A Deep Multitask Learning Framework Coupling Semantic Segmentation and Fully Convolutional LSTM Networks for Urban Change Detection

Maria Papadomanolaki, Maria Vakalopoulou, and Konstantinos Karantzalos, *Senior Member, IEEE*

***Abstract*— In this article, we present a deep multitask learning framework able to couple semantic segmentation and change detection using fully convolutional long short-term memory (LSTM) networks. In particular, we present a UNet-like architecture (L-UNet) that models the temporal relationship of spatial feature representations using integrated fully convolutional LSTM blocks on top of every encoding level. In this way, the network is able to capture the temporal relationship of spatial feature vectors in all encoding levels without the need to downsample or flatten them, forming an end-to-end trainable framework. Moreover, we further enrich the L-UNet architecture with an additional decoding branch that performs semantic segmentation on the available semantic categories that are presented in the different input dates, forming a multitask framework. Different loss quantities are also defined and combined together in a circular way to boost the overall performance. The developed methodology has been evaluated on three different data sets, i.e., the challenging bitemporal high-resolution Office National d'Etudes et de Recherches Aérospatiales (ONERA) Satellite Change Detection (OSCD) Sentinel-2 data set, the very high-resolution (VHR) multitemporal data set of the East Prefecture of Attica, Greece, and finally, the multitemporal VHR SpaceNet7 data set. Promising quantitative and qualitative results demonstrated that the synergy among the tasks can boost up the achieved performances. In particular, the proposed multitask framework contributed to a significant decrease in false-positive detections, with the F1 rate outperforming other state-of-the-art methods by at least 2.1% and 4.9% in the Attica VHR and SpaceNet7 data set cases, respectively. Our models and code can be found at https://github.com/mpapadomanolaki/multi-task-L-UNet.**

***Index Terms*— Change detection, deep learning, long short-term memory networks (LSTMs), multitemporal, remote sensing, satellite, urban.**

I. Introduction

THE diversity, volume, and frequency of accessible satellite data have contributed decisively to numerous studies focusing on monitoring our environment based on multitemporal remote observations. Man-made and natural phenomena keep transforming the planet's structure, thus creating the need for effective monitoring methods. Urban growth is one of the most critical categories as the world's population keeps expanding at extremely fast rates occupying more and more of the earth's surface. The continuous spread of both residential and commercial areas has resulted in several problems, such as the diminishing of rural zones, the destruction of wildlife, and increased levels of land, water, and air contamination. For that reason, the systematic observation of urban sprawl, at high and very high spatial resolutions, becomes essential in order to fully comprehend future tendencies, take precautions, and design more appropriate city infrastructures.

Indeed, identifying changes between satellite image pairs has been an active field of research for a very long time [1]–[4] and for a wide variety of applications [5]–[8]. First, differences among remote sensing data were recognized mainly with manual, time-consuming approaches. Today, a diverse range of supervised and unsupervised change detection methods exist in the literature, such as Markov Random Fields [9], [10], kernels [11], graphical models [12], [13], and principal component analysis [14], [15]. Determining the exact timing of the change based on time-series data has also been an active field of research [16]–[18]. In addition, with the advances of machine learning during the last years, more and more techniques based on neural networks are emerging [19]–[25] aiming to create robust systems that can successfully tackle the change detection problem. Among the machine learning approaches, deep learning architectures are the ones that have captured most of the attention due to state of the art on numerous computer vision applications [26]–[29], including remote sensing [30]–[37]. Even though successful results have been achieved on the remote sensing domain, the further development of deep neural networks is still hindered due to insufficient data sets lacking multimodal diverse information. Several supervised and unsupervised frameworks have been proposed; however, the construction of robust networks is still an active research area, especially for models that fully exploit multitemporal information. Considering these, one can realize that much progress remains to be made until deep networks can account for fully operational and reliable tools for the remote sensing applications [38]. Despite these obstacles, research can still be conducted with the available resources, enriching the existing knowledge on several topics, such as semantic segmentation and change detection.

Manuscript received July 2, 2020; revised November 22, 2020 and December 22, 2020; accepted January 16, 2021. Date of publication February 10, 2021; date of current version August 30, 2021. This work was supported by the Research Committee of the National Technical University of Athens through Scholarship Grant. *(Corresponding author: Maria Papadomanolaki.)*

Maria Papadomanolaki is with Remote Sensing Laboratory, National Technical University of Athens, 15780 Zografou, Greece and also with the MICS Laboratory, CentraleSupélec, Université Paris-Saclay, 91190 Gif-sur-Yvette, France (e-mail: mar.papadomanolaki@gmail.com).

Maria Vakalopoulou is with the MICS Laboratory, CentraleSupélec, Université Paris-Saclay, 91190 Gif-sur-Yvette, France and also with Inria Saclay, 91120 Gif-sur-Yvette, France.

Konstantinos Karantzalos is with Remote Sensing Laboratory, National Technical University of Athens, 15780 Zografou, Greece.

Digital Object Identifier 10.1109/TGRS.2021.3055584

Change detection is, in most cases, associated with sequential data, making it necessary to evaluate temporal dynamics. Modeling the temporal relationship among features has been largely addressed by the computer vision community using recurrent neural networks [39], [40] that have proven to be very powerful for a wide range of applications, such as tracking [41] and action recognition [42]. Long short-term memory networks (LSTMs) [43] are also really effective for such problems [44], [45] since they moderate the vanishing gradient problem [46] when dealing with long-term dependencies. The combination of recurrent networks and deep learning architectures has also been adopted for time series tasks [47]–[49] in an attempt to produce more fruitful feature representations by extracting both spatial information and temporal information during the learning process.

Recent remote sensing studies have considerably integrated deep learning techniques toward more effective change detection approaches. In [50], the authors propose a patch-based framework examining two different architectures (Siamese and early fusion) based on the Office National d'Etudes et de Recherches Aérospatiales (ONERA) Satellite Change Detection (OSCD) bitemporal data set. In the Siamese case, the two bitemporal patches are processed by two distinct but identical branches of convolutional layers with shared weights. Then, the produced feature vectors are concatenated and fed to a series of fully connected layers. Regarding the early fusion case, the bitemporal patch pairs are concatenated along the channel dimension before being passed as a single input to several convolutional and fully connected layers.

In [51], the aforementioned network designs evolve into fully convolutional versions according to a U-Net like framework. More specifically, the fully convolutional early fusion (FC-EF) structure downsamples the concatenated bitemporal patch pair through the encoder, while the decoder upsamples it back to its original dimensions using also skip connections to enrich the feature attributes. For the fully convolutional Siamese concatenation case (FC-Siam-Conc), the network is comprised of two separate encoding branches with shared weights that receive as input the bitemporal pairs. In this approach, skip connections link the decoding steps with the concatenation of the two encoding parts' outputs. Finally, the third proposed architecture also consists of two encoding branches; only this time, skip connections associate the decoding parts with the absolute difference of feature vectors that result from the corresponding encoding parts (FC-Siam-Diff).

In [52], a recurrent network (ReCNN) is integrated into a convolutional architecture taking advantage of both spatial and temporal features under an end-to-end framework. Giving some more details, 5 × 5 patch pairs taken from corresponding pixels of bitemporal images are processed by a succession of parallel but identical dilated convolutional layers. Next, the produced pair of feature vectors are passed through a recurrent neural network, which calculates the temporal dependence between them. In the end, fully connected layers collect the temporal volume and decide if a change has occurred. Apart from change detection purposes, sequential satellite images have also been exploited for land-cover classification as in [53], where multitemporal Sentinel-2 agricultural parcels are transformed to unordered sets of pixels. Each set is passed through a pixel-set encoder resulting in a feature descriptor, which is then processed by a temporal attention network [54]. Land-cover classification purposes are also handled in [55] using convolutional recurrent layers that also mitigate the problem of cloud coverage [56].

Multitask learning schemes [57], [58] have also been adopted when dealing with the change detection problem since complementary assignments can provide the models with useful information during the training process, enhancing in this way the performances. In [59], urban change detection is coupled with the task of semantic segmentation on buildings using a fully convolutional siamese network, while a focal loss [60] is also utilized in order to ease the class imbalance problem. Daudt *et al.* [61] also employed the multitask learning approach, enhancing the architecture's ability to identify changes by performing simultaneously the task of land-cover mapping. Here, the optimal results are derived when the network is optimized in two phases: first, the training process is focused on the identification of the different land-cover semantic categories; second, the network is trained again for change detection using the land-cover semantic segmentation weights as initialization. Finally, Yang *et al.* [62] combined multitask learning with transfer learning to balance the distributions of labeled and nonlabeled data. Specifically, an encoder–decoder network performs change detection on the bitemporal labeled input images, while the difference of unlabeled data is concurrently reconstructed by the network enriching the extracted features during the training procedure. After pretraining, fine-tuning methods are exploited for unsupervised training on the unlabeled data according to region-based and boundary-based strategies. Although fully convolutional models have resulted in very promising results regarding the semantic segmentation task [63]–[65], little effort has been made to adjust such frameworks for change detection related topics. Especially, for LSTMs, the processing of multidimensional matrices remains a very challenging problem since, in most cases, satellite images need to be flattened in order to be imported to such networks.

In order to tackle the aforementioned challenges, we have designed, implemented, and validated a deep multitask learning framework able to couple semantic segmentation with fully convolutional LSTM networks for urban change detection applications. Regarding the fully convolutional LSTM structure, it has been designed by replacing the gating mechanisms with convolutional layers. Our main goal here was to combine spectral information and spatial information while taking advantage of the temporal relationship among the feature matrices avoiding the computationally expensive task of multiplying high-dimensional feature vectors. The fully convolutional LSTM blocks are placed on top of each encoding level of a UNet-like deep architecture, capturing in this way temporal information for all the different resolution levels. This current study is an extension of our previous work [66] in which fully convolutional LSTMs were utilized to semantically segment the OSCD data set. Here, the novel framework is further enriched by adding dropout layers to the hidden states of the LSTM blocks. In addition, an extra decoding branch is explored for the semantic segmentation of the available categories, providing the network with fruitful supplementary feature attributes during the training procedure. An ensemble of losses combined in a circular way is also employed for the optimization process. To sum up, this article makes the following contributions.

1) A UNet-like architecture (L-UNet) is proposed, which models the temporal relationship of spatial feature representations using integrated fully convolutional LSTM blocks on top of every encoding level. Each LSTM block operates on the given sequential input by defining the weights and biases of the gating mechanisms as convolutional layers, thus avoiding the multiplication of high-dimensional matrices. In this way, the network is able to capture the temporal relationship of spatial feature vectors in all encoding levels without the need to downsample or flatten them, creating an end-to-end trainable framework.
2) The L-UNet architecture has been further enriched with an additional decoding branch that performs semantic segmentation on the semantic categories that are presented in the different available input dates (multitask L-UNet). Under this multitask framework, different loss quantities are also defined and combined together in a circular way to boost the reported accuracy of the change detection task.

The rest of this article is outlined as follows. Section II introduces the methodology along with the implementation details, the benchmark data sets, and the employed quantitative metrics. Section III presents the experimental results and the qualitative and quantitative assessment, the comparison with the state of the art, and the performance of the different components. Finally, Section IV concludes this article.

II. Materials and Methods

A. Recurrent Neural Networks

Recurrent neural networks are commonly employed for problems that include time-dependent data as they are able to capture the temporal relationship among sequential features. In their simplest structure, such networks process data that come in the form of $X = [X_1, X_2, \ldots, X_T]$, where $X \in \mathbb{R}^N$ is a list containing information related to $t \in [1, \ldots, T]$ different time steps. In every time step t, the respective list element X_t is multiplied elementwise with an associated weight matrix W_x. At the same time, a representation of previous list elements, also known as the "hidden state," is multiplied with a weight matrix W_h. The sum of these quantities is then passed to a hyperbolic tangent function to produce the hidden state H_t of the current time step. This chain process is described as

$$H_t = \tanh(W_x \cdot X_t + W_h \cdot H_{t-1})$$

with W_x and W_h being shared across all time steps. It should be noted here that biases are omitted for convenience reasons. Since the weights are shared across all time steps, the structures of such a network are very much likely to suffer from the vanishing gradient problem [46]. More specifically, during backpropagation, the contributions of each time step are summed up to the gradient according to the chain rule of differentiation for composite functions. This results in a recursive derivative based on multiplicative dynamics that tends to zero if the gradients become very small or if there are several time steps [67]. LSTMs, first proposed by Hochreiter and Schmidhuber [43], mitigate this problem by introducing a memory cell, most commonly known as the "cell state," which exploits gating functions in order to filter the flowing information more efficiently. Unlike conventional RNNs that employ a single hyperbolic tangent layer at each time step, LSTMs refine the input volumes by introducing four interrelated layers known also as gates. In particular, the additional operations involve the forget (f) and input (i) gates as follows:

$$f_t = \sigma(W_f \cdot (X_t, H_{t-1}))$$
$$i_t = \sigma(W_i \cdot (X_t, H_{t-1}))$$

where σ is the sigmoid function, while W_f and W_i are weight matrices employed for each gating unit. The forget gate employs a sigmoid function in order to throw away ineffectual feature representations, while the input gate determines which information part is going to be utilized for the update of the cell state. Apart from f and i gates, every LSTM cell consists also of the cell gate (c_t) and the output gate (o_t) defined as

$$c_t = \tanh(W_c \cdot (X_t, H_{t-1}))$$
$$o_t = \sigma(W_o \cdot (X_t, H_{t-1}))$$

where $tanh$ is the hyperbolic tangent function, while W_c and W_o are the corresponding weight matrices. The cell gate utilizes the hyperbolic tangent function to regulate the data and produce possible candidate cell state values, while the output gate further filters the information determining the outcome of the network. After that, the network is ready to create the new cell state by forgetting irrelevant features from the previous cell state and keeping valuable ones for the current cell state

$$C_t = f_t \cdot C_{t-1} + i_t \cdot c_t. \quad (1)$$

Finally, the cell state is given to a hyperbolic tangent function, and it is also multiplied by the output gate result to produce the hidden state of current time step t

$$H_t = o_t \cdot \tanh(C_t). \quad (2)$$

By meticulously filtering the flowing information through the gates, dependencies are maintained, while the memory state is more properly conserved. Apart from that, backpropagation through time becomes more efficient since additive dynamics are integrated into the recursive multiplicative derivatives that take place when calculating the error gradients with regard to the recurrent weights. In this way, the vanishing gradient problem is less likely to appear.

