Title: Unsupervised Deep Change Vector Analysis for Multiple-Change Detection in VHR Images

This paper appears in: IEEE Transactions on Geoscience and Remote Sensing

Date of Publication: 10 January 2019
Author(s): S.Saha, F. Bovolo, L. Bruzzone
Volume: 57, Issue: 6
Page(s): 3677 - 3693
DOI: 10.1109/TGRS.2018.2886643

Unsupervised Deep Change Vector Analysis for Multiple-Change Detection in VHR Images

Sudipan Saha, *Student Member, IEEE,* Francesca Bovolo, *Senior Member, IEEE,*
and Lorenzo Bruzzone, *Fellow, IEEE*

***Abstract*—Change Detection (CD) in multi-temporal images is an important application of remote sensing. Recent technological evolution provided Very High spatial Resolution (VHR) multitemporal optical satellite images showing high spatial correlation among pixels and requiring an effective modeling of spatial context to accurately capture change information. Here we propose a novel unsupervised context-sensitive framework - Deep Change Vector Analysis (DCVA) -for CD in multitemporal VHR images that exploits Convolutional-Neural-Network (CNN) features. To have an unsupervised system, DCVA starts from a sub-optimal pre-trained multilayered CNN for obtaining deep features that can model spatial relationship among neighbouring pixels and thus complex objects. An automatic feature selection strategy is employed layerwise to select features emphasizing both high and low prior probability change information. Selected features from multiple layers are combined into a deep feature hypervector providing a multi-scale scene representation. The use of the same pre-trained CNN for semantic segmentation of single images enables us to obtain coherent multi-temporal deep feature hypervectors that can be compared pixelwise to obtain deep change vectors that also model spatial-context information. Deep change vectors are analyzed based on their magnitude to identify changed pixels. Then deep change vectors corresponding to identified changed pixels are binarized to obtain a compressed binary deep change vectors that preserve information about the direction (kind) of change. Changed pixels are analyzed for multiple-change detection based on the binary features, thus implicitly using the spatial information. Experimental results on multitemporal datasets of Worldview-2, Pleiades, and Quickbird images confirm the effectiveness of the proposed method.**

***Index Terms*—Change detection, Very High Resolution images, Multi temporal images, Deep features, Deep Change Vector Analysis, Remote Sensing.**

I. Introduction

Land-cover change detection is critical for monitoring phenomena like urbanization, industrial operations, natural disaster. There is a need to proper and timely monitoring such events as they can cause critical crisis in long run. Remotely sensed multi-temporal images are used for understanding such changes taking place on the Earth surface. Multi-temporal images have been used for a variety of applications that gain of the temporal dimension, including land cover monitoring [1], disaster management [2], urban planning [3]. Using the latest generation of satellite optical sensors, such as Pleiades, Ikonos, QuickBird, Spot-5, and Worldview, VHR images (up to 0.5 m) can be obtained. The availability of multi-temporal VHR images has increased the range of possible applications of change detection.

Both supervised and unsupervised change-detection techniques have been used in remote sensing. Supervised methods are preferred when an exact ”from-to” information is required. Post classification comparison [4] is a popular supervised CD method, which independently classifies multi-temporal images based on supervised classification. Single date classification results are compared to obtain the change map. The accuracy highly depends on the accuracy of each single-date classification. Another popular supervised method is Direct-multidate classification (DMC) [4] that represents pixels by stacking feature vectors corresponding to bi-temporal images. Each class transition is considered as a single class in DMC. Thus training data should have samples for each class transition. A more practical approach is Compound Classification (CC) that exploits the maximization of the posterior joint probability of classes [5], [6]. All supervised methods depend on availability of labeled samples for training data. Obtaining them in the context of multi-temporal analysis is difficult. Unsupervised methods do not require ground truth data. So, inspite of the fact that supervised methods may produce better results than unsupervised ones, in the literature significant attention has been paid to the unsupervised methods [7], [8], [9].

Most of the unsupervised change detection techniques for optical passive sensor images are based on the concept of Change Vector Analysis (CVA) [10] or difference image. In this paradigm, pixelwise difference of radiometry values is computed. Alternatively, pixelwise difference of features derived from images can be computed, e.g., vegetation indexes [4], Tasseled Cap Transformation (TCP) features [11]. The magnitude of the difference image is further analyzed to distinguish changed pixels from unchanged ones [7], [8], [9]. Additionally, a direction variable can be obtained from the multispectral difference image which is used to distinguish different kinds of change [12], [13], [14]. Though simple, such a framework based on image comparison is very effective for low/medium-resolution multi-temporal images where one pixel generally represents a large geographical extent and may cover one or more objects on the ground. Neighbouring pixels can be assumed to be independent and hence a pixelwise comparison is practical.

Assumption of pixel independence does not hold for VHR images, thus it is necessary to model the spatial context information [15]. In VHR images, often radiometric changes are not enough to represent changes occurred on the ground due to many factors. One is the geometry of acquisition.

Sudipan Saha and Francesca Bovolo are with Fondazione Bruno Kessler, Trento, Italy. E-mail: saha@fbk.eu

Sudipan Saha and Lorenzo Bruzzone are with the Department of Information Engineering and Computer science, Trento, Italy.

VHR images are often acquired at off-nadir angles. Change in viewing angle causes change in geometrical properties of the images even if no changes occurred on the ground. Moreover, seasonal variations of the solar ray incidence angle cause shadow differences that are not associated to the changes on the ground. Inspite of co-registering the VHR multi-temporal images, a correspondence between pixels is difficult to be achieved. Thus object level information and spatial context of pixels need to be exploited to benefit from VHR information. Following this idea, some CD algorithms have been recently proposed for VHR images, which exploit the spatial context of a pixel [16], [17], [18], [19], [20], [21]. Thonfeld *et al.* [16] introduced a version of CVA for binary CD, known as Robust CVA (RCVA) developed to mitigate pixel neighbourhood effects. RCVA considers a neighbourhood around each pixel to mitigate effects of poor co-registration between multi-temporal images, but it does not capture object-level information. Li *et al.* [20] proposed an object-oriented CVA (OCVA) method for binary change detection that segments the multi-temporal images and compares each segment instead of each pixel. Bovolo [17] proposed a parcel CVA (PCVA) that is based on independent hierarchical segmentation of multi-temporal images to encode the spatial context of pixels. Accuracy of such approaches depends on the performance of the image segmentation algorithm. Moreover, despite a priori object detection, the compared features are usually shallow and do not properly capture spatial context object complexity. Morphological profiles and morphological attribute profiles [18], [19] have been exploited for CD in VHR images because of their non-linear nature. Markov Random Fields (MRFs) have also been exploited due to their ability to integrate the spatial and temporal context information [7], [22]. Lv *et al.* [21] proposed a binary CD algorithm using multi-feature probabilistic ensemble conditional random field that captures structural properties of objects in high resolution images. In [23], Lv *et al.* combined object based methods with random field based methods for binary CD. Despite differences between the mentioned methods, they emphasize the importance of using spatial context information, object level information, and complex non-linear features. Drawbacks of the existing algorithms are two-fold. First, they have limited capability in capturing spatial context information and complex visual features. Second, most of them are focused on binary CD (i.e., they distinguish presence/absence of change only) and there is still limited work on multiple CD in which the change class is further divided into different kinds of change.

Recently deep learning, especially CNN has drastically improved performance in image understanding tasks [24], [25], including remote sensing image understanding [26]. Deep learning algorithms have been applied to different remote sensing image processing tasks, including semantic labelling [27], hyperspectral image classification [28], and target detection [29]. Deep learning based framework is suitable to extract high-level visual features that are semantically rich [30]. They are effective in capturing rich information about objects or image parts. However, deep learning methods are data hungry and training a deep learning based algorithm usually requires enormous amount of training data, which are not available for multi-temporal remote sensing images [31], [26]. So there are only few works exploiting deep learning for change detection [32], [33], most of them are supervised and deal with binary change detection only. Zhan *et al.* [34] proposed a supervised CD method for optical aerial images based on the deep Siamese network. Lyu *et al.* [35] proposed a supervised CD method based on recurrent neural network. Geng *et al.* [36] proposed a supervised binary CD method based on contractive autoencoders. To reduce the need of labeled training samples, some methods use pre-classification schema to obtain a coarse initial change map that is used to further train the change detection model [37], [38]. Zhang *et al.* [37] proposed a binary CD method that exploits coarse initial change map to identify most unlikely pairs that are used to learn a mapping neural network. Gao *et al.* [38] proposed a binary CD method for SAR images that identifies pixels having high chance of being changed/unchanged using wavelet based features and uses patches centered at those pixels to train a neural network. Xu *et al.* [39] proposed a binary CD method using autoencoder that learns correspondence between pre-change image and post-change image. Such methods have limited reusability as the model needs to be trained/tuned for individual datasets.

However, it has been shown in the literature that deep learning based feature extraction shows excellent generality properties that enable transfer learning capability [40]. A deep network trained with images of a certain domain can become useful to treat images of other domains. Such feature representation, known as deep features, has exhibited stronger domain invariance capability than shallow features representation for remote sensing classification task [41], [42]. Moreover deep features capture the spatial context information effectively. CNN is capable of learning complex features by using non-linear activation functions in multi-layer network configuration. In many application domains pre-trained CNN networks are used as a black-box feature extractor to obtain deep features from images, which can be further processed for change information extraction [43].

There are many available pre-trained CNN architectures in the literature [44], [45], [46]. For CD, we need to choose a pre-trained CNN suitable for feature extraction from each pixel of multi-temporal images. Originally CNN architectures were designed for image classification [44], [45]. Such architectures accept an RGB image as input and provide as output a class or class probability that best describes the input image. Such architectures generally consist of a series of convolutional layers followed by a series of fully connected layers. Convolutional layers form the basis of CNN. Core operation of training is performed by learning the weights of the convolutional layers. The training process is further enhanced by the presence of the other layers including pooling, non-linear activation. Fully connected layers are used in the last stages of the CNN to construct the desired number of outputs. However, limitations of such CNN architectures in the field of CD are as following:

- Pre-trained CNNs are usually trained targeting RGB image classification. They are trained by the backpropagation method where an error is calculated at the final layer and is propagated back through the layers [24]. In case of

CNN architectures designed for image classification, the error is calculated based on the label of the entire image and not on pixelwise label. Thus such networks are not trained for finer inference.

