Anomaly Feature Learning for Unsupervised Change Detection in Heterogeneous Images: A Deep Sparse Residual Model

Redha Touati, Max Mignotte, and Mohamed Dahmane

***Abstract*—In this article, we propose a novel and simple automatic model based on multimodal anomaly feature learning in a residual space, aiming at solving the binary classification problem of temporal change detection (CD) between pairs of heterogeneous remote sensing images. The model starts by learning from image pairs the *normal* existing patterns in the before and after images to come up with a suitable representation of the normal (nonchange) class. To achieve this, we employ a stacked sparse autoencoder trained on a large number of temporal image features (training data) in an unsupervised manner. To classify pixels of new unseen image-pairs, the built anomaly detection model reconstructs the input from its representation in the latent space. First, the probe (new) image (i.e., the bitemporal heterogeneous image pair as the input request) is encoded in this compact *normal* space from a stacked hidden representation. The reconstruction error is computed using the *L2* norm in what we call the residual normal space. In which, the nonchange patterns are characterized by small reconstruction errors as a normal class while the change patterns are quantified by high reconstruction errors categorizing the abnormal class. The dichotomic (changed/unchanged) classification map is generated in the residual space by clustering the reconstructed errors using a Gaussian mixture model. Experimental results on different real heterogeneous images, reflecting a mixture of imaging and land surface CD conditions, confirm the robustness of the proposed anomaly detection model.**

***Index Terms*—Anomalous patterns, change detection (CD), deep learning, feature space reconstruction, heterogeneous remote sensing, multimodal anomaly detector, reconstruction error, sparse autoencoder.**

I. Introduction

NOWADAYS, detecting changes between images of the same geographical area over time is still an active topic in remote sensing image processing. A less explored problem is the multimodal change detection (CD) which is a challenging task that can be viewed as the generalization of the classical monomodal CD problem [1]–[5]. This research area became active with the launch of new satellite generations with different sensor characteristics. Definitely, the exploitation of heterogeneous multimodal data is important to increase the accuracy of any CD system. The existing monomodal systems are not usable as is and need to be adapted to solve the CD problems for environmental monitoring, deforestation, geological resources survey, disaster localization and quantification, and urban planning, to name a few.

Multimodal CD [6] is a data analysis procedure seeking directly to locate area of change that may have occurred between two heterogeneous satellite images acquired in the same region of interest at different times. Practical and technical advantages of this recent CD procedure have generated a growing interest, in the remote sensing research community since it should be more robust to natural changes due to environmental variables such as humidity or phenological state. The issue caused by the environmental variables can be avoided when comparing images coming from different sources (i.e., multimodal images) CD based on multimodal images (heterogeneous) generally refers to differences in two imaging modes in which acquired images are represented in two distinct feature spaces that do not share the same statistical properties. It is a nontrivial problem since it is subject to less stringent requirements about the source and characteristics of the acquired data, hence, leading to radically different image statistics that cannot be compared directly from traditional CD techniques.

To date, the multimodal CD issue has been addressed by few works, that can be grouped into five categories in which we can find parametric models [7]–[10] that use a set of parametric multidimensional distributions (mixture), nonparametric methods [11] which aim to minimize an energy model to satisfy an overdetermined set of constraints, algorithms based on operators using spatial and temporal similarity measures as in [12]–[14], projection-based techniques that try to map the two heterogeneous images to a common feature space where traditional monomodal CD can be applied [15]–[18], and, finally, machine learning methods [19]–[22].

1) In the first category of parametric methods, a set of multivariate distributions is used for common modeling

Manuscript received January 4, 2019; revised March 18, 2019, June 20, 2019, and November 15, 2019; accepted December 16, 2019. Date of publication January 22, 2020; date of current version February 19, 2020. This work was supported in part by the Computer Research Institute of Montreal (CRIM) and in part by the Ministry of Economic Science and Innovation (MESI) of the Government of Québec. *(Corresponding author: Redha Touati.)*

R. Touati is with the Vision Laboratory of the Département d'Informatique et de Recherche Opérationnelle (DIRO), Faculté des Arts et des Sciences, Université de Montréal, Montreal, QC H3C 3J7, Canada, and also with the R&D Vision Département, Centre de Recherche Informatique de Montréal (CRIM), Montreal, QC H3N 1M3, Canada (e-mail: touatire@iro.umontreal.ca).

M. Mignotte is with the Vision Laboratory of the Département d'Informatique et de Recherche Opérationnelle (DIRO), Faculté des Arts et des Sciences, Université de Montréal, Montreal, QC H3C 3J7, Canada (e-mail: mignotte@iro.umontreal.ca).

M. Dahmane is with the R&D Vision Département, Centre de Recherche Informatique de Montréal (CRIM), Montreal, QC H3N 1M3, Canada (e-mail: mohamed.dahmane@crim.ca).

Digital Object Identifier 10.1109/JSTARS.2020.2964409

dependencies between the two heterogeneous images acquired from different types of multimodality [7]–[10].

2) In the nonparametric methods, we can mention the energy-based model in the least-squares sense designed to satisfy an overdetermined constraint scenario which models each pair of pixels from the before and after multimodal images [11].
3) The methods of the third family try, first, to estimate the correspondence between the same existing points in the before and after heterogeneous images, and then to identify and detect the areas of change between the two multimodal images, using invariant similarity measures by imaging modality (such as correlation and mutual information) [12]–[14].
4) In the fourth category, projection techniques are used to transform the two multimodal images into a new common feature space, in which the before and after heterogeneous images share the same statistical properties, and on which classical monomodal CD methods can then be exploited [15]–[18].
5) In the last category, relying on machine learning methods, Merkle *et al.* [19] used an unsupervised generative adversarial network consisting of a generator stream that produces a binary map and a discriminator stream that tries to discriminate between the result of the generator and the result of a binarization algorithm. Liu *et al.* [20] try, first, to train a couple of convolutional neural networks in order to transform the two multimodal images in a new feature space allowing to calculate a difference map. In the second step, they apply a thresholding image processing algorithm to detect changes vs. nonchanges area from the resulting difference map. In the same vein as [20], Zhao *et al.* [22] proposed to employ a symmetric neural network composed of a restricted Boltzmann machine, whose parameters are updated based on the clustering result. Zhang *et al.* [21] based their approach on a denoising autoencoder network and used selected features of the difference image to build the network.

Let us note that the *parametric techniques* by construction the parametric models suffer from the fact that they are not easily generalizable for other pairs of different sensors. They have been specially designed via specific distributions for a given type of multimodal sensors (e.g., optical/SAR). Whereas the *nonparametric techniques* have the ability to process a wide variety of imaging modalities but they are possibly less accurate than specific heterogeneous CD models that deal with specific types of multimodality. The third and fourth family of techniques are the simplest, mathematically speaking, and also the most local, in terms of modeling. Because of their modeling (modeling associated with a neighborhood), they may have the disadvantage of making it more difficult to conceive of possible improvements and to understand how these changes would improve them. The efficiency of machine learning-based CD models depends on the availability of an adequate massive amounts of representative training data, sometimes manually selected and carefully chosen.

As the most advanced form of machine learning, deep learning was used for feature-based learning. For instance, a deep autoencoder neural network has been proposed to realize unsupervised feature learning in order to learn discriminative and effective features from a large amount of unlabeled data. The sparse autoencoders have been widely studied for feature-based deep learning methods [23], [24], as it is highly effective for finding high-level representations of complex data. In our case, the multimodal CD problem can be viewed as a binary classification task in which the *change* class or region refers to a set of pixel pairs (or instances), extracted from heterogeneous image pairs, that stand out as being different from all others. Such instances can be seen as anomalies that are indicative of a particular underlying process under the assumption that there are no errors generated from the sensor. Hence, the change class refers, practically, to different semantic regions from the same geographical area that is seen through two different imaging modalities. This anomaly detection problem can be efficiently solved using sparse autoencoder since it has the appealing ability to uncover potential anomalies in unlabeled data [25].