B. L-UNet

Even if the previous formulations capture successfully temporal relations on sequential data sets, they become a bit inefficient, augmenting significantly the number of parameters, in the case of high-dimensional data, such as remote sensing images $I = [I_1, I_2, ..., I_T]$, where $I_t \in \mathbb{R}^{Ch \times Hh \times Wd}$ with

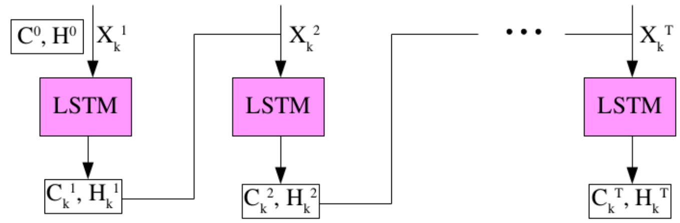

Fig. 1. Process based on which the temporal relationship among the spatial features is calculated. Cell state C_k^t, hidden state H_k^t, and input X_k^t are given as input to the LSTM at each time step t until the final hidden state H_k^T is produced, where T stands for the number of employed dates and k for the encoding level.

Ch denoting the number of spectral channels and Hh and Wd the spatial dimensions of image I_t. In such a case, we would end up with multiplications of immense multidimensional matrices, making the training process computationally expensive and hindering the model's convergence.

To deal with this predicament, weight matrices W_f, W_i, W_c, and W_o have been replaced with single strided convolutional layers comprised of 3 × 3 kernels and padding equal to one. In the proposed UNet-like framework, temporal image volumes of T different dates, each one in the form of ($Bs \times Ch \times Hh \times Wd$), with Bs denoting the batchsize, are passed to the model. Each of the I_t images is processed separately by the encoding layers using shared convolutional weights, with one LSTM network being placed on top of every encoding level. Fig. 1 shows a graphical description on how the temporal relationship among spatial features is computed after each encoding level k, with cell state and hidden state being initialized as zero matrices of shape ($Bs \times Ch \times Hh \times Wd$). Every illustrated box represents all the interior gating operations that take place inside the LSTM cell. By replacing the weight matrices with convolutional layers, each gating mechanism can now be defined as

$$G_k^t = \Phi\big(W_{G_k^t} * \big(X_k^t, H_k^{t-1}\big)\big) \qquad (3)$$

where G_k^t is the *forget*, *input*, *output*, or *cell* gate at time step t of encoding level k, Φ is an activation function, and $W_{G_k^t}$ is a convolutional layer applied on the concatenation of the current input X_k^t and the previous hidden state H_k^{t-1} along the channel dimension. The filtered information outcome is then utilized for calculating the current cell state and hidden state using 1 and 2. The final hidden state of every encoding level k is collected in order to be later concatenated with the output of the respective decoding part. In this way, the LSTM block acts as a skip connection for the L-UNet.

As far as optimization is concerned, we use a standard cross-entropy loss

$$\text{Loss}_{\text{CE}} = -\sum_{l=0}^{n} y_{s,l} \log(p_{s,l}). \qquad (4)$$

In the above expression, n is the number of classes, and $y_{s,l}$ 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. For the L-UNet case, the total loss for the optimization of the change detection task can be described as

$$\text{Loss}_{\text{ch}} = -\sum_{l=0}^{n} y_{s,l} \log(p_{\text{ch(s,l)}}) \qquad (5)$$

where $p_{\text{ch(s,l)}}$ denotes the probability that observation s belongs to class l, with l indicating each of the available *change* semantic categories. Also, the variable n here indicates the different types of change that may be available.

C. Multitask L-UNet

Apart from calculating the temporal relationship among the data, features related to the semantic segmentation of the available categories can be also utilized by further customizing the proposed scheme with an auxiliary decoding branch. This branch performs semantic segmentation for the various input dates, with skip connections concatenating the spatial feature vectors of each encoding level. For our investigation, we chose to perform the training process using only the semantic maps of the first and last dates. This selection was based on the fact that the available *change* ground truth of the data sets describes the changes that have occurred between the first and last dates. However, it is possible to use more dates or semantic categories depending on the application and the available computational resources. An overview of the proposed architecture is presented in Fig. 2, summarizing the multitask learning framework. It should be noted here that, in our experiments, the semantic segmentation task is performed on two different categories (buildings/nonbuildings) since these are the only available annotations that we have for the available dates. However, our method is modular, and it can be adjusted to any number of available semantic and change categories, as our formulation is based on multiclass cross-entropy loss for both segmentation and change detection tasks.

For the optimization of the proposed scheme, we utilize an ensemble of loss quantities based on cross-entropy; however, any other kind of loss function can be employed. Five different loss entities are used in total during the training process, which are also combined together in a circular way in an attempt to reduce false-positive (FP) detections. In particular, we use cross-entropy loss Loss_{ch}, as described in Section II-B, for the change detection task, as well as two more loss quantities for the building semantic maps of the first and last available dates

$$\text{Loss}_{\text{seg}}^t = -\sum_{l=0}^{m} y_{s,l} \log\left(p_{\text{seg}(s,l)}^t\right). \qquad (6)$$

In the above equation, $t = \{1, \ldots, T\}$ and $p_{\text{seg(s,l)}}^t$ hold the probability that observation s belongs to the l semantic category for time t. In addition, cross-entropy is employed for the definition of one more loss, Loss_{ch2}, that focuses on change detection by exploiting the features produced by the last convolutional layer of the semantic segmentation decoding branch. If we denote with F_{seg}^1 and F_{seg}^T these features for the first and last dates accordingly, then the features for the change detection can be defined as $F_{\text{ch2}} = F_{\text{seg}}^T - F_{\text{seg}}^1$. In particular, we subtract the features of the first date from the features of the last date. This way, Loss_{ch2} corresponds to a cross-entropy loss similar to 5 using the resulting F_{ch2} features. In the same

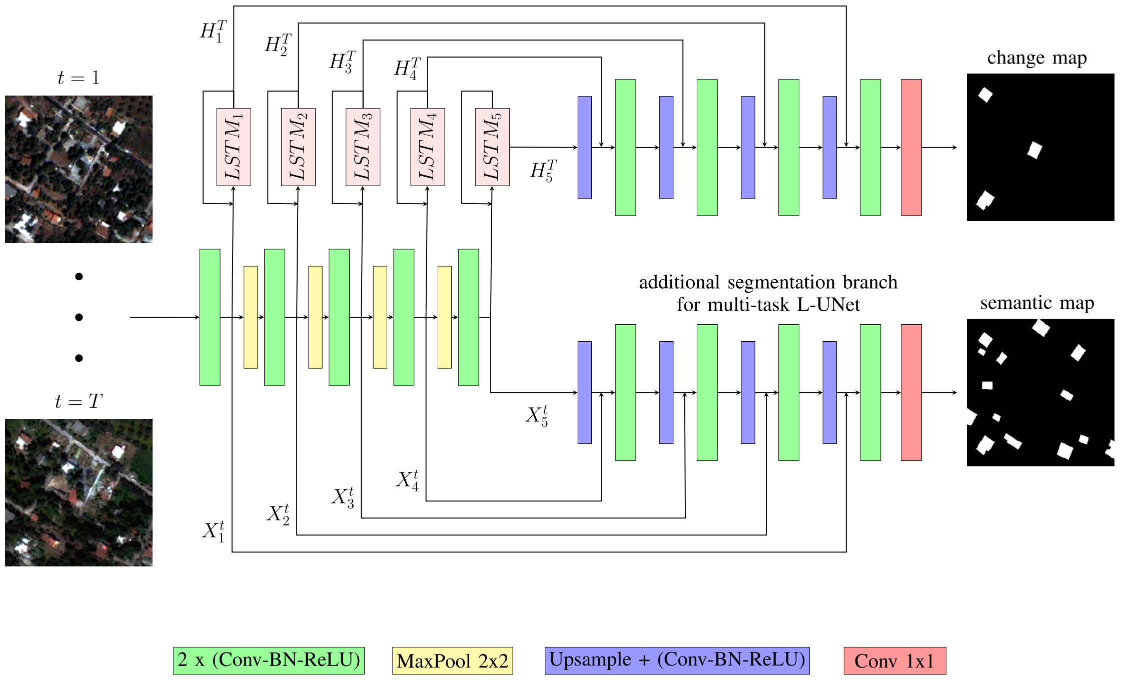

Fig. 2. Proposed multitask L-UNet architecture consisting of five encoding levels. The upper decoding branch is responsible for the *change* detection task, concatenating the final hidden states of the corresponding LSTM blocks. The lower decoding branch performs *semantic segmentation* on the available semantic labels that, for our experiments, include the building footprints for the first and last employed dates ($t = 1$ and $t = T$), each time concatenating the corresponding spatial feature vector. If the lower decoding branch is removed, then the architecture is considered a plain L-UNet, as in [66].

TABLE I

TO AUGMENT THE BITEMPORAL INFORMATION OF THE PUBLICLY AVAILABLE OSCD DATA SET AND GO BEYOND IMAGE PAIRS, WE COLLECTED AND INTEGRATED THREE ADDITIONAL INTERMEDIATE DATES. THE OBTAINED DATES ARE ILLUSTRATED FOR SOME OF THE DATA SETS CITIES. FROM TOP TO BOTTOM, THE FIRST AND LAST DATES ARE THE ALREADY PROVIDED ONES BY THE OSCD DATA SET

Abudhabi	Beirut	Chongqing	Dubai	Hong Kong	Milano	Paris	Rio
2016/01/20	2015/08/20	2017/04/14	2015/12/11	2016/09/27	2016/12/28	2016/11/30	2016/04/24
2016/09/29	2015/12/08	2017/07/23	2016/06/08	2017/01/25	2017/05/27	2017/02/15	2017/02/18
2017/03/18	2016/04/26	2017/09/16	2016/11/05	2017/04/02	2017/08/15	2017/04/09	2017/05/09
2017/09/09	2017/04/21	2018/01/14	2017/06/03	2017/10/22	2017/11/18	2017/08/29	2017/07/28
2018/03/28	2017/10/03	2018/04/02	2018/03/30	2018/03/23	2018/01/22	2017/11/07	2017/10/11

manner, we calculate $\text{Loss}_{\text{seg2}}^{T}$ by combining the final features from the semantic segmentation and change detection branches as $F_{\text{seg2}}^{T} = F_{\text{seg}}^{1} + F_{\text{ch}}$. That is, we add the features resulting from the semantic segmentation of the first date (F_{seg}^{1}) with the features resulting from the last convolutional layer of the change detection decoding branch (F_{ch}). As all these features are produced by the different branches, they fully exploit the representations that are produced by the multitask L-UNet in a circular way. For the final optimization of the network, we use the weighted sum of all these losses as

$$\text{Total\_Loss} = w_1\text{Loss}_{\text{ch}} + w_2\text{Loss}_{\text{seg}}^{1} + w_3\text{Loss}_{\text{seg}}^{T} + w_4\text{Loss}_{\text{ch2}} + w_5\text{Loss}_{\text{seg2}}^{T} \quad (7)$$

where the sum of the weights $(w_1, w_2, w_3, w_4, w_5)$ is 1.

D. Data Sets and Implementation Details

The conducted experiments were based on three multispectral data sets: the high-resolution OSCD, the very high-resolution (VHR) Attica VHR, and the very high-resolution SpaceNet7 data set. We should highlight here that, for all data sets, the change detection task is performed using two classes, namely, *change* and *no change*. As a result, in 5, n is equal to 2. For the Attica VHR and SpaceNet7 data sets, the semantic segmentation task is performed for the first and last available dates using two classes, namely *building* and *nonbuilding*. Hence, in 6, $m = 2$, and $t = \{1, T\}$. Further details for each of the data sets are provided in the next paragraphs.