- To obtain pixelwise features for an input image with spatial dimension $R \times C$, we have to evaluate the network $R \times C$ times, each time centering on a pixel and taking a window around it as input to the CNN. Such a strategy is not computationally efficient. A possible solution to this problem is to ignore the fully connected layers and use only the pipeline of convolutional layers when extracting features. In this way it is possible to obtain pixelwise features from input of any size at a single run. Such a solution is sub-optimal as we are ignoring visual concept learned by fully connected layers when discarding them completely during feature extraction process.
- Since fully connected layers can only deal with input of a fixed size, such architectures are restricted to accept input of pre-defined size only [47].
- Most of pre-trained CNNs are trained to accept RGB input images whereas satellite optical VHR images also have a NIR channel that is important for change detection, especially for vegetation analysis.

Recently some works have been proposed to shift the paradigm of CNN from image classification to finer interpretation of images. Some examples are bounding box object detection [48] and learning correspondence [49]. A step forward in this direction is the work by Long *et. al.* [47] for CNN based semantic segmentation. They proposed a new kind of architecture where all the learnable layers are convolutional. It consists of a series of convolutional, pooling, and activation layers followed by a series of deconvolutional and activation layers and can additionally have other layers like batch normalization and dropout [47], [46]. Such an architecture can accept input of any spatial dimension X_i and produce pixelwise output X_o for the entire image, effectively encoding the spatial context information of each pixel. This architecture has been extended to remote sensing applications too [50], [46].

In the proposed CD framework, we exploit a multi-layered CNN designed for semantic segmentation and trained on aerial optical images (thus accepting NIR input) to obtain pixelwise multi-layer deep features implicitly modeling the spatial context information of each pixels. Obtained deep features are sub-optimal as they are obtained from a CNN trained on different datasets and tasks. However, they reasonably capture object level information and the pre-trained CNN weights are still useful even if they are derived on other kinds of images. We exploit those sub-optimal deep features in a novel CD architecture. An automatic variance-ranking based feature selection strategy layerwise selects change-relevant deep features on spatial sub-splits to ensure that changes having low prior probability are retained. Selected deep features from multiple layers of CNN are combined to form a deep feature hypervector that aggregates multiple-scale abstractions. The use of the same pre-trained network on the pre-change and the post-change images enables us to obtain an unsupervised pseudo-Siamese architecture [51], [52], thus enabling a pixelwise comparison of deep features obtained from multi-temporal VHR images to obtain a deep change (hyper)vector. Assuming that unchanged pixels yield similar deep hypervectors, we follow the paradigm of CVA to discriminate between changed and unchanged pixels by analyzing the magnitude of deep change vectors (i.e., we define a Deep Change Vector Analysis - DCVA). Moreover, semantically rich deep features extracted from pre-trained CNN effectively capture information related to different types of change. To this end we binarize the deep change vectors to compress and simplify the deep change vectors, while retaining the information related to direction of changes. Then we further analyze the detected changed pixels based on binarized deep change vectors in order to extract multiple change information. Such a deep-feature-based CD framework preserves the simplicity of pixelwise comparison, while capturing spatial context information which is essential for VHR images understanding. Such a simplicity is important as many applications of remote sensing require (near) real-time processing capabilities. The proposed CD framework is not dependent on image segmentation or explicit object detection. Object level information is captured implicitly but effectively. CNN based feature-extraction strategy is robust to local spurious radiometric and geometric differences existing in multi-temporal VHR images. Deep change vectors contain semantically rich information relevant to both binary and multiple CD.

An ad hoc CNN for change detection could be trained that accounts for the information from multitemporal images simultaneously. This would require tens of thousands of patches of image pairs (pre-change, post-change) and corresponding change labels. However, multitemporal reference data are seldom available. Therefore we design an unsupervised approach and supervised methods are out of the scope of this work. In contrast to the state-of-the art CD methods based on deep learning, the proposed CD framework i) is completely unsupervised as it does not use any multi-temporal image for training the CNN, thus it does not require training overhead in terms of computational time/resources, ii) exploits the recently popular deep learning paradigm to obtain multiple-change information, thus solving the challenging task of multiple CD in VHR optical images, and iii) is reusable as new set of images does not need further training/tuning.

The rest of this paper is organized as follows. The problem statement and a brief synopsis of the proposed solution are presented in section II. Section III details the proposed method. We present datasets and results in section IV. Finally we conclude the paper and discuss scope of future research in section V.

II. Problem formulation and synopsis of the proposed solution

We aim to design a CD framework, specifically addressing the following aspects:

1) Let X_1, X_2 be two VHR optical images taken over the same geographical region at time t_1, t_2, respectively, using the same sensor. We aim to detect changes from the images in an unsupervised manner, dividing the set

of all pixels Ω into two subsets Ω_c and ω_{nc} corresponding to changed and unchanged pixels, respectively. Information in VHR images is complex and often redundant. Delineating change in such a scenario is not straightforward and can be subjective. If each pixel is considered individually, many pixels may have considerable radiometric dissimilarity between pre and post-change images. However, often such dissimilarities are either not meaningful from an object-based perspective or too small to be considered as a change of interest. Such small isolated radiometric dissimilarities are thus not considered as change.

2) Further analyzing the changed pixels (Ω_c) in an unsupervised way, we aim to separate different kinds of change $\omega_{c1}, \omega_{c2},..., \omega_{cK}$ which contain information related to different landscape characteristics. Here, $\bigcup_{k=1}^{K} \omega_{ck} = \Omega_c$ and $\omega_{ck1} \cap \omega_{ck2} = \emptyset$ $(\forall k1, k2 : 1 \leq k1 < k2 \leq K)$.

The proposed deep feature based unsupervised CD framework addresses the above aspects in effective manner. After geometric and radiometric pre-processing, both pre-change and post-change images are processed through a N -layered CNN pre-trained for semantic segmentation. Deep features are extracted from a set of layers L of the CNN where cardinality of L is $|L| \leq N$. Let us denote features extracted from layer $l \in L$ as f_l^1 and f_l^2 for the pre-change image and the post-change image, respectively. They are layerwise compared and a subset of features more informative for change detection is chosen to obtain a D-dimensional deep change vector G. Deep magnitude ρ is calculated from G and analyzed to distinguish changed pixels Ω_c from unchanged pixels ω_{nc}. Deep change vectors in G corresponding to detected changed pixels Ω_c are converted into binary deep change vectors G_{bin} and hierarchically clustered to distinguish different types of change ($\omega_{c1}, \omega_{c2}, ..., \omega_{cK}$). The proposed CD framework is shown in figure 1 and referred to as Deep CVA (DCVA).

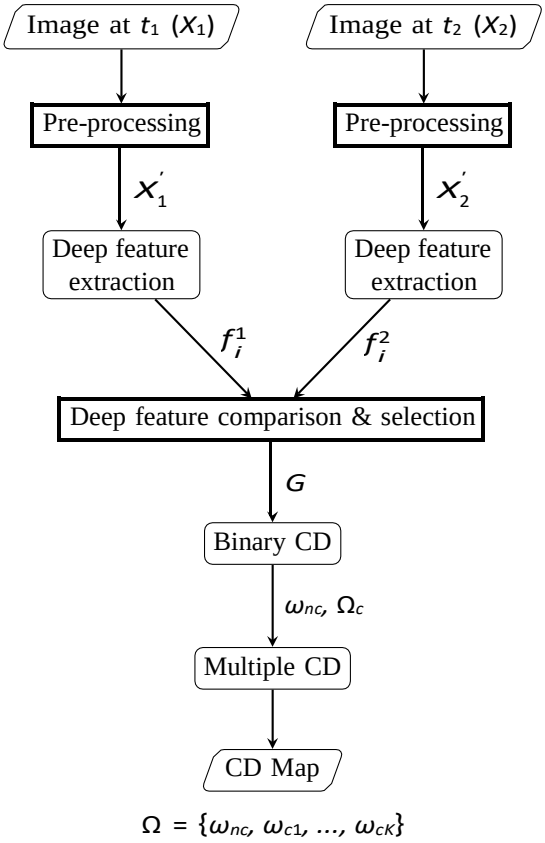


Fig. 1: Proposed Deep CVA technique

III. Proposed Deep Change Vector Analysis (DCVA)

DCVA is accomplished in the following steps: i) multi-temporal image pre-processing; ii) multi-temporal deep feature extraction; ii) deep feature comparison and selection; iii) binary CD; iv) multiple CD.

A. Multi-temporal image pre-processing

The input bi-temporal images X_1, X_2 are first pre-processed to remove distortions introduced by the atmosphere and other physical phenomena [11]. We apply radiometric normalization which involves conversion of Digital Number (DN) values into the corresponding ground reflectance values [53]. To achieve this, the digitalization process performed at the optical sensor during acquisition is inverted to obtain radiance values. This step is followed by atmospheric correction [53]. Images are further processed to mitigate geometric differences caused by different viewing angles of satellite sensors as well as misalignments caused by the impact of topography. Images are then processed through following subsequent steps - orthorectification, pansharpening, and coregistration. Following orthorectification, pansharpening is applied to integrate the multispectral bands, which have rich spectral but poor geometrical content, and the panchromatic band, which has rich geometrical content [54]. The result is a set of multispectral channels showing both high spatial and spectral resolution. It has been shown in the literature that change detection benefits from pansharpening [54], [11]. Generally in satellite optical VHR images, panchromatic bands have a geometric resolution 4 times higher than the multispectral bands. Thus we gain spatial resolution by 4 times after pansharpening. Here pansharpening is applied by Gram-Schmidt method [55]. Finally a co-registration step follows. After co-registration we obtain images X_1' X_2' which are then further processed to extract change information. Pre-processing steps help to mitigate radiometric and geometric differences in the multi-temporal images, but in real applications local differences still exist in the multi-temporal image set. As the proposed subsequent CD framework exploits a CNN based feature extraction strategy to encode spatial context information, we expect the proposed framework to be less affected by the residual local differences. This is because CNN based features are partially invariant to translation, distortion, and scaling [56] and exhibit more robust performance in description of data than the shallow features [57], [58].