In this article, we propose a new unsupervised CD model which belongs to the machine learning category, a deep learning solution that aims to define a model that tries to map, in a residual space, the changes as anomalies using a stacked layerwise sparse autoencoders which ensure the encoding and decoding stages with a well-adapted neural transfer function, for processing and detecting changes from multimodal remote sensing images coming from different sources and under different spatial image resolutions. Compared to the state-of-the-art methods, the proposed CD model is defined to be more robust to model the class changes as anomalies, thanks to its flexible learning architecture which is also well adapted to process new unseen pair of heterogeneous image inputs (as anomalies) in the absence of *annotated* data. Our proposal is to modelize in a residual space the changes as anomalies. More precisely, we propose an unsupervised anomaly-based heterogeneous CD modeling based on learning image features from deep sparse autoencoder neural network as a multimodal feature extractor to gather useful image features from the usual image patterns (nonchange or normal class) existing in the before and after multimodal images in the absence of labels. The built anomaly detection model utilizes a reconstruction error vector to perform anomaly detection. To analyze a new unseen image-pairs, the model projects the input into a new latent space from which it attempts to map the projected representation back to reconstruct the input. The residual difference between the original input and the reconstructed one defines our residual space. A Gaussian mixture model (GMM) is then used to model the extracted features in this space to separate normal from anomalous patterns corresponding, respectively, to nonchange and change class labels. The advantage of the proposed CD model lies in its flexibility to process a nonspecific source type, such as multisensor, multisource, or multilooking SAR image pairs, avoiding the drawbacks of parametric models which require knowledge of the conditional distributions; and the disadvantage of supervised machine learning models is that they often require labeled and well-balanced training data. The

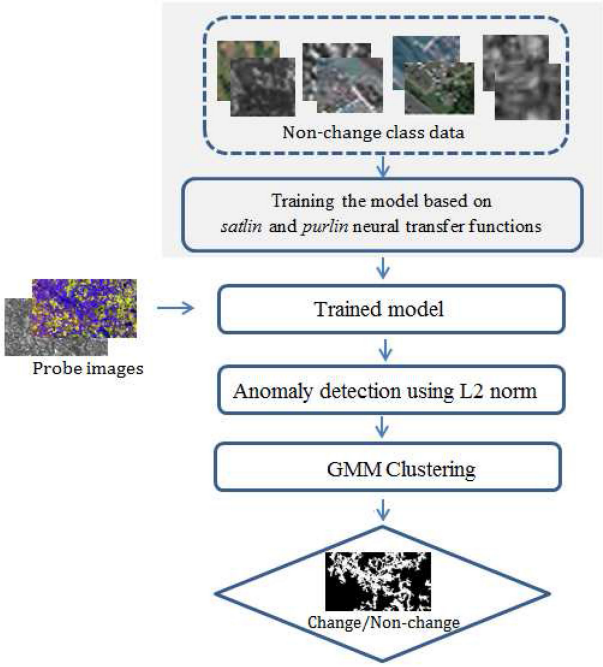

Fig. 1. Main steps of the proposed residual space-based change detection model.

main advantage of our model lies on its ability to learn the underlying latent space.

The rest of this article is organized as follows. Section II presents the proposed residual CD model and its architecture, which allows us to learn and to reconstruct a suitable representation (feature anomaly space), from which changed and unchanged areas are then identified as normal/abnormal classes. Section III describes the experimental framework used to evaluate the performance of the proposed CD model, and a set of experimental results compared to the state-of-the-art multimodal change detectors. Section IV concludes the article.

II. Proposed Change Detection Model

Let us assume two multimodal remote sensing images acquired before and after a given event in the same geographical area and also let us consider that the acquired images are coregistered. In order to estimate a binary change detection map which is supposed to represent the difference between the two temporal heterogeneous images, we rely on unsupervised reconstruction machine learning model designed especially to model the change class as anomalies in our CD problem in order to detect different possible change events such as floods and urban growing.

The proposed anomaly-based CD model takes as input a combination of a variety of multimodal remote sensing images, as a combination of two optical images, SAR/optical or optical/SAR images, or SAR images with different number of looks. The pixels in those images cannot be directly compared. The model is composed of two major parts: an unsupervised learning sparse-based modeling step, where a training phase is performed to learn a robust deep sparse change detector; and a binary clustering step, where a maximum *A Posteriori* criteria is used for data clustering (see Fig. 1). More precisely, in the training phase, the architecture of our CD model is based on stacked sparse autoencoder with a depth of two sparse layers, where each single sparse layer has an encoder layer with a corresponding decoder layer. Based on the proposed architecture, our CD model takes as input a temporal *normal* feature space and try to learn encoder–decoder layers using a layer-wise training technique in which each sparse layer is trained independently in an unsupervised manner. The internal and optimal values of the deep CD model parameters (prior) are predetermined using a grid search method (see Section III-C). The temporal *normal* feature space is fed to the first single-layer sparse autoencoder which was trained to extract low-level feature representations from its hidden layer. The lower level features are then used to train the second sparse autoencoder where high-level features are given by its hidden layer (the second layer) of the stacked sparse autoencoder. The encoder layer encodes the input in a compact representation, while the decoder layer ensures to predict the encodings in order to reconstruct an estimate of the original input. Once the training phase is accomplished, the built encoder–decoder layers ensure, respectively, the mapping of new input feature space in a compressed space and then the reconstruction of the original space from this compact representation. The reconstruction error between the input features and their reconstructed versions is then computed using the *L2 norm*. A clustering step is achieved in the residual space to generate, as output, two clusters of data (change versus nonchange) related to our bitemporal CD problem.

A. Unsupervised Learning Sparse Model

The anomaly-based CD problem aims at identifying the (usually rare) differences of ground features existing locally between two bitemporal heterogeneous images, acquired over the same geographical area, with two different imaging modalities (let us assume that the two remote sensing images are coregistered). It may be considered as a binary classification task in which the (small) local spatial changes, over the time, are potential indicatives of something that have truly changed in the area of interest and which can thus be identified as anomalies (i.e., different data seen through two different imaging modalities). More precisely, *anomalous patterns* are referred as patterns in the data that do not conform to a well-defined notion of normal behavior [26]. A common strategy to extract anomalies is to reduce the high-dimensional input space in lower-dimensional space and then apply a set of distance metrics within the reduced space in order to identify the anomalies [27].

To this end, supervised classification approaches require labeled and often well-balanced training data or, more generally, a preprocessing stage such as data augmentation to train a classifier model. In heterogeneous CD problem, especially in remote sensing imagery, training data are generally less available, unlabeled, and often highly unbalanced. Besides, data augmentation may be harder since the binary class *change* and *nonchange* are highly imbalanced over the whole acquired data.

In our CD problem, it is important to recall that the *changed* regions are smaller than the unchanged regions since a significant event (such as flooding and earthquake) occurs rarely and are thus very localized in time and space. Consequently, we have to

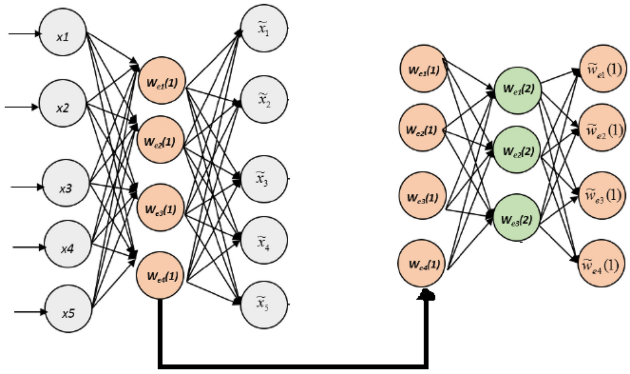

Fig. 2. Stacked autoencoder neural network composed of two layers of sparse autoencoders.

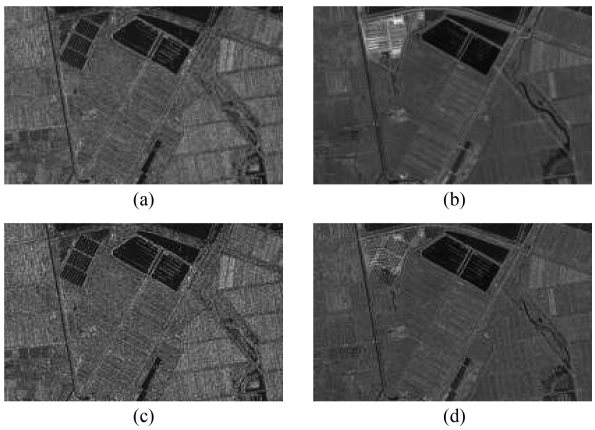

Fig. 3. (a) and (b) Original SAR/optical images. (c) and (d) Reconstructed images.

rely on machine learning-based binary classification method in which the training phase is only performed on patterns belonging to the predominant class (the *nonchange majority* class label in our case) while keeping robust to detect the minority class, i.e., the rare events belonging to the *change* class as anomalies during the test phase.

Among the existing machine learning-based strategies, the reconstruction-based methods, using sparse autoencoders, seem particularly well adapted to our heterogeneous CD problem. Its main ability is to learn, in the least square sense, a compressed representation minimizing the reconstruction error of the two imaging modalities in the residual space and to estimate within this space the reconstruction error of each bitemporal input patterns from local gray-level distribution as a reliable anomaly score. This score can then be exploited to identify the abnormal (rare) patterns caused by a given event (defining the *change* label) and the normal unchanged patterns belonging to the *nonchange* class label.

