Computer Vision and Image Understanding
journal homepage: www.elsevier.com

High resolution semantic change detection

Rodrigo Caye Daudt[a,b,**], Bertrand Le Saux[a], Alexandre Boulch[a], Yann Gousseau[b]

[a] *DTIS, ONERA, Université Paris-Saclay, FR-91123 Palaiseau, France*
[b] *LTCI, Télécom ParisTech, FR-75013 Paris, France*

ABSTRACT

Change detection is one of the main problems in remote sensing, and is essential to the accurate processing and understanding of the large scale Earth observation data available through programs such as Sentinel and Landsat. Most of the recently proposed change detection methods bring deep learning to this context, but openly available change detection datasets are still very scarce, which limits the methods that can be proposed and tested. In this paper we present the first large scale high resolution semantic change detection (HRSCD) dataset, which enables the usage of deep learning methods for semantic change detection. The dataset contains coregistered RGB image pairs, pixel-wise change information and land cover information. We then propose several methods using fully convolutional neural networks to perform semantic change detection. Most notably, we present a network architecture that performs change detection and land cover mapping simultaneously, while using the predicted land cover information to help to predict changes. We also describe a sequential training scheme that allows this network to be trained without setting a hyperparameter that balances different loss functions and achieves the best overall results.

1. Introduction

One of the main purposes of remote sensing is the observation of the evolution of the land. Satellite and aerial imaging enables us to keep track of the changes that occur around the globe, both in densely populated areas as well as in remote areas that are hard to reach. That is why change detection is a problem so closely studied in the context of remote sensing. Change detection is the name given to the task of identifying areas that have experienced changes between acquisitions. Change detection can be done using pairs or sequences of coregistered images of a given region at two different moments. Changes can be of several different types depending on the desired application. Examples of changes that are of interest in remote sensing images are natural disasters (e.g. fires, floods), urban expansion, and deforestation. In this paper we treat change detection as a dense classification problem, aiming to predict a label for each pixel in an input image pair, i.e. semantic segmentation.

The search for ever more accurate change detection comes from the value of surveying large amounts of land and analysing its evolution over a period of time. Detecting changes manually is a slow and laborious process. This is why the problem of automatic change detection using image pairs or sequences is a problem that has been studied for many decades. The history of change detection algorithms and overviews of the most important methods are described in the reviews written by Singh (1989) and Hussain et al. (2013). Throughout the years, computer vision and image processing techniques were applied to solve the problem of change detection. With the advances of these areas, so did the change detection techniques evolve.

In the recent years, machine learning techniques have been proved to be excellent at solving a wide range of problems related to image understanding. The rise of these techniques is explained by three main factors. First, the hardware required for doing the large amounts of calculations that are often required for machine learning techniques is becoming cheaper and more powerful. Second, new methods are being proposed to exploit the data in innovative ways. Finally, the amount of available data is increasing, which is essential for many machine learning techniques. In the context of remote sensing, examples of such large amounts of data are ESA's Copernicus and USGS's Landsat satellites which are able to image most of Earth's landmasses once every few days.

**Corresponding author: Tel.: +33-1-80-38-65-55;
e-mail: rodrigo.daudt@onera.fr (Rodrigo Caye Daudt)

In this paper we propose a versatile method to perform change detection from image pairs based on state-of-the-art computer vision ideas. The ideas presented here are extensions of the ones first published in Daudt et al. (2018a), where fully convolutional networks were used in the context of change detection for the first time. The proposed method is able to perform both binary and semantic change detection. Binary change detection simply attempts to identify which pixels correspond to areas where changes have occurred, whereas semantic change detection attempts to further identify the type of change that has occurred at each location. The proposed method is also able to perform change detection using both low and high resolution images. As was described by Hussain et al. (2013), high resolution change detection involves several extra challenges, but it is a more desirable outcome.

A new high resolution semantic change detection dataset of unprecedented size is also presented in this paper. This dataset will be released publicly to serve as a benchmark and as a research tool for researchers working on change detection. The methods used to create this dataset, as well as the limitations of the available data, will be described later on. Until now, the most advanced ideas from computer vision could not be brought to change detection due to the lack of large annotated datasets. This dataset will enable the application of more sophisticated machine learning techniques that were heretofore too complex for the amount of change detection data available.

2. Related work

Hussain et al. (2013) splits change detection algorithms into two main groups: pixel based and object based change detection. The former are attempts to identify whether or not a change has occurred at each pixel in the image pair, while the latter methods attempt to first group pixels that belong to the same object and use information such as the object's colour, shape and neighbourhood to help determine if that object has been changed between the acquisitions. In this work we focus on the first group of algorithms.

Most traditional and many modern change detection algorithms comprise two main steps. First, a difference metric is proposed so that a quantitative measurement of the difference between corresponding pixels can be calculated. The image generated from this step is usually called a difference image, although it may not represent the literal mathematical difference between the two pixel values. Second, a thresholding method or decision function is proposed to separate the pixels into "change" and "no change" based on the difference image. These two steps are usually independent. It is also commonplace to apply pre-processing and post-processing to the images, such as radiometric and atmospheric correction of the input images or filtering of the output classification result to avoid salt and pepper noise. Co-registration of the images is usually considered a separate problem. Most papers on change detection propose either a novel image differencing method (Bovolo and Bruzzone, 2005; El Amin et al., 2016, 2017; Zhan et al., 2017) or a novel decision function (Bruzzone and Prieto, 2000; Celik, 2009; Le Saux and Randrianarivo, 2013).