B. Deep-feature extraction

CNNs are multi-layered deep architectures that capture levels of increasing abstraction and complexity throughout

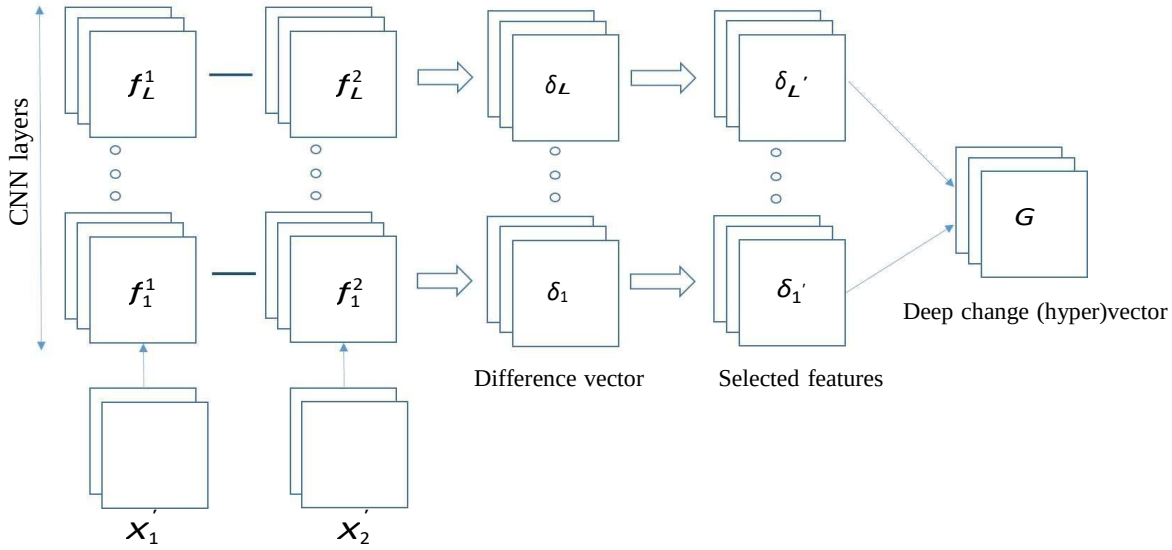


Fig. 2: Deep change (hyper)vector (G) generation

the feature hierarchy and can learn powerful discriminative features. Our objective is to effectively exploit such a multi-layered CNN to extract features to be used for change detection. Towards this goal, we obtain multi-temporal deep features by passing the pre-processed images (X_1', X_2') separately as input to a pre-trained CNN and extracting features from certain layers of the CNN. A CNN architecture consists of many layers (N) and each layer in turn consists of many features. Each feature has learned some complex visual concepts during the training process. The challenge for us is to effectively use those features to design a CD framework. Though we use a pre-trained CNN network like a blackbox feature extractor, it is important to select a suitable pre-trained CNN for the CD framework and to choose suitable layers to extract features from the CNN. Note that our challenge in the development of the unsupervised method is to define an approach for exploiting sub-optimal deep features for an accurate change detection task. CNN models trained on natural image datasets can accept only three channel RGB input, whereas optical satellite images generally have 4 (red, green, blue, Near InfraRed (NIR)) channels. Previous works on remote sensing with those architectures used only RGB input [41], [59], thus loosing a large amount of information. Currently some CNN architectures are available which are trained on remote sensing images [46], [27].

For the proposed DCVA, we can consider any possible architecture which can model remote sensing images to obtain pixelwise output. Features can be obtained from any of the N layers in the CNN. However, there is significant difference in characteristics of features depending on the layer from which they are extracted. As shown in [60], the initial layers of the CNN capture low-level visual concepts like edges, curves and color patches. As we go deeper, filters capture more complex concepts by combining lower level features of the previous layers. Such complex visual features are useful to analyze VHR images that are characterized by high local spatial variation. But as we go deeper in the CNN, features lose generalization capability to new inputs and also lack spatial fineness. A balance needs to be striked when choosing layers to extract features. In the literature, it has been shown that obtaining features from multiple layers of CNN is more effective than obtaining features from a single layer [61], [43] as it allows reasoning at multiple levels of abstraction and scales. Following that, we propose to combine features from multiple layers of the architecture to form a hypervector of features (the concept is similar to hypercolumn in [61]) which enables us to obtain a multi-scale representation. The hypervector of features is obtained by choosing a set of layers L from the total number of layers N. The hypervector of features derived from CNN is based on capturing the essence of both the deeper layers that model semantic information but lack spatial fineness and the initial layers that model the spatial properties of the image in a better fashion. It combines the responses of multiple CNN layers in a concatenated format. CNN learns higher level representation of data by means of convolutions. Convolutional layers are generally followed by ancillary layers like batch normalization, non-linear activation and pooling layers. They help in imposing regularization effects on the learned network [62], incorporating non-linearity in network and preventing overfitting. However they are just deterministic functions applied to the outputs of the convolutional layers and thus do not contain any additional information for feature hypervector. Thus, we choose a set of layers L from the convolutional layers and the deconvolutional layers. Features from each layer l in L are upsampled using bilinear interpolation to the spatial size of the input image to obtain f_l^1 and f_l^2 corresponding to the pre-change and the post-change images, respectively.

In this work, we chose the pre-trained model provided by Volpi and Tuia [46]. This network is trained on a remote sensing aerial image dataset for semantic labelling. The given architecture accepts 5 channel input, Red, Green, Blue, NIR,

and Digital Surface Model (DSM). Obtaining DSM data at very high resolution is difficult and in the context of multi-temporal applications we do not expect they provide a significant contribution. Thus we modify the first layer of the architecture to exclude DSM input. Though exclusion of the DSM input may impact the purity of the CNN network as a classifier, it does not significantly impact its use in the proposed CD architecture. DCVA exploits the sub-optimal CNN network in a quasi-Siamese style to pass the pre and post-change images through the same embedding function. The CNN is pre-trained for a classification task and adapted to solve an unsupervised change detection task. Though the weights learned by the CNN are sub-optimal for the change detection task, they generate similar outputs for pre-change and post-change images in case of no-change whereas significant different behaviours are expected in case of change. Thus the deep feature comparison in following steps allows us to to effectively extract multiple change information.

C. Deep-feature comparison and selection

The number of features obtained from each layer is high (up to 512 for the model provided in [46]). Some of these features may carry information relevant for CD whereas others do not. To identify relevant features, we use a feature selection strategy based on a variance measurement inspired by [63]. Assuming a fixed set of layers L, feature selection is done layerwise, i.e., individually for each $l \in L$. A deep change vector G is obtained as concatenation of selected features from each layer in L.

For a given layer l, a deep layerwise difference vector δ_l is computed by subtracting f_l^1 from f_l^2. In δ_l, there will be a subset $\delta_{l'}$ of features more sensitive to change information. Here we use the variance as an index of sensitivity to change information. We assume that features containing potentially relevant change information have higher variability than those less affected by changes as feature values strongly vary between changed and unchanged pixels. To select $\delta_{l'}$ ensuring that the features that model changes having low prior probability are considered, a spatial-split-based approach is used [63]. Features are spatially divided into S splits. For a given split s, feature variance ($\sigma^2_{l,s}$) is calculated for all features in δ_l. Features having higher $\sigma^2_{l,s}$ values are assumed to have potentially relevant change information. Thus, features in δ_l are sorted as per the descending order of $\sigma^2_{l,s}$ values. A subset δ_{ls} is selected by retaining a certain percentile of sorted δ_l. All the selected features $\delta_{l'}$ for layer l are obtained by taking features selected on each split:

$$\delta_{l'} = \bigcup_{s=1}^{S} \delta_{ls} \tag{1}$$

Selected features from each layer in L are concatenated to obtain deep change (hyper)vector (G):

$$G = (\delta_{1'}, ..., \delta_{l'}, ...\delta_{L'}) \tag{2}$$

G is a D— dimensional vector ($D = |\delta_{1'}| + |\delta_{2'}| + ... + |\delta_{L'}|$) with each component represented by g^d ($d = 1, ..., D$). A simplified block diagram of obtaining G is shown in figure 2.

D. Binary Change Dectection

In this step, we discriminate between unchanged (ω_{nc}) and changed (Ω_c) pixels based on the assumption that unchanged pixels yield similar deep features whereas changed pixels yield dissimilar deep features. Accordingly, components g^d of deep change vector G have higher value for changed pixels than for unchanged ones. For a comprehensive comparison of changed and unchanged pixels, deep magnitude ρ is computed for each pixel (r, c) following the popular technique of Change Vector Analysis (CVA) [64]:

$$\rho(r,c) = \sqrt{\sum_{d=1}^{D} (g^d)^2} \tag{3}$$

ρ maps D—dimensional G into a 1—dimensional index, while preserving the main properties of the changes despite this strong compression of the feature vector size. In lieu of ρ, any other possible feature reduction strategy that preserves the properties of the changes could be used. ρ is expected to assume larger values in case of changed pixels compared to unchanged ones. So, $\rho(r, c)$ are divided into two groups using a decision boundary or threshold value (T) to obtain two sets Ω_c($\rho \geq T$) for changed pixels and ω_{nc}($\rho < T$) for unchanged pixels.