To build our abnormal pattern-based model, we propose to learn a stacked constrained neural network model which can be trained with a layer-wise training procedure [28] in order to find a good representation for the input space [29], [30] and also to better reconstruct the normal patterns based on the learned multimodal imaging representation [31] (see Figs. 2 and 3). More precisely, we propose to use a stacked sparse autoencoder, which offers an unsupervised reconstruction framework consisting of multiple layers of sparse autoencoders, and which turns out to be robust to discover interesting structures from input image data. This allows us to build a robust anomaly CD model to identify with a high error the unusual and abnormal features (see Fig. 3). Let us note that deep learning methods, including deep autoencoder, have been applied to learn cross-modality and multimodal features. In particular, AECs (autoencoders) are able to fuse highly heterogeneous pairs of data types, such as text mixed with images, or audio-linked with video, and even combining facial expressions with sound, to name a few [27], [30], [32]–[43]. Hence, this work defines a novel application of deep networks to learn from heterogeneous normal patterns, a common space representation, and also an appealing strategy to reconstruct or fuse different imaging modalities within an unsupervised feature-based learning strategy [27], [32], [38].

It is important to remember that the intrinsic problems of the standard autoencoder model make it inefficient [44], [45]. Sparse autoencoder is a constrained model that can learn relatively sparse features by introducing a sparse penalty term inspired by the sparse coding [44] into the autoencoder. Putting constraints on the autoencoder neural network aims to encourage the sparsity of the model [44], [46], and can improve the performance relative to the traditional autoencoders [44], [45]. This can be simply achieved by adding a sparse penalty term to the cost function of the hidden layer to control the number of *active* neurons. Hence, the cost function we used in our case for training the anomaly-based deep sparse model is composed from [45]:

1) Sparsity Regularization Term: Sparsity regularization tends to create specialized neurons that focus on particular subset from the training data by increasing the number of inactive neurons. The average activation of each hidden neuron $\hat{\rho}_i$ is expected to be close to a small value, and each hidden neuron activation is expected to be close to zero, and thus the neurons of the hidden layer becomes *inactive*. To achieve this, the sparsity term is added to the objective function that penalizes $\hat{\rho}_i$ if it deviates significantly from a predefined small number ρ. The sparsity penalty term Ω_{sparsity} is employed as in [47], which attempts to impose a constraint on the sparsity of the output from the hidden layer. It is defined by

$$\Omega_{\text{sparsity}} = \sum_{i=1}^{D} \rho \log\left(\frac{\rho}{\hat{\rho}_i}\right) + (1-\rho)\log\left(\frac{1-\rho}{1-\hat{\rho}_i}\right) \quad (1)$$

where $\hat{\rho}_i$ is the average activation value for the ith hidden layer unit and D represents the number of neurons in the hidden layer. The sparsity penalty term constrains the value of $\hat{\rho}_i$ to be close to ρ according to the Kullback–Leibler divergence. This penalty function possesses the property that Kullback–Leibler divergence $\text{KL}(\rho_{\|}\hat{\rho}_i) = 0$ if $\hat{\rho}_i{=}\rho$. Otherwise, it increases monotonically as $\hat{\rho}_i$ diverges from ρ.

2) L2 Regularization Term: The L2 regularization term Ω_{weights} is added to keep the weight magnitudes small during the feature learning stage in order to prevent overfitting. It is defined as follows:

$$\Omega_{\text{weights}} = \frac{1}{2}\sum_{l}^{L}\sum_{j}^{N}\sum_{i}^{k}\left(\omega_{ji}^{(l)}\right)^2 \quad (2)$$

where $\omega_{ji}^{(l)}$ represents the weight, L is the number of hidden layers, N is the number of observations, and k is the number of variables in the training data.

3) Cost Function: The anomaly CD model is based on training an unsupervised sparse neural network whose cost function is an adjusted mean squared error function defined by (3) [45]. In our work, we propose to use a more robust encoding–decoding neural transfer functions (4) and (5) that better mitigate the convergence problem, and improve the performance of our CD model.

$$E = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{k}(x_{kn} - \hat{x}_{kn})^2 + \lambda \cdot \Omega_{\text{weights}} + \beta \cdot \Omega_{\text{sparsity}} \quad (3)$$

where x_{kn} is the input vector and $\hat{x}_{kn}$ is an estimate of the input vector x_{kn}. The coefficients λ and β control, respectively, the importance of the regularization and the sparsity terms.

4) Transfer Functions: To make our change detector more effective for anomaly detection, we make use of the positive saturating linear transfer function for the encoding stage, and the linear transfer function for the decoding stage. Each encoder layer has a corresponding decoder layer:

$$f_{\text{Enc}}(z) = \begin{cases} 0, & \text{if} \quad z \leq 0 \\ z, & \text{if} \quad 0 < z < 1 \\ 1, & \text{if} \quad z \geq 1 \end{cases} \quad (4)$$

$$f_{\text{Dec}}(z) = z. \quad (5)$$

The encoder maps the input representation x to another encoded representation as follows:

$$z_{\text{Enc}} = f_{\text{Enc}}^{(l)}(W^{(l)}x + b^{(l)}) \quad (6)$$

where $W^{(l)}$ is a weight matrix, and $b^{(l)}$ is a bias vector of the encoding layer.

The decoder maps the encoded representation z_{Enc} to reconstruct an estimate of the original input representation by

$$\hat{x} = f_{\text{Dec}}^{(l)}(W^{(l)}z_{\text{Enc}} + b^{(l)}) \quad (7)$$

where $W^{(l)}$ is a weight matrix, and $b^{(l)}$ is a bias vector of the decoding layer.

B. Binary Clustering

In this approach, we have formulated the heterogeneous CD problem into a learning-based reconstruction problem in which the learned constrained stacked sparse model uses its stacked hidden representation to map or reconstruct each new input image pattern. Given a new heterogeneous remote sensing image pair, we have thus to, first, compute the reconstruction error for each pixel (or for each feature vector centered on this pixel) occurring at the same position in the before and after image pair. The reconstruction error, between the feature vector expressed in the input feature space and the reconstructed space, is then measured in the L_2 norm sense and the pixels belonging to the *change* class label are then simply identified by their high abnormal reconstruction error.

Based on the reconstruction error, the automatic clustering of the residual space can be performed by a thresholding technique or a k-means-based classification strategy ($k = 2$). Another strategy, less sensitive to false alarms or the a priori assumption of two spherical class label datasets with the same radius (in the case of the k-means procedure), consists in estimating the parameters of a mixture of two Gaussians in the residual space with the EM algorithm. The MAP rule based on these mixture parameters is used as final binary decision to assign a normal class label to the *nonchange* class *and* the abnormal class label to the *change* class. Algorithm 1 shows the predictions of the CD model on the new unseen data.

Algorithm 1: Prediction Steps of the CD Model.

Step 1:

- $\hat{x} \leftarrow$ reconstruct a new input feature space (test) x using the built deep sparse model with the optimal parameter

foreach $\hat{x}_i \in$ *reconstructed space* $\hat{x}$ **do**

- $e_i \leftarrow$ compute the reconstruction error between x_i and $\hat{x}_i$ using the L2 norm

end

Step 2:

- Perform a clustering stage on e_i

III. Experimental Results

In order to validate and to show the strength of the proposed model to process both different imaging modality cases and CD conditions along with different spatial resolutions, we have conducted our study on 11 real heterogeneous image pairs with different kinds of modalities, namely multisensor (heterogeneous optical images), multisource (optical and SAR images), and multilooking (heterogeneous SAR images), in which the change mask (ground-truth) is provided for each heterogeneous dataset by a photo-interpreter.

In our application, we use the leave-one-out test scenario to evaluate the performance of the proposed CD model. In this well-known procedure, we remove one entire dataset from the 11 heterogeneous datasets and we train the model on the remaining heterogeneous datasets. The output of the trained model is then used to classify the removed dataset. We repeated this process 11 times, and at each time, we resort to the two heterogeneous images to be our test example.

A. Heterogeneous Dataset Description

- The first multimodal dataset is a pair of SAR/optical satellite images (Toulouse, France), with size 4404×2604 pixels, before and after construction. The SAR image was taken by the TerraSAR-X satellite (Feb. 2009) and the optical image by the Pleiades (High-Resolution Optical Imaging Constellation of CNES, Centre National d'Etudes Spatiales) satellite (July 2013). The TSX image was coregistered and resampled by [49] with a pixel resolution of 2 m to match the optical image.

