DETECTING URBAN CHANGES WITH RECURRENT NEURAL NETWORKS FROM MULTITEMPORAL SENTINEL-2 DATA

Maria Papadomanolaki¹,²*, Sagar Verma²,³*, Maria Vakalopoulou², Siddharth Gupta³, Konstantinos Karantzalos¹

¹ Remote Sensing Laboratory, National Technical University of Athens, Greece
² CVN, CentraleSupélec, Université Paris-Saclay and INRIA Saclay, France
³ Granular AI, MA, USA

ABSTRACT

Change detection is one of the most investigated topics in remote sensing, aiming accurate and efficient ways to continuously monitor the earth’s surface. In this paper, we present a method for urban change detection using Sentinel-2 multitemporal high-resolution data. A novel deep learning framework is proposed combining state-of-the-art fully convolutional networks (like U-Net) for feature representation and powerful recurrent networks (like LSTMs) for temporal modeling. We report our results on the recently publicly available bi-temporal Onera Satellite Change Detection (OSCD) dataset, adding even more dates to enhance the temporal information. Moreover, we evaluate the performance of the recurrent networks as well as the use of the additional dates on the unseen test set using an ensemble cross-validation strategy. All the implemented models score overall accuracy more than 95%, while the use of LSTMs and additional temporal information, boost the recall of the change class by more than 2%.

Index Terms— change detection, fully-convolutional, urban, recurrent networks, multi-temporal modeling.

1. INTRODUCTION

Change detection is a critical issue for the remote sensing community as it provides an effective way of monitoring the globe. A thorough understanding of the earth’s land usage and land cover (LULC) in time can be made by modeling the changes that occur owing to man-made structures and natural phenomena. As far as human intervention on earth is concerned, change detection techniques offer valuable information on a variety of topics such as urban sprawl, water and air contamination levels, illegal constructions, etc. In this way, we can fully comprehend the future LULC tendencies, take precautions and design more appropriate city infrastructures.

However, even if nowadays we can have access to a large amount of multitemporal datasets provided by satellites such as Landsat and Sentinel, the problem of change detection is very challenging. Traditional methods, summarised in surveys such as [1], use handcrafted techniques which heavily rely on pre-processing and post-processing making them not so easily adaptive to images that cover large areas. Change detection is a non-trivial problem as the accuracy of a method is highly influenced by registration errors [2] and illumination changes that do not really correspond to semantic changes.

Recently, with the advances in deep learning-based methods in different fields, a variety of change detection techniques have been proposed. In particular, in [3] a deep patch-based architecture is proposed where bi-temporal patches are processed in parallel by a series of dilated convolutional lay-
ers generating features which are then fed to a recurrent sub-
network to learn sequential information. In the end, fully-
connected layers are used to create the change prediction map. 
Although patch-based techniques produce promising results, 
they are time-consuming since they need to process every sin-
gle pixel of the image individually. While in the semantic 
segmentation domain fully-convolutional architectures domi-
nate the literature results, change detection methods lack such 
kind of experiments. Recently, Daudt et al. [4] suggested 
three different fully-convolutional siamese approaches based 
on the U-Net architecture [5] aiming to classify successfully the 
changes. However, these models require a large amount 
of data for training.

In this paper, we investigate the use of recurrent networks 
and in particular Long Short-Term Memory (LSTMs) [6] lay-
ers for pixel-wise urbanisation detection from multi-temporal 
high-resolution data. The proposed deep learning model uses 
U-Net to compute spatial features from multi-date inputs and 
then multiple LSTM layers are used to learn the tem-
poral change pattern by receiving the fused encoder outputs 
of multiple dates. For our experiments we use the recently 
publicly available Onera Satellite Change Detection dataset 
(OSCD) [7] after enriching it with more temporal information 
for each of the provided cities. Our final aim is to study the 
behaviour and prospects of such a model when multitemporal 
data are available eliminating at the same time the need for 
any fully-connected layers. The rest of the paper is organized 
as follows. In section 2 the proposed architecture is described, 
in section 3 dataset and experiments are presented while in 
section 4, quantitative and qualitative results are presented. 
Lastly, in section 5 a final conclusion is made.

2. METHODOLOGY

The architecture that we employed shares similarities with the 
fully-convolutional siamese concatenation proposed in [4], al-
though we used five convolutional blocks in the encoder and 
the decoder applying five max-pooling operations in order to 
bring the input volume down to a very low-resolution. All 
convolution operations apply 3x3 filters with both stride and 
padding being equal to 1. As far as the encoder is concerned, 
the first convolutional block increases the depth to 16, while 
the height and width of the input volume are downsampled in 
half their size owing to the 2x2 max pooling operation at 
the end of the block. The next three convolutional blocks 
follow the same pattern resulting in an output of size batch-
size x 128 x 2 x 2 while the last convolutional block keeps the 
depth unchanged producing a volume of batchsize x 128 x 1 x 1. 
Finally, a dropout layer of threshold 0.4 is also added at the 
end of the encoder.

After that, six LSTM layers follow which take as input the 
concatenation of the encoder’s outputs for the different dates. 
The encoding part of the proposed architecture results in a 
single-date output volume of size batchsize x 128 x 1 x 1, which 
is then reshaped to batchsize x 1 x 128 in order to be concate-
nated with the rest of the dates along the second dimension that represents the sequence of the data. Thus, in the three 
dates case the LSTM sub-network, whose hidden size is equal 
to 128, process a volume of batchsize x 3 x 128 and produces 
an output of batchsize x 3 x 128. The LSTM output is then fed 
to the decoder directly without the need to modify its shape 
through fully connected layers.