Any suitable thresholding technique can be employed to obtain the threshold value T. In our work, we have explored two techniques: a global technique and a local adaptive one. Otsu's global thresholding is a popular method to determine decision boundary between changed and unchanged pixels [43], [16]. Using this method, the decision boundary value T_{otsu} applies to the whole area. The set of pixels $\rho(r, c)$ is classified into Ω_c, ω_{nc} according to the following rule:

$$\rho(r,c) \in \begin{cases} \Omega_c, \text{ if } \rho(r,c) \geq T_{otsu} \\ \omega_{nc} \text{ otherwise} \end{cases} \tag{4}$$

An alternative to global thresholding is local adaptive thresholding which computes $T_{local}(r, c)$ as a function of spatial position (r, c). Local adaptive thresholding may capture the strong local variation in the VHR images and furthermore contextual characterization by deep features. Low pass filtering is used for local adaptive decision boundary determination [65], [66]. Here, we have chosen Gaussian filtering to decide local adaptive threshold. Such a filtering scheme generates context dependent threshold value for each pixel by producing a weighted average of each pixel neighbourhood. The idea is similar in essence to the idea of image split based thresholding proposed in [63]. In our case each pixel is a split and a window around each pixel is used to decide the threshold for it. Here we obtain a decision boundary $T_{local}(r, c)$ that is function of the spatial position (r, c). Pixels are classified into Ω_c, ω_{nc} according to the following rule:

$$\rho(r,c) \in \begin{cases} \Omega_c, \text{ if } \rho(r,c) \geq T_{local}(r,c) \\ \omega_{nc} \text{ otherwise} \end{cases} \tag{5}$$

A parameter to decide context dependent threshold using Gaussian filtering is the neighbourhood size of the filter (α). Changed objects having different sizes can be captured using

different neighbourhood sizes. The neighbourhood size can be varied based on the application and fixed if a priori knowledge on the expected size of change is available. As an alternative a multiscale approach can be designed that iteratively increases the neighbourhood size (α) and captures changed objects of different sizes. The final change map is obtained as a set union of the change maps obtained by different values of α.

E. Multiple Change Dectection

Changed pixels in Ω_c are further analyzed to separate different kinds of change. This is done in an unsupervised way without any training set or apriori knowledge about the different kinds of change. Our goal is to group different changed pixels into different clusters. Deep change hypervector G is a high dimensional vector and hence clustering is challenging due to curse of dimensionality [67]. Very few multiple CD methods have the capability to handle high dimensional vectors [14], [68]. High dimensional spaces are inherently sparse due to the small number of data samples compared to the number of dimensions. Thus the inter-point distances become less informative. Discretization reduces number of values of continuous feature and thus simplifies the clustering task. In this context, binarization of the direction information has been found to be effective [68]. Following this, we have devised a simple yet effective approach for clustering G based on feature binarization and hierarchical clustering which allows to identify features that are descriptive of clusters.

Considering changed pixels properties we are likely to have components of G which are either positive or negative, and different kinds of change are likely to show different behaviours on the g^d ($d = 1, ..., D$) components of G. Feature binarization is an effective approach to simplify the information in G by preserving information relevant to change directions [68]. We binarize G with all components greater than 0 set to 1 and all components smaller than 0 set to 0. Thus, we obtain a binary vector G_{bin} having the same dimension D as G. Each component of G_{bin} is represented by g^d_{bin} ($d = 1, ..., D$), where:

$$g^d_{bin} = \begin{cases} 1, \text{ if } g^d \geq 0 \\ 0 \text{ otherwise} \end{cases} \quad (6)$$

Thus each kind of change is expected to correspond to one specific binary signature. Following this, we hierarchically cluster all pixels in Ω_c based on G_{bin} to obtain different classes $\omega_{c1}, \omega_{c2}, ..., \omega_{cK}$ by constructing a decision tree [69]. Decision tree selects the most informative binary feature and groups the pixels into two classes based on the value of that feature. The most informative feature is selected inspired by the redundancy criteria [70]. It is plausible to assume that many features exhibit similar binary signatures and hence are redundant for discriminating different types of change. We define that a feature g^j_{bin} (j^{th} feature) is redundant given a feature g^i_{bin} (i^{th} feature) if they are equal for a large number of pixels in Ω_c. To measure redundancy, Hamming distance $H(i, j)$ between two features is used, which is defined as number of pixels in Ω_c for which the i^{th} and the j^{th} features differ. We define an index R_d for each feature d ($d = 1, ..., D$) that measures the informativeness of an individual feature by summing up Hamming distances with all features:

$$R_d = -\sum_{j=1}^{D} H(d, j) \quad (7)$$

The most informative feature (d^*) is defined as the feature given which most other features are considered redundant, i.e., which maximizes R_d:

$$d^* = \arg\max_d R_d \quad (8)$$

After using d^* to group pixels in Ω_c into two classes and discarding the features made redundant by it, the next most informative feature is selected and pixels in Ω_c are further clustered based on it. In this way, by selecting K' features, pixels in Ω_c are grouped into $2^{K'}$ classes. Subsequently, the number of desired classes ($K \leq 2^{K'}$) is obtained by agglomerating classes based on similarity of binary signatures. Though simple, this approach is effective demonstrating that deep features contain information that can effectively discriminate between different types of change.

IV. Experiments and Results

Section IV-A describes validation datasets, and section IV-B illustrates our choice of layers L for deep feature selection. Section IV-C presents the tenets of an experiment that shows that the variance/ standard deviation is an effective variable to choose features containing change information. Following that, section IV-D presents binary CD results and section IV-E presents multiple CD results.

A. Description of datasets

In order to validate the proposed CD framework, three bi-temporal image pairs acquired by three sensors have been used. The images are from Trento in North Italy. They show a quasi-urban area containing both the typical urban setup and patches of vegetation. They also contain vertical structures of different heights. Thus, effect of shadows is prominent as tall vertical structures generate different shadows in off-nadir VHR images. Such high elevations and variation in shadows are prone to cause false alarms in CD on such VHR images in contrast to the case of low/medium resolution images.

The Worldview-2 image pair was acquired in August 2010 and May 2011. Pre and post-change images were acquired with 18°and 12.9°off-nadir angle (figure 3(a) and (b)), respectively. They have 0.5 meter/pixel spatial resolution and a size of 1200 $\times$ 1200 pixels. Worldview-2 has 8 spectral bands in the spectral range 400-1040 nm. A reference CD map (figure 3(c)) has been obtained primarily using photo-interpretation combined with prior knowledge on the evolution of the analyzed geographic area. Changed pixels have been grouped into the reference map based on their characteristics. Three kinds of change were identified: ω_{c1} denotes a significant increase of vegetation amount (green, 8524 pixels, and 2 objects), ω_{c2} denotes the formation of white patches in place of semi-vegetation areas (blue, 5926 pixels, and 2 objects), and

ω_{c3} indicates the formation of colored structure (e.g., building) in place of white patches (red, 6332 pixels, and 2 objects).

The Pleiades image pair was acquired in August 2012 and September 2013. The pre-change image (figure 4(a)) was acquired with 16.9°off-nadir angle and the post-change image (figure 4(b)) with 20.9°off-nadir angle. They have 0.5 meter/pixel resolution and a size of 1400 $\times$ 1400 pixels. Pleiades has 4 spectral bands in the spectral range 450-900 nm. A reference change map is shown in figure 4(c). We demarcated two kinds of change: ω_{c1} denotes change of vegetation amount (green, 23236 pixels, and 5 objects) and ω_{c2} denotes changes related to the color of urban structure (red, 11945 pixels, and 2 objects).

The Quickbird image pair was acquired in October 2005 (figure 5(a)) and July 2006 (figure 5(b)). Pre and post-change images were acquired with 9°and 14.1°off-nadir angle, respectively. They have of 0.6 meter/pixel resolution and a size of 800 $\times$ 800 pixels. Effect of seasonal change and shadow is very prominent and hence this dataset can be considered more complex than the others. Quickbird has 4 spectral bands in the spectral range 485-830 nm. A reference CD map is shown in figure 5(c). We demarcated three kinds of change: ω_{c1} denotes increment in the greenness of the vegetation (green, 17133 pixels, and 2 objects), ω_{c2} denotes sharp reduction of vegetation (blue, 8432 pixels, and 1 object), and ω_{c3} indicates change in urban structures (red, 19040 pixels, and 2 objects).

B. Analysis on the choice of CNN and L

To extract deep features, we apply the pre-trained CNN proposed in [46], which has 33 layers: convolutional, batch normalization, (ReLu) activation, pooling, dropout, convolutional, batch normalization, activation, pooling, dropout, convolutional, batch normalization, activation, pooling, dropout, convolutional, batch normalization, activation, dropout, deconvolutional, batch normalization, activation, dropout, deconvolutional, batch normalization, activation, dropout, deconvolutional, batch normalization, activation, dropout, convolutional, and softmax prediction. CNN Dropout becomes a pass-through layer during feature extraction, thus in reality it does not apply to our processing. The network is trained on a remote sensing aerial optical image dataset and accepts R, G, B, NIR, and DSM input. As discussed in section III-B, DSM input is excluded by modifying the weights of first layer of CNN from 7×7 (spatial size of filter) $\times$ 5 (number of input channels) $\times$ 64 (number of output features) to $7 \times 7 \times 4 \times 64$. Worldview-2 images have 8 channels: red, green, blue, NIR, NIR-2, red edge, coastal blue, yellow. For Worldview-2 dataset, first layer of the CNN is modified to $7 \times 7 \times 8 \times 64$ by assigning to the additional input channels the same weights as those of the nearest channel in terms of wavelength, i.e., NIR-2 gets the same weight as NIR input. Though such modification of first layer is not optimal for using the CNN network as a classifier, note that this it does not significantly impact its use in the proposed CD architecture that does not exploit the network as a classifier but uses it in a pesudo-Siamese fashion to pass the pre and post-change images through the same embedding function.

As described in section III-B, layers like pooling, ReLu are excluded from hypervector, as they are simply deterministic functions applied to the outputs of the convolutional or deconvolutional layers and thus do not contain any additional information for feature hypervectors. The convolutional layers learn the semantics of the image at a degraded resolution and the deconvolutional layers mainly learn to reconstruct the spatial arrangements [46]. Based on this, we chose more convolutional layers than deconvolutional ones to form the hypervector. First convolutional layer is excluded as it learns very primitive features that are significantly noisy. Key layers of the pre-trained CNN with feature dimension is denoted in table I. Deep feature hypervectors can be built with several combinations of layers. DCVA demonstrated to be robust to the minimal variations in layer selection resulting in good and similar performance for several different combinations. For sake of brevity, we show results for $L = \{28, 16, 11, 6\}$, i.e., the 3rd deconvolutional, 4th, 3rd, and 2nd convolutional layers of the CNN and omit other combinations.

