URBAN CHANGE DETECTION BASED ON SEMANTIC SEGMENTATION AND FULLY CONVOLUTIONAL LSTM NETWORKS

Maria Papadomanolaki[1,2]*, Maria Vakalopoulou[2,3], Konstantinos Karantzalos[1]

[1] Remote Sensing Laboratory, National Technical University of Athens, Zographos, Greece
[2] Université Paris-Saclay, CentraleSupélec, MICS Laboratory, Gif-sur-Yvette, France
[3] Université Paris-Saclay, CentraleSupélec, Inria, Gif-sur-Yvette, France

Commission II, WG II/6

KEY WORDS: change detection, buildings segmentation, multi-task learning, deep learning, fully convolutional LSTMs, very high resolution imagery

ABSTRACT:

Change detection is a very important problem for the remote sensing community. Among the several approaches proposed during recent years, deep learning provides methods and tools that achieve state of the art performances. In this paper, we tackle the problem of urban change detection by constructing a fully convolutional multi-task deep architecture. We present a framework based on the UNet model, with fully convolutional LSTM blocks integrated on top of every encoding level capturing in this way the temporal dynamics of spatial feature representations at different resolution levels. The proposed network is modular due to shared weights which allow the exploitation of multiple (more than two) dates simultaneously. Moreover, our framework provides building segmentation maps by employing a multi-task scheme which extracts additional feature attributes that can reduce the number of false positive pixels. We performed extensive experiments comparing our method with other state of the art approaches using very high resolution images of urban areas. Quantitative and qualitative results reveal the great potential of the proposed scheme, with F1 score outperforming the other compared methods by almost 2.2%.

1. INTRODUCTION

Urban change detection is one of the most studied topics in remote sensing since it provides useful insights concerning the cities' growing patterns and future tendencies. Low air quality, water contamination and limited greenery spaces are only some of the environmental issues that arise from the continuous urban growth. Moreover, many other social problems can be raised from extending urban areas, like poverty and increased crime rates. It is therefore reasonable to understand and study thoroughly such expansion trends in different spatial scales so as to create better city infrastructures and prevent situations that can be extremely dangerous both for the environment and humanity.

During the last decades, the high availability of earth observation data has enabled the remote sensing community to collect multimodal, multitemporal satellite images laying in this way the foundation for constructive research studies. To this day, manual change detection approaches have been replaced with automatic supervised and unsupervised algorithms such as graphical models and Markov Random Fields (Singh et al., 2014, Benedek et al., 2015, Vakalopoulou et al., 2016, Vakalopoulou et al., 2015, Karantzalos, 2015), kernels (Volpi et al., 2012), as well as Principal Component Analysis (Li, Yeh, 1998, Deng et al., 2008). With the advent of neural networks, recent works are more and more oriented to deep learning approaches, producing state of the art results and setting promising prospects for the urban change detection task. In (Caye Daudt et al., 2018b), a patch-based framework is suggested where two different architectures (Siamese and Early Fusion) are examined using the Onera Satellite Change Detection bi-temporal dataset. (Caye Daudt et al., 2018a) then transforms these approaches to fully convolutional versions based on a UNet-like scheme. Multi-task learning methods involving supplementary tasks mainly including semantic segmentation (Liu et al., 2019, Daudt et al., 2018) have also been employed, since they can benefit greatly the training procedure by providing additional fruitful feature representations.

Furthermore, since the problem of change detection involves sequential data, the need to calculate temporal dynamics emerges, leading to the employment of Recurrent Neural Networks (Hopfield, 1982, Rumelhart et al., 1986). Such models have been largely employed by the computer vision community on a wide variety of applications like tracking (Milan et al., 2016), action recognition (Singh et al., 2016), *etc*. Long Short Term Memory Networks (LSTMs) (Hochreiter, Schmidhuber, 1997) are also widely adopted for such tasks (Byeon et al., 2015, Ehsani et al., 2018) since they moderate the vanishing gradient problem (Hochreiter, 1998) when dealing with long-term dependencies. As far as remote sensing is concerned, (Mou et al., 2019) incorporates a recurrent network on top of a convolutional architecture combining in this way spectral, spatial and temporal information in an end-to-end framework. Moreover, fully convolutional LSTMs have been utilized in (Papadomanolaki et al., 2019) where recurrent blocks are integrated into every encoding level of a UNet-like architecture (Ronneberger et al., 2015), thus capturing temporal relationships at different resolutions in a fully convolutional manner. That way pixel level maps of changed areas can be provided and analysed.