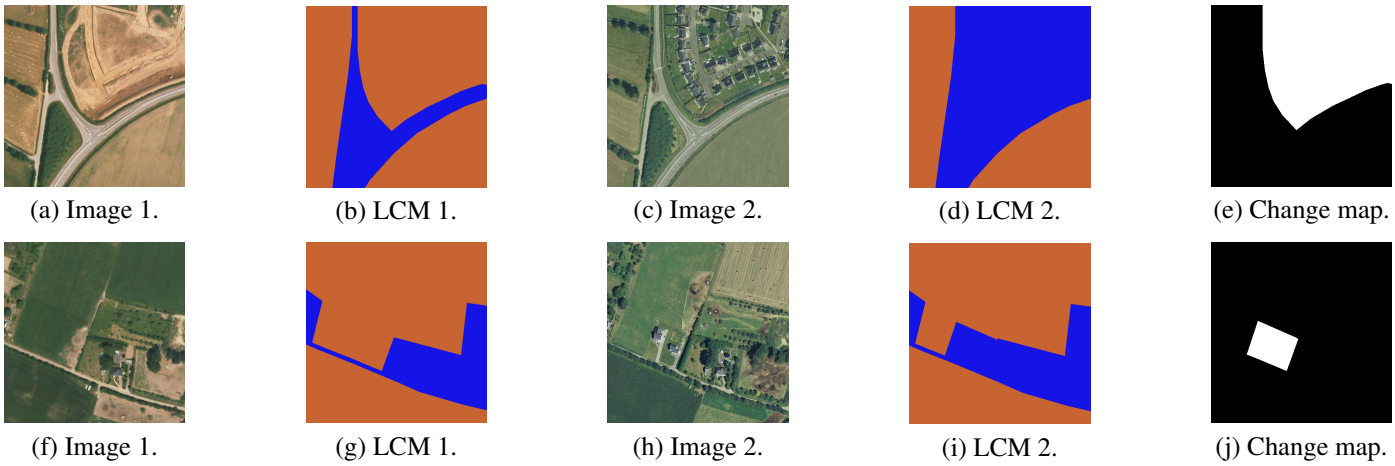
A well established family of change detection methods is change vector analysis (Lambin and Strahlers, 1994). This method creates a vector that describes the spectral-temporal profile of an area and compares it to a vector created from equal acquisition one or more years later. This method may help characterise the type of change that has taken place at a given pixel, but it has tight constraints regarding the input images. Several acquisitions are needed on both considered years, and the acquisitions must be of near-anniversary dates.

As noted by Hussain et al. (2013), change detection on low resolution images and on high resolution images face different challenges. In low resolution images, pixels frequently contain information about several objects contained within its area. In such cases, changes in one surface in a pixel may be masked by the other unchanging surfaces. High resolution images are more susceptible to problems such as parallax, high reflectance variability for objects of the same class, and co-registration problems. It follows that algorithms that perform change detection on high resolution images must be aware of not only a given pixel's values, but also of information about its neighbourhood.

Machine learning techniques have been widely used for image analysis for many years. Most notably, convolutional neural networks (CNNs) are a family of algorithms that are especially suited for working with images. The usage of CNNs for comparing image pairs has already been studied before (Chopra et al., 2005; Zagoruyko and Komodakis, 2015). The usage of such techniques usually requires large amounts of training data. Larger amounts of training data allow more elaborate models to be trained, which often leads to better results. In the context of change detection, the amounts of data available for training and testing are extremely scarce and often kept private, which discourages fair comparisons between different methods proposed by different authors.

Fully convolutional neural networks (FCNNs) are subgroup of CNNs that are especially suited for dense prediction of labels (Long et al., 2015). Unlike traditional CNNs, which output a single prediction for each input image, FCNNs are able to predict labels for each pixel independently and efficiently. Ronneberger et al. (2015) proposed a simple and elegant addition to FCNNs that aims to improve the accuracy of the final prediction results. The proposed idea is to connect directly layers in earlier stages of the network to layers at later stages to recover accurate spatial information of region boundaries. FCNNs currently achieve state-of-the-art results in semantic segmentation problems.

Various ways to use CNNs to perform change detection have been proposed. The vast majority of these methods avoid the problem of the lack of data by using transfer learning techniques, i.e. using networks that have been pre-trained for a different purpose on a large dataset. While transfer learning is a valid solution, it is also limiting. Firstly, end-to-end training tends to achieve the best results for a given problem when possible. Transfer learning also assumes all images have the same nature. As most large scale datasets contain RGB images, this means that extra bands contained in multispectral images must be ignored. The value of using all available multispectral bands for change detection has already been shown (Daudt



(a) Image 1. (b) LCM 1. (c) Image 2. (d) LCM 2. (e) Change map.

(f) Image 1. (g) LCM 1. (h) Image 2. (i) LCM 2. (j) Change map.

Fig. 1. Examples of image pairs, land cover mapping (LCM) and associated pixel-wise change maps from the HRSCD dataset.

et al., 2018b).

Daudt et al. (2018b) was the first work that trained end-to-end CNNs to perform change detection. Previously, several works have used CNNs to generate the difference image that was described earlier, followed by traditional thresholding methods on those images. El Amin et al. (2016, 2017) proposed using the activation of pre-trained CNNs to generate descriptors for each pixel, and using the Euclidean distance between these descriptors to build the difference image. Zhan et al. (2017) trained a network to produce a 16-dimensional descriptor for each pixel which were similar for pixels with no change and dissimilar for pixels that experienced change. Liu et al. (2016) used deep belief networks to generate pixel descriptors from heterogeneous image pairs, then the Euclidean distance is used to build a difference image. Gong et al. (2016) proposes a deep belief network that takes into account the context of a pixel to build its descriptor. Mou et al. (2018) proposes using patch based recurrent CNNs to detect changes in image pairs. CNNs for change detection has also been studied outside the context of remote sensing, such as surface inspection (Stent et al., 2015).

Most methods that propose image differencing techniques followed by thresholding assume that a threshold is chosen for each image, i.e. the image goes trough an adaptive thresholding method. This means there is an assumption that in every image pair given as input, there is a fraction of the pixels that have changed. This assumption does not scale to large datasets where images may contain no change at all, or contain any number of changed pixels. Some methods avoid this problem by setting a fixed threshold during training and maintaining that threshold value regardless of the properties of the test images. Hussain et al. (2013) and Rosin and Ioannidis (2003) noted that the performance of such algorithms is scene dependent.

Fully convolutional networks trained from scratch to perform change detection were proposed for the first time in Daudt et al. (2018a). Both Siamese and early fusion architectures were compared, expanding on the ideas proposed earlier by Chopra et al. (2005) and Zagoruyko and Komodakis (2015). To our knowledge, the only other time a fully convolutional Siamese network has been proposed was by Bertinetto et al. (2016) with the purpose of tracking objects in image sequences. Nevertheless, this work shares little with the presented method in terms of purpose, architecture and conception.

He et al. (2016) proposed using residual modules in CNNs, where convolutional layers attempt to learn residual functions instead of direct transformation. The motivation behind this idea is that when training deep neural networks it is easier to learn residual identity transformations rather than direct identity transformations. They have showed that deep neural networks may sometimes perform worse than their shallower counterparts due to the difficulty of training these deep networks. These results are not due to overfitting, as the same tendency could be observed on the performance on the training sets. Residual blocks have been shown to lead to performance improvements when training such deep networks, and have allowed networks of over 1000 convolutional layers to be trained.

3. Dataset

Research on the problem of change detection is hindered by a lack of open datasets. Such datasets are essential for a methodical evaluation of different algorithms. Benedek and Szirányi (2009) created a binary change dataset with 13 aerial image pairs split into three regions called the Air Change dataset. A dataset of 24 multispectral satellite image pairs is presented in Daudt et al. (2018b), called OSCD dataset. Both of these datasets allow for simple machine learning techniques to be applied to the problem of change detection, but with these small amounts of images overfitting becomes one of the main concerns even with relatively simple models. The Aerial Imagery Change Detection (AICD) dataset contains synthetic aerial images with artificial changes generated with a rendering engine(Bourdis et al., 2011).