1) ONERA Satellite Change Detection (OSCD): The OSCD data set [50] consists of bitemporal Sentinel-2 satellite images depicting 24 different cities around the world; 13 spectral channels are available for each image pair, while ground-truth information is related to urban change and provided for 14 cities. Our setup follows the submission system guidelines where the 14 image pairs are used for training and the rest for testing. We further enriched the OSCD data set with additional Sentinel-2 images depicting the provided cities at different times. The additional dates include Sentinel-2 images captured between the provided bitemporal dates, adapting them as much as possible to cover different periods of the season. In Table I, we present the timestamps of Sentinel-2 images used for eight

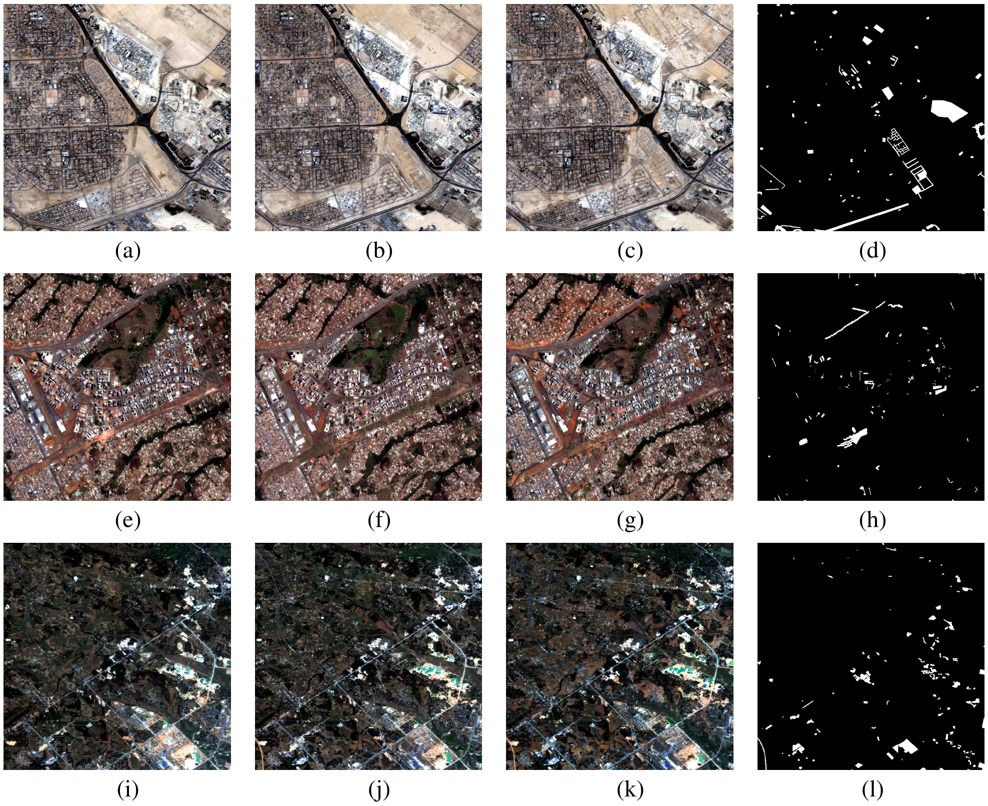

Fig. 3. Indicative images from the OSCD training data set for three different dates (in the form of dd/mm/yyyy) along with the corresponding *change* ground truth. From top to bottom, the illustrated cities are Abu Dhabi, Águas Claras, and Beihai. (a) 20/01/2016. (b) 18/03/2017. (c) 28/03/2018. (d) Ground truth. (e) 16/09/2015. (f) 18/01/2017. (g) 15/10/2017. (h) Ground truth. (i) 09/02/2016. (j) 19/12/2017. (k) 09/03/2018. (l) Ground truth.

different cities. It should be mentioned here that all extra images were coregistered according to the provided OSCD bitemporal data set. In Fig. 3, a part of the training images is illustrated for three different cities. The percentage of the *change* pixels for the OSCD collection constitutes only the 2.3% of the training data set.

As far as the training process is concerned, patches of size 32 × 32 were extracted with a stride of either 6 in case *change* pixels were included or 32 in case *no change* pixels were contained exclusively. This strategy was applied as a stratified sampling approach to enrich the training samples that involve *change* features. In addition, more data augmentation techniques mainly used by the computer vision community, namely flipping in all possible angles proportional to 90°, were implemented for patches whose number of *change* pixels exceeded the threshold of 5% for the entire patch. A total of approximately 32 000 patches containing both *change* and *no change* pixels resulted from the 14 training cities, while 8000 were intended for validation purposes. It should be mentioned here that, for the experiments, we utilized the four high-resolution channels of the Sentinel-2 satellite: red, green, blue, and near-infrared. Finally, this data set was used to evaluate the L-UNet architecture alone since semantic annotations are available only for the *change* samples. Thus, with no semantic annotations for the different available dates, the multitask L-UNet cannot be implemented.

In order to take advantage of the entire data set, our final predictions were produced by an ensemble of different trained models following a cross-validation scheme. Giving some more details, the training patches were divided into five equal parts, and the same model was optimized five times using all possible combinations of the training data set partitions. Then, predictions for the testing images were produced from all five models, with the final result being formulated by averaging the five model outcomes. It should be mentioned here that since the testing ground truth is not publicly available for this data set, the quantitative results are obtained by submitting our predictions online.[1]

2) Attica VHR: This data set includes five multispectral VHR images, illustrating a 9-km^2 region in the East Prefecture of Attica, Greece. All images were acquired by Quickbird and WorldView-2 between the years of 2006 and 2011. Specifically, the images of 2006 and 2007 were captured by Quickbird, while images of 2009, 2010, and 2011 were captured by the WorldView2 satellite. Every sample is pansharpened and atmospherically corrected, while the ground truth has been manually annotated by remote sensing experts after an attentive and time-demanding photointerpretation. For this data set, ground truth is provided for the building changes and the building footprints of every single available date. It should be mentioned here that since we have two different sensors, Quickbird images were resized to the WorldView2 resolution, which is approximately 8000 by 7000 pixels. In Fig. 4, a train-

[1] http://dase.grss-ieee.org

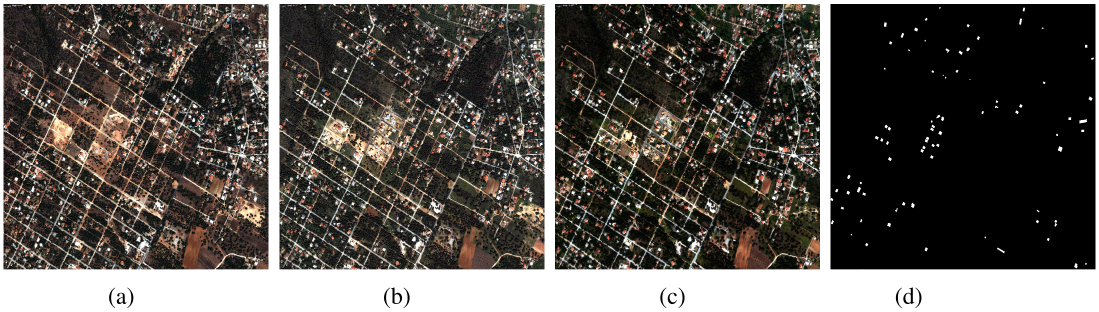

Fig. 4. Training data from the Attica VHR data set for three different years along with the corresponding *change* ground truth. (a) 2006. (b) 2009. (c) 2011. (d) Ground truth.

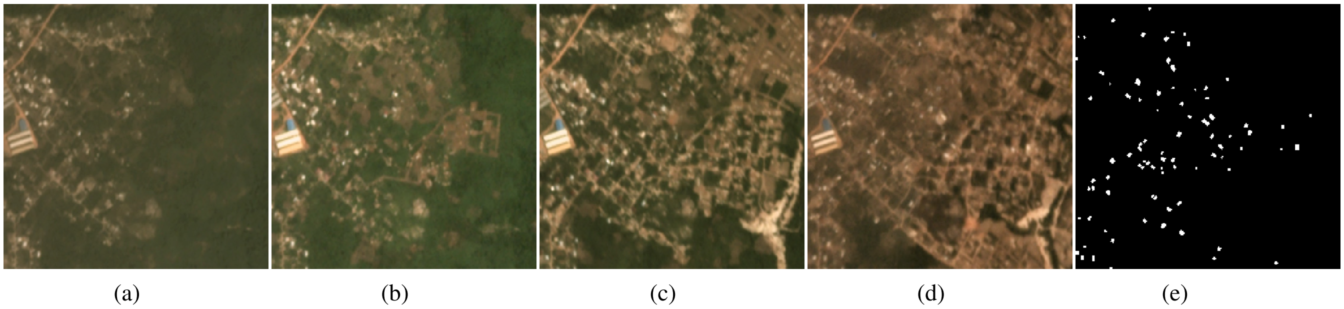

Fig. 5. Training data from the SpaceNet7 data set for four different dates along with the corresponding *change* ground truth. (a) 01/2018. (b) 04/2018. (c) 04/2019. (d) 01/2020. (e) Ground truth.

ing area for three different years along with the corresponding *change* ground truth is presented.

The whole region was divided into 36 equal nonoverlapping subregions of approximate size 1100 × 1300 pixels; 28 of them were used for training, four for validation, and four for testing. For the training process, patches of size 64 × 64 were produced with a stride of either 32 in case *change* pixels were included or 64 in case the patch did not include any *change* pixels at all. This strategy was applied as a data augmentation approach to enrich the *change* semantic category since it is extremely scarce compared to the *no change* one. Specifically, the percentage of *change* pixels in relation to the whole data set is only 1.2%. In addition, patches whose number of *change* pixels exceeded the threshold of 3% were randomly flipped in all possible angles proportional to 90°, while their brightness, contrast, and saturation levels were also randomly altered. Approximately 20 200 patches containing both *change* and *no change* pixels were backpropagated through the models for training, while 4000 were employed for validation.

3) SpaceNet7: This data set was recently released for multitemporal building detection in one of NeurIPS 2020 challenges. It consists of multitemporal satellite image cubes at the 4-m resolution, illustrating regions from all six continents of the earth. Red, green, blue, and near-infrared are the available spectral channels, while each image is approximately 1024 × 1024 pixels. Each region is depicted at different months spanning across the years of 2018, 2019, and 2020, with the largest data cube containing 24 dates; 60 image cubes are intended for training, while 20 of them are used as a test, and they are evaluated through an online submission process. The ground truth for this data set includes building footprints for each of the available dates, with the aim to monitor urban development. Competition participants are supposed to track the building locations for all the different dates, showing, in this way, the urban extension for each region.

In order to provide an additional evaluation benchmark for this work, we employed the 60 available training image cubes splitting them into 40 items for training, ten items for validation, and ten items for testing. We produced the change ground truth by subtracting the building footprints of the first date from the building footprints of the last date. In Fig. 5, one can observe a training sample from the SpaceNet7 data set for four different dates. It should be mentioned here that SpaceNet7 contains additional preprocessed versions of the available RGB-near infrared response (NIR) images where the clouds have been masked. For our investigation, however, we used the raw images so that the explored models can be evaluated on real conditions. In our experiments, the time-series related methods were examined using ten dates. Patches of size 32 × 32 were produced with a stride of either 16 in case *change* pixels were included or 32 in case the patch did not include any *change* pixels at all. Approximately 56 000 patches were produced and intended for training purposes, while 10 000 were used for validation. Similar to the previous data sets, *change* and *no change* samples are again extremely uneven, with *change* samples making up only 0.94% of the whole data set.

E. Optimization

Hyperparameters were similar for both data sets, picking Adam optimizer with a learning rate of 10^{-4}. Batchsize was 32 for the OSCD data set, 10 for the Attica VHR data set, and 2 for the SpaceNet7 data set. Early stopping criteria were employed for every adopted approach in order to cease the training process and pick the optimal network weights. All applied methods needed less than 60 epochs to converge, while all experiments were implemented using the PyTorch deep learning library [68] on a single NVIDIA GeForce GTX TITAN with 12 GB of GPU memory. Each class was associated with a weight inversely proportional to the total pixel number included in it for the cross-entropy loss. Furthermore, for the evaluation of multitask L-UNet, a grid search was employed to determine the weight values for 7. Specifically, regarding the Attica VHR data set, w_1 was equal to 0.6, and the rest of the weights equals 0.1. In the SpaceNet7 case, w_1 was equal to 0.8, while the rest of the weights were equal to 0.05. For the OSCD data set, as we have already mentioned, we evaluated only L-UNet since there is no available ground truth for the semantic categories of the different dates. Hence, $w1$ was equal to 1, while the rest of the weights were equal to 0. Finally, it should be clarified here that the employed ground truth in each data set case is related to the changes that have occurred between the first and last dates.

F. Quantitative Evaluation Metrics

To assess the quality of the results, we employed five different evaluation metrics: precision, recall, F1 score, and balanced accuracy (BA). They can be derived from the calculated true positives (TPs), FP, true negatives (TNs), and false negatives (FNs) values. If we have observations belonging in l different categories, then TP is the number of observations that have been correctly classified as l. FP is the number of observations that have been wrongly classified as l. TN is the number of observations that have been rightly recognized as not belonging to l. Finally, FN represents the observations that belong to l, but the model has associated them to another category

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \quad \text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$
$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}$$
$$\text{Balanced\_Accuracy} = \frac{\frac{\text{TP}}{\text{TP+FN}} + \frac{\text{TN}}{\text{TN+FP}}}{2}.$$

III. Experimental Results and Discussion

In this section, we evaluate each of the components of the proposed formulation benchmarking their performance. Moreover, we provide a comparison with state-of-the-art deep learning-based methods for change detection.

A. Comparison With State-of-the-Art Methods

Experiments were conducted using a various number of dates, while the results were compared with fully convolutional networks [51], multitask learning methods described in [59] and [61], and the methods proposed for time-series data sets [55], [69]. All these literature methods were adjusted, using the same backbone U-Net-like architecture to ensure reliable comparison. In the following subsections, details are given for every investigated framework.