In this paper, we tackle the problem of building change detection for very high resolution satellite images by further evolving the already existing framework in (Papadomanolaki et al., 2019). More specifically, the proposed architecture is enriched with

* mar.papadomanolaki@gmail.com

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-2-2020-541-2020 | © Authors 2020. CC BY 4.0 License.

an additional decoding branch that performs building semantic segmentation, providing the network with ancillary feature attributes during the training process. The attained quantitative and qualitative results indicate the great potential of the suggested scheme which is also compared with state of the art fully convolutional approaches, namely fully convolutional Early Fusion (FC-EF), Siamese Concatenation (FC-Siam-Conc) and Siamese Difference (FC-Siam-Diff) (Caye Daudt et al., 2018a).

The remaining of the paper is organized as follows. In Section 2, we describe the proposed fully convolutional, multi-task framework as well as the employed dataset. In Section 3 we present and discuss the experimental results while in Section 4 we make a conclusion and examine potential future directions.

2. METHODOLOGY

2.1 Multi-task L-UNet

The proposed scheme is based on the widely known UNet architecture (Ronneberger et al., 2015) consisting of one encoding branch with n levels and two decoding branches. Firstly, D input image volumes in the form of (*Batchsize* x *Channels* x *Height* x *Width*) are passed to the model, where D stands for the employed number of dates. Each of the D images is processed separately by the encoding layers using shared convolutional weights. More specifically, every encoding level E_i with $i \in \{1, 2, .., n\}$ produces spatial feature vectors X_i^t for $t \in \{1, 2, .., D\}$. These feature vectors are then fed to a Long Short Term Memory (LSTM) block which is added as a skip connection on top of every encoding level, determining the temporal attributes using a gating mechanism (Hochreiter, Schmidhuber, 1997). Here, instead of multiplying the spatial feature vectors X_i^t with high dimensional weight matrices, we perform convolution operations in order to calculate the hidden and cell states. In this way, any gating process is configured as

$$G_i^t = \Phi(W_{G_i^t} * (X_i^t, H_i^{t-1})), \tag{1}$$

where G_i^t is the general form for each of the *forget (f_i^t), input (in_i^t), output (o_i^t)* or *cell (c_i^t)* gates at time step t of encoding level i, Φ is the activation function and $W_{G_i^t}$ is a single strided convolutional layer with 3×3 kernels and padding equal to 1. Next, cell state C_i^t is obtained as

$$C_i^t = f_i^t \cdot C_i^{t-1} + in_i^t \cdot c_i^t, \tag{2}$$

where f_i^t is the forget gate, in_i^t is the input gate and c_i^t is the cell gate at time step t of encoding level i. Finally, hidden state H_i^t is calculated as

$$H_i^t = o_i^t \cdot tanh(C_i^t), \tag{3}$$

where o_i^t is the output gate.

After the last encoding level E_n, hidden state H_n^D is upsampled by the corresponding decoding level and concatenated with the information stored in H_{n-1}^D. This upsampling procedure continues in the same way until the last decoding level where the feature vectors are back to their original dimensions.

For a better comprehension, the left part of Figure 1 illustrates the previously described approach for the case of $D = 5$. In every encoding level there are two sets of convolutional, batch normalization and rectified linear unit layer (*Conv-BN-ReLU*) successions with a convolutional LSTM block on top of them. At the first encoding level the input image planes are increased to 16, while in the following ones the depth planes rise to twice their size reaching in this way the number of 256 at the end of the last encoding level. After that, the decoding levels decrease the planes from 256 to 128, 64, 32, 16 and finally the probabilities are produced after a $log\ softmax$ layer. All the convolutional layers adopt 3×3 filter kernels with stride and padding equal to 1, while the pooling layers reduce the spatial resolution of the images by 2.

The proposed scheme is further enriched with an additional decoding branch, depicted on the right part of Figure 1, which performs *building* detection for the input dates. This time skip connections involve concatenations not with temporal, but with spatial feature vectors that have resulted from the different encoding levels. The building segmentation maps can be produced for all the input dates or for some of them depending on the application and the computational complexity allowed. For this implementation, we decided to train the network using only the first and last, out of the multiple date inputs.

