in a semi-supervised fashion via consistency regularization is introduced. The components of the proposed methods are described in the following two sections.

2.3.1. Network Architecture

The underlying architecture for the proposed method is a Siam-diff architecture extended with the dual-task concept [17,26] (Figure 3). The basic units of the Siam-diff dual-task architecture are encoder and decoder blocks based on the U-Net architecture [18]. Several change detection studies using S1 and/or S2 data have proposed CNN network architectures that employ U-Net-based encoders and decoders as building blocks [17,32,33,44]. The Siam-diff dual-task network processes images separately using two encoders with shared weights (red arrows) to extract corresponding features (f_1–f_5) from images $t1$ and $t2$. The temporal features are then forwarded via skip connections (black arrows) to the respective level of the change decoder, where they are subtracted from one another before being passed through subsequent layers of the network. Finally, a change prediction p_c ($\in [0, 1]$) is obtained from the extracted feature map via a 1×1 convolution operation followed by the sigmoid activation function. In addition, two decoders with shared weights are used to generate building predictions for image $t1$ ($p_s^{t1} \in [0, 1]$) and image $t2$ ($p_s^{t2} \in [0, 1]$), using the features extracted by the respective encoder.

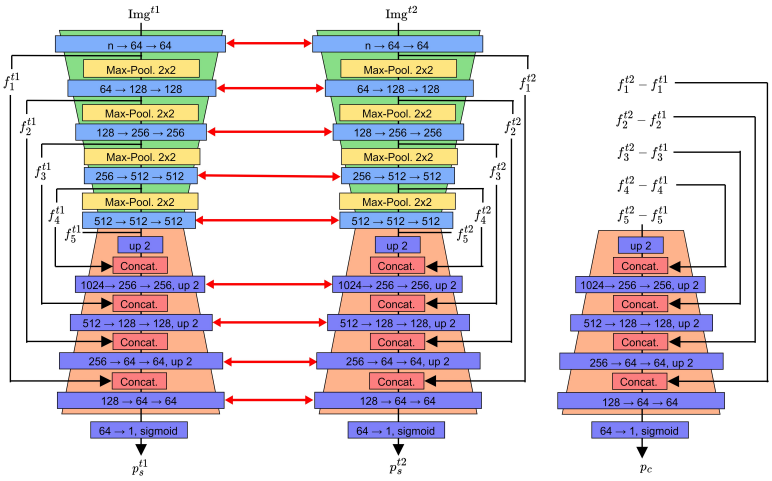

Figure 3. Diagram of the Siam-diff dual-task network for urban change detection. The diagram style was adopted from [17], where blue, yellow, red, and purple blocks denote the operations convolution, max pooling, concatenation, and transpose convolution, respectively. Red arrows illustrate shared weights. The number of input channels is denoted by n.

The proposed network architecture for urban change detection and building segmentation, using multi-modal S1 SAR and S2 MSI data, is visualized in Figure 4. It is a multi-modal version of the Siam-diff dual-task architecture (i.e., multi-modal Siam-diff dual-task network) consisting of two pairs of encoders with shared weights to separately extract feature maps from the S1 and S2 images for $t1$ and $t2$. Two decoders, one for each modality, are converting the subtracted multi-temporal feature maps into a multi-modal feature map, containing the change information extracted from the S1 change decoder and the S2 change decoder. Change predictions are obtained from the multi-modal feature map via a 1×1 convolution followed by the sigmoid activation function. Similar to the Siam-diff dual-task network, building predictions are obtained from the S1 image pair and the S2 image pair, using the respective semantic decoders with shared weights. Two additional building predictions are obtained from the concatenated S1S2 features extracted with the respective semantic decoders. Consequently, the proposed network produces three building predictions (S1, S2, and S1S2) for both timestamps ($t1$ and $t2$), in addition to the change prediction.

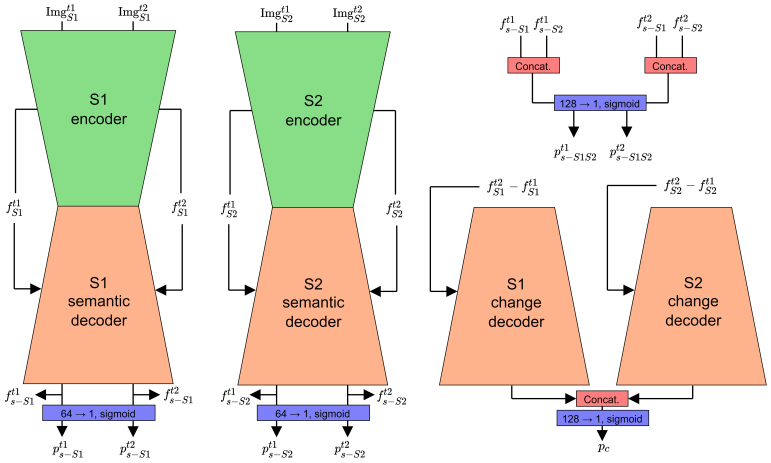

Figure 4. Diagram of the proposed network architecture. Green and orange blocks correspond to the encoder and decoder, respectively. For brevity, the skip connections showing the flow of feature maps from the encoder to the decoder for different network depths are summarized into a single connection.

2.3.2. Training Process

The network is trained in a semi-supervised fashion using a loss function composed of two supervised terms for labeled samples, namely for the urban change detection task ($\mathcal{L}_c$) and the building segmentation task ($\mathcal{L}_s$), and an unsupervised term for unlabeled samples ($\mathcal{L}_{cons}$). For all loss terms, the power Jaccard loss [52], denoted by $J(,)$, is used. The power Jaccard loss is defined as follows:

$$J(p,y) = 1 - \frac{(p \cdot y) + \varepsilon}{(p^2 + y^2 - p \cdot y) + \varepsilon}, \tag{1}$$

where y and p denote label and network prediction, respectively, and ε is a very small number (i.e., $1 \cdot 10^{-6}$) to prevent a division by zero.