1) Method in [59]: Based on a UNet-like architecture, change detection and building semantic segmentation were performed simultaneously on the input dates, using two different decoding branches. In addition, two separate loss quantities were employed: cross-entropy for the building semantic maps of the first and last dates and the focal loss [60] for the change detection task.

2) Method in [61]: For the method in [61], we first trained the plain UNet-like model on the building semantic segmentation task using the first and last available dates. Then, using these weights as initialization, we trained the network once again on the change detection task, using as skip connections the absolute difference of the produced encoding feature vectors.

3) FC-Siam-Conc: In the FC Siamese concatenation case [51] (FC-Siam-Conc), the encoder was comprised of several encoding streams depending on the number of processed dates. The streams had shared weights and identical configuration, with all the feature vectors produced after each encoding step being concatenated with the feature vector of the respective decoding step.

4) FC-EF: As far as the FC early fusion [51] (FC-EF) approach is concerned, the different temporal input volumes were concatenated along the channel dimension before being passed through the network. Skip connections concatenated the encoding outputs with the corresponding decoding ones.

5) FC-Siam-Diff: The FC Siamese difference method [51] (FC-Siam-Diff) followed the same principles as FC-Siam-Conc although this time the concatenation of skip connections was performed using the absolute difference of the resulting encoding feature volumes.

We should mention here that FC-Siam-Conc and FC-Siam-Diff were also evaluated in a multitask setting, by adding a supplementary decoding branch for the semantic segmentation task. In the FC-EF case, semantic segmentation could not be performed simultaneously with change detection 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.

6) LSTM [55]: In this work, sequential recurrent encoders were employed to perform the task of land-cover classification based on image time series. Specifically, the multitemporal volume was given as input to a convolutional layer, the output of which was split into four equal parts, representing the four different gates of the LSTM structure. This forward pass was implemented in a bidirectional way, namely, the input dates were passed to the encoder in sequential and reverse orders. The final cell states were concatenated and transformed to softmax-normalized activations so that the prediction map could be produced.

7) Method in [69]: Finally, the authors here have proposed a network for crop type classification from multitemporal data. Specifically, the sequential input was passed through a succession of convolutional layers that downsample and upsample it back using also skip connections. The produced

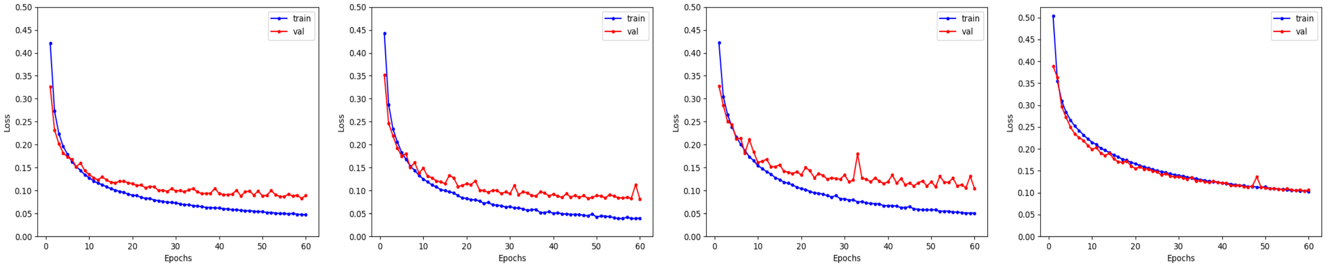

Fig. 6. Training and validation loss curves for different models trained with five dates on the OSCD data set. From left to right: FC-Siam-Conc, FC-EF, FC-Siam-Diff, and the proposed L-UNet.

TABLE II

QUANTITATIVE EVALUATION OF THE PROPOSED L-UNET ON THE TESTING PART OF THE OSCD DATA SET. PRECISION, RECALL, AND F1 RATES ARE ASSOCIATED WITH THE *Change* CLASS, WHILE BA IS ALSO PROVIDED. ALL THE ROWS DEMONSTRATE RESULTS USING THE RGB-NIR BANDS WITH THE LAST COLUMN INDICATING THE TIME NEEDED BY EACH METHOD TO PRODUCE ANNOTATIONS FOR A TESTING IMAGE OF DIMENSIONS 550 × 550

Models	**Dates**	**Precision**	**Recall**	**F1**	**BA**	**Time(sec)**
FC-Siam-Conc [51]	2	59.32	54.73	56.93	76.34	≈ 4
	3	62.51	50.12	55.63	74.24	≈ 4
	5	60.45	50.46	55.01	74.33	≈ 5
FC-EF [51]	2	45.46	**71.42**	55.56	**83.37**	≈ 3
	3	53.05	59.74	56.20	78.43	≈ 4
	5	59.09	55.56	57.27	76.73	≈ 5
FC-Siam-Diff [51]	2	60.86	54.07	57.26	76.09	≈ 3
	3	57.13	50.63	53.68	74.28	≈ 4
	5	55.04	57.92	56.44	77.67	≈ 5
LSTM [55]	5	45.05	54.24	49.22	75.32	≈ 5
Method in [69]	5	62.62	49.35	55.20	73.87	≈ 7
Ours (L-UNet)	2	54.29	63.05	58.34	80.08	≈ 4
	3	63.49	55.01	**58.94**	76.64	≈ 4
	5	**64.42**	53.09	58.21	75.75	≈ 5

feature vector was then given to a convolutional LSTM so that the temporal relationship between the dates could be computed.

B. Evaluation of the L-UNet

In this section, we evaluate the use of fully convolutional LSTMs, integrated as skip connections on the UNet-based architecture. To benchmark their performance, we report results on the three outlined data sets, and we compare them with other change detection approaches and, in particular, methods proposed in [51], [55], and [69].

Looking at the results on the OSCD data set (see Table II), one can observe that L-UNet and FC-EF methods result in higher precision rates as additional temporal information is integrated, demonstrating their important contribution to the lessening of FPs. On the other hand, recall rates increase only in the FC-Siam-Diff case, which means that even though, in most of the methods, the number of FP detections is ameliorated, FN pixels continue to exist in a large quantity. L-UNet produces the best precision and F1 scores, meaning that the total number of FN and FP *change* pixels is smaller compared to the rest of the methods. The FC-EF bitemporal approach attains the highest recall and BA metrics when employing two dates. However, in this case, the precision rate is lower than 50%, indicating a high number of FP values, which is one of the main problems in change detection. In addition, the highest F1 score for FC-EF is attained in the case of five dates, and it is approximately 1.7% lower than L-UNet. As far as time-series approaches are concerned, the method proposed in [69] produces a higher F1 score compared to the LSTM case [55]. Both approaches, however, result in lower accuracy metrics compared to the rest of the methods, except the method presented in [69], which attains the second best precision score, after L-UNet.

Moreover, training and validation curves are presented in Fig. 6 for some of the investigated approaches using five dates. All the evaluated methods converge, however, the training of the proposed model seems to be smoother and more stable without very high variations between the training and validation performances.

We should mention here that since the ground truth for the testing images is not publicly available, we could not provide a thorough qualitative evaluation with a proper illustration of TP, FP, and FN regions. Nevertheless, in Fig. 7, some advantages of the proposed L-UNet can be observed for the methods that attained the highest F1 scores according to Table II. In the first row, there is an example where no change has taken place between the different dates; however, all compared methods get disorientated by the existing cloud. In the case of L-UNet, we can see that even though FP pixels exist, they are less than the rest of the approaches. The proposed method also seems to be robust on changes that are not related to urbanization, reducing FP detections. Specifically, in the second row, our formulation is the only one that does not highlight agricultural changes as changed areas. Finally, in the third row, we can observe that L-UNet has detected successfully urban changes, reporting fewer FPs (regions indicated with red circles).

Additional experiments on the Attica VHR data set for the change detection task are presented in Table III. One can observe that the highest recall and BA rates have been achieved by the LSTM [55], while the best precision and F1 scores have been attained by the L-UNet approach. In the LSTM [55] case, however, the precision rate is exceptionally low, which means that FP predictions have dramatically raised. The much higher F1 score in the L-UNet case demonstrates the finer balance of FP and FN predictions, especially with the integration of more temporal information. The FC-Siam-Conc model is also boosted when additional dates are employed. FC-EF reports poor results, especially if we consider that precision rates

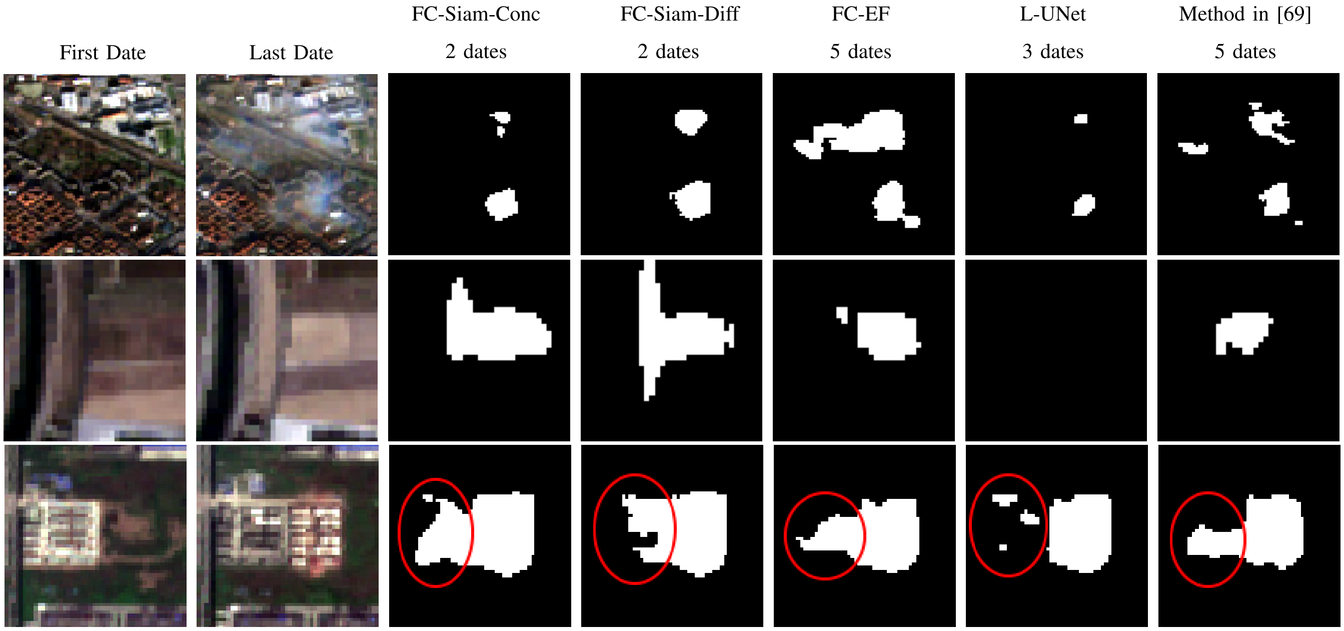

Fig. 7. Qualitative evaluation of the proposed L-UNet on zoomed testing regions of the OSCD data set. (First Column) RGB images of the first available date. (Second Column) RGB images of the last available date. (Third Column) FC-Siam-Conc with two dates. (Fourth Column) FC-Siam-Diff with two dates. (Fifth Column) FC-EF with five dates. (Sixth Column) L-UNet with three dates. (Seventh Column) Method in [69] with five dates.

TABLE III

Quantitative Evaluation of the Proposed L-UNet on the Testing Part of Attica VHR Data Set. Precision, Recall, and F1 Rates Are Associated With the *Change* Class, While BA Is Also Provided. All the Rows Demonstrate Results Using the RGB-NIR Bands With the Last Column, Indicating the Time Needed by Each Method to Produce Annotations for a Testing Image of Dimensions 1200 × 1300

Models	**Dates**	**Precision**	**Recall**	**F1**	**BA**	**Time(sec)**
	2	42.47	56.52	48.49	78.00	≈ 13
FC-Siam-Conc [51]	3	43.24	56.94	49.15	78.21	≈ 14
	5	46.62	59.03	52.09	79.28	≈ 16
	2	41.10	55.53	47.24	77.49	≈ 13
FC-EF [51]	3	45.50	52.32	48.67	75.94	≈ 14
	5	43.85	52.95	47.97	76.24	≈ 15
	2	45.67	56.80	50.63	78.17	≈ 13
FC-Siam-Diff [51]	3	46.90	54.18	50.28	76.88	≈ 14
	5	41.45	40.59	41.02	70.10	≈ 15
LSTM [55]	5	31.42	**68.04**	42.98	**83.51**	≈ 19
Method in [69]	5	40.63	61.22	48.84	80.31	≈ 20
	2	47.25	55.21	50.92	77.39	≈ 14
Ours (L-UNet)	3	43.15	61.08	50.57	80.26	≈ 15
	5	**47.96**	60.19	**53.38**	79.87	≈ 17

remain always below 46% revealing the existence of many FP detections. Finally, in the FC-Siam-Diff case, even though the F1 score reaches approximately the rate of 50%, low precision levels reveal a large number of FPs.