During the training process of the previously described multi-task learning framework we define different loss quantities using cross entropy for the optimization of each one of them

$$Loss_{CE} = -\sum_{l=0}^{1} y_{s,l} log(p_{s,l}), \tag{4}$$

where $y_{s,l}$ is a binary indicator that shows if class l is the correct answer for observation s, while $p_{s,l}$ holds the probability that observation s belongs to class l.

Five different loss values are involved in our training scheme, which are also combined together in a circular way so as to achieve better performances. In particular, we use cross entropy loss $Loss_{ch}$ for the building change detection map, as well as $Loss_{seg}^1$ and $Loss_{seg}^D$ for the building semantic maps. For this study, we provide the building segmentation maps for the first and last date only. Additionally, cross entropy is used to provide one more loss, $Loss_{ch2}$, that focuses on the building change detection by mixing together the final feature outputs resulting from the building segmentation branches. If s denotes the final network output, then $Loss_{ch2}$ is defined by calculating the cross entropy for feature vector $s_{ch} = s_{seg}^D - s_{seg}^1$, where s_{seg}^D and s_{seg}^1 are the output feature vectors resulting from the building segmentation branch for the first and last date respectively. Moreover, we also establish $Loss_{seg2}^D$ for the building segmentation map of the last date, which is computed using the feature vector $s_{seg}^D = s_{seg}^1 + s_{ch}$, namely the addition of feature outputs resulting from the building map of the first date and the building change map. These two additional loss functions are integrated to reduce the number of false positive values in the building change detection map, combining the predicted probabilities of the network's both decoding branches in a circular way.

For the final optimization of the network we use the weighted sum of all the previous losses choosing a weight equal to 0.6 for the building change detection branch due to the limited number of changed pixels in the dataset. In all the rest employed losses we assign a weight equal to 0.1.

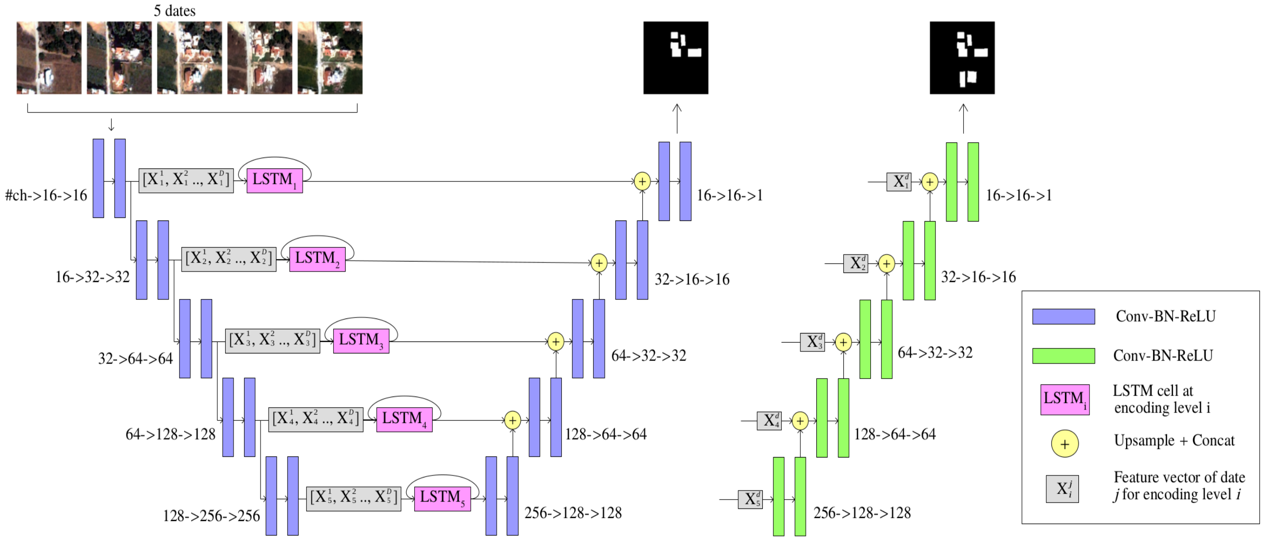

Figure 1. Graphical illustration of the proposed multi-task deep architecture. The fully convolutional LSTM network depicted on the left is further enriched with the additional building segmentation decoding branch depicted on the right. Every encoding level i results in five spatial feature vectors which are fed to the LSTM cell so that the hidden state can be calculated and concatenated with the corresponding decoding level. Spatial feature vectors of certain dates d, in our case $d \in \{1, D\}$, are also concatenated with the corresponding feature vectors of the building segmentation decoding branch.