C. Validation of the feature selection techniques

In section III-B, we postulated that variance/standard deviation based ranking is a simple but effective approach to select a subset of features being sensitive to change information. Since deep feature selection is done layerwise (see section III-C), validation was conducted for each layer in L independently. For sake of brevity, we demonstrate this for one layer (layer 28) and we omit the other layers since results are similar. To illustrate this process, we chose two test regions from Pleiades multi-temporal images where change occurred (blue, yellow in figure 6) and two test regions from same couple where change did not occur (red, green in figure 6). Pre and post-change images marked with the four regions are shown in figure 6(a) and 6(b). All regions are of size 200 pixel $\times$ 200 pixel, i.e. 100 meter $\times$ 100 meter. Deep features for the 28th layer (f_{28}) are evaluated for both the pre and the post-change images to obtain δ_{28}. The number of features from δ_{28} is 512. Standard deviation is computed (σ_{28}) for all the 512 features for each of the four regions individually. We sort the σ_{28} values in descending order for each region (figure 6). We observe that the σ_{28} for both changed regions are much higher than those for the unchanged regions. This confirms that variance is a good index for layerwise selection of subset of features that are more sensitive to change information.

Plain concatenation of features from $L = \{28, 16, 11, 6\}$ would have produced a deep change vector (G) of dimension 960, whereas the dimension of G after feature subset selection is $d = 63$ for the Worldview-2 images, $d = 85$ for the Pleiades images, and $d = 86$ for the Quickbird images. Features have been selected with a split size of 400 $\times$ 400 pixels for Worldview-2 and Quickbird dataset, 350 $\times$ 350 pixels for Pleiades dataset (slight variation to obtain integer valued number of splits). Larger split size causes omission of features that model changes having low prior probability, thus increasing missed detection. Smaller split size allows to capture features that increase false alarm.

D. Binary CD results

Experiments were conducted to achieve the following goals:

- *To compare the performance of global and local adaptive thresholding methods for the ρ variable.* For local adaptive thresholding, we have used α values (as discussed in section III-D) of 50 meter, 100 meter and 150 meter. As our test images are quasi urban, we expect all changes to be captured by the selected neighbourhood sizes.
- *To evaluate if formation of hypervector of layers is more effective than choosing a single layer to extract deep features.* We demonstrate the performance of DCVA by using $L = \{28, 16, 11, 6\}$ (as a hypervector of layers) versus $L = \{28\}$ and $L = \{6\}$ (as single individual layers).
- *To demonstrate that DCVA outperforms a mere post-classification CD setup.* We demonstrate performance of DCVA by using $L = \{28, 16, 11, 6\}$ versus $L = \{33\}$ (the final layer).
- *To compare results obtained by DCVA to the state of the art methods like RCVA, OCVA, and PCVA.* Those methods calculate change magnitude followed by discrimination into changed and unchanged pixels using a decision boundary. For a fair comparison, the state of the art methods are modified to equip them with the same decision boundary determination scheme as used for DCVA. Comparison is shown for all the three considered datasets.

Results are provided in terms of sensitivity and specificity. Sensitivity measures the proportion of changed pixels that are correctly identified as such. Specificity measures the proportion of unchanged pixels that are correctly identified as such [71]. Additionally, number of correctly identified and false changed objects are provided.

Worldview-2 dataset: Change detection results for Worldview-2 bi-temporal images are shown in figure 3. The proposed DCVA using $L = \{28, 16, 11, 6\}$ and a context dependent adaptive decision boundary T_{local} obtains a sensitivity of 0.86 and a specificity of 0.98. It detects all the 6 changed objects and 4 false objects as shown in figure 3(d). The proposed DCVA using same layers and the global decision boundary determination obtains a sensitivity of 0.94 and a specificity of 0.89. It detects all the true changed objects and many false objects as shown in figure 3(e). Thus, context dependent adaptive decision boundary scheme outperforms the global decision boundary one.

As described in section III-B and section IV-B, we used features from multiple layers ($L = \{28, 16, 11, 6\}$) of the CNN to form a hypervector of features. To demonstrate that this strategy is effective, we show the results obtained using the proposed DCVA but using only $L = \{6\}$ (i.e., 2nd convolutional layer) in figure 3(f) and only $L = \{28\}$ (i.e., 3rd deconvolutional layer) in figure 3(g). As $L = \{6\}$ is an initial layer of the CNN, it captures primitive features and hence it is prone to false alarms caused by edges. As $L = \{28\}$ is a deeper layer, it learns more high-level abstract features. However it fails to extract some change information specially those related to vegetation patches. Moreover, detected boundaries of changed objects are less accurate compared to those in figure 3(d) as deeper layers have reduced spatial accuracy.

We further show the result obtained using DCVA but using the final layer of the CNN, i.e. $L = \{33\}$ in figure 3(h). It fails to capture most of the change information demonstrating the difference of DCVA from a mere post-classification CD approach.

Results obtained by using RCVA are shown in figure 3(i). It detects 5 out of 6 changed objects and many false objects. OCVA and PCVA detect 3 and 4 changed objects, respectively, and many false objects (figures 3(j) and 3(k)). Results demonstrate that the proposed method is more effective than the considered state-of-the-art ones. The proposed method produces more accurate boundaries for changed objects and it is less prone to error due to edges and shadows. However, there are still few false alarms due to high elevations and shadows. This can be attributed to the large difference of acquisition angle of the pre and post-change images. There are also false alarms due to inaccurate decision boundary demarcation in the vegetation area. Indeed, decision boundary determination in such inhomogeneous images with high local and global variance is challenging. Table II shows quantitative results. An analysis of the table clearly shows that DCVA outperforms state-of-the-art methods in terms of sensitivity (accuracy in detecting changed pixels) and produces similar or better specificity (accuracy in detecting unchanged pixels).

Pleiades dataset: CD results for the Pleiades data are shown in figure 4. The proposed DCVA using $L = \{28, 16, 11, 6\}$ and a context dependent adaptive decision boundary T_{local} achieves a sensitivity of 0.84 and specificity of 0.96 and detects 6 out of the 7 changed objects as shown in figure 4(d). DCVA using global decision boundary determination achieves a sensitivity of 0.89 and specificity of 0.86 and detects 6 changed objects and many false objects as shown in figure 4(e). Proposed DCVA using $L = \{6\}$, $L = \{28\}$, or $L = \{33\}$ (figure 4(f-h)) exhibits similar characteristics as in Worldview-2 dataset. RCVA, OCVA, and PCVA detect 5, 3, and 4 changed objects respectively and many false objects as shown in figures 4(i-k). It is evident that DCVA outperforms the considered state-of-the-art methods in terms of sensitivity as shown in Table III. Concerning specificity, DCVA outperforms OCVA and obtains similar result to RCVA and PCVA. There are some false alarms due to high elevation and shadow, which can be attributed to a quite large difference of acquisition angle of images (4°).

Quickbird dataset: CD results for the Quickbird multi-temporal images are shown in figure 5. Despite the strong effect of shadow and seasonal changes in Quickbird images, DCVA performs similarly as on other datasets. DCVA achieves a sensitivity of 0.87 and specificity of 0.93 and detects all 5 changed objects as shown in figure 5(d). DCVA using global decision boundary determination achieves a sensitivity of 0.86 and specificity of 0.91 and detects 4 out of 5 changed objects as shown in figure 5(e). Proposed DCVA using $L = \{6\}$, $L = \{28\}$, or $L = \{33\}$ (figure 5(f-h)) exhibits similar characteristics as in the other two datasets. RCVA, OCVA, and PCVA detect 2, 0, and 1 changed objects only, respectively. Results obtained using RCVA, OCVA, and PCVA methods

TABLE I: Key structure of the pre-trained CNN [46]

Layer number	Layer type	Feature dimension
6	convolutional	64
11	convolutional	128
16	convolutional	256
20	deconvolutional	512
24	deconvolutional	512
28	deconvolutional	512
32	convolutional	6
33	prediction	6

TABLE II: Binary CD results (Worldview-2 images)

Method	Sensitivity	Specificity
Proposed DCVA	0.86	0.98
DCVA (T_{otsu})	0.94	0.89
DCVA ($L = 6$)	0.76	0.97
DCVA ($L = 28$)	0.36	0.98
DCVA ($L = 33$)	0.28	0.98
RCVA	0.49	0.97
OCVA	0.25	0.95
PCVA	0.49	0.96

are shown in figures 5(i-k), respectively. Specificity value of DCVA slightly decreases for the Quickbird images compared to the other two datasets. This is due to a slight increment of false alarms. Some of the false alarms, especially those adjacent to the roads, are indeed real radiometric changes and denote growth of tiny green patches besides road. Some other false alarms are caused by shadow and edges of buildings. However, a stronger impact is recorded on the performance of the state-of-the-art methods that all show a significantly lower sensitivity compared to their performance on other datasets. Table IV presents detailed quantitative results.

It is evident from the presented results that DCVA strongly outperforms the state-of-the-art methods as the contrast in acquisition (difference of acquisition angle, difference in acquisition season) of pre and post-change images increases.

TABLE III: Binary CD results (Pleiades images)

Method	Sensitivity	Specificity
Proposed DCVA	0.84	0.96
DCVA (T_{otsu})	0.89	0.86
DCVA ($L = 6$)	0.80	0.97
DCVA ($L = 28$)	0.50	0.95
DCVA ($L = 33$)	0.10	0.87
RCVA	0.52	0.97
OCVA	0.52	0.87
PCVA	0.64	0.96

TABLE IV: Binary CD results (Quickbird images)

Method	Sensitivity	Specificity
Proposed DCVA	0.87	0.93
DCVA (T_{otsu})	0.86	0.91
DCVA ($L = 6$)	0.83	0.92
DCVA ($L = 28$)	0.67	0.95
DCVA ($L = 33$)	0.54	0.90
RCVA	0.34	0.85
OCVA	0.12	0.91
PCVA	0.34	0.91

E. Multiple CD results

Multiple CD experiments aim at demonstrating the ability of the deep features in separating different kinds of change. For all the three datasets, the set Ω_c after binary CD is further analyzed (section III-E) by setting K equal to the number of kinds of change demarcated in the reference map. For quantitative demonstration, we use the confusion matrix. As multiple change detection is dependent on the outcome of the binary change detection, error propagates from the binary CD to the multiple CD maps, and are shown in shaded cells of confusion matrices. The proposed hierarchical clustering approach is compared to C²VA approach [14] applied on deep change hypervectors (G).