In Fig. 8, some qualitative outcomes are illustrated on zoomed regions from the Attica VHR testing areas. Identifying changes on buildings without additional information for the building class is quite challenging and can be sensitive to illumination changes. In the first row, one can notice that only L-UNet with five dates identifies that the depicting images do not contain any change related to the building class. Rooftop alterations in the second row have disoriented FC-Siam-Diff with two dates, FC-EF with five dates, and the method in [69], whereas FC-Siam-Conc with five dates and L-UNet with five dates have addressed them successfully. Continuing with the third row, we can observe that the proposed approach has correctly identified that the swimming pool does not correspond to a building change category, in contrast to the rest of the methods. The total number of FN and FP pixels seems to be lower for L-UNet in the fourth row, while, in the fifth row, the most successful detections have been attained by FC-Siam-Conc with five dates and L-UNet with five dates. In general, L-UNet and the addition of the LSTMs on the different encoding levels seem to capture semantic changes in a better way.

As far as inference time is concerned, we can observe that L-UNet does not require more time to produce annotations compared with the rest of the methods.

C. Evaluation of the Multitask L-UNet

In Table IV, we provide a comparison of the proposed multitask L-UNet with state-of-the-art change detection methods. In particular, we compare our architecture with the models in [51] transformed into a multitask setting, as well as the multitask methods in [61] and [59]. On the left part of Table IV, the evaluation of the change detection task is demonstrated, while, on the right part, different metrics for the building semantic segmentation are provided.

Starting with the methods described in [51], compared to Table III, one can observe that, in the FC-Siam-Conc approach, the integration of building semantic segmentation ameliorates the F1 score especially in the case of five dates where the precision rate also rises above 50%. Regarding FC-Siam-Diff, although recall rate and BA reach their peak value, precision

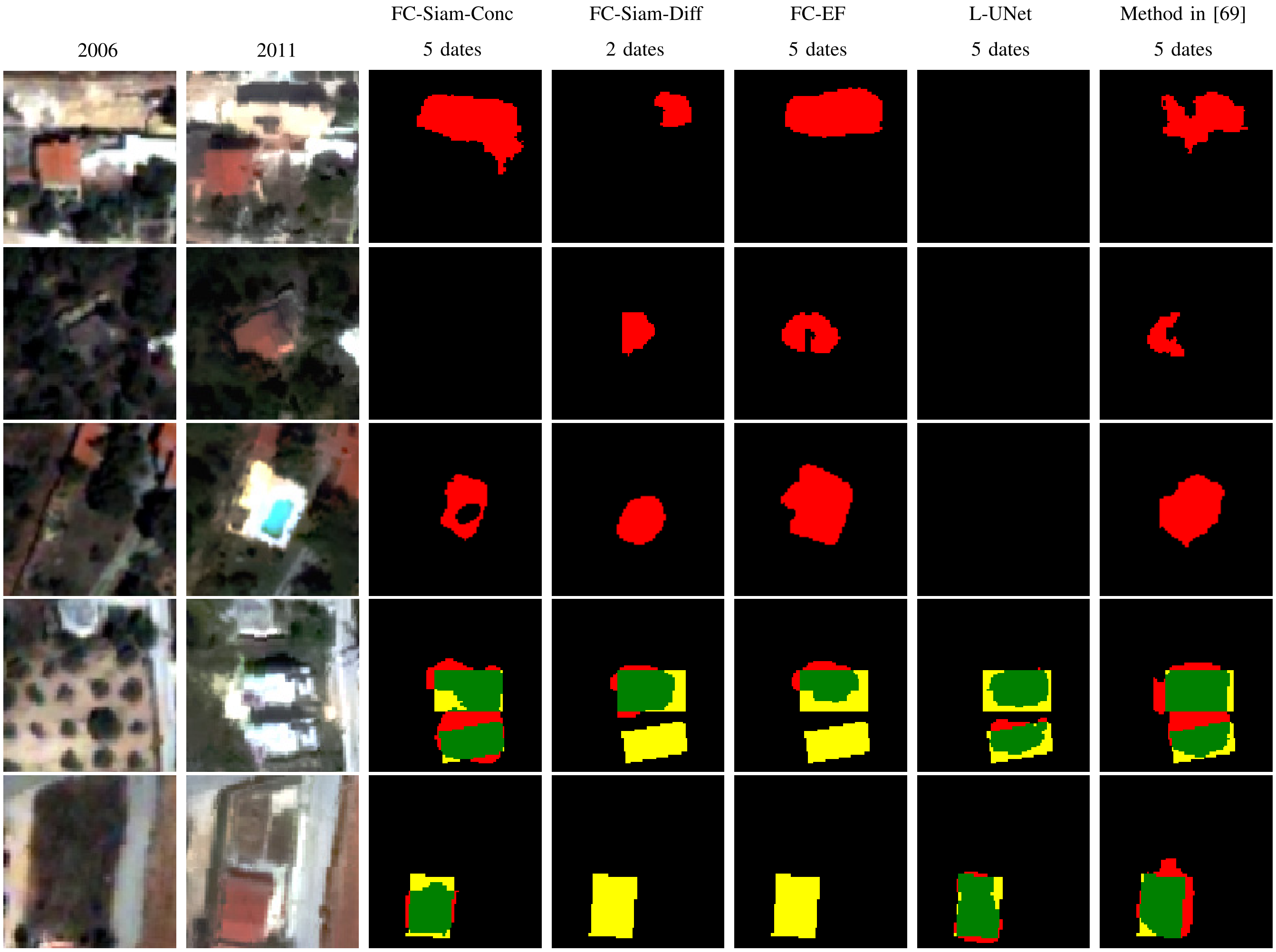

Fig. 8. Qualitative evaluation of the proposed L-UNet on zoomed regions of the Attica VHR testing areas for the *change* detection task. (First Column) RGB images of 2006. (Second Column) RGB images of 2011. (Third Column) FC-Siam-Conc with five dates. (Fourth Column) FC-Siam Diff with two dates. (Fifth Column) FC-EF with five dates. (Sixth Column) L-UNet with five dates. (Seventh Column) Method in [69] with five dates (*Green*: TPs, *Black*: TNs, *Red*: FPs, and *Yellow*: FNs).

TABLE IV

QUANTITATIVE EVALUATION OF THE PROPOSED MULTITASK L-UNET ON THE TESTING PART OF ATTICA VHR DATA SET. ON THE LEFT PART, THE EVALUATION OF THE *Change* DETECTION TASK IS PRESENTED, WHILE, ON THE RIGHT PART, METRICS FOR THE BUILDING SEMANTIC SEGMENTATION ARE PROVIDED. PRECISION, RECALL, AND F1 RATES ARE ASSOCIATED WITH THE *change* CLASS AND THE *Building* CLASS FOR 2006, WHILE BA IS ALSO PROVIDED. ALL THE ROWS DEMONSTRATE RESULTS USING THE RGB-NIR BANDS WITH THE LAST COLUMN, INDICATING THE TIME NEEDED BY EACH METHOD TO PRODUCE ANNOTATIONS FOR A TESTING IMAGE OF DIMENSIONS 1200 × 1300

		Building Change Detection				**Building Semantic Segmentation for 2006**				-
Models	**Dates**	**Precision**	**Recall**	**F1**	**BA**	**Precision**	**Recall**	**F1**	**BA**	**Time(sec)**
multi-task FC-Siam-Conc [51]	2	44.70	59.07	50.89	79.28	74.41	65.92	69.91	82.36	≈ 14
	3	46.96	57.74	51.79	78.65	75.15	63.14	68.63	81.02	≈ 14
	5	50.23	57.68	53.70	78.65	**78.78**	57.12	66.22	78.15	≈ 15
multi-task FC-Siam-Diff [51]	2	44.09	**62.11**	51.57	**80.79**	75.90	63.50	69.15	81.21	≈ 14
	3	45.18	59.72	51.44	79.61	75.39	64.00	69.23	81.44	≈ 14
	5	41.87	48.92	45.12	74.23	73.79	66.81	70.13	82.78	≈ 15
Method in [59]	2	47.02	56.83	51.46	78.20	76.34	63.38	69.25	81.17	≈ 14
	5	40.67	58.96	48.14	79.19	75.11	63.77	68.98	81.32	≈ 15
Method in [61]	2	51.97	49.61	50.76	74.65	74.41	66.07	70.00	82.43	≈ 14
	5	42.46	42.82	42.64	71.21	74.41	66.07	70.00	82.43	≈ 15
Ours multi-task L-UNet	2	44.53	61.39	51.62	80.44	67.38	**75.54**	**71.23**	**86.80**	≈ 15
	3	46.89	61.07	53.05	80.30	75.48	61.48	67.77	80.21	≈ 15
	5	**52.42**	59.68	**55.82**	79.65	76.08	61.52	68.03	80.24	≈ 18

rates remain very low, which means that even though FN pixels are more limited, many FP pixels continue to exist.

Continuing with the bitemporal multitask learning approaches in [59] and [61], it seems that they report higher

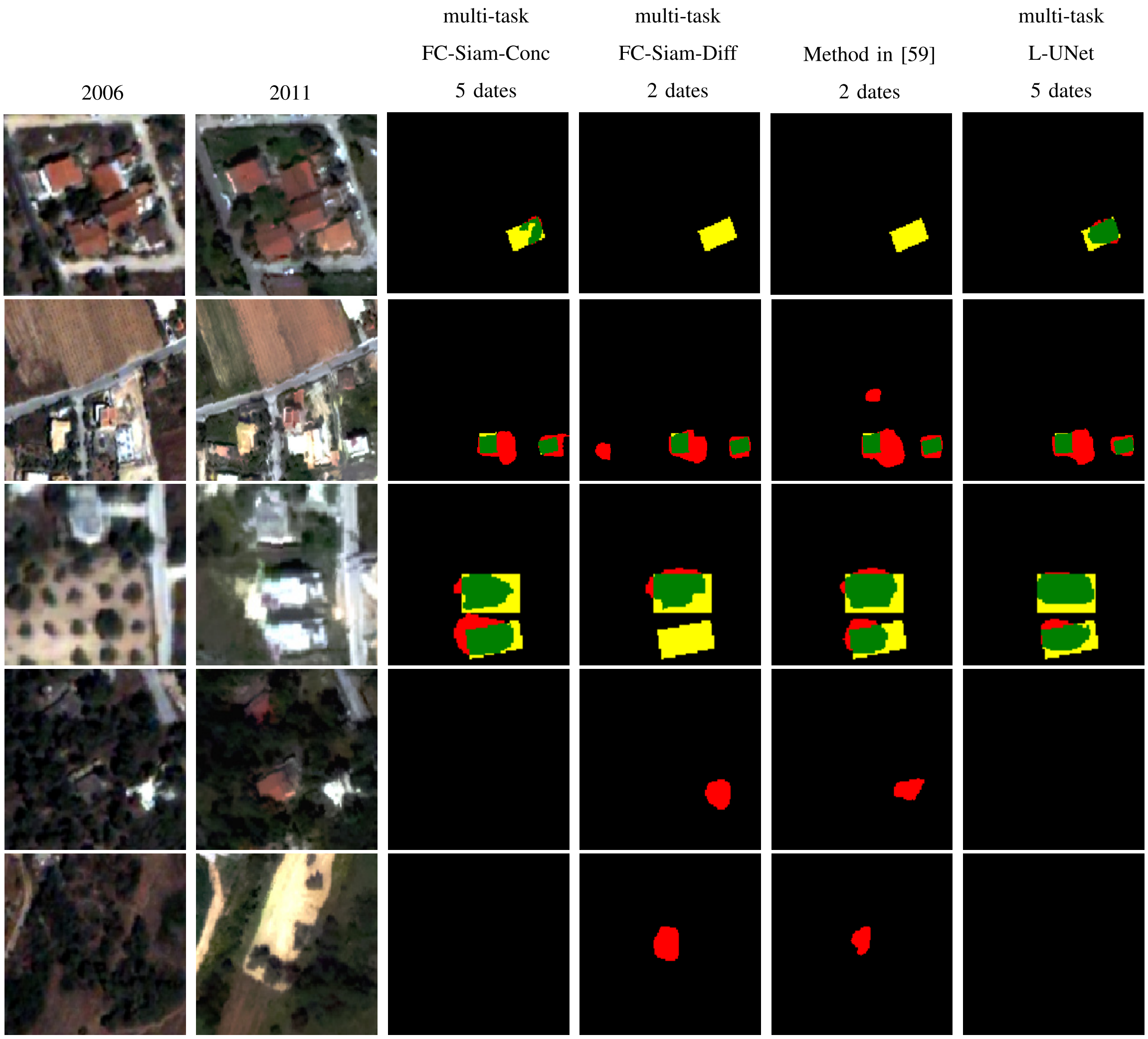
2006
2011
multi-task FC-Siam-Conc 5 dates
multi-task FC-Siam-Diff 2 dates
Method in [59] 2 dates
multi-task L-UNet 5 dates

Fig. 9. Qualitative evaluation of the proposed multitask L-UNet on zoomed regions of the Attica VHR testing areas for the *change* detection task. (First Column) RGB images of 2006. (Second Column) RGB images of 2011. (Third Column) multitask FC-Siam-Conc with five dates. (Fourth Column) Multitask FC-Siam Diff with two dates. (Fifth Column) Method in [59] with two dates. (Sixth Column) Multitask L-UNet with five dates (*Green*:TPs, *Black*: TNs, *Red*: FPs, and *Yellow*: FNs).

precision rates compared with the corresponding bitemporal cases of [51], with [59] achieving a higher F1 score than [61]. As the dates rise, however, these methods result in lower performance, as shown in the five dates case. The best performance concerning FP rates is attained by the proposed multitask L-UNet. Specifically, the precision rate reaches 52.42% in the case of five dates, exceeding the second highest precision score of multitask FC-Siam-Conc by approximately 2.2%. In addition, the F1 score becomes equal to 55.82%, which is also 2.2% higher than the corresponding F1 rate in the multitask FC-Siam-Conc case. In the proposed approach, the F1 score remains always above 50%, which means that temporal attributes in combination with the additional features of the semantic segmentation task boost greatly the performance of the network attaining a more balanced total number of FP and FN pixels. Looking back at Table III, we can further realize the great benefits of the multitask setting for the L-UNet since, in Table IV, precision and F1 rates have been raised by 4.5% and 2.4%, respectively.