2.2 Dataset and Implementation Details

All the experiments were based on the Attica VHR dataset, which involves 5 multispectral very high resolution images illustrating a 9 km^2 region in the East Prefecture of Attica, Greece, for five different years. In particular, there are images acquired in 2006, 2007 and 2009 which were captured by Quickbird satellite, while there are also images for the years of 2010 and 2011, which were captured by WorldView-2. Every sample is pansharpened and atmospherically corrected, with the available groundtruth of both *buildings* and *change* of buildings having been manually annotated by remote sensing experts after an attentive and time demanding photo-interpretation. Also, Quickbird images were resized to the WorldView-2 resolution which is approximately 8000 by 7000 pixels. It should be mentioned here that all experiments were conducted using the four similar spectral bands of both sensors; Red, Green, Blue and Near Infra-Red.

The whole region was divided into 36 equal subregions of approximate size 1100 by 1300 pixels; 28 of them were used for training, 4 for validation and 4 for testing. We tried to split the dataset parts as wisely as possible so that there is sufficient information during the training process as well as adequate testing features in order to draw reliable conclusions. For the training process, patches of size 64x64 were produced with a stride of either 32 in case *building change* pixels were included, or 64 in case the patch did not include any *building change* pixels at all. This strategy was applied as a data augmentation approach to enrich the *building change* semantic category since it is extremely scarce compared to the *no change* one. In addition, patches whose number of *building change* pixels exceeded the threshold of 3% were randomly flipped in all possible angles proportional to 90 degrees while their brightness, contrast and saturation levels were also randomly altered. Lastly, each class was associated with a weight inversely proportional to the total pixel number included in it.

As far as hyperparameters are concerned, Adam optimizer was picked with a learning rate equal to 10^{-4} and a batchsize of 10. The dataset split seemed to work properly since the training was conducted successfully without overfitting signs. Early stopping criteria were employed for every adopted approach in order to cease the training process and pick the optimal network weights. The applied methods needed less than 50 epochs to converge, while all experiments were implemented using the PyTorch deep learning library (Paszke et al., 2017) on a single NVIDIA GeForce GTX TITAN with 12 GB of GPU memory.

3. EXPERIMENTAL RESULTS AND DISCUSSION

In this section we present the experimental results along with a comparative study.

3.1 Quantitative Evaluation

For the quantitative evaluation, precision, recall, F1 score and balanced accuracy metrics have been employed.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{7}$$

$$BA = \frac{\frac{TP}{TP+FN} + \frac{TN}{TN+FP}}{2} \tag{8}$$

They are all expressed through the calculated TP (True Positives), FP (False Positives), TN (True Negatives) and FN (False Negatives). If we have a class l, then TP is the number of pixels that have been correctly classified as l. FP is the number of

Methods	Dates	Building Change Detection				Building Semantic Segmentation				Time ($mins$)
		Precision	Recall	F1	BA	Precision	Recall	F1	BA	
(Caye Daudt et al., 2018a) *FC-Siam-Conc*	2	42.47	56.52	48.49	78.00	-	-	-	-	≈ 1
	2	44.70	59.07	50.89	79.28	74.41	65.92	69.91	82.36	≈ 1.5
	5	46.62	59.03	52.09	79.28	-	-	-	-	≈ 2
	5	***50.23***	57.68	***53.70***	78.65	**78.78**	57.12	66.22	78.15	≈ 2.5
(Caye Daudt et al., 2018a) *FC-EF*	2	41.10	55.53	47.24	77.49	-	-	-	-	≈ 1
	5	43.85	52.95	47.97	76.24	-	-	-	-	≈ 2
(Caye Daudt et al., 2018a) *FC-Siam-Diff*	2	45.67	56.80	50.63	78.17	-	-	-	-	≈ 1
	2	44.09	**62.11**	51.57	**80.79**	75.90	63.50	69.15	81.21	≈ 1.5
	5	41.45	40.59	41.02	70.10	-	-	-	-	≈ 2
	5	41.87	48.92	45.12	74.23	73.79	***66.81***	***70.13***	***82.78***	≈ 2.5
Proposed *L-UNet*	2	47.25	55.21	50.92	77.39	-	-	-	-	≈ 2
	2	44.53	***61.39***	51.62	***80.44***	67.38	**75.54**	**71.23**	**86.80**	≈ 3
	5	47.96	60.19	53.38	79.87	-	-	-	-	≈ 4
	5	**52.42**	59.68	**55.82**	79.65	***76.08***	61.52	68.03	80.24	≈ 5