Worldview-2 dataset: Multiple CD results obtained by the proposed DCVA are shown in figure 3(d) and illustrated in the confusion matrix shown in Table V(a). The increase of vegetation (green, ω_{c1}) has clearly been identified as a kind of change. Though there is a little clutter, ω_{c2} and ω_{c3} are also correctly identified. C²VA confused a fraction of ω_{c2} with ω_{c3} and vice-versa (figure 3(l) and Table V(b)).

Pleiades dataset: Multiple CD results obtained by the propsoed DCVA are shown in figure 4(d) and illustrated in the confusion matrix shown in Table VI(a). We observe that the discrimination of changes into two kinds, ω_{c1} and ω_{c2} (shown in green and red), has been satisfactory. C²VA confuses a major part of ω_{c2} as ω_{c1} as shown in figure 4(l) and illustrated in the confusion matrix in Table VI(b).

Quickbird dataset: Multiple CD results obtained by the proposed DCVA are shown in figure 5(d) and illustrated in the confusion matrix shown in Table VII(a). The increment in the greenness of the vegetation (ω_{c1}, green change) has been identified. Some parts of the urban change are erroneously grouped together with sharp reduction of vegetation (ω_{c2}, blue). The other changes related to the urban setup have been correctly grouped together (ω_{c3}, red). Multiple CD results obtained by using C²VA are shown in figure 5(l). As shown in the confusion matrix shown in Table VII(b), C²VA confuses significant part of ω_{c3} as ω_{c2}.

V. Conclusions

In this paper, a CNN based unsupervised technique for detecting changes in multi-temporal VHR optical images has been proposed. VHR images are highly complex and pixels have high spatial correlation in a neighbourhood. Moreover, inspite of being acquired by the same sensor, images often show strong differences in characteristics as angle of acquisition, season of acquisition, atmospheric condition. To mitigate these problems, we propose an unsupervised CD technique that exploits sub-optimal (due to the lack of training samples) deep features extracted from a pre-trained multi-layer CNN in a novel CD architecture. Recent results in both remote sensing and general computer vision have demonstrated that CNN based feature extraction has better generalization capability than traditional hand-crafted features or pixel radiance values and thus tends to be reasonably invariant to the differences in acquisition conditions. Moreover, such deep learning based features are suitable to capture contextual information. The

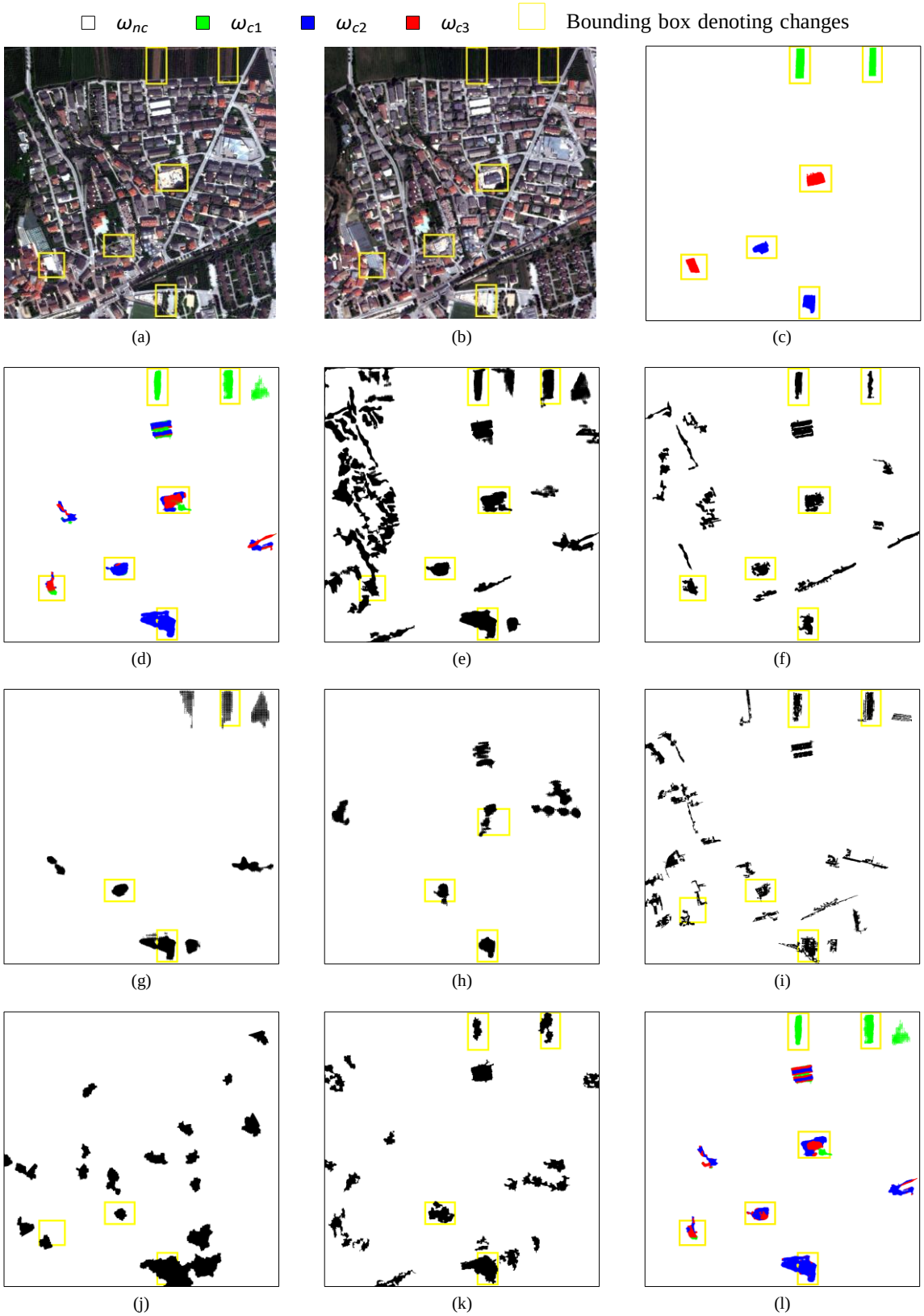


Fig. 3: Worldview-2 bi-temporal images: (a) Pre-change image (RGB), (b) Post-change image (RGB), (c) Reference change map. (d) CD map: Proposed DCVA. Binary CD map: (e) DCVA with global thresholding, (f) DCVA with $L = \{6\}$, (g) DCVA with $L = \{28\}$, (h) DCVA with $L = \{33\}$, (i) RCVA, (j) OCVA, (k) PCVA. (l) Multiple CD map: C^{2}VA [14]

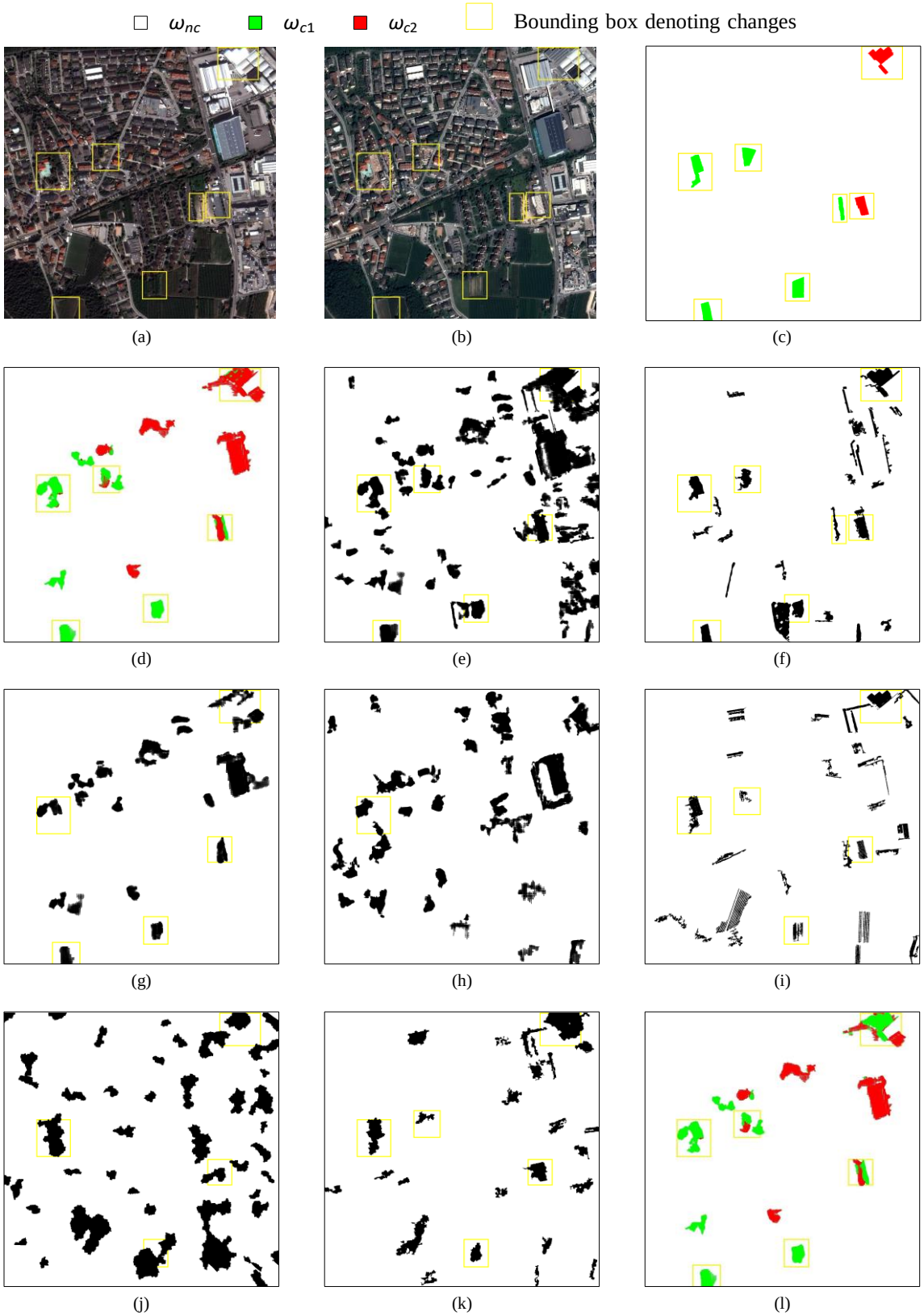


Fig. 4: Pleiades bi-temporal images: (a) Pre-change image (RGB), (b) Post-change image (RGB), (c) Reference change map. (d) CD map: Proposed DCVA. Binary CD map: (e) DCVA with global thresholding, (f) DCVA with $L = \{6\}$, (g) DCVA with $L = \{28\}$, (h) DCVA with $L = \{33\}$, (i) RCVA, (j) OCVA, (k) PCVA. (l) Multiple CD map: C^{2}VA [14]