- The second one is a pair of optical/SAR satellite images (Gloucestershire region, in southwest England, near Gloucester), with size 2325×4135 pixels, before and after a flooding taking place in an urban and a rural area. The optical image comes from the Quick Bird 02 (QB02) VHR satellite (15 July 2006) and the SAR image was acquired by the TerraSAR-X satellite (July 2007). The TSX image presents a resolution of 7.3 m and the QB02 image (with resolution of 0.65 m and 0% cloud cover) was coregistered and resampled by [49] to match this resolution.
- The third dataset shows two heterogeneous optical images acquired in Toulouse (Fr) area by different sensor specifications (size 2000×2000 pixels with a resolution of 0.5 m). The *before* image is acquired by the Pleiades sensor in May 2012 before the beginning of the construction work, and the *after* image is acquired by WorldView2 satellite from three (Red, Green, and Blue) spectral bands (11 July 2013) after the construction of a building. The WorldView2 VHR-image was coregistered by [49] to match the Pleiades image.
- The fourth dataset [8] is a pair of SAR/SAR satellite images (Gloucester, U.K.) before and during a flood event caused by intense and prolonged rainfall, overwhelming the drainage capacity, on urban and agricultural/rural areas, with size 762×292 pixels, acquired by RADARSAT satellite with different number of looks. The number of looks for the before SAR image is 1-look image (Sep. 2000) and the number of looks for the after image is 5-looks (Oct. 2000). These two SAR images have a resolution of about 40 m.
- The fifth dataset [51], [53] consists of one multispectral image and one SAR image showing the area of Gloucester (U.K.), with a size of 1318×2359 pixels. The multispectral image is taken by the Spot VHR satellite on Sep. 1999 before a flooding event. The SAR image is captured by the European Remote Sensing (ERS) satellite (around Nov. 2000) during the flooding event. The resolution of these two images are about 10 m [53].
- The sixth dataset consists of one SAR image and one SPOT image with the same size of 330×590 pixels. The ERS image is acquired on Nov. 16, 1999 before the flood in Gloucester, U.K., and the optical image combined with three bands is acquired on Oct. 21, 2000 during the flood in Gloucester U.K.
- The seventh dataset is composed of two heterogeneous optical images. It shows the changes of the Mediterranean in Sardinia area (Italy). This dataset is acquired by different sensor specifications, and consists of one TM image and one optical image. The TM image is the near-infrared band of the Landsat-5 (Sep. 1995 with a spatial resolution of 30 m). The optical image comes from Google Earth (RGB, Jul. 1996, Landsat-5) with a spatial resolution of 4 m. After coregistration, they are of the same pixel-resolution 412×300 pixels.
- The eighth dataset shows two heterogeneous optical images from another area in the south campus of Hubei province of China were, respectively, acquired by the QuickBird satellite in May 2002 and the IKONOS satellite in July 2009, with a size of 240×240 pixels. The images after preprocessing have the same spatial resolution of 3.28 m.
- The ninth dataset is a pair of SAR/optical satellite images with a size of 291×343 pixels. The before image is acquired by RADARSAT-2 in June 2008 over the River of China. The optical image comes from Google Earth (Sep. 2010), acquired after a flooding event, and which integrates imagery from both Quickbird US VHR satellite and SPOT5 satellite. After, coregistration, they are of the same spatial resolution of 8 m.
- The tenth dataset shows two heterogeneous optical images covering the campus of Wuhan University in Hubei province of China. They were, respectively, acquired by the QuickBird satellite in April 2005 and the IKONOS satellite in July 2009, and correspond to four bands (red, green, blue, and NIR band) with a size of 400×400 pixels. The resolution of these images is of 2.44 and 3.28 m. After resampling, the after image have the same spatial resolution as the before image, 2.44 m.
- The eleventh data set consists of one SAR image and one RGB optical image. It shows a piece of the Dongying City in China, before and after a new building construction. The SAR image is acquired by RADARSAT-2 (Jun. 2008) with a spatial resolution of 8 m. The optical image comes from Google Earth image (Sep. 2012) with a spatial resolution of 4 m [20]. After coregistration, they are of the same pixel-resolution to give a size of 921×593 pixels.

B. Results and Evaluation

In our anomaly-based CD problem, we first convert the multibands image to a grayscale image; the temporal feature image space is simply done by collecting the local gray-level intensities using a squared window of size Sw ($Sw = 9$ in our case).

We have used an architecture composed of a stacked sparse autoencoder and consisting of two layers of sparse autoencoders, where each encoder layer has a corresponding decoder layer, a deep sparse autoencoder with a number of hidden layers $L_h = 2$, that takes a bitemporal feature vector input of dimension $D_{\text{inp}} = 162 (= 2 \times 9 \times 9)$. The learned encoder layers compress the input space into a low-dimensional representation, first into a number of dimensions $d_{hl_1} = 80$ and then into a number of dimensions $d_{hl_2} = 40$. The reconstruction of this compact representation of dimension $d_{hl_2} = 40$ is done by using the two previously learned decoder layers, respectively, from $d_{hl_2} = 40$ to $d_{hl_1} = 80$, and from 80 to the original input dimension $\hat{D}_{\text{inp}} = 162$. We recall that in this learning architecture, we use the *satlin* function for the encoding stage and the *purelin* function for the decoding stage.

Our anomaly-based CD model can be optimized via a layer-wise training technique [28], using a scaled conjugate gradient descent algorithm [54], by starting to train the first layer to learn to encode the normal representation $D^{(N)}$ to $d^{(hl_1)}$ and to decode $D^{(N)}$ from $d^{(hl_1)}$, and then to train the second layer to learn to encode $d^{(hl_1)}$ to $d^{(hl_2)}$ and to decode $d^{(hl_2)}$ from $d^{(hl_1)}$.

In our application, the coefficients λ and β for the $L2$ regularization and the sparsity regularization terms were fixed,

TABLE I

ACCURACY RATE OF CHANGE DETECTION ON THE 11 HETEROGENEOUS DATASETS OBTAINED BY THE PROPOSED METHOD AND THE STATE-OF-THE-ART MULTIMODAL CHANGE DETECTORS (FIRST UPPER PART OF EACH TABLE) AND MONOMODAL CHANGE DETECTORS (SECOND LOWER PART OF EACH TABLE)

SAR/Optical Dataset	Accuracy (%)
Proposed method	**0.892**
Prendes *et al.* [48]	0.844
Correlation [48]	0.670
Mutual Inf. [48]	0.580

Optical/SAR Dataset	Accuracy (%)
Proposed method	**0.961**
Prendes *et al.* [10], [49]	0.918
Prendes *et al.* [9]	0.854
Copulas [7], [9]	0.760
Correlation [7], [9]	0.688
Mutual Inf. [7], [9]	0.768
Pixel Dif. [9], [50]	0.782
Pixel Ratio [9], [50]	0.813

Optical/Optical Dataset	Accuracy (%)
Proposed method	**0.880**
Prendes *et al.* [48], [49]	0.844
Correlation [48], [49]	0.679
Mutual Inf. [48], [49]	0.759
Pixel Dif. [49], [50]	0.708
Pixel Ratio [49], [50]	0.661

SAR 1-look / SAR 5-looks Dataset	Accuracy (%)
Proposed method	**0.814**
Chatelain *et al.* [8]	0.732
Correlation [8]	0.521
Ratio edge [8]	0.382

VHR Optical/SAR Dataset	Accuracy (%)
Proposed method	**0.780**
Gregoire *et al.* [51]	0.70

ERS/Spot Dataset	Accuracy (%)
Proposed method	**0.836**
Liu *et al.* [17]	0.818
Liu *et al.* [17]	0.655

SAR/Optical Dataset	Accuracy (%)
Proposed method	0.767
PCC [20]	0.961
SCNN without pre-training [20]	0.958
SCNN with 1 coupling layer [20]	0.964
SCNN with 2 coupling layer [20]	0.969
SCNN with 3 coupling layer [20]	**0.977**
Zhao *et al.* [22]	0.974

SAR/Optical Dataset	Accuracy (%)
Proposed method	**0.980**
Zhao *et al.* [22]	0.979
Liu *et al.* [20]	0.976
SCNN [22]	0.952
PCC [20]	0.821

Optical(NIR band)/Optical Dataset	Accuracy (%)
Proposed method	0.929
Zhang *et al.* [21]	**0.975**
PCC [21]	0.882

Quickbird/IKONOS Dataset	Accuracy (%)
Proposed method	0.847
Yuqi *et al.* [52]	0.986
Multiscale [52]	**0.991**

Quickbird/IKONOS Dataset	Accuracy (%)
Proposed method	0.817
Yuqi *et al.* [52]	0.959
Multiscale [52]	**0.966**