Table 1. Quantitative evaluation of the proposed framework for the testing part of Attica VHR dataset. Precision, recall and F1 rates are associated to the *building change* class as well as the *building* class for the image of 2006, while Balanced Accuracy (BA) is also provided. All the rows demonstrate results using the RGB-NIR bands with the last column indicating the computational time needed for each method to complete one training epoch. With bold we indicate the best performance and with bold and italic the second best performance per evalaution metric.

pixels that have been wrongly classified as l. TN is the number of pixels that have been rightly recognized as not belonging to l. Finally, FN represents the pixels that belong to l but the model has associated them to some other class.

In Table 1 we provide the quantitative outcomes of the proposed method *with* and *without* semantic segmentation of *buildings*. We also compare them with all the methods described in (Caye Daudt et al., 2018a). The estimation of the accuracy metrics was carried out on the testing part of the Attica VHR dataset after a post processing phase where objects with areas smaller than 150 pixels were discarded.

As one can observe, the integration of the *building* semantic segmentation decoding branch boosts the F1 score in all approaches. Starting with the FC-Siam-Conc method, F1 rate has increased by 1.6% in the case of 5 dates, while precision has also raised by 3.6% which indicates that the multi-task learning process contributes much to the lessening of false positive pixels. On the contrary, for the FC-Siam-Diff method, the precision rates remain very low not only in the 2 dates case but also in the 5 dates case, with or without multi-task learning. Regarding FC-EF, the numerical results are slightly better in the 5 dates case, although F1 score does not exceed the level of 48% in neither of the two experiments. It should be mentioned here that for the FC-EF method we could not perform the task of *building* semantic segmentation simultaneously since the different dates are fused along the channel dimension before being passed through the model, preventing in this way the construction of separate spatial feature vectors for each individual date. As far as the proposed framework is concerned, it appears that it yields the most successful results regarding false positive pixels since the precision rate is 52.42% in the case of 5 dates, exceeding the next best precision score of multi-task FC-Siam-Conc by approximately 2.2%. In addition, the F1 score reaches the value of 55.82% which is also 2.2% higher than the corresponding F1 rate in the multi-task FC-Siam-Conc case. For the L-UNet approach, we notice that the F1 rate is always above 50% which means that when temporal dynamics are calculated, the attained total number of false positive and false negative pixels seems to be more balanced. Finally, the highest balanced accuracy score was established by the multi-task FC-Siam-Diff method of 2 dates, where the highest recall rate has also been achieved. Nevertheless, the precision rate is particularly low which means that even though false negative pixels are more limited, false positive pixels continue to exist. As a whole, FC-Siam-Conc as well as L-UNet approaches achieve almost the same balanced accuracy rates with precision and F1 scores outperforming the FC-Siam-Diff cases. Regarding building semantic segmentation, the provided accuracy metrics are related to the year of 2006, with multi-task L-UNet with 2 dates reporting the best performances. Lastly, in the last column of Table 1 we provide the approximate computational time needed by the different employed methods to complete one training epoch. L-UNet requires twice the time needed in all the methods presented in (Caye Daudt et al., 2018a), while time demands increase further when building semantic segmentation is integrated.

In general, one can notice that in all applied methods accuracy results demonstrate that the networks are prone to many errors, especially if we consider that precision rates for the *building change* category never go above 53%. This is probably caused by two main problems that exist in change detection datasets. The first one is related to registration and parallax errors that disorientate the network's learning process. Secondly, the appearance of certain building roofs may be differentiated through time, resulting in a variation of provided spectral values for certain areas that do not really undergo any semantic *change* on the buildings. One critical issue also lies in the fact that the *building change* semantic category is greatly scarce compared to the *no change* one. In the case of Attica VHR dataset, the total number of *no change* pixels for the training dataset is almost 84 times larger than the number of *change* ones. Nevertheless, despite these obstacles the extraction of temporal features seems to handle the available information in a better and more constructive way, especially when it is coupled simultaneously with *building* semantic segmentation.