Regarding semantic segmentation on buildings, the provided accuracy metrics are related to the year of 2006. The highest precision rate is achieved by multitask FC-Siam-Conc with five dates, while the rest of the accuracy scores are better in the case of multitask L-UNet with two dates. In order to assess the performance of multitask L-UNet on the semantic segmentation task and in particular on the building footprint detection, we compare it with the performance of a standard U-Net architecture that is commonly used for this problem. The evaluation of the standard U-Net on the testing images of 2006 resulted in 80.15%, 60.91%, 69.22%, and 80.05% for precision, recall, F1, and BA, respectively. Looking at Table IV that summarizes the evaluation for the building footprints of 2006 using the multitask networks, one can notice that all evaluation metrics except precision can achieve higher values when combined with the change detection task. This indicates that our formulation boosts the performance of each individual task by fusing together useful features from each problem. From a qualitative perspective, the building

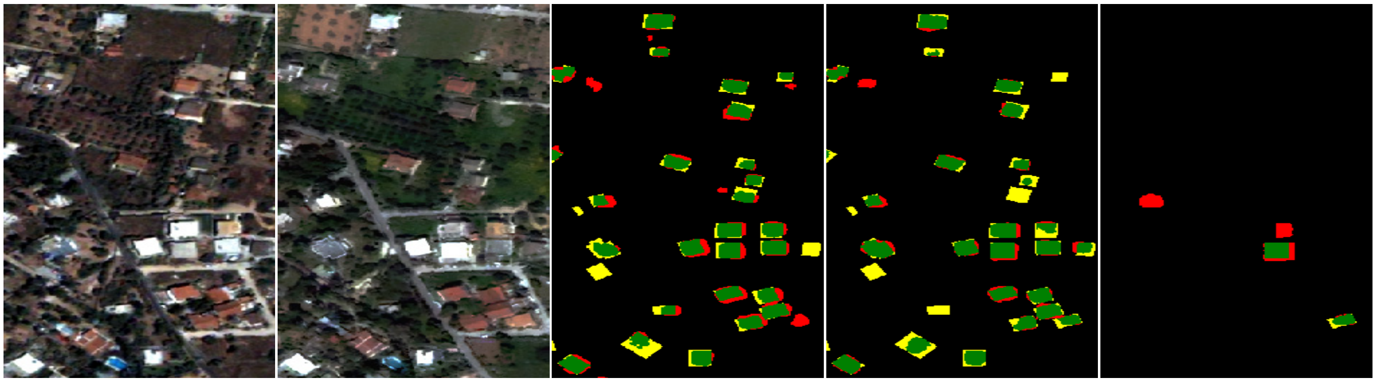

Fig. 10. Qualitative results of multitask L-UNet with five dates, for a region of Attica VHR testing part. From left to right: RGB image of 2006, RGB image of 2011, *building* predictions of 2006, *building* predictions of 2011, and *change* predictions. (*Green*: TPs, *Black*: TNs, *Red*: FPs, and *Yellow*: FNs).

predictions resulting from the multitask framework are quite similar to those resulting from the semantic segmentation task alone. In Fig. 11, we can observe building predictions of 2006 using the multitask L-UNet with five dates and the standard U-Net.

As a whole, precision values never exceed the rate of 53% for the change detection task, indicating the more challenging nature of the complex VHR images compared to the high-resolution ones. Two are the principal reasons that constitute this problem: registration and parallax errors that perplex the learning procedure and the different types of change that are included in the satellite images. The variety of changes (e.g., land-use diversification and alterations in vegetation) results in a wide range of spectral values for certain areas where *urban change* does not take place. Another fundamental problem that hinders the successful learning process is that *change* and *no change* categories are greatly disproportionate. For the Attica VHR data set, the total number of *no change* pixels for the training data set is almost 85 times larger than the number of *change* ones. Notwithstanding these difficulties, the L-UNet method seems to fully exploit the available information.

Continuing with the qualitative evaluation, in the first row of Fig. 9, one can notice that the proposed approach has detected more accurately the additional building compared to the rest of the methods. Continuing with the second row, we can observe that even though all methods are confused by certain rooftop illumination changes, the approaches employing five dates have resulted in fewer FP values. The third row demonstrates a case where the total number of FP and FN pixels is lower for multitask L-UNet with five dates. Finally, the last two rows display instances where methods utilizing additional dates deal with rooftop and vegetation alterations in a more constructive way.

One problem that is evident from the qualitative evaluation is that of inconsistent building boundaries. All employed methods regardless of their level of success in identifying the *urban changes* usually fail to provide accurate boundaries resulting in many FP pixels along the perimeter of buildings. This issue can also be noticed in Fig. 10 where the *building* predictions of multitask L-UNet for both 2006 and 2011 are provided along with the corresponding *change* for a larger testing region of the Attica VHR data set.

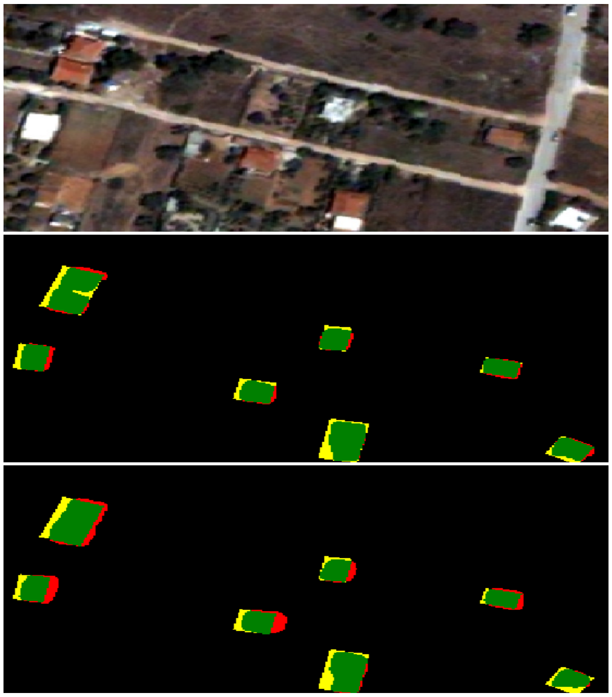

Fig. 11. Qualitative evaluation on building predictions of 2006. From top to bottom: RGB image of 2006, building semantic segmentation with standard U-Net, multitask L-UNet with five dates (*Green*: TPs, *Black*: TNs, *Red*: FPs, and *Yellow*: FNs).

As far as the SpaceNet7 data set is concerned, numerical results of the conducted experiments are outlined in Table V. Starting with the methods presented in [51], FC-Siam-Conc has achieved the highest F1 score when employing two dates, while FC-EF attained the best BA in the case of ten dates. The corresponding multitask results seem to be better only for the FC-Siam-Conc case since the F1 score has raised from 44.80% to 45.16%. On the contrary, accuracy metrics for FC-Siam-Diff did not benefit much neither from the additional dates nor from the multitask setting. In the FC-EF case, all accuracy metrics except precision are ameliorated when more dates are used. Continuing with the multitask approaches proposed in [59] and [61], one can notice that they have produced many FP values since the precision rates are very low. The highest F1 score has resulted from the method in [61] when utilizing ten dates. Regarding the LSTM [55] and the

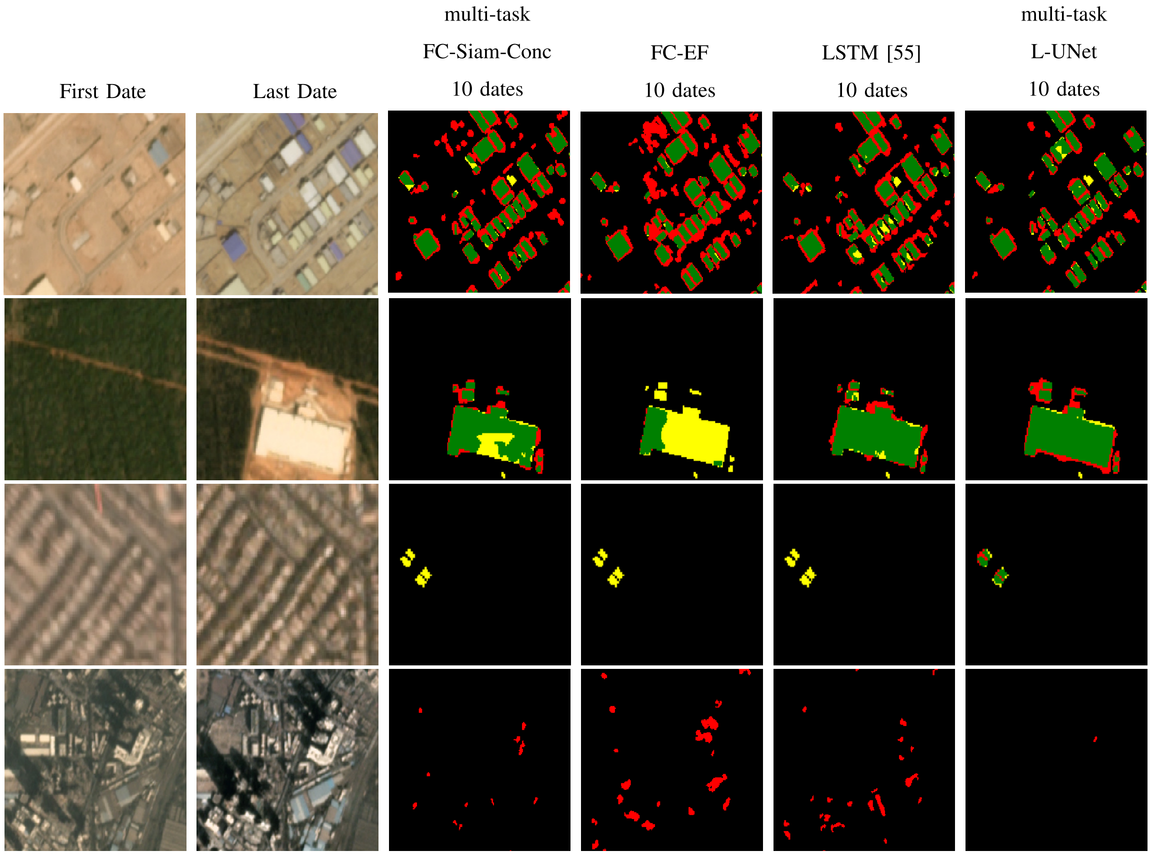

Fig. 12. Qualitative evaluation of the proposed multitask L-UNet on zoomed regions of the SpaceNet7 testing areas for the *change* detection task. (First Column) RGB images of the first date. (Second Column) RGB images of the last date. (Third Column) Multitask FC-Siam-Conc with ten dates. (Fourth Column) FC-EF with ten dates. (Fifth Column) LSTM [55] with ten dates. (Sixth Column) Multitask L-UNet with ten dates (*Green*: TPs, *Black*: TNs, *Red*: FPs, and *Yellow*: FNs).

TABLE V

Quantitative Evaluation of L-UNet and Multitask L-UNet on the Testing Part of SpaceNet7 Data Set. Precision, Recall, and F1 Rates Are Associated With the *Change* Class, While BA Is Also Provided. All the Rows Demonstrate Results Using the RGB-NIR Bands With the Last Column, Indicating the Time Needed by Each Method to Produce Annotations for a Testing Image of Dimensions 1024 × 1024

Models	Dates	Precision	Recall	F1	BA	Time(sec)
FC-Siam-Conc [51]	2	39.07	52.50	44.80	75.84	≈ 8
	10	42.22	38.97	40.53	69.22	≈ 10
multi-task FC-Siam-Conc [51]	2	34.41	55.78	42.56	77.36	≈ 8
	10	38.70	54.20	45.16	76.67	≈ 10
FC-Siam-Diff [51]	2	29.74	48.00	36.73	73.43	≈ 8
	10	31.84	53.11	39.81	75.99	≈ 10
multi-task FC-Siam-Diff [51]	2	29.78	46.79	36.40	72.84	≈ 8
	10	22.66	24.59	23.59	61.88	≈ 10
FC-EF [51]	2	42.96	42.68	42.82	71.06	≈ 8
	10	34.29	57.71	43.02	78.30	≈ 10
Method in [59]	2	21.19	60.72	31.41	79.23	≈ 8
	10	26.15	60.99	36.61	79.63	≈ 10
Method in [61]	2	24.19	61.84	34.77	79.95	≈ 8
	10	31.97	55.85	40.66	77.33	≈ 10
LSTM [55]	10	38.75	48.76	43.18	74.00	≈ 13
Method in [69]	10	34.87	54.72	42.59	76.85	≈ 13
L-UNet	2	36.42	56.46	44.28	77.74	≈ 9
	10	44.83	53.82	48.92	76.58	≈ 12
multi-task L-UNet	2	32.92	**62.41**	43.11	**80.57**	≈ 9
	10	**47.71**	52.75	**50.11**	76.09	≈ 12

method in [69] that have been proposed for time-series data sets, we can observe that accuracy rates are quite similar between the two approaches. Compared with the rest of the approaches, the time-series methods provide better results than the multitask ones [59], [61] but similar results with the methods in [51]. Finally, L-UNet boosts the precision and F1 rates when we take advantage of more dates. In the multitask L-UNet framework, recall rate and BA reach the highest level in the case of two dates, while precision and F1 scores become optimal in the case of ten dates. As a whole, in this data set, the proposed multitask L-UNet outperforms all other methods for all accuracy metrics. Especially, when more temporal information is integrated, precision and F1 scores benefit the most by the suggested method, exceeding the rest of the approaches by at least 4.9%. As far as the results on building semantic segmentation are concerned, we evaluated the multitask models on the first date of SpaceNet7. The highest F1 score and BA resulted from the [59] and were equal to 47.14% and 77.30%, respectively. The F1 score attained by the multitask L-UNet was similar and equal to 46.79%, while the BA was 73.24%.