In the supervised case, the training objective for multi-modal instances (x^{t1}_{S1}, x^{t2}_{S1}, x^{t1}_{S2}, x^{t2}_{S2}) with labels (y^{t1}_{s}, y^{t2}_{s}) is to minimize the two loss terms, defined as follows:

$$\begin{aligned}\mathcal{L}_c =& J(p_c, y_c)\\ \mathcal{L}_s =& J(p^{t1}_{s-S1}, y^{t1}_{s}) + J(p^{t2}_{s-S1}, y^{t2}_{s}) + J(p^{t1}_{s-S2}, y^{t1}_{s}) + J(p^{t2}_{s-S2}, y^{t2}_{s}) + J(p^{t1}_{s-S1S2}, y^{t1}_{s}) + \\ & J(p^{t2}_{s-S1S2}, y^{t2}_{s}),\end{aligned} \tag{2}$$

where change and semantic variables are sub-scripted with c and s, respectively. The first supervised loss term, $\mathcal{L}_c$, measures the similarity between the urban change label (y_c) and the change prediction (p_c). On the other hand, the second supervised loss term, $\mathcal{L}_s$, measures the similarities between the building labels at $t1$ (y^{t1}_s) and $t2$ (y^{t2}_s) with the corresponding semantic predictions obtained from the S1 SAR inputs (p^{t1}_{s-S1} and p^{t2}_{s-S1}) and the S2 MSI inputs (p^{t1}_{s-S2} and p^{t2}_{s-S2}), as well as the semantic predictions obtained from the multi-modal features (p^{t1}_{s-S1S2} and p^{t2}_{s-S1S2}).

The unsupervised term exploits unlabeled data via consistency regularization [41,42]. Consistency regularization has the goal of learning more robust features by training networks to produce similar outputs for realistic perturbations of the same sample [40]. Since different data modalities can be exploited as natural perturbations [53,54], we apply a consistency loss ($\mathcal{L}_{cons}$) across predictions obtained from different sensor modalities, i.e., multi-modal consistency regularization. Consequently, inconsistencies between the building predictions obtained from the S1 and S2 semantic decoders for $t1$ and $t2$ are penalized during training using the unsupervised loss term below:

$$\mathcal{L}_{cons} = J(p^{t1}_{s-S1}, p^{t1}_{s-S2}) + J(p^{t2}_{s-S1}, p^{t2}_{s-S2}) \tag{3}$$

During training, mini-batch gradient descent is used, where a mini-batch can consist of labeled and unlabeled data. Consequently, the cost for a mini-batch is computed by determining the loss for each sample in the mini-batch separately according to Equation (4), before adding them together. Hyperparameter λ was added as a weight factor to regulate the impact of the consistency term on the final loss.

$$\mathcal{L}_{sample} = \begin{cases} \mathcal{L}_c + \mathcal{L}_s, & \text{if } y \text{ exists} \\ \lambda \cdot \mathcal{L}_{cons}, & \text{otherwise} \end{cases} \tag{4}$$

2.4. Experimental Setup

The following sections describe the experimental setup of this study. The experiments are implemented in Python using Facebook's deep learning framework PyTorch [55], and code is available at https://github.com/SebastianHafner/SemiSupervisedMultiModalCD.git .

2.4.1. Comparison Experiments

The proposed method was compared to several change detection methods. Specifically, for unimodal change detection, the three commonly used supervised methods U-Net

early fusion [17], Siam-diff [17], and Siam-diff dual-task [25,26] were considered, alongside the semi-supervised methods, Siamese SSL [44] and SemiCD [43]. All supervised unimodal methods were separately tested with S1 data and S2 data. On the other hand, the unimodal semi-supervised methods were only tested with S2 data since they employ perturbations that were designed specifically for optical data [43,44]. For multi-modal data, the two supervised methods, dual-stream U-Net [33] and multi-modal Siam-diff [32], were considered. This resulted in a total of ten input data-method combinations (S1 U-Net, S1 Siam-diff, S1 Siam-diff dual-task, S2 U-Net, S2 Siam-diff, S2 Siam-diff dual-task, S2 Siamese SSL, S2 SemiCD, S1S2 dual-stream U-Net, and S1S2 multi-modal Siam-diff) that were considered for the comparison with the proposed method. The benchmark methods are described below:

1. U-Net early fusion [17], a classical U-Net that concatenates bitemporal image pairs along the channel axis, also referred to as early fusion.
2. Siam-diff [17], which uses two U-Net encoders with shared weights to extract features from the images separately. The extracted bitemporal feature pair is subtracted and subsequently fed to a U-Net decoder via skip connections.
3. Siam-diff dual-task [25,26], which adds a second decoder to the Siam-diff network for the semantic segmentation of buildings. The Siam-diff dual-task network (Figure 3) is trained using a supervised loss for change, as well as two supervised losses for the semantics at t1 and t2.
4. Siamese SSL [44], which also uses the Siam-diff dual-task network but an unsupervised loss is employed to enforce consistency between the outputs of the change decoder and change predictions derived from the bitemporal buildings predictions obtained from the semantic decoder.
5. SemiCD [43], which employs an encoder with shared weights to extract features from bitemporal image pairs. Then, consistency is enforced between the change prediction obtained from decoding the subtracted features and a change prediction obtained from adding small perturbations to the subtracted features by using a separate decoder. It should be noted that while the original paper used several different perturbations, we only considered random feature noise since the ablation study in [43] showed that adding additional perturbations had little effect on the performance of the model.
6. Dual0stream U-Net [33], which processes the S1 and S2 image pairs in separate U-Nets using early fusion, before fusing the extracted change features at the decision level.
7. Multi-modal Siam-diff [32], which is a multi-modal version of the Siam-diff network, consisting of two encoders to separately extract features from the the S1 and S2 image pair. A single decoder is used to detect changes by concatenating the multi-modal features.

2.4.2. Training Setup

Training samples from the prepared dataset were generated on the fly by randomly selecting two timestamps from the time series of a site. The building labels for these timestamps, obtained from rasterizing the building footprints (10 m spatial resolution), were used to compute the change label. To account for the fact that the occurrence of change is usually considerably less frequent than no change [56], change areas were oversampled during network training. For a given site, twenty patches of size 128 $\times$ 128 pixels were randomly cropped from the change label, before assigning each patch a probability according to its change pixel percentage, including a base probability for patches with no change pixels. A single patch was chosen based on those probabilities. In order to enhance the training dataset, we applied two common data augmentation operations, namely rotations and flips, which can improve model performance in remote sensing scene classification [57]. During model training, images and labels were randomly rotated by an angle of $k \cdot 90°$, where $k \in \{0, 1, 2, 3\}$ and randomly horizontally or vertically flipped with a probability of 50%. For validation and testing, on the other hand, only the first and the last image of a time series were selected and no data augmentation was applied. For each model, hy-

perparameters were tuned empirically on the validation set using grid search. Specifically, an exhaustive search with three learning rates ($1 \cdot 10^{-5}$, $5 \cdot 10^{-5}$, $1 \cdot 10^{-5}$) and two batch sizes (8, 16) was performed to determine the optimum values of hyperparameters. For the proposed method, two values for hyperparameter λ ($1 \cdot 10^{-2}$, $1 \cdot 10^{-1}$), controlling the impact of the consistency loss term, were added to the grid search. By drawing one hundred samples from each site per epoch, models were trained for 100 epochs on NVIDIA GeForce RTX 3090 graphics cards. Early stopping with patience 10 was added to prevent models from overfitting to the training set. AdamW was used as the optimizer [58].