3.2 Qualitative Evaluation

In Figure 2, we provide predictions that resulted from some of the investigated methods on the Attica VHR testing regions for the *building change* detection task. Green colour stands for true positive pixels, black for true negatives, red for false positives and yellow for false negatives. In the first row, the additional building of 2011 is detected successfully only by the multi-task L-UNet with 5 dates. Multi-task FC-Siam-Conc with 5 dates has only partly detected the building change, whereas multi-task FC-Siam-Diff with 2 dates and FC-EF with 5 dates have failed completely to recognize it. In the second row, all methods seem to have identified the building changes, with multi-task L-Unet with 5 dates having attained the lowest number of

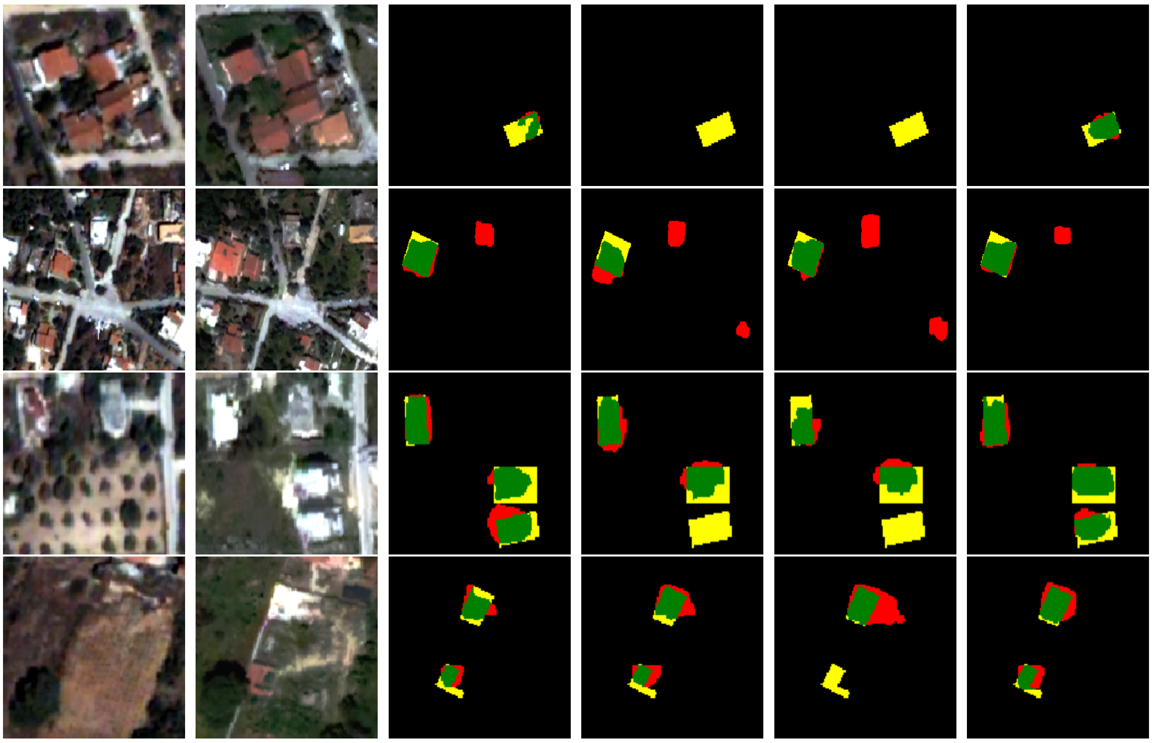

Figure 2. Qualitative evaluation on zoomed regions of the Attica VHR testing areas for the *building change* detection task. 1st column: RGB images of 2006, 2nd column: RGB images of 2011, 3rd column: multi-task FC-Siam-Conc with 5 dates, 4th column: multi-task FC-Siam-Diff with 2 dates, 5th column: FC-EF with 5 dates, 6th column: multi-task L-UNet with 5 dates [*Green*: True Positives, *Black*: True Negatives, *Red*: False Positives, *Yellow*: False Negatives]

false positive pixels. Regarding the third row, multi-task FC-Siam-Diff with 2 dates and FC-EF with 5 dates have failed to identify certain building changes while finally, in the last row all methods have detected the building changes adequately except FC-EF with 5 dates which also includes the largest number of false positive pixels.