TABLE II

CONFUSION MATRIX IN TERMS OF NUMBER OF PIXELS AND PERCENTAGE FOR THE 11 MULTIMODAL DATASETS, I.E., [TSX/PLEIADES] (4404×2604 PIXELS), [QB02/TSX] (2325×4135 PIXELS), [PLEIADES/WORLDVIEW 2] (2000×2000 PIXELS), [SAR 1-LOOK / SAR 5-LOOKS] (762×292 PIXELS), [SPOT VHR/ ERS] (1318×2359 PIXELS), [ERS/SPOT] (330×590 PIXELS), [MS (NIR)/MS] (412×300 PIXELS), [QB02 /IKONOS] (240×240 PIXELS), [SAR/OPTICAL] (291×343 PIXELS), [QB02 /IKONOS] (400×400 PIXELS), [SAR/OPTICAL] (921×593 PIXELS)

Multimodal image pairs	**TP**	**TN**	**FP**	**FN**
TSX/Pleiades	440211 (48.2%)	9791031 (92.8%)	764001 (7.2%)	472773 (51.8%)
QB02/TSX	419342 (68.0%)	8819894 (98.0%)	177191 (2.0%)	197448 (32.0%)
Pleiades/WorldView 2	339464 (56.0%)	3183160 (93.8%)	210542 (6.2%)	266834 (44.0%)
SAR 1-look/SAR 5-looks	26544 (68.1%)	154679 (84.3%)	28871 (15.7%)	12410 (31.9%)
VHR Spot/ERS	480846 (70.4%)	1946913 (80.2%)	479675 (19.8%)	201728 (29.6%)
ERS/spot	13703 (57.2%)	149187 (87.4%)	21555 (12.6%)	10255 (42.8%)
MS (NIR band) /MS	6353 (83.9%)	108577 (93.6%)	7451 (6.4%)	1219 (16.1%)
Quickbird/IKONOS	4689 (54.3%)	44096 (90.1%)	4863 (9.9%)	3952 (45.7%)
SAR/Optical	2317 (73.4%)	74217 (76.8%)	22440 (23.2%)	839 (26.6%)
QuickBird /IKONOS	13450 (52.2%)	117384 (87.4%)	16876 (12.6%)	12290 (47.8%)
SAR/Optical	14746 (66.3%)	520632 (99.4%)	3286 (0.6%)	7489 (33.7%)

respectively, to 0.01 and 4.0. The value of the sparsity proportion ρ was set to 0.10 and the maximum number of training epochs for each of the sparse autoencoder architecture was set to 1000 and 400 epochs.

In order to discuss the obtained results, from the conducted experiments, we compare our results to the state-of-the-art methods in terms of classification rate, i.e., the accuracy that measures the percentage of the correct changed and unchanged pixels.

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} \tag{8}$$

where TP and TN denote the number of pixels that are correctly classified, and FN and FP denote the number of misclassified pixels

Table I summarizes the different CD accuracy rates obtained by our approach and draws a comparison with both supervised and unsupervised state-of-the-art approaches.

Based on the leave-one-out evaluation strategy, we can notice that the accuracy rate of the proposed method outperforms the most state-of-the-art approaches and remains comparable to the other supervised and unsupervised state-of-the-art methods. The strength of our model is its ability to process a wide variety of satellite imaging modalities, i.e., multisources, multisensor, and multilooking SAR images, under different resolutions. The method can effectively process images corrupted by different noise types and different degradation levels (see Fig. 6 where SAR images are corrupted by different speckle noise levels).

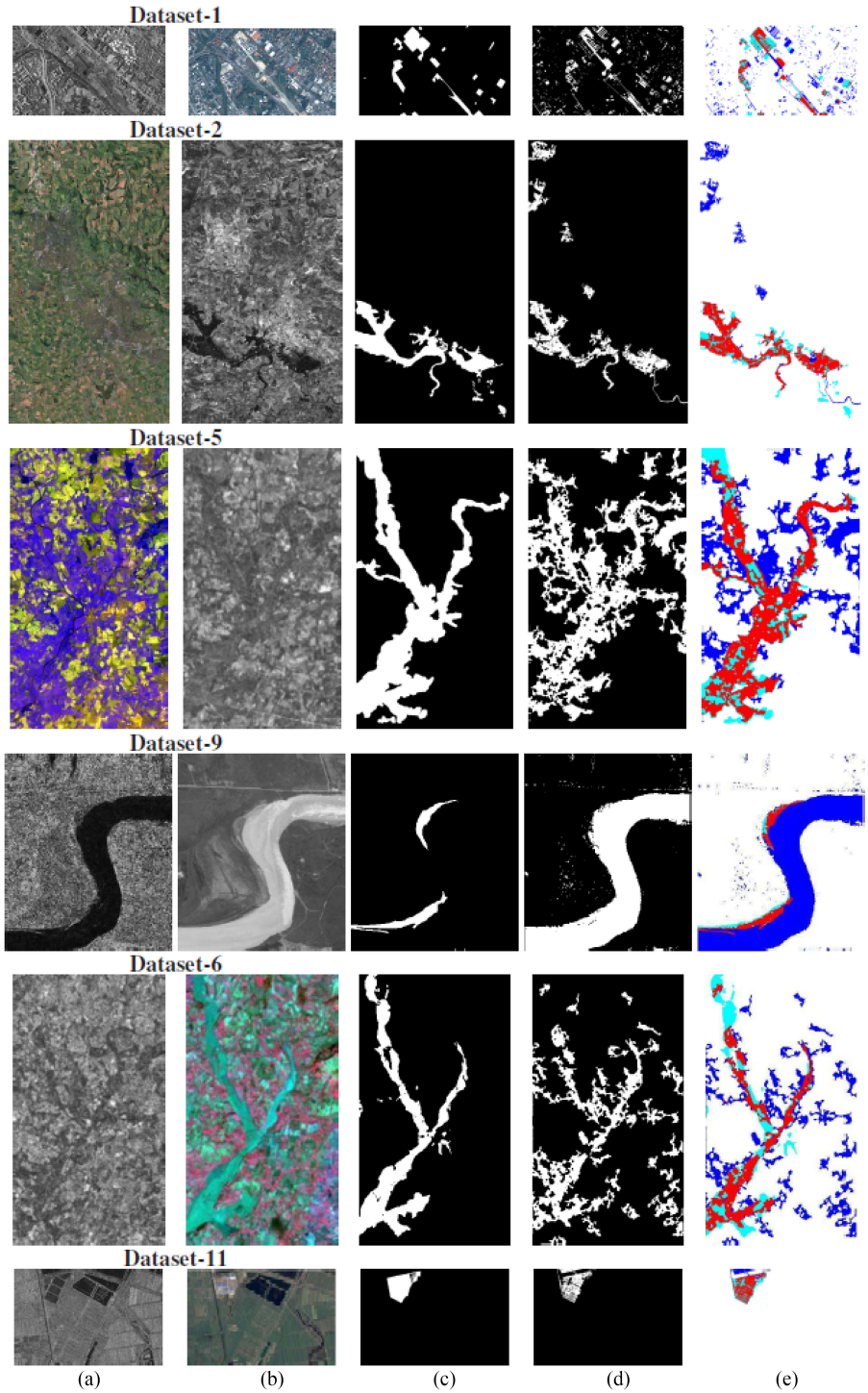

Fig. 4. Heterogeneous (multisource) optical/SAR and SAR/optical datasets. (a)–(c) Image t_1, t_2, ground truth. (d) and (e) Final (changed–unchanged) clustering result and confusion map (white: TN; red: TP; blue: FP; Cyan: FN) obtained by the proposed approach.

From Table II, we can see also that the changed and unchanged area are well detected and that the different resulting binary maps match fairly the different regions shown in the ground truth for the different satellite imagery sources (see Figs. 4–6).

The global accuracy rate obtained by our unsupervised anomaly detection model, over the 11 heterogeneous image pairs, using the leave-one-out evaluation scenario, is 0.863%.

C. Architecture Configuration and Experimental Settings

In all our experiments, we choose the best architecture as the one having the least mean reconstruction error (MSE) on the validation set containing only normal patterns. For parameters settings, we note that our training/validation dataset is a subset of each multimodal pair image and having dimension $d_s = 162$.