2.4.3. Accuracy Metrics

Two accuracy metrics were used for the quantitative assessment of predicted changes: F1 score and intersection over union (IoU). The combination of the F1 score and IoU is commonly used for performance assessments in change detection studies, e.g., [45]. Formulas for the metrics are given in Equations 5 and 6, where TP, FP, and FN represent the number of true positive, false positive, and false negative pixels, respectively.

$$\text{F1 score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{5}$$

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{6}$$

3. Results

3.1. Change Detection Results

Table 2 lists the quantitative change detection results obtained on the test set from training the models on limited fractions of the labeled training set, i.e., 40% ($n = 12$), 20 % ($n = 6$), and 10% ($n = 3$). The last column of Table 2 lists the results obtained from training the models on the entire labeled training set (i.e., 100%), even though it is generally assumed in semi-supervised learning that the size of the unlabeled dataset is considerably larger than that of the labeled dataset (e.g., [43,45]). However, this column was added to test whether the proposed method manages to perform on par with the supervised method under no label scarcity. It is apparent that under this condition, all models achieved their best performance in terms of both accuracy metrics. The multi-modal models achieved similar F1 scores (0.554–0.559) and IoU values (0.384–0.388) when having access to all labeled data. In comparison, the accuracy values of the unimodal models trained on S2 MSI data are slightly worse; nevertheless, they all exceed 0.520 (F1 score) and 0.350 (IoU). The lowest values under the 100% labeled condition were obtained by the unimodal models trained on S1 SAR data (F1 scores < 0.420 and IoU values < 0.270). In both unimodal cases, the Siam-diff dual-task network outperformed the Siam-diff and U-Net early fusion networks. However, if supervised models are only given access to a limited amount of labeled data during training, their performance decreases greatly. This is particularly well apparent in Table 2 for the multi-modal networks. For example, the dual-stream U-Net network trained on 10% of the labeled data suffered a performance decrease of 0.250 and 0.205 in terms of the F1 score and IoU, respectively, compared to the 100% case. In contrast, the performance of the proposed semi-supervised change detection method decreased by only 0.064 (F1 score) and 0.059 (IoU). Although the unimodal semi-supervised methods also outperformed all supervised methods (unimodal and multi-modal) under the condition of very limited access to labeled data (i.e., 10% and 20%), the proposed method achieved considerable performance gains across all tested label fraction conditions. Therefore, the proposed method surpassed not only uni and multi-modal supervised learning methods under label-scarce conditions but also semi-supervised learning using optical data.

Table 2. Quantitative change detection results under different label fraction conditions. Values were obtained on the test set. The best and second-best performances are highlighted in red and blue, respectively. Semi-supervised methods are denoted by †.

Input	Network	Fraction of the Labeled Training Set Used							
		10%		20%		40%		100%	
		F1	IoU	F1	IoU	F1	IoU	F1	IoU
S1	U-Net EF	0.291	0.170	0.339	0.204	0.357	0.217	0.363	0.222
	Siam-diff	0.182	0.100	0.368	0.226	0.359	0.219	0.410	0.246
	Siam-diff DT	0.267	0.154	0.341	0.206	0.363	0.222	0.414	0.261
S2	U-Net EF	0.266	0.153	0.429	0.273	0.466	0.303	0.520	0.351
	Siam-diff	0.347	0.210	0.447	0.288	0.522	0.353	0.522	0.353
	Siam-diff DT	0.350	0.212	0.459	0.298	0.484	0.319	0.551	0.380
	Siamese SSL †	0.387	0.240	0.478	0.314	0.496	0.330	0.515	0.347
	SemiCD †	0.402	0.252	0.467	0.305	0.506	0.338	0.513	0.345
S1S2	DS U-Net	0.309	0.183	0.397	0.248	0.522	0.354	0.559	0.388
	MM Siam-diff	0.383	0.237	0.458	0.297	0.500	0.333	0.554	0.383
	Proposed †	0.491	0.325	0.501	0.335	0.537	0.367	0.555	0.384

Qualitative results comparing our change predictions with those obtained from a unimodal semi-supervised method (SemiCD S2) and the supervised multi-modal methods (dual-stream U-Net S1S2 and multi-modal Siam-diff S1S2) are visualized in Figures 5–7 for a selection of sites located in the United States, India, and Australia (in order). Correctly detected changes (TP) and no changes (true negatives) are colored white and black, respectively. On the other hand, incorrectly detected changes (FP) are colored green, and undetected changes (FN) are colored magenta. In addition to the predictions, S2 images in true color (red: B4, green: B3, blue: B2) for $t1$ (pre-change) and $t2$ (post-change) are shown. In general, all multi-modal methods accurately detect urban changes when the entire labeled training set is utilized for supervision (i.e., 100%); consequently, the FP and FN pixel appearances are mainly limited to the borders of urban change areas. However, reducing the amount of labeled data to a fraction of 40% of the training set resulted in an increase in undetected urban changes (FN) for both supervised methods for the United States site (Figure 5), as well as for the India site for the multi-modal Siam-diff network (Figure 6). In contrast, the quality of the results obtained with the proposed method decreased for neither site. Further reducing the fraction of the labeled data used for training to 20% and 10% resulted in even more undetected urban change areas for the supervised methods. Notably, these methods completely failed to detect new built-up areas on the left side of the Australia site (Figure 7) at 20% and 10%, while both methods still achieved good performance in these areas at 40%. In comparison, SemiCD retained more of its change detection accuracy when lowering the fraction of the labeled training set for the sites in the United States and Australia. However, across all sites, the best change detection results under label-scarce conditions were achieved by the proposed method.