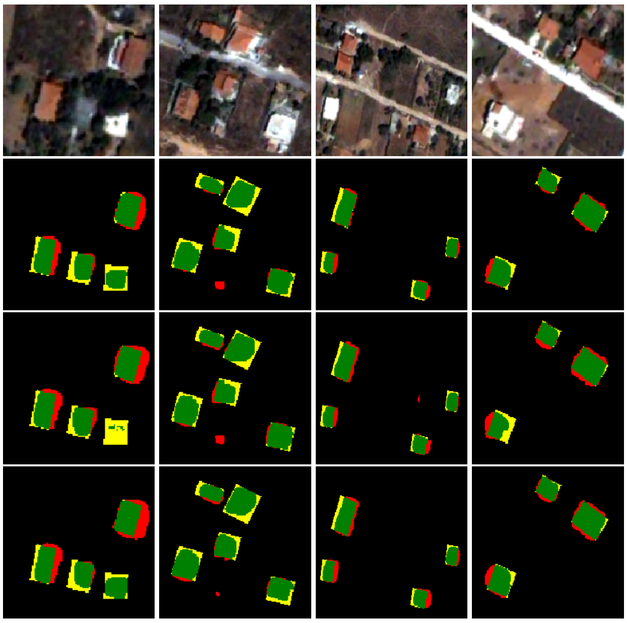

Figure 3. Building predictions on testing areas of Attica VHR dataset for the year of 2006. 1st row: RGB images, 2nd row: multi-task FC-Siam-Conc with 5 dates, 3rd row: multi-task FC-Siam-Diff with 2 dates, 4th row: multi-task L-UNet with 5 dates. [*Green*: True Positives, *Black*: True Negatives, *Red*: False Positives, *Yellow*: False Negatives]

In Figure 3 there are also several illustrations on *building* predictions from Attica VHR testing regions of 2006. With a closer look, one can observe that all multi-task methods appear to handle the building semantic segmentation problem in a similar manner. In every approach however it is visible that the building boundaries are not very precise most of the times. This problem can also be observed in Figure 4 where all network outcomes resulting from the proposed multi-task scheme are demonstrated, for a region of Attica VHR testing part.

4. CONCLUSION

Throughout this work, we have evaluated and compared a fully convolutional multi-task deep architecture which takes advantage of temporal dynamics as well as building footprint features among the different dates in order to deal with the building change detection problem. Results show that the exploitation of temporal dynamics alone can boost the model's performance compared to other state of the art architectures which are based exclusively on spatial feature representations. Accuracy metrics are even further ameliorated when the task of building semantic segmentation is performed simultaneously for the first and last date of the dataset. Still, urban change detection remains a remarkably challenging problem since the *building change* semantic category is extremely scarce compared to the *no change* one. Apart from that, spectral rooftop alterations and registration errors tend to disorientate greatly the networks during the training process, resulting in a large number of false positive pixels. In the future, we aim to investigate further the possible combinations of multi-task deep frameworks as well as tackle

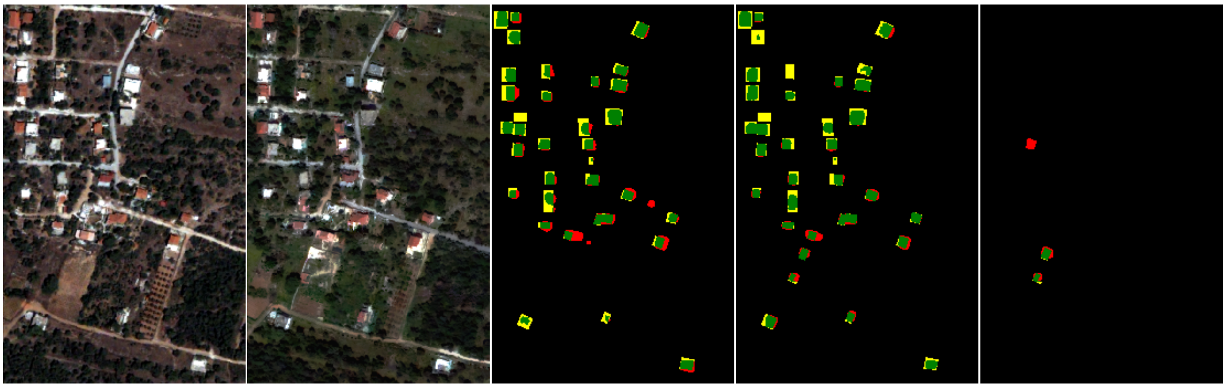

Figure 4. Qualitative results of multi-task L-UNet with 5 dates, for a region of Attica VHR testing part. From left to right: RGB image of 2006, RGB image of 2011, *building* predictions of 2006, *building* predictions of 2011, *building change* predictions. [*Green*: True Positives, *Black*: True Negatives, *Red*: False Positives, *Yellow*: False Negatives]

the issue of imprecise boundaries in an attempt to produce even more accurate segmentation maps. In addition, we plan to explore if the trained models can generalize well when tested on other very high resolution datasets depicting cities with different city infrastructures.