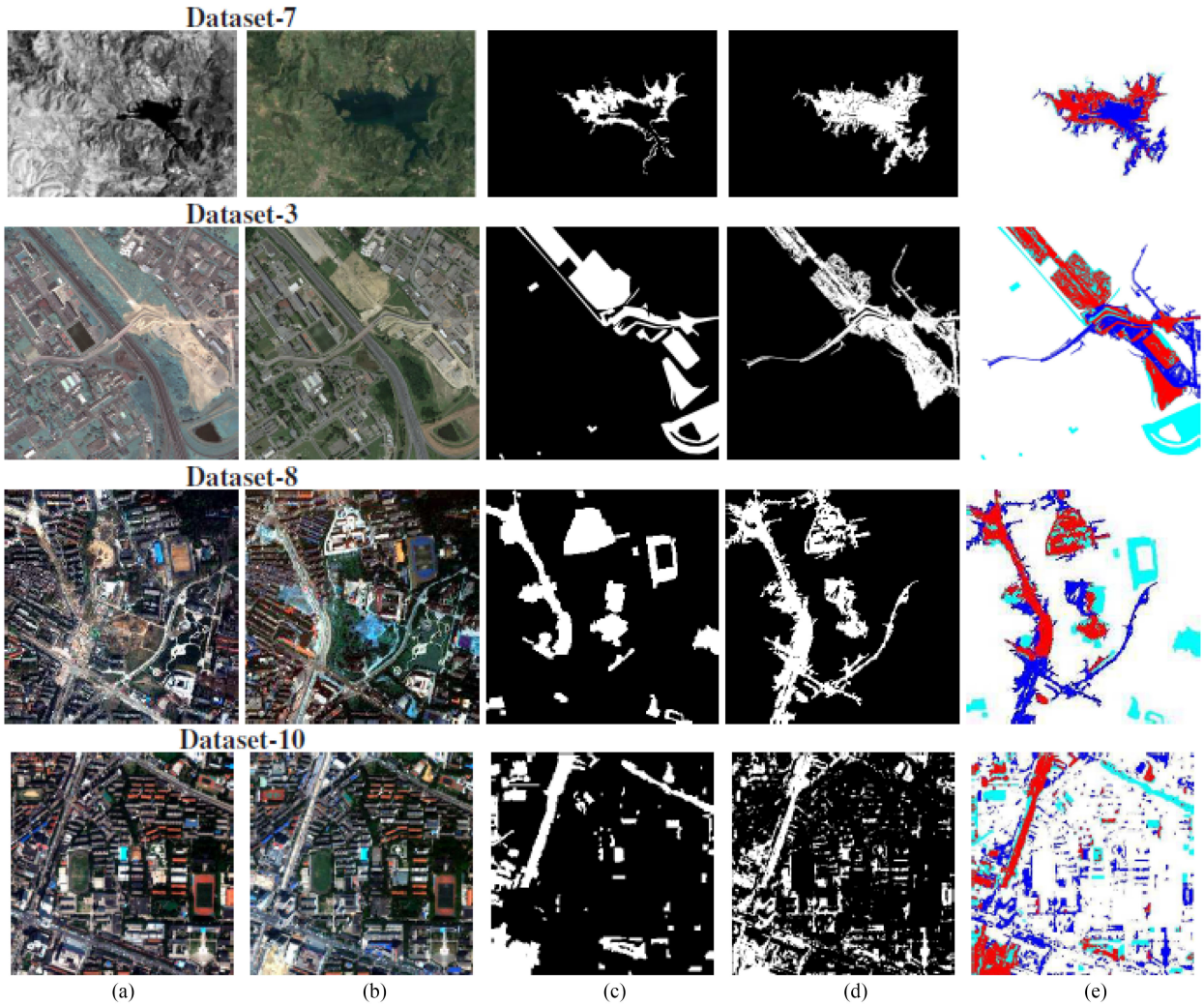

Fig. 5. Heterogeneous (multisensor) optical/optical dataset. (a)–(c) Image t_1, t_2, ground truth. (d) and (e) Final (changed/unchanged) clustering result and confusion map (white: TN; red: TP; blue: FP; cyan: FN) obtained by the proposed approach.

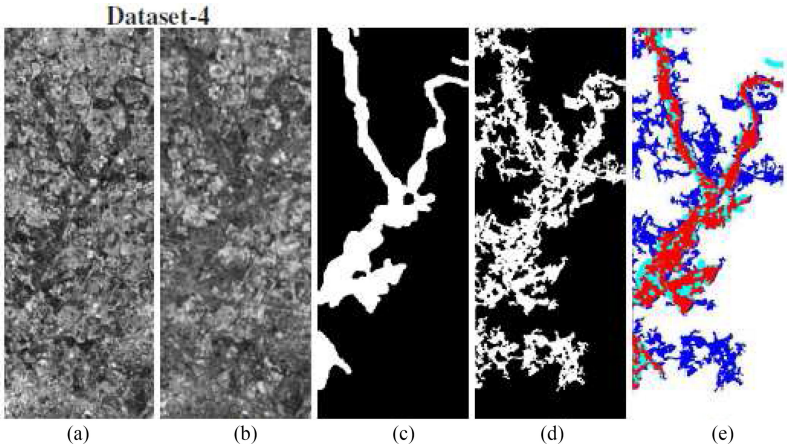

Fig. 6. Heterogeneous (multilooking) SAR/SAR datasets:. (a)–(c) Image t_1, t_2, ground truth. (d) and (e) Final (changed/unchanged) clustering result and confusion map (white: TN; red: TP; blue: FP; cyan: FN) obtained by the proposed approach.

TABLE III
PARAMETERS OF THE STACKED SPARSE AUTOENCODER

Parameter name	Min	Step	Factor	Max
Hidden layer 1	80	10	-	120
Hidden layer 2	30	10	-	50
Hidden layer 3	10	5	-	20
Hidden layer 4	3	2	-	7
ρ	0.00625	-	2	0.8
λ	0.0001	-	10	0.1
β	0.5	-	2	8

Algorithm 2: Grid Search Based Hyper-parameter Optimization of the Proposed CD Model.

Step 1:

- Set of hyperparameters from a defined space
- Normal training and validation subsets

Step 2:

foreach *combination of the model parameters* $\in$ *defined space* **do**
- Train the first and second layers of the sparse AEC model using (Eq.3)
- Compute the MSE on the validation subset.

end

- Optimal hyper-parameters outputs with the least squares.

The dataset is randomly subdivided into two subsets: (2/3) for the training set and (1/3) for the testing set. We inject in our normal training dataset a proportion of 3.0% anomalous (change) patterns to form the final training dataset.

We present empirical results produced by our anomaly CD model on this data subset. We use a simple stacked sparse neural network model with normal class. The network parameter settings are described in Table III. In order to fix the neural network architecture and to find optimal hyperparameters, we rely on a grid search method performed in a defined space with a fixed step/factor and using the following hyperparameters space: number of hidden units per layer for the first, second, third, and fourth hidden layers; the coefficient of the sparsity term β; the coefficient of the regularization term λ; and the sparsity proportion ρ.

Once a layer-wise training strategy was adopted, each layer was trained independently from the others and the parameter values (ρ, λ, β) were varied by exploring different combinations of optimization parameters for each of the four layers with the corresponding number of hidden units. We gradually increased the hidden layer number starting from two layers and chose the architecture, giving the best parameter values that minimized the MSE. Algorithm 2 shows the estimation step (with a grid search-based optimization technique) of the internal parameters of the stacked sparse neural network reconstruction model.

When the number of hidden layers was set to 3 and 4, the mean-squared error is, respectively, 0.1385% and 0.1409%, which are greater than the MSE value of 0.0640% obtained only with two hidden layers. Therefore, the number of the hidden layers in our anomaly-based CD model was set to 2 in our application. Table IV shows the optimal parameters and the MSE obtained by the grid search method for different architectures depth.

D. Discussion

Before all, it is important to recall that this type of deep autoencoder will necessarily be well adapted to our multimodal CD detection task, since this one has already proven its efficiency to learn and fuse highly heterogeneous pairs of data types in a common space representation [29], [32]–[43] and also has proven to be effective in modeling/fusing highly heterogeneous data/sources supported in the multimedia domain (such as words/images [33], [34], speech/images [35], [38], [39], audio/video [32], facial expressions/sound [36], [37], or multimodal DCE/MRI medical images [29], and two MRI medical images modalities [43]). In this study (the first study to our knowledge), we confirm the relevance of this type of deep autoencoder in dealing/fusing heterogeneous data (or heterogeneous imaging modalities) used in remote sensing.

We now discuss the influence of the different parameter settings for our anomaly CD model on 11 benchmark multimodal datasets using the leave-one-out evaluation strategy. To this end, we vary the parameter to be evaluated and fix the others to their optimal values (see Table IV), and quantify the average accuracy.

In our application, the parameter ρ plays a crucial role because it conditions the level of sparsity which may affect considerably our analysis. More precisely, ρ is used to optimize false alarm rates in our unsupervised anomaly CD detection problem and its tuning is based only on normal class images. Indeed, a small ρ induces an over classification of many normal class patterns as anomalous/outliers. In the opposite case, a large ρ discourages normal data patterns from being classified as anomalous/outliers. Thereby, a bad choice of the value of ρ classifies many normal patterns as anomalous and increases the false-positive rate or classify many abnormal patterns as normal and increases the false negative rate, which decreases the performance of the anomaly CD model (see Table V). Accordingly, the optimal ρ (in our case, $\rho = 0.1$) balances both false-positive and false-negative rates (see Table II).