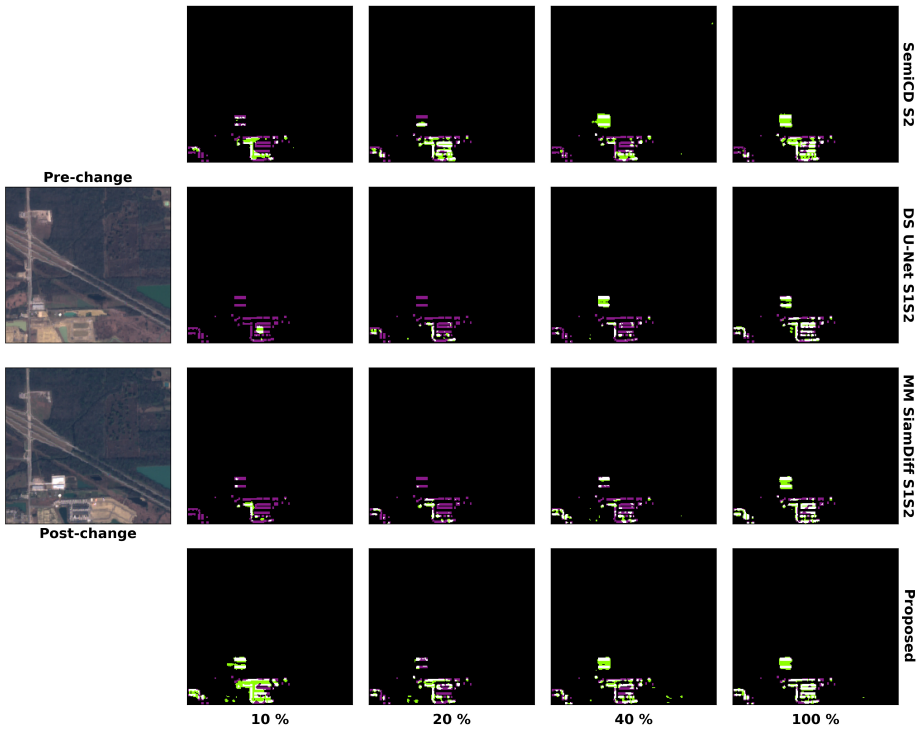

Figure 5. Qualitative change detection results for a test site located in the United States. Pre and post-change Sentinel-2 images visualized in true color (B4, B3, B2) are shown in the outermost left column. The following columns show network predictions under varying label fraction conditions. The colors white, green, magenta, and black represent TP, FP, FN, and true negative pixels, respectively.

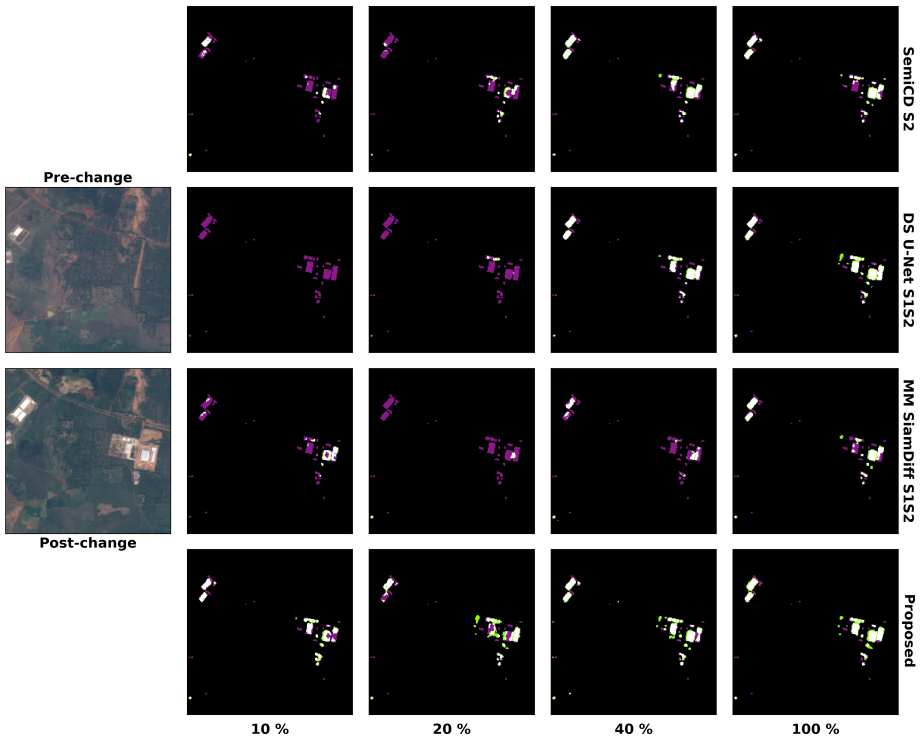

Figure 6. Qualitative change detection results for a test site located in India. Pre and post-change Sentinel-2 images visualized in true color (B4, B3, B2) are shown in the outermost left column. The following columns show network predictions under varying label fraction conditions. The colors white, green, magenta, and black represent TP, FP, FN, and true negative pixels, respectively.