For this reason, we have created the first large scale dataset for semantic change detection, which we present in this section. The High Resolution Semantic Change Detection (HRSCD) dataset will be released to the scientific community to be used as a benchmark for semantic change detection algorithms and to open the doors to the usage of state-of-the-art deep learn-

ing algorithms in this context. The dataset contains not only information about where changes have taken place, but also semantic information about the imaged terrain in all images of the dataset. Examples of image pairs, land cover maps (LCM) and change maps taken from the dataset are depicted in Fig. 1.

3.1. Images

The dataset contain a total of 291 RGB image pairs of 10000x10000 pixels. These are mosaics of aerial images taken by the French National Institute of Geographical and Forest Information (IGN)[1]. The image pairs contain an earlier image acquired on 2005 or 2006, and a second image acquired on 2012. They come from a database named *BD ORTHO* which contains orthorectified aerial images of several regions of France from different years at a resolution of 50 cm per pixel. The 291 selected images are all the images in this database that satisfy the conditions for the labels, which will be described below. The images cover a range of urban and countryside areas around the French cities of Rennes and Caen.

The dataset contains more than 3000 times more annotated pixel pairs than either OSCD or Air Change datasets. Also, unlike these datasets, the labels contain information about the types of change that have occurred. Finally, labels about the land cover of the images in the dataset are also available. This is much more data than was previously available in the context of change detection and opens the doors for many new ideas to be tested.

3.2. Labels

The labels in the dataset come from EEA's Copernicus Land Monitoring Service - Urban Atlas project[2]. It provides "reliable, inter-comparable, high-resolution land use maps" for functional urban areas in Europe with more than 50000 inhabitants. These maps were generated for the years of 2006 and 2012, and a third map is available containing the changes that took place in that period. Only the images in the regions mapped in the Urban Atlas project and with a maximum temporal distance of one year were kept in the dataset.

The available land cover maps are divided in several semantic classes, which are in turn organised in different hierarchical levels. By grouping the labels at different hierarchical levels it is possible to generate maps that are more coarsely or finely divided. For example, grouping the labels with the coarsest hierarchical level yields five classes (plus the "no information" class) shown in Table 1. This hierarchical level will henceforth be referred to as L1.

These maps are openly available in vector form online. We have used these vector maps and the georeferenced *BD ORTHO* images to generate rasters of the vector maps that are aligned with the rasters of the images. These rasters allow us to have ground truth information about each pixel in the dataset.

[1] http://www.ign.fr/

[2] https://www.eea.europa.eu/data-and-maps/data/urban-atlas

Table 1. Urban Atlas land cover mapping classes at hierarchical level L1

Code	Class
0	No information
1	Artificial surfaces
2	Agricultural areas
3	Forests
4	Wetlands
5	Water

It is important to note that there are slight differences in the semantic classes present in Urban Atlas 2006 and in Urban Atlas 2012. These differences do not affect the L1 hierarchical grouping and therefore had no consequence in the work presented later in this paper. It may nevertheless affect future works done with the data. We leave it up to the users how to best interpret and deal with these differences. More information will be in the dataset files.

3.3. Distribution rights

The *BD ORTHO* images provided by IGN are available for free for research purposes, but not all images can be redistributed by the users. That is the case for the images taken in 2005 and 2006. Nevertheless, we will make available all the data for which we have the rights of redistribution and the rasters that we have generated for semantic change detection and land cover mapping. The dataset will also contain instructions for downloading the remaining images that are necessary for using the dataset directly from IGN's website.

3.4. Dataset analysis

Despite its unprecedented size and qualities, we acknowledge in this section the dataset's limits and challenges. Nevertheless, we will show later in this paper that despite these limitations, the dataset allows for the boundaries of the state-of-the-art in semantic change detection through machine learning to be pushed.

One of the most apparent issues with the images in the dataset are the boundaries resulting from the mosaicing of the images. Since the available images are the result of a mosicing of several acquisitions, the boundaries where two different images touch is sometimes very clear as can be seen in Fig. 2 (j)-(l). These artefacts are strong in some cases and add to the difficulty of using these images with any computer algorithms, including change detection.

Another issue is the accuracy of the labels contained in the Urban Atlas vector maps with respect to the *BD ORTHO* images. We do not have access to the images used to build the Urban Atlas vector maps, nor to the exact dates of their acquisitions, nor to the dates of acquisition of the images in *BD ORTHO*. Hence, there are discrepancies between the information in the vector maps and in the images. Furthermore, EEA only guarantees a minimum label accuracy of 80-85% depending on the considered class. Most of the available data is accurate, but it is important to consider that the labels in the dataset are noisy and not flawless. Examples of false negatives and false positives can be see in Fig. 2 (d)-(f) and Fig. 2 (g)-(i), respectively.

Table 2. Change class imbalance at hierarchical level L1. Row number represent class in 2006, column number represent class in 2012.

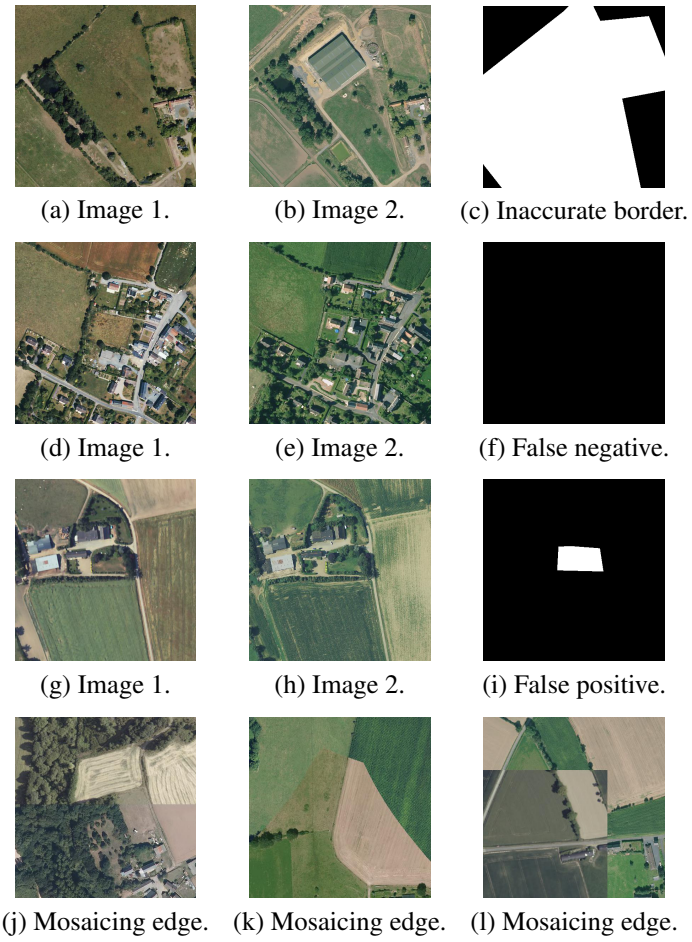
	1	2	3	4	5
1	0%	0.011%	0%	0.001%	0.001%
2	0.653%	0%	0.001%	0%	0.077%
3	0.014%	0.002%	0%	0%	0%
4	0%	0%	0%	0%	0%
5	0.001%	0.004%	0%	0.004%	0%
No change			99.232%		