Next, the decoder receives the LSTM output to upsample 
it back to its original dimensions. For this purpose, five con-
volutional blocks similar to the encoder are used, this time 
applying 2x2 upsampling operations instead of max-pooling 
one. In addition, the resulted feature map of each upsam-
pling operation is concatenated with the feature map of the 
symmetrical block existing in the encoder part for each of the 
different dates. In this way, higher resolution information is 
combined with lower resolution information producing more 
sophisticated features and maintaining spatial and temporal 
knowledge. Finally, at the end of the model, a 1x1 convo-
olution operation is applied to compute the final probability 
heatmap detecting areas of urbanisation.

2.1. Dataset and Implementation Details

All the experiments were conducted using the Onera Satel-
lite Change Detection dataset (OSCD) [7] which consists of 
Sentinel-2 satellite images depicting 24 different cities around 
the world for two distinct dates. 13 spectral channels are 
available for each image pair with ground truth information 
provided in 14 cases. Our setup follows the submission sys-
tem guidelines where the 14 image pairs accompanied by 
ground truth information are used for training and the rest for 
testing.

As mentioned earlier, Sentinel-2 images depicting the 
provided cities in different dates were acquired to further en-
rich the temporal information of the OSCD dataset. The dates 
of some images are shown in Table 1. The first and last rows include the before and after dates that are already provided in the 
OSCD dataset. In general, we tried to obtain dates corre-
sponding to similar seasons for every city, adapting them as 
 much as possible to the two existed OSCD dates.

For the training process, patches of size 32x32 were pro-
duced with a stride of either 6 in case change pixels were 
included, or 32 in case the patch did not include any change 
pixels. This strategy was applied as a data augmentation ap-
proach to enrich the training samples that contain change pix-
els. In addition, more data augmentation techniques mainly 
used from the computer vision community, namely flipping 
in all possible angles proportional to 90 degrees, were imple-
mented for patches whose number of change pixels exceeded 
a certain threshold. Lastly, each class was associated with 
a weight inversely proportional to the total pixel number in-
cluded in it. A total of 32421 patches containing both change 

http://dase.grss-ieee.org
and non-change pixels resulted from the 14 training cities and was feedforwarded to the proposed architecture for training. To provide more robust results our final predictions were produced by an ensemble of different trained models following a cross-validation. Giving some more details, the training patches were divided into five equal parts and the same model was trained five times using all possible combinations of the dataset partitions. Then, predictions for the testing images were produced from all five models and in the end the final predictions were formulated by averaging the five model outcomes. This pipeline was necessary for the effective use of all patches during training. Regarding hyperparameter details, the chosen optimizer was Adam while batch size and learning rate values were equal to 64 and 0.0001 respectively. All experimental setups were conducted using the PyTorch deep learning library [8] on a single NVIDIA GeForce GTX TITAN with 12 GB of GPU memory, with the training time for each model being approximately 70 minutes.

### 3. EXPERIMENTAL RESULTS AND DISCUSSION

Different model combinations were evaluated on the testing images, the results of which are shown in Table 2. Precision, recall, F1 score and Overall Accuracy rates have resulted from the confusion matrix of the testing images for each different approach while all quantitative numbers except OA are associated to the change class. Beginning with the plain U-Net architecture we can observe that all evaluation metrics are higher for the 2/4 dates/channels combination. Even when 13 channels are exploited in the 2 dates case, the reported accuracy is slightly lower despite the additional spectral information. This can be explained by the fact that R-G-B-NIR channels provide the highest spatial resolution in Sentinel-2 images. This means that the integration of the remaining lower resolution channels to the training process requires the use of more sophisticated pansharpening methods than only simple resizing. Based on this fact and since 13 channel volumes require much more computational power we based our experiments on the four higher resolution channels. Continuing with the quantitative analysis, one can notice that the more the dates are added to the simple U-Net architecture, the lower the accuracy levels become. In fact, when more temporal information is provided the model tends to fail detecting the change class as more than half of it is classified as non-change. This is obvious from the recall rates which fall below 50%. Such a result indicates that simple architectures that do not maintain temporal information fail to integrate properly multi-temporal inputs.

On the other hand, the change class seems to be detected more successfully when LSTM layers are integrated in the deep architectures. Although Overall Accuracy rates do not increase, precision rates are over 50% in all Dates-Channels combinations. In Figure 2, we can also see some qualitative results from the best model combinations. Red and green parts in the predictions’ visualization represent false positive and false negative areas that we have manually annotated, only for the qualitative analysis of our method, since no ground truth is available for the testing part of the dataset. As we can observe, multi-date methods have detected changes related to smaller buildings more successfully as shown in the Chongqing case in the second row. In addition, bi-temporal approaches tend to confuse urban changes with changes related to bare soil lands, whereas multi-date approaches seem to better distinguish these different types of change. This is obvious in the last row of Figure 2 where false positives are reduced when more dates are exploited. In the Dubai case however, bare-soil-related false positives continue to exist in the 5-dates case, whereas they are eliminated in the 3-dates case. Such false positive predictions may be a result of various factors like clouds, improper registration etc.

### 4. CONCLUSION

In this paper, we investigated the behaviour of LSTM layers integrated into fully-convolutional deep architectures for ur-
ban change detection. Several experiments were implemented using various combinations of architectures and inputs. Results on the OSCD dataset indicate that the use of recurrent networks can boost the accuracy of the change class by more than 2%. Our future steps include the investigation of other types and combinations of recurrent and fully convolutional architectures on high-resolution images. Moreover, we will also investigate the use of these models for change detection on very high-resolution satellite imagery.

5. REFERENCES