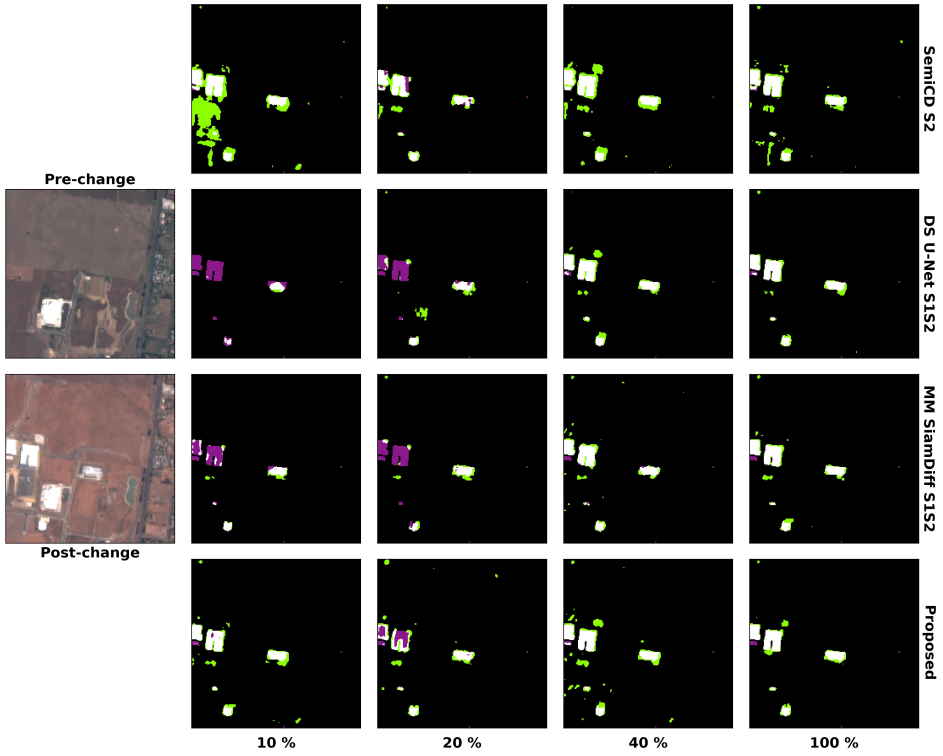

Figure 7. Qualitative change detection results for a test site located in Australia. Pre and post-change Sentinel-2 images visualized in true color (B4, B3, B2) are shown in the outermost left column. The following columns show network predictions under varying label fraction conditions. The colors white, green, magenta, and black represent TP, FP, FN, and true negative pixels, respectively.

3.2. Semantic Segmentation Results

In addition to urban change prediction, the proposed method produces building predictions for $t1$ and $t2$. The quantitative building semantic segmentation results are presented in Table 3. Specifically, Table 3 compares the building predictions obtained from the Siam-diff dual-task networks for both data modalities to those obtained by our network as part of the S1 semantic decoder, the S2 semantic decoder, and the concatenated features extracted by the two semantic decoders. The individual semantic decoder predictions of the proposed method outperform the respective unimodal methods in all limited labeled data scenarios, except for the 10% scenario where the Siamese SSL trained on S2 data achieved marginally better results (+0.002 for F1 score and IoU). However, the fusion prediction of the proposed method showed the best performance in all scenarios, including when all labeled training data were used. Another observation is that S2-based predictions are consistently better than S1-based predictions. Furthermore, F1 scores and IoU values generally decrease when using a lower fraction of the labeled training set. However, accuracy values for the proposed method only dropped considerably when using less than 20% of the labeled training data, while the Siam-diff dual-task network suffered large performance drops below 40%.

3.3. Ablation Study

Since the proposed method combines a new network architecture with a new loss function, we ran an ablation study to investigate the contribution of multi-modal consistency regularization to model performance in terms of change detection, as well as semantic segmentation of buildings. To that end, we trained the multi-modal Siam-diff dual-task network in a fully supervised fashion (i.e., without consistency loss) and compared its performance with the proposed method that trains the same network in a semi-supervised fashion (i.e., with consistency loss). The results of the ablation study are visualized in Figure 8. Adding multi-modal consistency regularization improves performances over

fully supervised training for both tasks and under all limited label conditions. It is also noteworthy that multi-modal consistency regularization is particularly effective when only very few labeled samples are available. On the other hand, when all labeled data are included in model training, semi-supervised learning does not greatly improve change detection performance, and it even has a slightly negative effect on the semantic task. However, the effectiveness of semi-supervised methods is typically only demonstrated under severe label scarcity, since it is assumed that the size of the labeled training set is considerably smaller than that of the unlabeled training set [43,45]. That being noted, it is also possible that enforcing consistency between S1 and S2 data during training can have negative effects on the building segmentation performance of the network due to the difference in spatial resolution between the sensors or the fact that the contextual information in SAR data is lower than in optical data [54].

Table 3. Quantitative test results for semantic segmentation. The best and second-best performances are highlighted in red and blue, respectively. Semi-supervised methods are denoted by †.

Input	Network	Fraction of the Labeled Training Set Used							
		10%		20%		40%		100%	
		F1	IoU	F1	IoU	F1	IoU	F1	IoU
S1	Siam-diff DT	0.302	0.178	0.399	0.249	0.504	0.337	0.480	0.316
	Proposed †	0.361	0.221	0.475	0.312	0.492	0.327	0.486	0.321
S2	Siam-diff DT	0.356	0.216	0.488	0.323	0.524	0.355	0.589	0.417
	Siamese SSL †	0.416	0.263	0.473	0.310	0.515	0.346	0.524	0.355
	Proposed †	0.414	0.261	0.559	0.388	0.578	0.406	0.591	0.420
S1S2	Proposed †	0.526	0.356	0.590	0.418	0.586	0.415	0.612	0.441

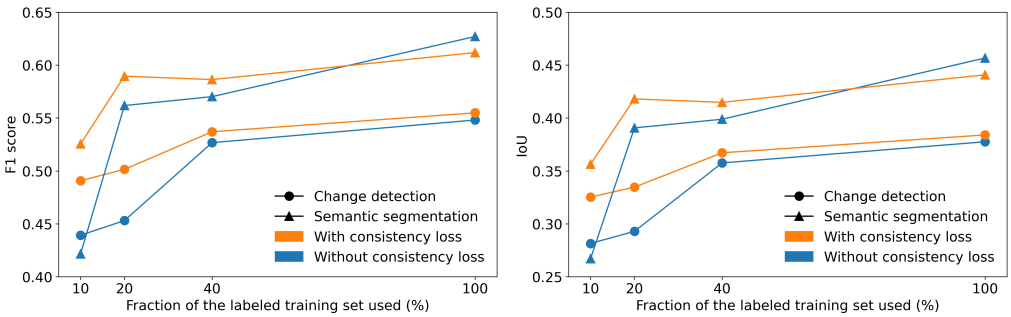

Figure 8. Ablation study showing how adding multi-modal consistency regularization to the training of the multi-modal Siam-diff dual-task network contributes to urban change detection and building segmentation in terms of the F1 score (**left**) and IoU (**right**).

4. Discussion

4.1. Fusion of SAR and Optical Data

The underlying idea of fusing S1 and S2 data is to exploit the complementary information in SAR and optical data to improve urban change detection results upon unimodal methods. In our supervised experiments, SAR-optical data fusion improved network performance compared to unimodal methods in the 100% labeled training data case (Table 2). These findings are in line with other works using supervised deep learning and S1S2 data fusion for urban change detection [32–34]. However, Table 2 also reveals that when labeled training data are limited, supervised networks trained on S2 data perform similarly to supervised networks trained on multi-modal data. In the 20% case, the Siam-diff dual-task network trained on S2 data even achieved the best performance apart from the semi-supervised methods. These findings indicate that deep learning methods using multi-modal data require a substantial amount of labeled training data in order to outperform unimodal methods in a fully supervised setting. On the other hand, S1-based networks

seem to require little training data. We attribute this to the fact that vertical construction generally results in increased backscatter values due to the steady increase of backscatter values with building height [59]. Furthermore, learned representations of built-up areas from SAR data are generally more robust than their optical counterparts, meaning that SAR-based models generalize better across regions; in contrast, models trained on optical models are prone to suffer from distribution shifts due to geographical changes [54]. However, it should be noted that S1-based predictions may lack spatial details due to the 20 m spatial resolution of S1 SAR data, whereas optical predictions are based on input data with a finer, 10 m, spatial resolution (i.e., the blue, green, red, and near-infrared S2 MSI bands).