Qualitative samples from the testing part of the SpaceNet7 data set are delineated in Fig. 12 for some of the investigated approaches. In the first row, we can see that even though all methods produce FP pixels between the building boundaries, multitask L-UNet has resulted in a more clear separation of the buildings. In the second row, the LSTM network [55] and the proposed multitask L-UNet have identified better the main building footprint of the image. Regarding the third row, here, the multitask L-UNet has

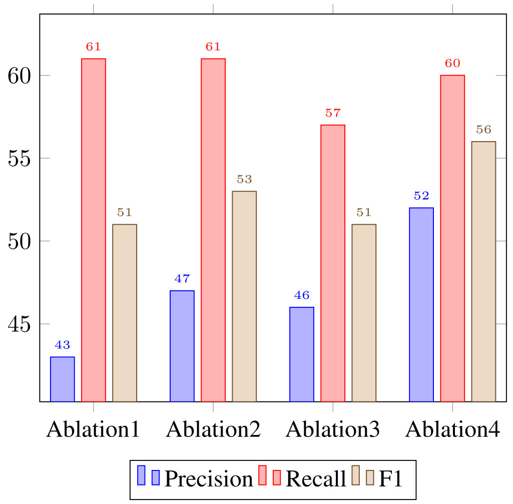

Fig. 13. Accuracy metrics for multitask L-UNet with five dates using different combinations of losses. The provided metrics have resulted from the testing part of Attica VHR data set. Ablation1 [$\text{Loss}_{\text{ch}}, \text{Loss}^1_{\text{seg}}, \text{Loss}^T_{\text{seg}}$], Ablation2 [$\text{Loss}_{\text{ch}}, \text{Loss}^1_{\text{seg}}, \text{Loss}^T_{\text{seg}}, \text{Loss}_{\text{ch2}}$], Ablation3 [$\text{Loss}_{\text{ch}}, \text{Loss}^1_{\text{seg}}, \text{Loss}^T_{\text{seg}}, \text{Loss}^T_{\text{seg2}}$], and Ablation4 [$\text{Loss}_{\text{ch}}, \text{Loss}^1_{\text{seg}}, \text{Loss}^T_{\text{seg}}, \text{Loss}_{\text{ch2}}, \text{Loss}^T_{\text{seg2}}$].

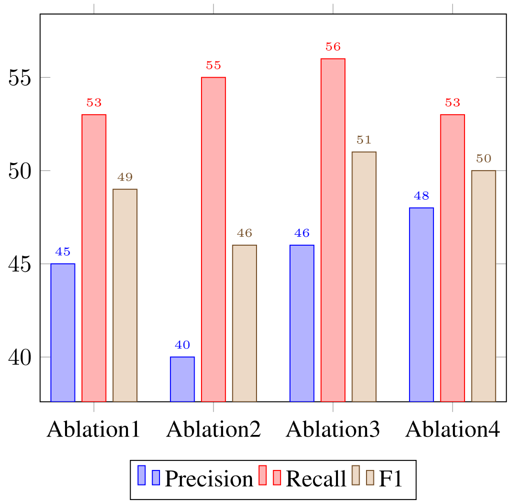

Fig. 14. Accuracy metrics for multitask L-UNet with ten dates using different combinations of losses. The provided metrics have resulted from the testing part of SpaceNet7 data set. Ablation1 [$\text{Loss}_{\text{ch}}, \text{Loss}^1_{\text{seg}}, \text{Loss}^T_{\text{seg}}$], Ablation2 [$\text{Loss}_{\text{ch}}, \text{Loss}^1_{\text{seg}}, \text{Loss}^T_{\text{seg}}, \text{Loss}_{\text{ch2}}$], Ablation3 [$\text{Loss}_{\text{ch}}, \text{Loss}^1_{\text{seg}}, \text{Loss}^T_{\text{seg}}, \text{Loss}^T_{\text{seg2}}$], and Ablation4 [$\text{Loss}_{\text{ch}}, \text{Loss}^1_{\text{seg}}, \text{Loss}^T_{\text{seg}}, \text{Loss}_{\text{ch2}}, \text{Loss}^T_{\text{seg2}}$].

managed to detect the construction of some small buildings. Finally, the last row shows an example where the proposed approach includes the fewer FP pixels caused by building shadows and illumination differences.

Regarding inference time, similar to L-UNet, we can notice that multitask L-UNet does not need much more time to produce annotations compared with the rest of the methods.

D. Evaluation of the Different Segmentation and Change Detection Loss Components

In this section, we conduct an ablation study, and we discuss the importance of each of the loss components for the coupling of semantic segmentation and change detection tasks. The performance of these components is reported on the Attica VHR and SpaceNet7 data sets, on which annotations for both the building class and the urban change class are available. For this evaluation, we chose the best performing model (multitask L-UNet with five dates for the Attica VHR data set and multitask L-UNet with ten dates for the SpaceNet7 data set) and trained it using different loss compositions, as described in the caption of Fig. 13. Giving some more details, Ablation1 represents multitask L-UNet with five dates using three losses: one for the change detection task and two for the building semantic segmentation of the first and last available dates. In Ablation2, the additional Loss_{ch2} is employed, while, in Ablation3, $\text{Loss}^T_{\text{seg2}}$ is used as the additional fourth loss component. Finally, Ablation4 includes all five losses, as described in Section II-C, resulting in our proposed framework. After training each of the different frameworks, the evaluation metrics were calculated on the test part of Attica VHR and SpaceNet7 data sets. Looking at Fig. 13 for the Attica VHR data set, we can observe that recall rates are almost the same in Ablation1, Ablation2, and Ablation4, while, in Ablation3, a lower value is attained, meaning that the integration of $\text{Loss}^T_{\text{seg2}}$ alone produces a deteriorated result regarding FN detections. As far as precision rates are concerned, they increase from 43% to 47% when Loss_{ch2} or $\text{Loss}^T_{\text{seg2}}$ are incorporated into the training process, indicating the contribution of the additional circular losses to the lessening of FP values. The amelioration of FP pixels is even more obvious when both circular losses are integrated, with the precision rate reaching the value of 52%. In this case, the F1 score also reaches the highest level, becoming equal to 56%.

Regarding the SpaceNet7 data set, the ablation study results are shown in Fig. 14. In this setting, we can observe that Ablation1 has a lower precision rate than Ablation4, meaning that Loss_{ch2} and $\text{Loss}^T_{\text{seg2}}$ contribute to the lessening of FP detections. In Ablation2, the recall rate has raised, but precision has become very low, leading also to a lower F1 score. Finally, in Ablation3, the results are similar with Ablation4 but with a wider difference between the precision and recall rates. Once again, in this case, we notice that Ablation4 results in the lowest number of FP detections, benefiting from the combination of all the loss components.

E. Discussion

Taking into consideration all the conducted experiments on all the different data sets, we can draw some conclusions about the investigated methods. Overall, the integration of LSTM networks as skip connections provides efficient aggregation strategies able to encode information presented on two or more temporal time stamps. Results indicated that such an approach can outperform other aggregation functions, such as concatenation and early fusion. Subtraction was the least efficient when more temporal information was incorporated. In the case of concatenation and subtraction, the multitask setting can boost the performances, however without exceeding the results of the proposed formulation. In fact, our method attains the highest precision and F1 rates in all data set cases, while, in the SpaceNet7 case, all accuracy metrics reach the highest scores. Concerning the compared bitemporal multitask methods, they seem to have difficulty in successfully processing the additional temporal information, whereas the suggested method benefits greatly not only from the supplementary temporal features but also from the multitask setting, outperforming, at the same time, the rest of the time-series methods. Our method reports stable and higher performance on the precision rate and F1 score, contributing a lot to the lessening of FP detections.

Change detection applications suffer greatly from the numerous FP detections that arise from registration errors, illumination differences, or other types of change unrelated to the problem of interest. Hence, the development of proper formulas that overcome this obstacle becomes necessary. In the L-UNet case, the multitask setting contributes greatly to the amelioration of the results compared with the single-task change detection framework. Specifically, precision and F1 scores are greatly improved, benefiting both from the supplementary segmentation features and the circular losses. The superiority of the method has been proven in three completely different data sets with different spatial and temporal resolutions. In the future, we would like to investigate also the impact of more than two semantic categories on the proposed formulation both in terms of model complexity and evaluation performance.

One interesting direction that could be further investigated in the multitemporal setting is the integration of the time that the change has occurred in the time series. Determining the time point that an urban change first emerges can be very useful for tracking purposes, providing more thorough information for the frequency of urbanization. For our experiments, the *change* annotation of all the employed data sets describes the changes that have occurred between the first and last dates without exploiting additional information about the exact time of the change. Such a setup may lead to concerns as to whether a deep learning-based architecture can handle, in a constructive way, this irregular distribution of the changes through the different timestamps. Our extensive experiments indicated that the proposed approach boosts the performance when more dates are employed. This encourages us to draw the conclusion that the integration of LSTMs in the different encoding parts can provide meaningful temporal feature vectors independent of the specific time of change.

Given a time-series data set, if we want to determine the exact point of change using LSTMs, one solution can be to pass the multitemporal images to a convolutional LSTM encoder, as in [55], and produce detection scores for every LSTM cell output h_t [70], that is, passing h_t to a convolutional layer for classification and then applying a softmax function to extract the semantic probability map. If we can properly optimize such an approach, we will end up with distinct semantic maps for every single date, meaning that we can compose a tracking map and determine when each object appears for the first time. Another interesting approach is to concentrate on the internal activations of the LSTM cell, namely the operations of the internal gates and the information stored in the cell state. According to [71], the cell state is able to keep important information across consecutive time steps, concerning the changed parts of the current sequential image pair. Hence, with proper fine-tuning, it may be able to recognize the changes after every time step. For example, in [55], the activations of the cell state and the internal gates have helped the model to identify and discard the cloud regions, operating as filter mechanisms.

IV. Conclusion

In this article, a novel multitask learning framework for urban change detection is proposed where fully convolutional LSTM blocks are integrated on top of every encoding level of a U-Net, such as deep architecture, while an additional decoding branch is utilized for the semantic segmentation of the first and last employed dates. By using such a framework, temporal relationships are calculated for feature vectors of various resolutions without the need to downsample or flatten them. At the same time, the extra decoder provides supplementary information concerning semantic categories. Quantitative and qualitative analyses indicated that even though the problem of change detection can be very challenging due to illumination differences and registration errors, more constrained formulations and multitask deep learning frameworks based on LSTMs can provide very good tools for it. In fact, the proposed method contributed greatly to the lessening of FP values boosting significantly the precision and F1 rates in all data set cases. In the future, we plan to further evolve the proposed formula by performing simultaneously image registration and change detection, in order to eliminate parallax errors that tend to disorientate the learning process during training. In addition, we will explore ways to determine the change timestamps based on the operations of the LSTM networks. Finally, we will try to ameliorate and preserve the shape of the detected objects.

References

[1] A. Singh, “Review article digital change detection techniques using remotely-sensed data,” *Int. J. Remote Sens.*, vol. 10, no. 6, pp. 989–1003, Jun. 1989, doi: 10.1080/01431168908903939.

[2] K. Karantzalos, “Recent advances on 2D and 3D change detection in urban environments from remote sensing data,” in *Computational Approaches for Urban Environments* (Geotechnologies and the Environment), vol. 13. Cham, Switzerland: Springer, Nov. 2015, pp. 237–272, doi: 10.1007/978-3-319-11469-9_10.

[3] R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, “Image change detection algorithms: A systematic survey,” *IEEE Trans. Image Process.*, vol. 14, no. 3, pp. 294–307, Mar. 2005.

[4] L. Gomez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, “Multimodal classification of remote sensing images: A review and future directions,” *Proc. IEEE*, vol. 103, no. 9, pp. 1560–1584, Sep. 2015.

[5] N. Longbotham *et al.*, “Multi-modal change detection, application to the detection of flooded areas: Outcome of the 2009–2010 data fusion contest,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 1, pp. 331–342, Feb. 2012.

[6] J. Yang, P. J. Weisberg, and N. A. Bristow, “Landsat remote sensing approaches for monitoring long-term tree cover dynamics in semi-arid woodlands: Comparison of vegetation indices and spectral mixture analysis,” *Remote Sens. Environ.*, vol. 119, pp. 62–71, Apr. 2012.

[7] B. Liang and Q. Weng, “Assessing urban environmental quality change of Indianapolis, United States, by the remote sensing and GIS integration,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 4, no. 1, pp. 43–55, Mar. 2011.