It is also worth noting that the labels have been created using previously known vector maps, mostly by labelling correctly each of the known regions. This means a single label was given to each region, and this led to inaccurate borders in many cases. Often, a single change inside a predefined region will result in the whole region being marked as a change, i.e. the change class is often over labelled. This can be clearly seen in Fig. 2 (a)-(c).

One of the main challenges involved in using this dataset for supervised learning is the extreme label imbalance. As can be seen in Table 2, 99.232% of all pixels are labelled as no change, and the largest class is from agricultural areas to artificial surfaces (i.e. class 2 to class 1), which accounts for 0.653% of all pixels. These two classes together account for 99.885% of all pixels, which means all other change types combined account for only 0.115% of all pixels. Furthermore, many of the possible types of change have no examples at all in any of the images of the dataset. It is of paramount importance when using this dataset to take into account this imbalance. This also means that using the overall accuracy as a performance metric with this dataset is not a good choice, as it virtually only reflects how many pixels of the no change class have been classified correctly. Other metrics, such as Cohen's kappa coefficient or the Sørensen-Dice coefficient for binary cases, must be used instead.

Despite these challenges, we will show in Section 5 that it is possible to use this dataset for supervised training for change detection, although the quality of the labels has consequences in the quality of the results. Nevertheless, the problem of supervised learning using noisy labels has already been studied and evidence suggests that supervised learning with noisy labels is possible as long as a dataset of a large enough size is used (Rolnick et al., 2017). Other works attempt to explicitly deal with the noisy labels present in the dataset and prioritise the correct labels during training (Maggiolo et al., 2018).

Finally, we acknowledge how challenging it is to use hierarchical levels finer than L1. First, this would result in a massive increase in the number of possible changes. Second, the difference between similar classes becomes more abstract an context based. For example, the difference between the "Discontinuous Medium Density Urban Fabric" and the "Discontinuous Low Density Urban Fabric" classes defined in Urban Atlas depends not only in correctly identifying the surface at a given pixel (e.g. building or grass), but also by understanding the surroundings of the pixel and calculating the ratio between these two classes at a given neighbourhood that is not clearly defined.



(a) Image 1. (b) Image 2. (c) Inaccurate border.
(d) Image 1. (e) Image 2. (f) False negative.
(g) Image 1. (h) Image 2. (i) False positive.
(j) Mosaicing edge. (k) Mosaicing edge. (l) Mosaicing edge.

Fig. 2. Examples of: ((a)-(c)) overly large change markings, ((d)-(f)) failure to mark changes, ((g)-(i)) false positive, ((j)-(l)) mosaicing edges.

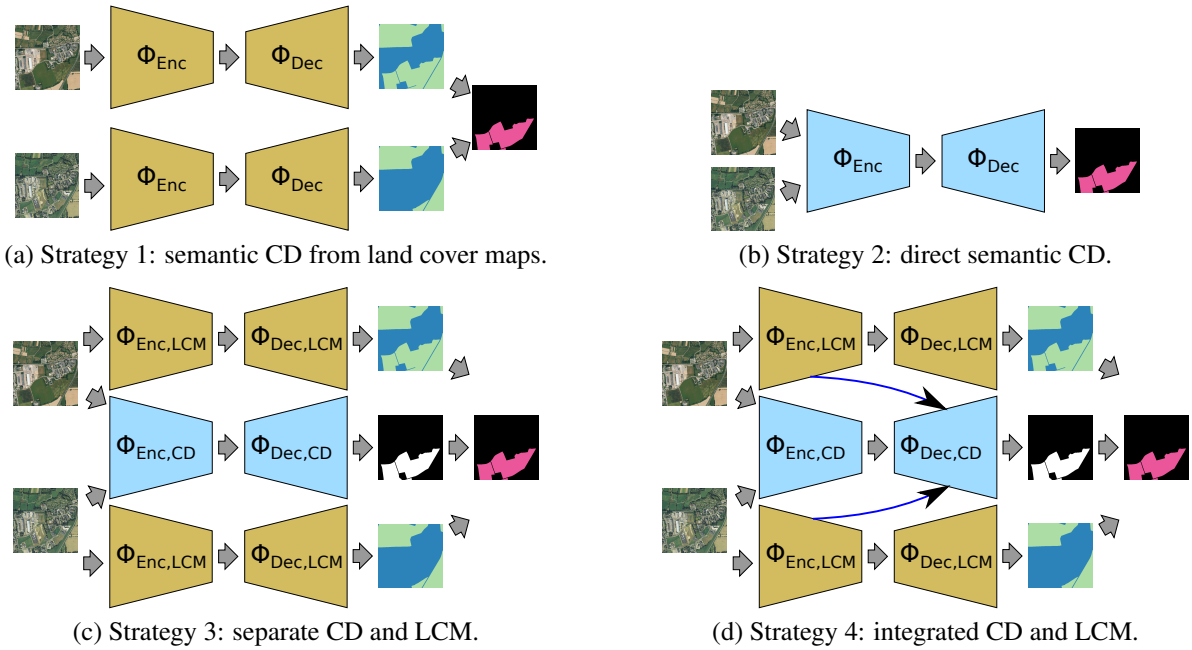
4. Methodology

4.1. Binary change detection

We have already showed in a previous work the efficacy of using three different architectures of fully convolutional neural networks for change detection (Daudt et al., 2018a). Chen et al. (2018) simultaneously proposed a fully convolutional architecture for change detection that is very similar to one of the three initially proposed architectures. In both of these works, FCNN architectures performed better than previous methods for change detection.