4.2. Multi-Modal Consistency Regularization

Urban change detection with SAR-optical data fusion is commonly investigated in the context of supervised deep learning [32–34]. Although label scarcity has recently led to the development of deep data fusion frameworks based on either semi-supervised learning [54] or self-supervised learning (i.e., contrastive learning) [53,60], research studies addressing label scarcity in change detection predominantly focus on unimodal data [43–45]. Here, we propose S1S2 data fusion using semi-supervised learning, more specifically multi-modal consistency regularization, to perform not only urban land cover mapping but also urban change detection with limited availability of labeled training data. The presented results demonstrate that consistency regularization performed across data modalities is an effective semi-supervised method to improve building segmentation over supervised methods, especially when labels are scarce (Table 3). Moreover, we empirically prove that the improved building segmentation is linked to the consistency loss imposed on unlabeled data (Figure 8). These results are in line with the findings of Hafner et al. [54] where multi-modal consistency regularization was proposed to overcome domain shifts in urban mapping. However, the main aim of this work is to improve change detection performance by combining data fusion and semi-supervised learning. Our experiments show that the addition of an unsupervised loss not only improved building segmentation but also urban change detection performance (Table 3). Therefore, we consider multi-modal consistency regularization effective for urban change detection.

4.3. Limitations and Perspective

An apparent limitation of the proposed semi-supervised urban change detection method is that it does not outperform multi-modal architectures that are trained using full supervision when labeled training data are not limited (see 100% case in Table 2). However, as previously mentioned, semi-supervised methods generally assume that the unlabeled dataset is considerably larger than the labeled one [43,45]. Consequently, it is unsurprising that all semi-supervised methods fail to achieve performance gains over their supervised counterparts when a large part of the training dataset is labeled. On the other hand, our ablation experiment shows that consistency loss also improves change detection performance in the 100% case, even though the improvement is smaller than when labeled training data are scarce and limited to the change detection task (Figure 8). We infer from these findings that, if trained in a fully supervised manner, the multi-modal Siam-diff dual-task architecture may be slightly less powerful than the multi-modal Siam-diff and dual-stream U-Net networks. However, one should also take into consideration that these networks only perform a single task, i.e., change detection, whereas the proposed network performs built-up area segmentation in addition to change detection.

The F1 scores and IoU values obtained in this study highlight that detecting urban changes from bitemporal S2 MSI images is a challenging task, especially in rapidly urbanizing regions where the SpaceNet 7 sites are located. In particular, the detection of newly constructed built-up areas with small extents may be difficult due to the limited spatial resolution of S1 and S2 imagery (i.e., 20 and 10 m, respectively). Other urban change detection studies using Sentinel imagery confirm the challenging nature of the task. For example, F1 scores achieved by supervised deep learning methods on the OSCD dataset (bitemporal S2 image pairs) typically do not exceed 0.600 [17], even if additional S2 scenes are added [27,61], or data fusion is considered by adding

S1 data [32,33]. Moreover, it should be taken into account that the urban change detection labels in the OSCD dataset were manually annotated based on S2 MSI imagery [16], whereas the urban change detection labels in this study were derived from building footprint annotations based on Planet imagery [46]. The urban change detection task posed by the SpaceNet 7 dataset may, therefore, be more challenging due to the presence of more detailed changes than in the OSCD dataset.

Finally, a limitation of the proposed method is that it was designed for change detection from bitemporal image pairs, while the high temporal frequency of image acquisitions provided by the S1 and S2 missions offer the potential to use dense time series of observations. For example, time series information can help to reduce negative effects due to cloud cover for urban change detection from S2 data [61]. Therefore, future work will explore the integration of S1 and S2 time series into existing urban mapping and change detection methods (i.e., multi-temporal change detection with multi-modal data). However, a particular challenge will be the fact that acquisition times may not correspond between SAR and optical data, and the optical modality may not always be available due to clouds, e.g., [34,62].

5. Conclusions

This research presents a novel semi-supervised urban change detection method that exploits S1 SAR and S2 MSI data via multi-modal consistency regularization. To demonstrate the effectiveness of the proposed method, we enrich the multi-temporal urban mapping dataset SpaceNet 7 with monthly mean S1 SAR images and cloud-free S2 MSI images and train the proposed network on different fractions of the labeled training set. While supervised multi-modal methods, as well as the proposed semi-supervised method, achieved good change detection performance when all labeled data were used for training (F1 scores > 0.550), the supervised methods performed poorly (F1 scores < 0.400) when the labeled data fraction was reduced to 10%. In contrast, the proposed method achieved an F1 score of 0.491. This is also a performance increase of more than 22.1% compared to the best semi-supervised method using optical data. Although performance differences were smaller under less severe label-scarce conditions (i.e., 20 and 40%), with F1 scores of 0.501 and 0.537, the proposed method also achieved performance gains of 4.8 and 2.9% compared to the second-best method for the 20 and 40% cases, respectively. Therefore, our experiments demonstrate that the proposed method improves change detection performance upon several supervised and semi-supervised methods in scenarios where labeled training data are limited. We successfully link these improvements to the consistency loss imposed upon the multi-modal built-up area outputs. Since the proposed method achieves good performance even when labeled training data are scarce, this research has the potential to contribute to the monitoring of urbanization in the Global South where labeled training data are lacking. Finally, since our findings highlight the challenging aspects of urban change detection from S1 SAR and S2 MSI data, our future work will aim to improve urban change detection performance by developing novel deep learning methods that incorporate dense time series of S1 and S2 observations.