We can notice that the weight decay λ and the regularization parameter β affect less the behavior of the autoencoder compared to the sparsity parameter ρ (see Table V). The performances depicted in Table V show that when the expected average of the neurons' activations is too low (i.e., 0.00625) or higher than (0.8), the performances decrease drastically to 0.579 and 0.715, respectively. This is, theoretically, in concordance with the fact that higher value of ρ leads to higher degree of sparsity, which means that the domain knowledge is not smoothly distributed with more specialized neurons. A solution could be to increase the number of neurons but the best strategy is to keep an equilibrium between specialization and sparsity. Moreover, Table V does not show any effect of the L_2 regularization sparsity

TABLE IV
STACKED SPARSE AUTOENCODER HYPERPARAMETERS OBTAINED ON THE SUBSET MULTIMODAL DATASET WITH THE MEAN SQUARED RECONSTRUCTION ERROR (MSE)

Number of layers	ρ	λ	β	Size of first layer	Size of second layer	Size of third layer	Size of fourth layer	MSE
2	0.1	0.01	4	80	40	-	-	0.0640
3	0.05	0.01	2	110	50	10	-	0.1385
4	0.05	0.01	4	100	40	10	3	0.1409

TABLE V
AVERAGE CLASSIFICATION ACCURACY AND THE STACKED SPARSE AUTOENCODER HYPERPARAMETERS USED WITH THE FIRST AND SECOND HIDDEN LAYERS

ρ	λ	β	Average accuracy (%)
0.00625	0.01	4	0.579
0.05	0.01	4	0.822
0.4	0.01	4	0.764
0.8	0.01	4	0.715
0.1	0.0001	4	0.801
0.1	0.001	4	0.808
0.1	0.1	4	0.823
0.1	0.01	0.5	0.830
0.1	0.01	1	0.837
0.1	0.01	2	0.829
0.1	0.01	8	0.832

TABLE VI
IMPACT OF THE SQUARE WINDOW SIZE ON THE AVERAGE CLASSIFICATION ACCURACY

Sw	Average accuracy (%)
9	0.863
11	0.849
13	0.838
15	0.844

on the generalization performance since we are not at risk of overfitting as our models were trained on a huge amount of data patches with diversified content that were taken only from the nonchange subsets. Also, an unlimited set of training patches could be generated by data augmentation. In addition, experiments conducted on different numbers of hidden layers show that augmenting the number of layers does not effectively increase the average classification accuracy. The average classification rate obtained using three and four hidden layers with a number of nodes set to 10 and 3 are, respectively, equal to 0.847% and 0.845% which are lower than our average classification rate 0.863% that corresponds to the optimal number $(=2)$ of layers. Varying the number of nodes of the hidden layers also does not enhance necessarily the average accuracy. Different combinations were tested giving very close values to the optimal average accuracy which is obtained by 80 nodes for the first and 40 nodes in the second hidden layer.

In the same way, the impact of the squared window size (Sw) is assessed by a comparison study done on the average classification accuracy of the anomaly CD model using different sizes. Table VI demonstrates that the average classification accuracy is not much significantly influenced by the size (Sw).

To conclude, the results obtained from different experiments have shown that the choice of the optimization hyperparameters is a crucial task in the features network setting, particularly the ρ parameter which is the key parameter of the network, contrary to the other parameters such as the depth of the network that does not significantly influence the anomaly CD model performances.

The main quality of our model is that it achieves a better classification rate accuracy under different CD conditions, reflecting a variety of imaging modalities with different noise types and levels, where the sensitivity of different parameters is analyzed (see Table V). This justifies the fact that it can also be less accurate than some specific supervised/unsupervised multimodal CD models, dealing only with a specific type of noise and a specific imaging modalities such as PCC and SCNN methods [20] which also use denoising algorithms to reduce the speckle noise of the SAR images and/or the Gaussian noise of the optical images [particularly when the SAR images are too much corrupted by the multiplicative speckle noise degrading their quality and creating for each texture class a kind of macro texture with grainy patterns (see dataset-9 Fig. 4)].

IV. CONCLUSION

In this article, we have proposed a new anomaly-based CD model for heterogeneous remote sensing image pairs. This model exhibits quite interesting properties. First, the proposed model is based on unsupervised training stage in which a stacked multimodal sparse autoencoder model employing a *satlin* and *purlin* neural transfer functions is trained to learn and infer a suitable latent representation of the normal image patterns existing in the before and after multimodal images. The training is done in order to identify and to disentangle from the normal image patterns (belonging to the *nonchange* class label) the change class as unusual from abnormal feature patterns in the residual space; the trained anomaly-based CD model tries to reconstruct the feature space for each new unseen image pair by encoding and decoding the image pair inputs using its stacked hidden representation. The reconstruction error between the original input feature and its reconstruction is quantified to generate (for each pixel) an anomaly-based error score that highlights the usual and unusual (rare) patterns that belong to the abnormal class (*change* class label) or to the normal class (*nonchange* class label). Finally, a GMM assigns a class label to each pixel (change vs. nonchange) in the MAP sense. The different experimentation conducted on the proposed CD model, in the leave-one-out test scenario, demonstrates its effectiveness in processing new unseen input heterogeneous image pairs. Besides, the model seems to be flexible enough to process heterogeneous image pairs with both different spatial resolutions, covering different heterogeneous CD conditions (as multisource, multisensor, and multilooking image pairs). It accurately determines different kinds

of natural and/or man-made changes (e.g., major urban construction and changes resulting from different types of natural phenomenon).

ACKNOWLEDGMENT

The authors would like to acknowledge all other researchers who made at our disposal the CD dataset in order to validate the proposed anomaly CD model.

REFERENCES

[1] X. Huang, L. Zhang, and T. Zhu, "Building change detection from multitemporal high-resolution remotely sensed images based on a morphological building index," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 1, pp. 105–115, Jan. 2014.

[2] P. Du, S. Liu, P. Gamba, K. Tan, and J. Xia, "Fusion of difference images for change detection over urban areas," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 4, pp. 1076–1086, Aug. 2012.

[3] A. A. Nielsen, K. Conradsen, and H. Skriver, "Change detection in full and dual polarization, single- and multi-frequency sar data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 8, pp. 4041–4048, Aug. 2015.

[4] O. Yousif and Y. Ban, "Improving SAR-based urban change detection by combining map-MRF classifier and nonlocal means similarity weights," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 10, pp. 4288–4300, Oct. 2014.

[5] R. Hedjam, M. Kalacska, M. Mignotte, H. Z. Nafchi, and M. Cheriet, "Iterative classifiers combination model for change detection in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 6997–7008, Dec. 2016.

[6] 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.*, vol. 5, no. 1, pp. 331–342, Feb. 2012.

[7] G. Mercier, G. Moser, and S. Serpico, "Conditional copulas for change detection in heterogeneous remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1428–1441, May 2008.

[8] F. Chatelain, J. Y. Tourneret, and J. Inglada, "Change detection in multisensor sar images using bivariate gamma distributions," *IEEE Trans. Image Process.*, vol. 17, no. 3, pp. 249–258, Mar. 2008.

[9] J. Prendes, M. Chabert, F. Pascal, A. Giros, and J. Tourneret, "A new multivariate statistical model for change detection in images acquired by homogeneous and heterogeneous sensors," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 799–812, Mar. 2015.

[10] J. Prendes, M. Chabert, F. Pascal, A. Giros, and J.-Y. Tourneret, "Change detection for optical and radar images using a Bayesian nonparametric model coupled with a Markov random field," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Apr. 2015, pp. 1513–1517.

[11] R. Touati and M. Mignotte, "An energy-based model encoding non-local pairwise pixel interactions for multi-sensor change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 1046–1058, Jan. 2018.

[12] V. Alberga, "Similarity measures of remotely sensed multi-sensor images for change detection applications," *Remote Sens.*, vol. 1, no. 3, pp. 122–143, 2009.

[13] R. Touati, M. Mignotte, and M. Dahmane, "A new change detector in heterogeneous remote sensing imagery," in *Proc. 7th IEEE Int. Conf. Image Process. Theory Tools Appl.*, Dec. 2017, pp. 1–6.

[14] D. Brunner, G. Lemoine, and L. Bruzzone, "Earthquake damage assessment of buildings using VHR optical and SAR imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 5, pp. 2403–2420, May 2010.

[15] R. Touati, M. Mignotte, and M. Dahmane, "Change detection in heterogeneous remote sensing images based on an imaging modality-invariant MDS representation," in *Proc. 25th IEEE Int. Conf. Image Process.*, Oct. 2018, pp. 3998–4002.