Building on this previous work, we have modified the FC-EF architecture proposed in Daudt et al. (2018a) to use residual blocks (He et al., 2016). The resulting network is later referred to as FC-EF-Res. These residual blocks were used in an encoder-decoder architecture with skip connections to improve the spatial accuracy of the results (Ronneberger et al., 2015). These residual blocks were chosen to facilitate the training of the network, which is especially important for its deeper variations that will be discussed later.

When testing on the OSCD dataset (Section 5.1), the size of the network has been kept approximately the same as in Daudt et al. (2018a) to avoid overfitting. When using the proposed HRSCD dataset (Section 5.2), the larger amount of annotated



(a) Strategy 1: semantic CD from land cover maps.

(b) Strategy 2: direct semantic CD.

(c) Strategy 3: separate CD and LCM.

(d) Strategy 4: integrated CD and LCM.

Fig. 3. Schematics for all four proposed strategies for semantic change detection.

pixels allows us to use deeper and more complex models. In that case, the number of encoding levels and residual blocks per level has been increased, but the idea behind the network is the same as of FC-EF-Res.

4.2. Semantic change detection

As was mentioned earlier, the efficiency of the proposed architecture for binary change detection and the availability of the HRSCD dataset enables us to tackle the problem of semantic change detection. This problem consists of two separate but not independent parts. The first task is analogue to binary change detection, i.e. we attempt to determine whether a change has occurred at each pixel in a co-registered multi-temporal image pair. The second task is to differentiate between types of changes. In our case, this consists of predicting the class of the pixel in each of the two given images. The problem of semantic change detection lies in the intersection between change detection and land cover mapping.

Below we will describe four different intuitive strategies to perform semantic change detection using deep neural networks. Starting from the plain comparison of land cover maps, we then develop more involved strategies. These strategies vary in complexity and performance, as will be discussed in Section 5.

4.2.1. Direct comparison of land cover maps

The problem of automatic land cover mapping is a well studied problem. In particular, methods involving CNNs have recently been proposed, yielding good performances (Audebert et al., 2016). When the land cover information is available, as it is the case in the HRSCD dataset, the most intuitive method that can be proposed for semantic change detection would be to train a land cover mapping network and to compare the results for pixels in the image pair (see Fig. 3(a)).

The advantage of this method is its simplicity. In many cases we could assume changes occurred where the predicted class label differs between the two images, and the type of change is given by the predicted labels at each of the two acquisition moments. The weakness of this method is that it heavily depends on the accuracy of the predicted land cover maps. While modern FCNNs are able to map areas to a certain degree of accuracy, the results are not perfect. Furthermore, when comparing the results for two acquisitions the prediction errors would accumulate. This means the accuracy of this change detection algorithm would be lower than the land cover mapping network, and would likely predict changes in the borders between classes simply due to the inaccuracy of the network.

4.2.2. Direct semantic change detection

A second intuitive approach is to treat each possible type of change as a different and independent label, and treating semantic change detection as a simple semantic segmentation along the lines of what has been done to binary change detection in the past (Daudt et al., 2018a).

The weakness of this method is that the number of change classes grows proportionately to the square of the number of land cover classes that is considered. This, combined with the class imbalance problem that was discussed earlier, proves to be a major challenge when training the network.

4.2.3. Separate land cover mapping and change detection

Since it has been proven before that FCNNs are able to perform both binary change detection and land cover mapping, a third possible approach is to train two separate networks that together perform semantic change detection (see Fig. 3(c)). The first network performs binary change detection on the image pair, while the second network performs land cover mapping of

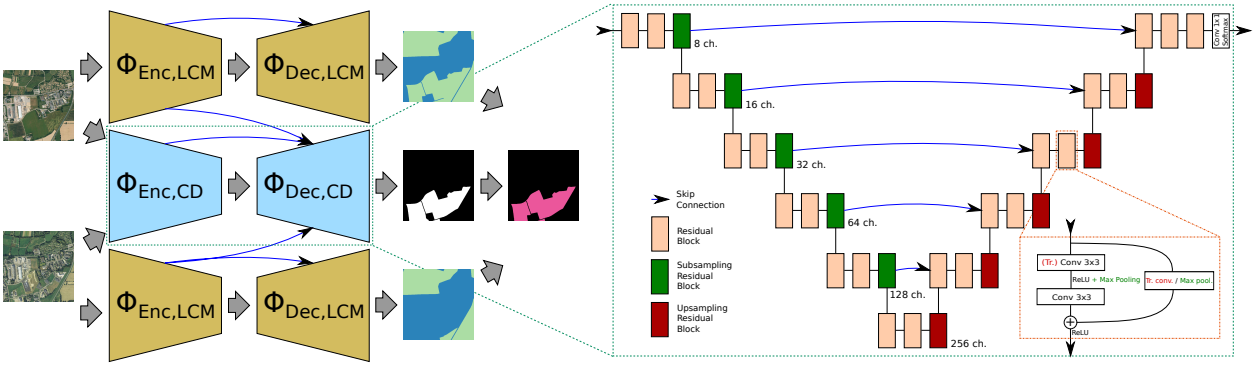


Fig. 4. Detailed schematics for the integrated change detection and land cover mapping network (Strategy 4). The encoder-decoder architecture is the same that was used for all 4 strategies.

each of the input images. The two networks are trained separately since they are independent.

In this strategy, the two input images produce three outputs: two land cover maps and a change map. At each pixel, the presence of change is predicted by the change map, and the type of change is defined by the classes predicted by the land cover maps at that location. This way the number of predicted classes is reduced relative to the previous strategy (i.e. the number of classes is no longer proportional to the square of land cover classes) without loss of flexibility. This helps with the class imbalance problem. It also avoids the problem of predicting changes at every pixel where the land cover maps differ, since the change detection problem is treated separately from land cover mapping.

We argue that such network may be able to identify changes of types it has not seen during training, as long as it has seen the land cover classes during training. For example, the network could in theory correctly classify a change from agricultural area to wetland even if such changes are not in the training set, as long as it has enough examples of those classes to correctly classify them in the land cover mapping branches. The combination of two separate networks allows us to split the problem into two, and optimise each part to maximise performance.

4.2.4. Integrated land cover mapping and change detection

The last of the proposed approaches is an evolution of the previous strategy of using two FCNNs for the tasks of binary change detection and land cover mapping. We propose to integrate the two FCNNs into a single multitask network (see Fig. 3(d) and Fig. 4) so that land cover information can be used for change detection. The combined network takes as input the two co-registered images and outputs three maps: the binary change map and the two land cover maps.

In the proposed architecture, information from the land cover mapping branches of the network is passed to the change detection branch of the network in the form of difference skip connections, which was shown to be the most effective form of skip connections for Siamese FCNNs (Daudt et al., 2018a). The weights of the two land cover mapping branches are shared since they perform an identical task, allowing us to significantly reduce the number of learned parameters.