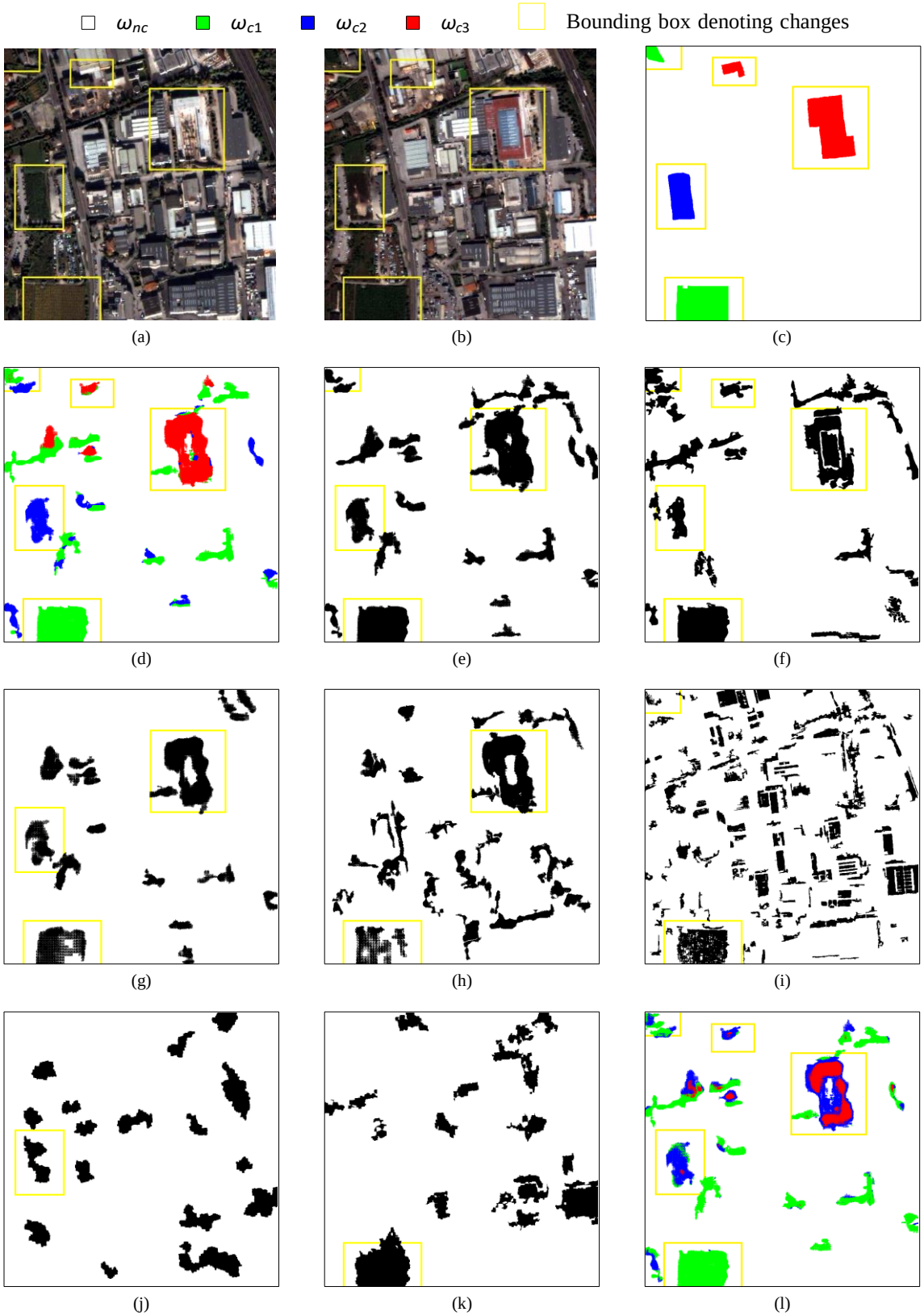


Fig. 5: Quickbird bi-temporal images: (a) Pre-change image (RGB), (b) Post-change image (RGB), (c) Reference change map. (d) CD map: Proposed DCVA. Binary CD map: (e) DCVA with global thresholding, (f) DCVA with $L = \{6\}$, (g) DCVA with $L = \{28\}$, (h) DCVA with $L = \{33\}$, (i) RCVA, (j) OCVA, (k) PCVA. (l) Multiple CD map: C^{2}VA [14]

proposed DCVA exploits these properties of deep features and processes those features through a layerwise feature selection mechanism that ensures that only change-relevant features are retained. A feature hypervector is subsequently formed by combining features from different layers of CNN that ensure the spatial context is captured at multiple level of abstractions. Pixelwise comparison of deep change vectors from the pre and post-change images enables us to obtain deep change vectors that are further analyzed to extract both binary and multiple-change information from multi-temporal VHR images. Binary CD is performed based on the magnitude of the deep change vectors. Multiple CD is performed by identifying the direction of changes after a compression of deep change vectors based on a binarization process and a subsequent clustering. Thus, on the one hand DCVA is effective in capturing the spatial context, on the other hand it preserves the simplicity of pixel based comparison. DCVA effectively exploits the recently popular CNN, without using any training data or supervision. Experiments have been conducted by using the proposed method on three datasets acquired by different sensors: Worldview-2, Pleiades, and Quickbird. Results demonstrated the effectiveness of the proposed approach in capturing change information. Despite images were highly complex consisting of quasi-urban areas, results demonstrate that the proposed DCVA is able to effectively capture spatial information and is more resilient than other state-of-the-art techniques to variations in acquisition conditions (e.g., acquisition angle).

As future development of this work we aim to improve the performance of binary change detection by developing better techniques for the decision boundary determination scheme to distinguish changed pixels from unchanged ones. The multiple change detection scheme can be improved by refining the hierarchical clustering technique for high dimensional deep change vector and automatically deciding the number of kinds of change.

As a final remark we point out that the proposed DCVA is a step forward in designing an effective unsupervised technique for multi-temporal VHR image analysis. Note that even if the DCVA is focused on the processing of bi-temporal images acquired by optical sensors, it can also be extended to active sensors (SAR) and image time-series.

REFERENCES

[1] G. Hulley, S. Veraverbeke, and S. Hook, "Thermal-based techniques for land cover change detection using a new dynamic modis multispectral emissivity product (MOD21)," *Remote Sensing of Environment*, vol. 140, pp. 755–765, 2014.

[2] S. Stramondo, C. Bignami, M. Chini, N. Pierdicca, and A. Tertulliani, "Satellite radar and optical remote sensing for earthquake damage detection: results from different case studies," *International Journal of Remote Sensing*, vol. 27, no. 20, pp. 4433–4447, 2006.

[3] D. Lu, E. Moran, and S. Hetrick, "Detection of impervious surface change with multitemporal landsat images in an urban–rural frontier," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 66, no. 3, pp. 298–306, 2011.

[4] A. Singh, "Review article digital change detection techniques using remotely-sensed data," *International journal of remote sensing*, vol. 10, no. 6, pp. 989–1003, 1989.

[5] L. Bruzzone, D. F. Prieto, and S. B. Serpico, "A neural-statistical approach to multitemporal and multisource remote-sensing image classification," *IEEE Transactions on Geoscience and remote Sensing*, vol. 37, no. 3, pp. 1350–1359, 1999.

TABLE V: Confusion matrix for multiple CD results (Worldview-2) using: (a) proposed hierarchical clustering and (b) C²VA

(a)

Reference	Predicted			
	ω_{nc}	ω_{c1}	ω_{c2}	ω_{c3}
ω_{nc}	1389054	8234	17546	4384
ω_{c1}	2017	6507	0	0
ω_{c2}	113	9	5777	27
ω_{c3}	696	736	239	4661

(b)

Reference	Predicted			
	ω_{nc}	ω_{c1}	ω_{c2}	ω_{c3}
ω_{nc}	1389054	6790	17592	5782
ω_{c1}	2017	6488	0	19
ω_{c2}	113	0	4859	954
ω_{c3}	696	236	1899	3501

TABLE VI: Confusion matrix for multiple CD results (Pleiades) using: (a) proposed hierarchical clustering and (b) C²VA

(a)

Reference	Predicted		
	ω_{nc}	ω_{c1}	ω_{c2}
ω_{nc}	1855218	19696	49905
ω_{c1}	5257	16793	1186
ω_{c2}	220	1421	10304

(b)

Reference	Predicted		
	ω_{nc}	ω_{c1}	ω_{c2}
ω_{nc}	1855218	23908	45693
ω_{c1}	5257	16503	1476
ω_{c2}	220	9907	1818