REFERENCES

Benedek, C., Shadaydeh, M., Kato, Z., Szirányi, T., Zerubia, J., 2015. Multilayer Markov Random Field Models for Change Detection in Optical Remote Sensing Images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 107, 22-37. https://hal.inria.fr/hal-01116609.

Byeon, W., Breuel, T. M., Raue, F., Liwicki, M., 2015. Scene labeling with lstm recurrent neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3547–3555.

Caye Daudt, R., Le Saux, B., Boulch, A., 2018a. Fully convolutional siamese networks for change detection. *IEEE International Conference on Image Processing (ICIP)*.

Caye Daudt, R., Le Saux, B., Boulch, A., Gousseau, Y., 2018b. Urban change detection for multispectral earth observation using convolutional neural networks. *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*.

Daudt, R. C., Saux, B. L., Boulch, A., Gousseau, Y., 2018. Multitask learning for large-scale semantic change detection. *Computer Vision and Image Understanding*, 187.

Deng, J., Wang, K., H. Deng, Y., J. Qi, G., 2008. PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data. *International Journal of Remote Sensing - INT J REMOTE SENS*, 29, 4823-4838.

Ehsani, K., Bagherinezhad, H., Redmon, J., Mottaghi, R., Farhadi, A., 2018. Who Let the Dogs Out? Modeling Dog Behavior from Visual Data. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4051-4060.

Hochreiter, S., 1998. The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6, 107-116.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. *Neural computation*, 9(8).

Hopfield, J. J., 1982. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, 79 8, 2554-8.

Karantzalos, K., 2015. Recent advances on 2d and 3d change detection in urban environments from remote sensing data.

Li, X., Yeh, A. G.-O., 1998. Principal component analysis of stacked multi-temporal images for the monitoring of rapid urban expansion in the pearl river delta.

Liu, Y., Pang, C., Zhan, Z., Zhang, X., Yang, X., 2019. Building Change Detection for Remote Sensing Images Using a Dual Task Constrained Deep Siamese Convolutional Network Model. *ArXiv*, abs/1909.07726.

Milan, A., Rezatofighi, S. H., Dick, A. R., Reid, I. D., Schindler, K., 2016. Online multi-target tracking using recurrent neural networks. *AAAI*.

Mou, L., Bruzzone, L., xiang Zhu, X., 2019. Learning Spectral-Spatial-Temporal Features via a Recurrent Convolutional Neural Network for Change Detection in Multispectral Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 57, 924-935.

Papadomanolaki, M. G., Verma, S., Vakalopoulou, M., Gupta, S., Karantzalos, K., 2019. Detecting Urban Changes with Recurrent Neural Networks from Multitemporal Sentinel-2 Data. *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 214-217.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in PyTorch.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. N. Navab, J. Hornegger, W. M. Wells, A. F. Frangi (eds), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Springer International Publishing, Cham, 234–241.

Rumelhart, D. E., Hinton, G. E., Williams, R. J., 1986. Learning representations by back-propagating errors. *Nature*, 323, 533-536.

Singh, B., Marks, T. K., Jones, M. J., Tuzel, O., Shao, M., 2016. A Multi-stream Bi-directional Recurrent Neural Network for Fine-Grained Action Detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1961-1970.

Singh, P., Kato, Z., Zerubia, J., 2014. A multilayer Markovian model for change detection in aerial image pairs with large time differences. IAPR, IEEE, Stockholm, Sweden. Accepted.

Vakalopoulou, M., Karantzalos, K., Komodakis, N., Paragios, N., 2015. Simultaneous registration and change detection in multitemporal, very high resolution remote sensing data. 61–69.

Vakalopoulou, M., Karantzalos, K., Komodakis, N., Paragios, N., 2016. Graph-Based Registration, Change Detection, and Classification in Very High Resolution Multitemporal Remote Sensing Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9, 1-12.

Volpi, M., Tuia, D., Camps-Valls, G., Kanevski, M. F., 2012. Unsupervised Change Detection With Kernels. *IEEE Geoscience and Remote Sensing Letters*, 9, 1026-1030.