This multipurpose network gives rise to a new issue during the training phase. Given that the network outputs three different image predictions, it is necessary to balance the loss functions from these results. Since two of the outputs have exactly the same nature (the land cover maps), it follows from the symmetry of these branches that they can be combined into a single loss function by simple addition. The question remains on how to balance the binary change detection loss function and the land cover mapping loss function to maximise performance.

We have proposed and tested two different strategies for training the network. The first and more naive approach to this problem is to minimise a loss function that is a weighted combination of the two loss functions. This loss function would have the form

$$\begin{aligned}\mathcal{L}_\lambda(\Phi_{\text{Enc,CD}},&\Phi_{\text{Dec,CD}},\Phi_{\text{Enc,LCM}},\Phi_{\text{Dec,LCM}})\\ &=\mathcal{L}(\Phi_{\text{Enc,CD}},\Phi_{\text{Dec,CD}})+\lambda\mathcal{L}(\Phi_{\text{Enc,LCM}},\Phi_{\text{Dec,LCM}})\end{aligned} \tag{1}$$

where Φ represents the network parameters (for the encoder and decoder parts of the change detection and land cover mapping branches), and $\mathcal{L}$ is a pixel-wise loss function. In this work, the pixel-wise cross entropy function was used as loss function as is traditional in semantic segmentation problems. The problem then becomes the search for the value of λ that leads to the best balance between the two loss terms. This can be found through a grid search, but the test of each value of lambda is done by training the whole network until convergence, which is a slow and costly procedure. This will later be referred to as Strategy 4.1.

To reduce the aforementioned training burden, we propose a second approach to train the network that avoids the need of setting the hyperparameter λ. We train the network in two stages. First, we consider only the land cover mapping loss

$$\begin{aligned}\mathcal{L}_1(\Phi_{\text{Enc,CD}},&\Phi_{\text{Dec,CD}},\Phi_{\text{Enc,LCM}},\Phi_{\text{Dec,LCM}})\\ &=\mathcal{L}(\Phi_{\text{Enc,LCM}},\Phi_{\text{Dec,LCM}})\end{aligned} \tag{2}$$

and train only the land cover mapping branches of the network, i.e. we do not train $\Phi_{\text{Enc,CD}}$ or $\Phi_{\text{Dec,CD}}$ at this stage. Since the change detection branch has no influence on the land cover mapping branches, we can train these branches to achieve the maximum possible land cover mapping performance with the given architecture and data. Next, we use a second loss function

Table 3. Change detection results of several methods on the OSCD dataset

Data	Network	Prec.	Recall	Tot. acc.	F1
3 ch.	FC-EF	44.72	53.92	94.23	48.89
	FC-Siam-conc	42.89	47.77	94.07	45.20
	FC-Siam-diff	49.81	47.94	94.86	48.86
	FC-EF-Res	**52.27**	**68.24**	**95.34**	**59.20**
13 ch.	FC-EF	**64.42**	50.97	**96.05**	56.91
	FC-Siam-conc	42.39	65.15	93.68	51.36
	FC-Siam-diff	57.84	57.99	95.68	57.92
	FC-EF-Res	54.93	**66.48**	95.64	**60.15**

based only on the change detection branch:

$$\begin{aligned}\mathcal{L}_2(\Phi_{\text{Enc,CD}},\Phi_{\text{Dec,CD}},\Phi_{\text{Enc,LCM}},\Phi_{\text{Dec,LCM}})\\ =\mathcal{L}(\Phi_{\text{Enc,CD}},\Phi_{\text{Dec,CD}})\end{aligned} \tag{3}$$

while keeping the weights for the land cover mapping $\Phi_{\text{Enc,LCM}}$ and $\Phi_{\text{Enc,LCM}}$ fixed. This way, the change detection branch learns to use the predicted land cover information to help to detect changes without affecting land cover mapping performance. This will later be referred to as Strategy 4.2.

5. Results

5.1. Multispectral change detection

We first evaluate the performance of the proposed FC-EF-Res network. As explained in Section 4.1, this network is an evolution of the convolutional architecture FC-EF proposed in Daudt et al. (2018a), to which residual blocks have been added in place of traditional convolutional layers.

The FC-EF-Res architecture was compared to the previously proposed FCNN architectures on the OSCD dataset for binary change detection. As expected, the residual extension of the FC-EF architecture outperformed all previously proposed architectures. The difference was noted on both the RGB and the multispectral cases. On the RGB case, the improvement was of such magnitude that the change detection performance on RGB images almost matched the performance on multispectral images. The results can be seen in Table 3. This corroborates the claims made by He et al. (2016) that using residual blocks improves the training performance of CNNs. For this reason, all networks that are tested with the HRSCD dataset use residual modules.

5.2. High resolution semantic change detection

To test the methods proposed in Section 4.2 we split the HRSCD images into two groups: 146 image pairs for training and 145 image pairs for testing. By splitting the train and test sets this way we can ensure that no pixel in the test set has been seen during training. Class weights were set inversely proportional to the number of training examples to counterbalance the dataset's class imbalance. The results for each of the proposed strategies can be seen in Table 4, and illustrative image results can be seen in Fig. 5.

In Strategy 1, which naively attempts to predict change maps from land cover maps, we can see that the network succeeds in accurately classifying the imaged terrains, but this is not enough to predict accurate change maps. The change detection kappa coefficient for this strategy is very low, which means this method is marginally better than chance for change detection.

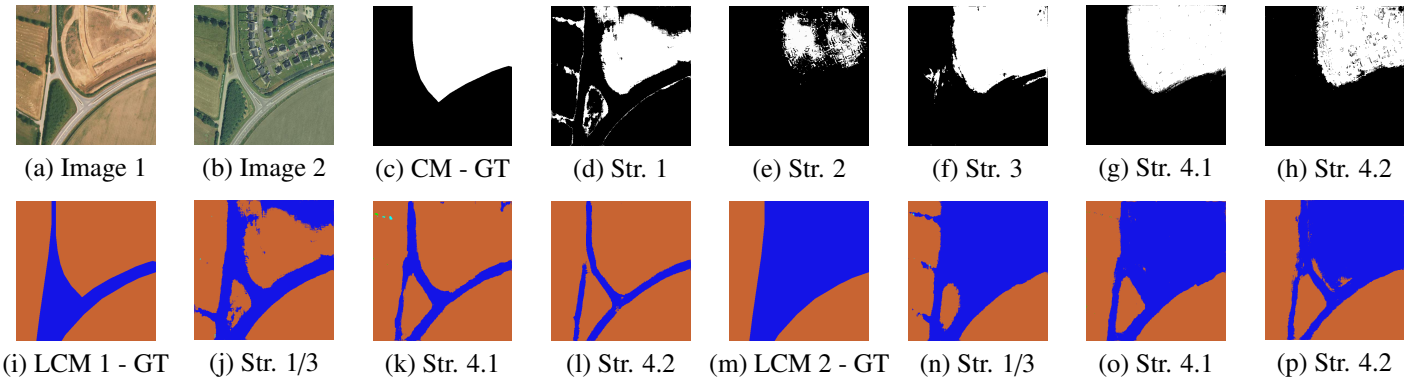
The results for Strategy 2 are a fair improvement over those of Strategy 1. The change detection Dice coefficient and the land cover mapping results for this method are not reported due to its nature, since Dice coefficients can only be calculated for binary classification problems, and this strategy bypasses the land cover mapping steps. Despite achieving a higher kappa coefficient, the network learned to always predict the same type of change where changes occurred. This means that despite using appropriately tuned class weights, the learning process did not succeed in overcoming the extreme class imbalance present in the dataset. In other words, the network learned to detect changes but no semantic information was present in the results.