[8] A. Taneja, L. Ballan, and M. Pollefeys, "City-scale change detection in cadastral 3D models using images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 113–120.

[9] P. Singh, Z. Kato, and J. Zerubia, "A multilayer Markovian model for change detection in aerial image pairs with large time differences," in *Proc. 22nd Int. Conf. Pattern Recognit.* Stockholm, Sweden: IEEE, Aug. 2014, pp. 924–929.

[10] C. Benedek, M. Shadaydeh, Z. Kato, T. Szirányi, and J. Zerubia, "Multilayer Markov random field models for change detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 107, pp. 22–37, Sep. 2015. [Online]. Available: https://hal.inria.fr/hal-01116609

[11] M. Volpi, D. Tuia, G. Camps-Valls, and M. Kanevski, "Unsupervised change detection with kernels," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 6, pp. 1026–1030, Nov. 2012.

[12] M. Vakalopoulou, K. Karatzalos, N. Komodakis, and N. Paragios, "Simultaneous registration and change detection in multitemporal, very high resolution remote sensing data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 61–69.

[13] M. Vakalopoulou, C. Platias, M. Papadomanolaki, N. Paragios, and K. Karantzalos, "Simultaneous registration, segmentation and change detection from multisensor, multitemporal satellite image pairs," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 1827–1830.

[14] J. S. Deng, K. Wang, Y. H. Deng, and G. J. Qi, "PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data," *Int. J. Remote Sens.*, vol. 29, no. 16, pp. 4823–4838, Aug. 2008.

[15] X. Li and A. G. O. Yeh, "Principal component analysis of stacked multi-temporal images for the monitoring of rapid urban expansion in the pearl river delta," *Int. J. Remote Sens.*, vol. 19, no. 8, pp. 1501–1518, Jan. 1998.

[16] X. Huang, Y. Cao, and J. Li, "An automatic change detection method for monitoring newly constructed building areas using time-series multi-view high-resolution optical satellite images," *Remote Sens. Environ.*, vol. 244, Jul. 2020, Art. no. 111802.

[17] X. Li, Y. Zhou, Z. Zhu, L. Liang, B. Yu, and W. Cao, "Mapping annual urban dynamics (1985–2015) using time series of landsat data," *Remote Sens. Environ.*, vol. 216, pp. 674–683, Oct. 2018.

[18] X.-P. Song, J. O. Sexton, C. Huang, S. Channan, and J. R. Townshend, "Characterizing the magnitude, timing and duration of urban growth from time series of landsat-based estimates of impervious cover," *Remote Sens. Environ.*, vol. 175, pp. 1–13, Mar. 2016.

[19] F. Pacifici and F. D. Frate, "Automatic change detection in very high resolution images with pulse-coupled neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 1, pp. 58–62, Jan. 2010.

[20] C. Pratola, F. D. Frate, G. Schiavon, and D. Solimini, "Toward fully automatic detection of changes in suburban areas from VHR SAR images by combining multiple neural-network models," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 4, pp. 2055–2066, Apr. 2013.

[21] S. Stent, R. Gherardi, B. Stenger, and R. Cipolla, "Detecting change for multi-view, long-term surface inspection," in *Proc. Brit. Mach. Vis. Conf.*, 2015, p. 127.

[22] J. Liu, M. Gong, K. Qin, and P. Zhang, "A deep convolutional coupling network for change detection based on heterogeneous optical and radar images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 3, pp. 545–559, Dec. 2016.

[23] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, "Change detection in synthetic aperture radar images based on deep neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 1, pp. 125–138, Jan. 2016.

[24] A. M. El Amin, Q. Liu, and Y. Wang, "Zoom out CNNs features for optical remote sensing change detection," in *Proc. 2nd Int. Conf. Image, Vis. Comput. (ICIVC)*, Jun. 2017, pp. 812–817.

[25] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep Siamese convolutional network for optical aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1845–1849, Aug. 2017.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[27] M. Vakalopoulou *et al.*, "Atlasnet: Multi-atlas non-linear deep networks for medical image segmentation," in *Proc. MICCAI*, 2018, pp. 658–666.

[28] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, "Very deep convolutional networks for text classification," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics, (EACLPapers)*, 2017, pp. 1–10.

[29] J. Y.-H. Ng, J. Choi, J. Neumann, and L. S. Davis, "ActionFlowNet: Learning motion representation for action recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1616–1624.

[30] S. Saha, Y. T. Solano-Correa, F. Bovolo, and L. Bruzzone, "Unsupervised deep transfer learning-based change detection for HR multispectral images," *IEEE Geosci. Remote Sens. Lett.*, May 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9089195

[31] S. Saha, L. Mou, X. X. Zhu, F. Bovolo, and L. Bruzzone, "Semi-supervised change detection using graph convolutional network," *IEEE Geosci. Remote Sens. Lett.*, Apr. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9069898

[32] M. Vakalopoulou, K. Karantzalos, N. Komodakis, and N. Paragios, "Building detection in very high resolution multispectral data with deep learning features," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 1873–1876.

[33] M. Papadomanolaki, M. Vakalopoulou, and K. Karantzalos, "A novel object-based deep learning framework for semantic segmentation of very high-resolution remote sensing data: Comparison with convolutional and fully convolutional networks," *Remote Sens.*, vol. 11, no. 6, p. 684, Mar. 2019.

[34] M. Papadomanolaki, M. Vakalopoulou, S. Zagoruyko, and K. Karantzalos, "Benchmarking deep learning frameworks for the classification of very high resolution satellite multispectral data," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 7, pp. 83–88, Jun. 2016.

[35] B. Huang *et al.*, "Large-scale semantic classification: Outcome of the first year of inria aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 6947–6950.

[36] N. Audebert, B. L. Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 20–32, Jun. 2018.

[37] M. Papadomanolaki, K. Karantzalos, and M. Vakalopoulou, "A multi-task deep learning framework coupling semantic segmentation and image reconstruction for very high resolution imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Yokohama, Japan, Jul. 2019, pp. 1069–1072.

[38] J. Li, X. Huang, and J. Gong, "Deep neural network for remote-sensing image interpretation: Status and perspectives," *Nat. Sci. Rev.*, vol. 6, no. 6, pp. 1082–1086, Nov. 2019, doi: 10.1093/nsr/nwz058.

[39] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proc. Nat. Acad. Sci. USA*, vol. 79, no. 8, pp. 2554–2558, 1982.

[40] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.

[41] A. Milan, S. H. Rezatofighi, A. R. Dick, I. D. Reid, and K. Schindler, "Online multi-target tracking using recurrent neural networks," in *Proc. AAAI*, 2016, pp. 4225–4232. [Online]. Available: https://dl.acm.org/doi/10.5555/3298023.3298181

[42] B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao, "A multi-stream bi-directional recurrent neural network for fine-grained action detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1961–1970.

[43] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[44] W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki, "Scene labeling with LSTM recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3547–3555.

[45] K. Ehsani, H. Bagherinezhad, J. Redmon, R. Mottaghi, and A. Farhadi, "Who let the dogs out? Modeling dog behavior from visual data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4051–4060.

[46] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 6, no. 2, pp. 107–116, Apr. 1998.

[47] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 677–691, Apr. 2017.

[48] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W. C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. NIPS*, 2015, pp. 1–12.

[49] J. Chen, L. Yang, Y. Zhang, M. S. Alber, and D. Z. Chen, "Combining fully convolutional and recurrent neural networks for 3D biomedical image segmentation," in *Proc. NIPS*, 2016, pp. 1–9.

[50] R. C. Daudt, B. L. Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral Earth observation using convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 2115–2118.

[51] R. C. Daudt, B. L. Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 4063–4067.

[52] L. Mou, L. Bruzzone, and X. X. Zhu, "Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 924–935, Feb. 2019.

[53] V. S. F. Garnot, L. Landrieu, S. Giordano, and N. Chehata, "Satellite image time series classification with pixel-set encoders and temporal self-attention," 2019, *arXiv:1911.07757*. [Online]. Available: http://arxiv.org/abs/1911.07757

[54] M. Rußwurm and M. Körner, "Self-attention for raw optical satellite time series classification," 2019, *arXiv:1910.10536*. [Online]. Available: http://arxiv.org/abs/1910.10536

[55] M. Rußwurm and M. Körner, "Multi-temporal land cover classification with sequential recurrent encoders," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 4, p. 129, Mar. 2018.

[56] M. Rußwurm and M. Körner, "Convolutional LSTMs for cloud-robust segmentation of remote sensing imagery," 2018, *arXiv:1811.02471*. [Online]. Available: http://arxiv.org/abs/1811.02471

[57] C. Tan, L. Zhao, Z. Yan, K. Li, D. Metaxas, and Y. Zhan, "Deep multi-task and task-specific feature learning network for robust shape preserved organ segmentation," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 1221–1224.

[58] D. Zhang, J. Han, L. Yang, and D. Xu, "SPFTN: A joint learning framework for localizing and segmenting objects in weakly labeled videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 475–489, Feb. 2020.

[59] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual task constrained deep Siamese convolutional network model," 2019, *arXiv:1909.07726*. [Online]. Available: http://arxiv.org/abs/1909.07726

[60] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[61] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Multitask learning for large-scale semantic change detection," *Comput. Vis. Image Understand.*, vol. 187, Oct. 2019, Art. no. 102783.

[62] M. Yang, L. Jiao, F. Liu, B. Hou, and S. Yang, "Transferred deep learning-based change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6960–6973, Sep. 2019.

[63] A. Radoi and M. Datcu, "Multilabel annotation of multispectral remote sensing images using error-correcting output codes and most ambiguous examples," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2121–2134, Jul. 2019.

[64] D. Marmanis, J. Wegner, S. Galliani, K. Schindler, M. Datcu, and U. Stilla, "Semantic segmentation of aerial images with an ensemble of CNSS," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 3, pp. 473–480, Jun. 2016.

[65] L. Mou, Y. Hua, and X. X. Zhu, "Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7557–7569, Apr. 2020.

[66] M. Papadomanolaki, S. Verma, M. Vakalopoulou, S. Gupta, and K. Karantzalos, "Detecting urban changes with recurrent neural networks from multitemporal Sentinel-2 data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Yokohama, Japan, Jul. 2019, pp. 214–217.

[67] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. 30th Int. Conf. Int. Conf. Mach. Learn. (ICML)*, vol. 28, 2013, pp. 1310–1318. [Online]. Available: http://dl.acm.org/citation.cfm?id=3042817.3043083

[68] A. Paszke *et al.*, "Automatic differentiation in PyTorch," in *Proc. NIPS Autodiff Workshop*, 2017, pp. 1–4.

[69] N. Teimouri, M. Dyrmann, and R. N. Jørgensen, "A novel spatio-temporal FCN-LSTM network for recognizing various crop types using multi-temporal radar images," *Remote Sens.*, vol. 11, no. 8, p. 990, Apr. 2019.

[70] Z. Ebrahimzadeh, M. Zheng, S. Karakas, and S. Kleinberg, "Deep learning for multi-scale changepoint detection in multivariate time series," 2019, *arXiv:1905.06913*. [Online]. Available: http://arxiv.org/abs/1905.06913

[71] H. Lyu, H. Lu, and L. Mou, "Learning a transferable change rule from a recurrent neural network for land cover change detection," *Remote Sens.*, vol. 8, no. 6, p. 506, Jun. 2016.

Maria Papadomanolaki received the master's degree in rural and surveying engineering from the National Technical University of Athens, Athens, Greece, in 2016, where she is pursuing the Ph.D. degree with Remote Sensing Laboratory.

She is also collaborating with the Laboratoire mathématiques et informatique pour la complexité et les systémes (MICS Laboratory), CentraleSupélec, Gif-sur-Yvette, France.

Maria Vakalopoulou received the Diploma degree in engineering and the Ph.D. degree in computer science and remote sensing from the National Technical University of Athens, Athens, Greece, in 2012 and 2017, respectively.

She has been an Assistant Professor of artificial intelligence with the Laboratory of Mathematics and Computer Science, CentraleSupélec, Gif-sur-Yvette, France, since 2019. She is with the School of Engineering, University Paris-Saclay, Gif-sur-Yvette. Her research interests include machine and deep learning, medical imagery, and remote sensing. The researcher has published her research in top-rank international journals and conferences.

Dr. Vakalopoulou has received a number of awards for her research contributions. She has served as an Area Chair in multiple conferences and as a Reviewer in multiple journals and conferences in the fields of medical imaging, remote sensing, and artificial intelligence.

Konstantinos Karantzalos (Senior Member, IEEE) received the Diploma degree in engineering from the National Technical University of Athens (NTUA), Athens, Greece, in 2000, and the Ph.D. degree from NTUA in collaboration with Ecole Nationale de Ponts et Chaussees (ENPC), Champs-sur-Marne, France, in 2007.

In 2007, he joined the Department of Applied Mathematics, Ecole Centrale de Paris (ECP), Gif-sur-Yvette, France as a Postdoc. He is an Associate Professor of remote sensing with the National Technical University of Athens. His teaching and research interests include geoscience and earth observation, geospatial data analytics, spectral data analysis, and machine learning with applications in, e.g., environmental monitoring and precision agriculture. He has several publications in top-rank international journals and conferences and a number of awards and honors for his research contributions. He serves on the Board of Directors of the Greek Space Center.