[16] C. Wu, B. Du, and L. Zhang, "Slow feature analysis for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2858–2874, May 2014.

[17] Z. g. Liu, G. Mercier, J. Dezert, and Q. Pan, "Change detection in heterogeneous remote sensing images based on multidimensional evidential reasoning," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 168–172, Jan. 2014.

[18] Z. Liu, G. Li, G. Mercier, Y. He, and Q. Pan, "Change detection in heterogenous remote sensing images via homogeneous pixel transformation," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1822–1834, Apr. 2018.

[19] N. Merkle, P. F. S. Auer, and R. Muller, "On the possibility of conditional adversarial networks for multi-sensor image matching," in *Proc. IEEE Geoscience Remote Sens. Soc.*, Fort Worth, TX, USA, Jul. 2017, pp. 1–4.

[20] 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, Mar. 2018.

[21] P. Zhang, M. Gong, L. Su, J. Liu, and Z. Li, "Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 116, pp. 24–41, 2016.

[22] W. Zhao, Z. Wang, M. Gong, and J. Liu, "Discriminative feature learning for unsupervised change detection in heterogeneous images based on a coupled neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 7066–7080, Dec. 2017.

[23] M. A. Ranzato, Y.-L. Boureau, and Y. LeCun, "Sparse feature learning for deep belief networks," in *Proc. 20th Int. Conf. Neural Inf. Process. Syst.* (ser. NIPS'07). Red Hook, NY, USA: Curran Associates Inc., 2007, pp. 1185–1192.

[24] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proc. IEEE*, vol. 98, no. 6, pp. 1031–1044, Jun. 2010.

[25] R. Chalapathy, A. K. Menon, and S. Chawla, "Anomaly detection using one-class neural networks," 2018, *arXiv:1802.06360*.

[26] R. Laxhammar and G. Falkman, "Online learning and sequential anomaly detection in trajectories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1158–1173, Jun. 2014.

[27] D. Park, Y. Hoshi, and C. C. Kemp, "A multimodal anomaly detector for robot-assisted feeding using an LSTM-based variational autoencoder," *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 1544–1551, Jul. 2018.

[28] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006. doi: 10.1162/neco.2006.18.7.1527.

[29] H. C. Shin, M. R. Orton, D. J. Collins, S. J. Doran, and M. O. Leach, "Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1930–1943, Aug. 2013.

[30] Y. Qi, Y. Wang, X. Zheng, and Z. Wu, "Robust feature learning by stacked autoencoder with maximum correntropy criterion," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2014, pp. 6716–6720.

[31] A. Droniou, S. Ivaldi, and O. Sigaud, "Deep unsupervised network for multimodal perception, representation and classification," *Robot. Auton. Syst.*, vol. 71, pp. 83–98, 2015.

[32] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. Int. Conf. Mach. Learn.*, Omnipress, 2011, pp. 689–696.

[33] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. 22nd ACM Int. Conf. Multimedia* (ser. MM'14). New York, NY, USA: ACM, 2014, pp. 7–16.

[34] W. Wang, B. C. Ooi, X. Yang, D. Zhang, and Y. Zhuang, "Effective multimodal retrieval based on stacked auto-encoders," *Proc. VLDB Endow.*, vol. 7, no. 8, pp. 649–660, Apr. 2014.

[35] H. T. Lu, Y. M. Liou, H. Y. Lee, and L.-S. Lee, "Semantic retrieval of personal photos using a deep autoencoder fusing visual features with speech annotations represented as word/paragraph vectors," in *INTERSPEECH*, ISCA, 2015, pp. 140–144.

[36] W. Liu, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal deep learning," in *Proc. Int. Conf. Neural Inf. Process.*, 2016, pp. 521–529.

[37] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2013, pp. 3687–3691.

[38] K. Leidal, D. Harwath, and J. R. Glass, "Learning modality-invariant representations for speech and images," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2017, pp. 424–429.

[39] W. Hsu, Y. Zhang, and J. R. Glass, "Learning latent representations for speech generation and transformation," 2017, *arXiv:1704.04222*.

[40] C. Hong, J. Yu, J. Wan, D. Tao, and M. Wang, "Multimodal deep autoencoder for human pose recovery," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5659–5670, Dec. 2015.

[41] N. Jaques, S. Taylor, A. Sano, and R. Picard, "Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact.*, Oct. 2017, pp. 202–208.
[42] P. Wu, S. C. Hoi, H. Xia, P. Zhao, D. Wang, and C. Miao, "Online multimodal deep similarity learning with application to image retrieval," in *Proc. 21st ACM Int. Conf. Multimedia* (ser. MM'13). New York, NY, USA: ACM, 2013, pp. 153–162.
[43] G. van Tulder and M. de Bruijne, "Learning cross-modality representations from multi-modal images," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 638–648, Feb. 2019.
[44] B. Olshausen and D. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.
[45] W. Sun, S. Shao, R. Zhao, R. Yan, X. Zhang, and X. Chen, "A sparse auto-encoder-based deep neural network approach for induction motor faults classification," *Measurement*, vol. 89, pp. 171–178, 2016.
[46] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?" *Vision Res.*, vol. 37, no. 23, pp. 3311–3325, 1997.
[47] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in *Proc. Conf. Neural Inf. Process. Syst.*, 2012, pp. 341–349.
[48] J. Prendes, M. Chabert, F. Pascal, A. Giros, and J. Tourneret, "Performance assessment of a recent change detection method for homogeneous and heterogeneous images," *Revue Française Photogrammétrie Télédétection*, vol. 209, pp. 23–29, 2015.
[49] J. Prendes, "New statistical modeling of multi-sensor images with application to change detection," Ph.D. dissertation, École supérieure d'électricité (Supélec), the Institut de Recherche en Informatique de Toulouse, Toulouse, France, 2015.
[50] O. D. Team, "The ORFEO toolbox software guide," 2014. [Online]. Available: http://orfeo-toolbox.org/
[51] G. Mercier, G. Moser, and S. Serpico, "Conditional copula for change detection on heterogeneous SAR data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2007, pp. 2394–2397.
[52] Y. Tang and L. Zhang, "Urban change analysis with multi-sensor multispectral imagery," *Remote Sens.*, vol. 9, no. 3, p. 252, 2017.
[53] G. Mercier and J. Inglada, "Change detection with misregistration errors and heterogeneous data through the Orfeo Toolbox," Lab-STICC(TB) – Laboratoire en sciences et technologies de l'information, de la communication et de la connaissance (UMR CNRS 6285 - Télécom Bretagne – Université de Bretagne Occidentale - Université de Bretagne Sud – ENSTA Bretagne - Ecole Nationale d'ingénieurs de Brest), CNES - Centre national d'études spatiales, Tech. Rep., 2008.
[54] M. F. Möller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Netw.*, vol. 6, no. 4, pp. 525–533, 1993.

Redha Touati received the M.Sc. degree in computer science from the Department of Computer Science and Operations Research (DIRO), University of Montreal, Montreal, QC, Canada, in 2014 . He is currently working toward the Ph.D. degree at the Vision Laboratory of the University of Montreal (DIRO), in collaboration with the Imaging and Vision Department of the Computer Research Institute of Montreal (CRIM), Montreal, QC, USA.

His research interests include statistical methods and applied mathematics in video imaging and remote sensing imagery.

Max Mignotte received the DEA (postgraduate degree) in digital signal, image, and speech processing from the INPG University, Grenoble, France, in 1993, and the Ph.D. degree in electronics and computer engineering from the University of Bretagne Occidentale (UBO), Brest, France, and the Digital Signal Laboratory (GTS) of the French Naval academy, France, in 1998.

He was an INRIA Postdoctoral Fellow with the University of Montreal (DIRO), Montreal, QC, Canada, from 1998 to 1999. He is currently with the Computer Vision and Geometric Modeling Laboratory, University of Montreal, as a Professor. He is also a member of LIO (Laboratoire de recherche en imagerie et orthopedie, Centre de recherche du CHUM, Hopital Notre-Dame) and a Researcher with the CHUM. His current research interests include statistical methods, Bayesian inference, and energy-based models for solving diverse large-scale high-dimensional ill-posed inverse problems in imaging.

Mohamed Dahmane received the graduation degree from the Université de Montréal, Montreal, QC, USA, in 2012, and the two master's degrees, one in the area of satellite imagery and the other in video surveillance from the Department of Computer Science and Operations Research (DIRO), University of Montréal, QC, Canada.

He joined the Computer Research Institute of Montreal (CRIM), Montreal, QC, USA, in 2012, as an Expert in development of algorithms for facial expression recognition, the main subject of his doctoral thesis. He is currently a Researcher with the CRIM. He is also an Associate Professor with the École de Technologie Supérieure (ÉTS) Université du Québec, Montreal, QC, USA.