For Strategy 3, the land cover mapping network that was used was the same as that of Strategy 1, which achieved good performance. A binary change detection network was trained to be used for masking the land cover maps. The performance of this network was better than that of Strategy 1 but worse than that of Strategy 2. The results show that this is due to an overestimation of the change class. This shows once again how challenging dealing with the extreme class imbalance is.

The results of Strategy 4 are the best ones overall. The simultaneous training strategy (Str. 4.1) achieves excellent performance in both land cover mapping and change detection, proving the viability of this strategy. The reported results were obtained with $\lambda = 0.05$, which is a value that prioritises the training of the change detection branch of the network. We then see that the same network trained with sequential training (Str. 4.2) obtained even better results in both change detection and land cover mapping without needing to search for an adequate parameter λ. This, according to our results, is the best semantic change detection method. By comparing the results for Strategies 3 and 4 we can see the improvements that result directly from integrating the change detection and land cover mapping branches of the networks. In other words, Strategy 4.2 allows us to maximise the change detection performance without reducing the land cover mapping accuracy.

The best performing land cover mapping method was the single purpose network that was trained and used for Strategies 1 and 3. The fact that it achieves a better kappa coefficient than Strategy 4.2 is merely due to the randomness of the initialisation and training of the network, as the land cover mapping branches of Strategy 4.2 are identical to those used in Strategies 1 and 3. This also explains why their results are so similar. By comparing these results to those of Strategy 4.1 it emphasises once again the fact that attempting to train the network shown in Fig. 4 all at once damages performance in both change detection and land cover mapping.

In Fig. 5 we can see the results of the proposed networks on a pair of images from the dataset. Note the amount of false detections by Strategy 1 due to the lack of accuracy of prediction of the land cover maps on region boundaries. The second row shows the predicted classes at each pixel for each image. The semantic information about the changes comes from compar-



(a) Image 1 (b) Image 2 (c) CM - GT (d) Str. 1 (e) Str. 2 (f) Str. 3 (g) Str. 4.1 (h) Str. 4.2

(i) LCM 1 - GT (j) Str. 1/3 (k) Str. 4.1 (l) Str. 4.2 (m) LCM 2 - GT (n) Str. 1/3 (o) Str. 4.1 (p) Str. 4.2

Fig. 5. Illustrative images of the obtained results: (a)-(b) multitemporal image pair; (c) ground truth change map; (d)-(h) predicted change maps; (i)-(l) ground truth and predicted land cover maps for image 1; (m)-(p) ground truth and predicted land cover maps for image 2.

Table 4. Change detection (CD) and land cover mapping (LCM) results of all four of the proposed strategies on the HRSCD dataset

	CD			LCM	
	Kappa	Dice	Tot. acc.	Kappa	Tot. acc.
Str. 1	3.99	5.56	86.07	**71.92**	87.22
Str. 2	21.54	-	**98.30**	-	-
Str. 3	12.48	13.79	94.72	**71.92**	87.22
Str. 4.1	19.13	20.23	96.87	67.25	85.74
Str. 4.2	**25.49**	**26.33**	98.19	71.81	**89.01**

Table 5. Change detection results on Eppalock lake test images

		ReCNN-LSTM	EF
Binary CD	OA	98.67	**99.35**
	Kappa	97.28	**98.67**
	No change	98.83	**99.47**
	Change	98.46	**99.19**
Semantic CD	OA	**98.70**	98.48
	Kappa	**97.52**	97.10
	No change	**98.49**	97.73
	City exp.	84.72	**100**
	Soil change	**100**	86.07
	Water change	99.25	**99.93**

ing these two predictions. For example, comparing the images in Fig. 5 (k) and (o) we can say that the changes predicted in (g) were from the "Agricultural areas" class to the "Artificial surfaces" class.

In our tests we observed that the trained networks had the tendency to overestimate the size of the detected changes. It is likely that this happens simply due to the nature of the data that was used for training. The labels in the HRSCD dataset, which come from Urban Atlas, mark as a change the whole terrain where a change of class happened. This means that not only the pixels associated with a given change are marked as change, but the neighbouring pixels that are in the same parcel are also marked as change. This leads to the networks learning to overestimate the boundary of the detected changes in an attempt to also correctly classify the pixels surrounding the detected change. This once again reflects the challenges of the HRSCD dataset.

Finally, it is important to note that the label imperfections in the HRSCD dataset occur not only in the training images, but also in the test images. This means that the performance of the proposed methods may be even higher than the numbers suggest, since some of the disagreements between prediction and ground truth data are actually due to errors in the ground truth data.

5.3. Eppalock lake images

Very few of the change detection algorithms proposed to date aim to use large datasets, which were heretofore unavailable. Furthermore, most of the proposed algorithms do not have open source implementations, often are not described in enough detail in the papers to be accurately reimplemented and are usually tested on private data. This makes accurate comparison between different algorithms very difficult. One of the methods worth of comparison here was proposed by Mou et al. (2018), which used recurrent convolutional neural networks for change detection. The authors of the paper have been kind to share one of the image sets used in the paper for this comparison, the Eppalock lake images. In that work, pixels were randomly split into train and test sets. We believe that this split leads to overfitting since neighbouring pixels contain redundant information. This is especially true when using CNNs, which take as inputs patches centred on the considered pixels, meaning the network sees the same information for training and testing. The claim that overfitting happens is further corroborated by the fact that an accuracy of over 98% is achieved by using only 1000 labelled pixels to train a network with 67500 parameters (for their LSTM architecture, which performed the best). The data consists of a single image pair of 631x602 pixels only partially annotated, with a total of 8895 annotated pixels which is much less data than what is required for deep learning methods. In other words, the HRSCD dataset contains over 3 million times more labelled pixels than the Eppalock lake image pair. Despite our disagreeing with the merits of this testing scheme, we have followed it to achieve a fair comparison between the methods.