Author Contributions: Conceptualization, data curation, methodology, visualization, validation, writing—original draft, writing—review and editing, S.H.; conceptualization, methodology, writing—review and editing, supervision, funding acquisition, resources, Y.B.; conceptualization, methodology, writing—review and editing, A.N. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Swedish National Space Agency (grant dnr 155/15); Digital Futures (under a grant for the EO-AI4GlobalChange project); the ESA-China Dragon 5 program under the EO-AI4urban project; and the EU Horizon 2020 HARMONIA project (agreement no. 101003517).

Data Availability Statement: The SEN12 Multi-Temporal Urban Mapping dataset presented in this paper is publicly accessible on Zenodo: https://doi.org/10.5281/zenodo.7794693.

Acknowledgments: The authors would like to acknowledge the European Space Agency (ESA) for providing the valuable Sentinel-1 SAR and Sentinel-2 MSI data used in this study.

Conflicts of Interest: The authors declare no conflicts 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.

References

1. Ban, Y.; Yousif, O. Change detection techniques: A review. *Multitemporal Remote Sens.* **2016**, 19 –43.
2. Paolini, L.; Grings, F.; Sobrino, J.A.; Jiménez Muñoz, J.C.; Karszenbaum, H. Radiometric correction effects in Landsat multi-date/multi-sensor change detection studies. *Int. J. Remote Sens.* **2006**, *27*, 685–704. [CrossRef]
3. Dekker, R. Speckle filtering in satellite SAR change detection imagery. *Int. J. Remote Sens.* **1998**, *19*, 1133–1146. [CrossRef]
4. Lu, D.; Mausel, P.; Brondizio, E.; Moran, E. Change detection techniques. *Int. J. Remote Sens.* **2004**, *25*, 2365–2401. [CrossRef]
5. Lv, Z.; Zhong, P.; Wang, W.; You, Z.; Shi, C. Novel Piecewise Distance based on Adaptive Region Key-points Extraction for LCCD with VHR Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–9. [CrossRef]
6. 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]
7. Bazi, Y.; Bruzzone, L.; Melgani, F. An unsupervised approach based on the generalized Gaussian model to automatic change detection in multitemporal SAR images. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 874–887. [CrossRef]
8. Bovolo, F.; Marin, C.; Bruzzone, L. A hierarchical approach to change detection in very high resolution SAR images for surveillance applications. *IEEE Trans. Geosci. Remote Sens.* **2012**, *51*, 2042–2054. [CrossRef]
9. Marin, C.; Bovolo, F.; Bruzzone, L. Building change detection in multitemporal very high resolution SAR images. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 2664–2682. [CrossRef]
10. Sun, Y.; Lei, L.; Guan, D.; Wu, J.; Kuang, G.; Liu, L. Image regression with structure cycle consistency for heterogeneous change detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**. [CrossRef]
11. Bruzzone, L.; Prieto, D.F. Automatic analysis of the difference image for unsupervised change detection. *IEEE Trans. Geosci. Remote Sens.* **2000**, *38*, 1171–1182. [CrossRef]
12. Hu, H.; Ban, Y. Unsupervised change detection in multitemporal SAR images over large urban areas. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 3248–3261. [CrossRef]
13. Cao, G.; Li, Y.; Liu, Y.; Shang, Y. Automatic change detection in high-resolution remote-sensing images by means of level set evolution and support vector machine classification. *Int. J. Remote Sens.* **2014**, *35*, 6255–6270. [CrossRef]
14. 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]
15. Jiang, H.; Peng, M.; Zhong, Y.; Xie, H.; Hao, Z.; Lin, J.; Ma, X.; Hu, X. A Survey on Deep Learning-Based Change Detection from High-Resolution Remote Sensing Images. *Remote Sens.* **2022**, *14*, 1552. [CrossRef]
16. Daudt, R.C.; Le Saux, B.; Boulch, A.; Gousseau, Y. Urban change detection for multispectral earth observation using convolutional neural networks. In Proceedings of the IGARSS 2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 2115–2118.
17. Daudt, R.C.; 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.
18. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
19. Jiang, H.; Hu, X.; Li, K.; Zhang, J.; Gong, J.; Zhang, M. Pga-siamnet: Pyramid feature-based attention-guided siamese network for remote sensing orthoimagery building change detection. *Remote Sens.* **2020**, *12*, 484. [CrossRef]
20. Fang, S.; Li, K.; Shao, J.; Li, Z. SNUNet-CD: A densely connected Siamese network for change detection of VHR images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
21. Zhou, H.; Zhang, M.; Hu, X.; Li, K.; Sun, J. A Siamese convolutional neural network with high–low level feature fusion for change detection in remotely sensed images. *Remote Sens. Lett.* **2021**, *12*, 387–396. [CrossRef]
22. Wang, X.; Du, J.; Tan, K.; Ding, J.; Liu, Z.; Pan, C.; Han, B. A high-resolution feature difference attention network for the application of building change detection. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102950. [CrossRef]
23. Basavaraju, K.; Sravya, N.; Lal, S.; Nalini, J.; Reddy, C.S.; Dell'Acqua, F. UCDNet: A Deep Learning Model for Urban Change Detection from Bi-temporal Multispectral Sentinel-2 Satellite Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–10. [CrossRef]
24. Lv, Z.; Zhong, P.; Wang, W.; You, Z.; Falco, N. Multi-scale Attention Network Guided with Change Gradient Image for Land Cover Change Detection Using Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5.
25. Daudt, R.C.; Le Saux, B.; Boulch, A.; Gousseau, Y. Multitask learning for large-scale semantic change detection. *Comput. Vis. Image Underst.* **2019**, *187*, 102783. [CrossRef]
26. Liu, Y.; Pang, C.; Zhan, Z.; Zhang, X.; Yang, X. Building Change Detection for Remote Sensing Images Using a Dual-Task Constrained Deep Siamese Convolutional Network Model. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 811–815. [CrossRef]