Using the EF architecture described in Daudt et al. (2018b),

we achieved excellent numeric results which discouraged the usage of more complex methods since these would lead to even more extreme overfitting. The results achieved by the EF network were better for binary change detection and equivalent for semantic change detection compared to ReCNN-LSTM. The results can be seen in Table 5. We have only considered here the ReCNN-LSTM results reported in the original paper, which is the best performing architecture.

6. Conclusion

The first major contribution presented in this paper is the first large scale high resolution semantic change detection dataset that will be released to the scientific community. This dataset contains 291 pairs of aerial images, together with aligned rasters for change maps and land cover maps. This dataset allows for the first time for deep learning methods to be used in this context in a fully supervised manner with minimal concern for overfitting. We have then proposed different methods for using deep FCNNs for semantic change detection. The best among the proposed methods is an integrated network that performs land cover mapping and change detection simultaneously, using information from the land cover mapping branches to help with change detection. We also proposed a sequential training scheme for this network that avoids the need of tuning a hyperparameter, which circumvents a costly grid search.

The automatic methods used to generate the HRSCD dataset resulted in noisy labels for both training and testing, and how to deal with this problem is still an open question. It would also be interesting to explore ways to explicitly deal with parallax problems which are present in high resolution images which sometimes lead to false positives due to the different points of view and the geometry of the scene.

Acknowledgments

This work is part of ONERA's project DELTA. We thank X. Zhu and L. Mou (DLR) for the Eppalock Lake images.

References

Audebert, N., Le Saux, B., Lefèvre, S., 2016. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks, in: Asian Conference on Computer Vision, Springer. pp. 180–196.

Benedek, C., Szirányi, T., 2009. Change detection in optical aerial images by a multilayer conditional mixed markov model. IEEE Transactions on Geoscience and Remote Sensing 47, 3416–3430.

Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H., 2016. Fully-convolutional siamese networks for object tracking, in: European conference on computer vision, Springer. pp. 850–865.

Bourdis, N., Denis, M., Sahbi, H., 2011. Constrained optical flow for aerial image change detection, in: 2011 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 4176–4179.

Bovolo, F., Bruzzone, L., 2005. A wavelet-based change-detection technique for multitemporal sar images, in: Analysis of Multi-Temporal Remote Sensing Images, 2005 International Workshop on the, IEEE. pp. 85–89.

Bruzzone, L., Prieto, D.F., 2000. Automatic analysis of the difference image for unsupervised change detection. IEEE Transactions on Geoscience and Remote sensing 38, 1171–1182.

Celik, T., 2009. Unsupervised change detection in satellite images using principal component analysis and k-means clustering. IEEE Geoscience and Remote Sensing Letters 6, 772–776.

Chen, Y., Ouyang, X., Agam, G., 2018. MFCNET: End-to-end approach for change detection in images, in: 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE. pp. 4008–4012.

Chopra, S., Hadsell, R., LeCun, Y., 2005. Learning a similarity metric discriminatively, with application to face verification, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, IEEE. pp. 539–546.

Daudt, R.C., Le Saux, B., Boulch, A., 2018a. Fully convolutional siamese networks for change detection, in: 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 4063–4067.

Daudt, R.C., Le Saux, B., Boulch, A., Gousseau, Y., 2018b. Urban change detection for multispectral earth observation using convolutional neural networks, in: International Geoscience and Remote Sensing Symposium (IGARSS), IEEE. pp. 2119–2122.

El Amin, A.M., Liu, Q., Wang, Y., 2016. Convolutional neural network features based change detection in satellite images, in: First International Workshop on Pattern Recognition, International Society for Optics and Photonics. pp. 100110W–100110W.

El Amin, A.M., Liu, Q., Wang, Y., 2017. Zoom out cnns features for optical remote sensing change detection, in: Image, Vision and Computing (ICIVC), 2017 2nd International Conference on, IEEE. pp. 812–817.

Gong, M., Zhao, J., Liu, J., Miao, Q., Jiao, L., 2016. Change detection in synthetic aperture radar images based on deep neural networks. IEEE transactions on neural networks and learning systems 27, 125–138.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Hussain, M., Chen, D., Cheng, A., Wei, H., Stanley, D., 2013. Change detection from remotely sensed images: From pixel-based to object-based approaches. ISPRS Journal of Photogrammetry and Remote Sensing 80, 91–106.

Lambin, E.F., Strahlers, A.H., 1994. Change-vector analysis in multitemporal space: a tool to detect and categorize land-cover change processes using high temporal-resolution satellite data. Remote sensing of environment 48, 231–244.

Le Saux, B., Randrianarivo, H., 2013. Urban change detection in sar images by interactive learning, in: Geoscience and Remote Sensing Symposium (IGARSS), 2013 IEEE International, IEEE. pp. 3990–3993.

Liu, J., Gong, M., Qin, K., Zhang, P., 2016. A deep convolutional coupling network for change detection based on heterogeneous optical and radar images. IEEE transactions on neural networks and learning systems .

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440.

Maggiolo, L., Marcos, D., Moser, G., Tuia, D., 2018. Improving maps from cnns trained with sparse, scribbled ground truths using fully connected crfs, in: International Geoscience and Remote Sensing Symposium (IGARSS), IEEE. pp. 2103–2103.

Mou, L., Bruzzone, L., Zhu, X.X., 2018. Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery. CoRR abs/1803.02642. URL: `http://arxiv.org/abs/1803.02642`.

Rolnick, D., Veit, A., Belongie, S.J., Shavit, N., 2017. Deep learning is robust to massive label noise. CoRR abs/1705.10694. URL: `http://arxiv.org/abs/1705.10694`.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Rosin, P.L., Ioannidis, E., 2003. Evaluation of global image thresholding for change detection. Pattern recognition letters 24, 2345–2356.

Singh, A., 1989. Review article digital change detection techniques using remotely-sensed data. International journal of remote sensing 10, 989–1003.

Stent, S., Gherardi, R., Stenger, B., Cipolla, R., 2015. Detecting change for multi-view, long-term surface inspection., in: BMVC, pp. 127–1.

Zagoruyko, S., Komodakis, N., 2015. Learning to compare image patches via convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4353–4361.

Zhan, Y., Fu, K., Yan, M., Sun, X., Wang, H., Qiu, X., 2017. Change detection based on deep siamese convolutional network for optical aerial images. IEEE Geoscience and Remote Sensing Letters 14, 1845–1849.