27. Papadomanolaki, M.; Vakalopoulou, M.; Karantzalos, K. A Deep Multitask Learning Framework Coupling Semantic Segmentation and Fully Convolutional LSTM Networks for Urban Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7651–7668. [CrossRef]
28. Chen, H.; Shi, Z. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sens.* **2020**, *12*, 1662. [CrossRef]
29. Chen, H.; Qi, Z.; Shi, Z. Remote sensing image change detection with transformers. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
30. Liu, W.; Lin, Y.; Liu, W.; Yu, Y.; Li, J. An attention-based multiscale transformer network for remote sensing image change detection. *Isprs J. Photogramm. Remote Sens.* **2023**, *202*, 599–609. [CrossRef]
31. Bandara, W.G.C.; Patel, V.M. A transformer-based siamese network for change detection. In Proceedings of the IGARSS 2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 207–210.
32. Ebel, P.; Saha, S.; Zhu, X.X. Fusing multi-modal data for supervised change detection. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2021**, *43*, 243–249. [CrossRef]
33. Hafner, S.; Nascetti, A.; Azizpour, H.; Ban, Y. Sentinel-1 and Sentinel-2 Data Fusion for Urban Change Detection Using a Dual Stream U-Net. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
34. Saha, S.; Shahzad, M.; Ebel, P.; Zhu, X.X. Supervised Change Detection Using Prechange Optical-SAR and Postchange SAR Data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 8170–8178. [CrossRef]
35. Yousif, O.; Ban, Y. Fusion of SAR and optical data for unsupervised change detection: A case study in Beijing. In Proceedings of the 2017 Joint Urban Remote Sensing Event (JURSE). Dubai, United Arab Emirates, 6–8 March 2017; pp. 1–4.
36. Saha, S.; Bovolo, F.; Bruzzone, L. Unsupervised deep change vector analysis for multiple-change detection in VHR images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 3677–3693. [CrossRef]
37. Saha, S.; Solano-Correa, Y.T.; Bovolo, F.; Bruzzone, L. Unsupervised deep transfer learning-based change detection for HR multispectral images. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 856–860. [CrossRef]
38. Kondmann, L.; Toker, A.; Saha, S.; Schölkopf, B.; Leal-Taixé, L.; Zhu, X.X. Spatial Context Awareness for Unsupervised Change Detection in Optical Satellite Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
39. Chapelle, O.; Scholkopf, B.; Zien, A. Semi-Supervised Learning (Chapelle, o. et al., Eds.; 2006) [Book reviews]. *IEEE Trans. Neural Netw.* **2009**, *20*, 542. [CrossRef]
40. Oliver, A.; Odena, A.; Raffel, C.; Cubuk, E.D.; Goodfellow, I.J. Realistic evaluation of deep semi-supervised learning algorithms. *arXiv* **2018**, arXiv:1804.09170.
41. Laine, S.; Aila, T. Temporal ensembling for semi-supervised learning. *arXiv* **2016**, arXiv:1610.02242.
42. Sajjadi, M.; Javanmardi, M.; Tasdizen, T. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *arXiv* **2016**, arXiv:1606.04586.
43. Bandara, W.G.C.; Patel, V.M. Revisiting consistency regularization for semi-supervised change detection in remote sensing images. *arXiv* **2022**, arXiv:2204.08454.
44. Hafner, S.; Ban, Y.; Nascetti, A. Urban change detection using a dual-task Siamese network and semi-supervised learning. In Proceedings of the IGARSS 2022–2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 1071–1074.
45. Shu, Q.; Pan, J.; Zhang, Z.; Wang, M. MTCNet: Multitask consistency network with single temporal supervision for semi-supervised building change detection. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *115*, 103110. [CrossRef]
46. Van Etten, A.; Hogan, D.; Martinez-Manso, J.; Shermeyer, J.; Weir, N.; Lewis, R. The Multi-Temporal Urban Development SpaceNet Dataset. *arXiv* **2021**, arXiv:2102.04420.
47. Ji, S.; Wei, S.; Lu, M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 574–586. [CrossRef]
48. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **2017**, *202*, 18–27. [CrossRef]
49. Chini, M.; Pelich, R.; Hostache, R.; Matgen, P.; Lopez-Martinez, C. Towards a 20 m global building map from Sentinel-1 SAR data. *Remote Sens.* **2018**, *10*, 1833. [CrossRef]
50. Hafner, S.; Ban, Y.; Nascetti, A. Exploring the Fusion of Sentinel-1 SAR and Sentinel-2 MSI Data for Built-Up Area Mapping Using Deep Learning. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; pp. 4720–4723.
51. Schmitt, M.; Hughes, L.H.; Qiu, C.; Zhu, X.X. Aggregating cloud-free Sentinel-2 images with Google earth engine. *Isprs Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2019**, *4*, 145–152. [CrossRef]
52. Duque-Arias, D.; Velasco-Forero, S.; Deschaud, J.E.; Goulette, F.; Serna, A.; Decencière, E.; Marcotegui, B. On power Jaccard losses for semantic segmentation. In Proceedings of the VISAPP 2021: 16th International Conference on Computer Vision Theory and Applications, Online, 8–10 February 2021.
53. Scheibenreif, L.; Hanna, J.; Mommert, M.; Borth, D. Self-supervised vision transformers for land-cover segmentation and classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1422–1431.

54. Hafner, S.; Ban, Y.; Nascetti, A. Unsupervised domain adaptation for global urban extraction using Sentinel-1 SAR and Sentinel-2 MSI data. *Remote Sens. Environ.* **2022**, *280*, 113192. [CrossRef]
55. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv* **2019**, arXiv:1912.01703.
56. Bovolo, F.; Bruzzone, L. The time variable in data fusion: A change detection perspective. *IEEE Geosci. Remote Sens. Mag.* **2015**, *3*, 8–26. [CrossRef]
57. Yu, X.; Wu, X.; Luo, C.; Ren, P. Deep learning in remote sensing scene classification: A data augmentation enhanced convolutional neural network framework. *Giscience Remote Sens.* **2017**, *54*, 741–758. [CrossRef]
58. Loshchilov, I.; Hutter, F. Fixing Weight Decay Regularization in Adam. *arXiv* **2018**, arXiv:1711.05101.
59. Koppel, K.; Zalite, K.; Voormansik, K.; Jagdhuber, T. Sensitivity of Sentinel-1 backscatter to characteristics of buildings. *Int. J. Remote Sens.* **2017**, *38*, 6298–6318. [CrossRef]
60. Chen, Y.; Bruzzone, L. Self-Supervised Change Detection by Fusing SAR and Optical Multi-Temporal Images. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; pp. 3101–3104.
61. Papadomanolaki, M.; Verma, S.; Vakalopoulou, M.; Gupta, S.; Karantzalos, K. Detecting urban changes with recurrent neural networks from multitemporal Sentinel-2 data. In Proceedings of the IGARSS 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 214–217.
62. Hafner, S.; Ban, Y. Multi-Modal Deep Learning for Multi-Temporal Urban Mapping With a Partly Missing Optical Modality. In Proceedings of the 2023 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Pasadena, CA, USA, 16–21 July 2023; pp. 6843–6846.