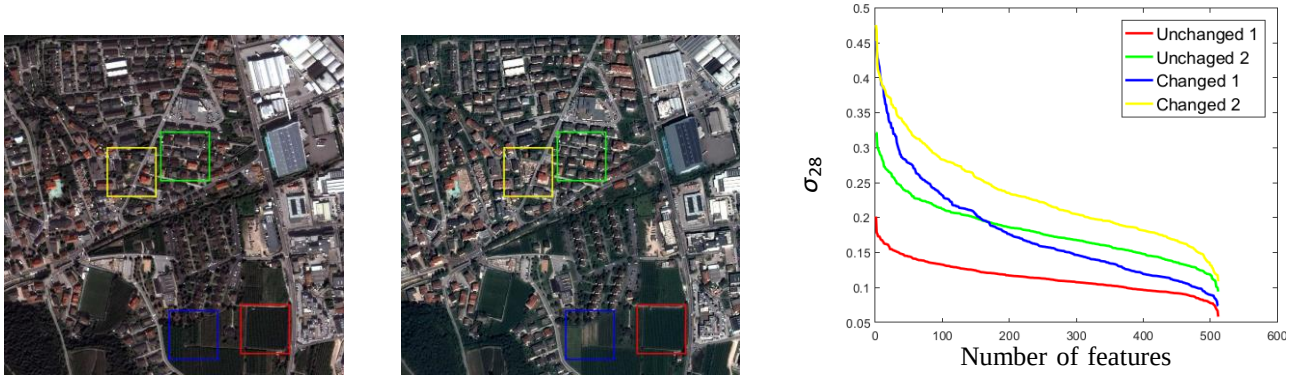
TABLE VII: Confusion matrix for multiple CD results (Quickbird) using: (a) proposed hierarchical clustering and (b) C²VA

(a)

Reference	Predicted			
	ω_{nc}	ω_{c1}	ω_{c2}	ω_{c3}
ω_{nc}	555445	23453	9945	6552
ω_{c1}	1428	15581	50	74
ω_{c2}	2267	0	6165	0
ω_{c3}	2111	505	413	16011

(b)

Reference	Predicted			
	ω_{nc}	ω_{c1}	ω_{c2}	ω_{c3}
ω_{nc}	555445	28105	9640	2205
ω_{c1}	1428	15127	578	0
ω_{c2}	2267	1070	4976	119
ω_{c3}	2111	100	8173	8656



(a) (b) (c)

Fig. 6: Pleiades dataset: standard deviation analysis for features extracted from the 28th layer for 2 changed region and 2 unchanged region: (a) Pre-change image (RGB), (b) Post-change image (RGB) - (2 changed regions are shown in blue and yellow rectangles, 2 unchanged regions are shown in red and green rectangles), (c) Difference standard deviation of features extracted from 28th layer (sorted in descending order)

[6] L. Bruzzone, R. Cossu, and G. Vernazza, "Detection of land-cover transitions by combining multidate classifiers," *Pattern Recognition Letters*, vol. 25, no. 13, pp. 1491–1500, 2004.

[7] L. Bruzzone and D. F. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Transactions on Geoscience and Remote sensing*, vol. 38, no. 3, pp. 1171–1182, 2000.

[8] L. Bruzzone and D. F. Prieto, "A minimum-cost thresholding technique for unsupervised change detection," *International Journal of Remote Sensing*, vol. 21, no. 18, pp. 3539–3544, 2000.

[9] T. Celik, "Unsupervised change detection in satellite images using principal component analysis and k-means clustering," *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 4, pp. 772–776, 2009.

[10] W. A. Malila, "Change vector analysis: an approach for detecting forest changes with Landsat," in *LARS symposia*, 1980, p. 385.

[11] Y. T. S. Correa, F. Bovolo, and L. Bruzzone, "Change detection in very high resolution multisensor optical images," in *SPIE Remote Sensing*, vol. 9244. International Society for Optics and Photonics, 2014.

[12] F. Bovolo and L. Bruzzone, "A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 1, pp. 218–236, 2007.

[13] F. Bovolo, S. Marchesi, and L. Bruzzone, "A nearly lossless 2d representation and characterization of change information in multispectral images," in *Geoscience and Remote Sensing Symposium (IGARSS), 2010 IEEE International*. IEEE, 2010, pp. 3074–3077.

[14] F. Bovolo, S. Marchesi, and L. Bruzzone, "A framework for automatic and unsupervised detection of multiple changes in multitemporal images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 6, pp. 2196–2212, 2012.

[15] G. Chen, G. J. Hay, L. M. Carvalho, and M. A. Wulder, "Object-based change detection," *International Journal of Remote Sensing*, vol. 33, no. 14, pp. 4434–4457, 2012.

[16] F. Thonfeld, H. Feilhauer, M. Braun, and G. Menz, "Robust change vector analysis (RCVA) for multi-sensor very high resolution optical satellite data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 50, pp. 131–140, 2016.

[17] F. Bovolo, "A multilevel parcel-based approach to change detection in very high resolution multitemporal images," *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 1, pp. 33–37, 2009.

[18] N. Falco, M. Dalla Mura, F. Bovolo, J. A. Benediktsson, and L. Bruzzone, "Change detection in VHR images based on morphological attribute profiles," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 3, pp. 636–640, 2013.

[19] N. Falco, G. Cavallaro, P. R. Marpu, and J. A. Benediktsson, "Unsupervised change detection analysis to multi-channel scenario based on morphological contextual analysis," in *Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International*. IEEE, 2016, pp. 3374–3377.

[20] L. Li, X. Li, Y. Zhang, L. Wang, and G. Ying, "Change detection for high-resolution remote sensing imagery using object-oriented change vector analysis method," in *Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International*. IEEE, 2016, pp. 2873–2876.

[21] P. Lv, Y. Zhong, J. Zhao, H. Jiao, and L. Zhang, "Change detection based on a multifeature probabilistic ensemble conditional random field model for high spatial resolution remote sensing imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 12, pp. 1965–1969, 2016.

[22] B. N. Subudhi, F. Bovolo, A. Ghosh, and L. Bruzzone, "Spatio-contextual fuzzy clustering with markov random field model for change detection in remotely sensed images," *Optics & Laser Technology*, vol. 57, pp. 284–292, 2014.

[23] P. Lv, Y. Zhong, J. Zhao, and L. Zhang, "Unsupervised change detection based on hybrid conditional random field model for high spatial resolution remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 7, pp. 4002–4015, 2018.

[24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[26] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 22–40, 2016.

[27] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "High-resolution semantic labeling with convolutional neural networks," *arXiv preprint arXiv:1611.01962*, 2016.

[28] X. Ma, J. Geng, and H. Wang, "Hyperspectral image classification via contextual deep learning," *EURASIP Journal on Image and Video Processing*, vol. 2015, no. 1, pp. 1–12, 2015.

[29] X. Chen, S. Xiang, C.-L. Liu, and C.-H. Pan, "Vehicle detection in satellite images by hybrid deep convolutional neural networks," *IEEE Geoscience and remote sensing letters*, vol. 11, no. 10, pp. 1797–1801, 2014.

[30] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in neural information processing systems*, 2014, pp. 487–495.

[31] F. Bovolo and L. Bruzzone, "The time variable in data fusion: a change detection perspective," *IEEE Geoscience and Remote Sensing Magazine*, vol. 3, no. 3, pp. 8–26, 2015.

[32] J. E. Ball, D. T. Anderson, and C. S. Chan, "Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community," *Journal of Applied Remote Sensing*, vol. 11, no. 4, p. 042609, 2017.

[33] Y. Zhong, A. Ma, Y. soon Ong, Z. Zhu, and L. Zhang, "Computational intelligence in optical remote sensing image processing," *Applied Soft Computing*, 2017.

[34] 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 Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1845–1849, 2017.

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

[36] J. Geng, H. Wang, J. Fan, and X. Ma, "Change detection of SAR images based on supervised contractive autoencoders and fuzzy clustering," in *Remote Sensing with Intelligent Processing (RSIP), 2017 International Workshop on*. IEEE, 2017, pp. 1–3.

[37] 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 Journal of Photogrammetry and Remote Sensing*, vol. 116, pp. 24–41, 2016.

[38] F. Gao, J. Dong, B. Li, and Q. Xu, "Automatic change detection in synthetic aperture radar images based on PCANet," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 12, pp. 1792–1796, 2016.

[39] Y. Xu, S. Xiang, C. Huo, and C. Pan, "Change detection based on auto-encoder model for VHR images," in *MIPPR 2013: Pattern Recognition and Computer Vision*, vol. 8919. International Society for Optics and Photonics, 2013, p. 891902.

[40] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 806–813.

[41] O. A. Penatti, K. Nogueira, and J. A. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 44–51.

[42] K. Nogueira, O. A. Penatti, and J. A. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognition*, vol. 61, pp. 539–556, 2017.

[43] A. M. El Amin, Q. Liu, and Y. Wang, "Convolutional neural network features based change detection in satellite images," in *First International Workshop on Pattern Recognition*. International Society for Optics and Photonics, 2016, pp. 100 110W–100 110W.

[44] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[46] M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 881–893, 2017.

[47] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[48] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[49] J. L. Long, N. Zhang, and T. Darrell, "Do convnets learn correspondence?" in *Advances in Neural Information Processing Systems*, 2014, pp. 1601–1609.

[50] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 645–657, 2017.

[51] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a siamese time delay neural network," in *Advances in Neural Information Processing Systems*, 1994, pp. 737–744.

[52] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 539–546.

[53] F. Pacifici, N. Longbotham, and W. J. Emery, "The importance of physical quantities for the analysis of multitemporal and multiangular optical very high spatial resolution images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 10, pp. 6241–6256, 2014.

[54] F. Bovolo, L. Bruzzone, L. Capobianco, A. Garzelli, S. Marchesi, and F. Nencini, "Analysis of the effects of pansharpening in change detection on VHR images," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 1, pp. 53–57, 2010.

[55] Y. Han, F. Bovolo, and L. Bruzzone, "An approach to fine coregistration between very high resolution multispectral images based on registration noise distribution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 12, pp. 6650–6662, 2015.

[56] Q. Guo, J. Xiao, and X. Hu, "New keypoint matching method using local convolutional features for power transmission line icing monitoring," *Sensors*, vol. 18, no. 3, p. 698, 2018.

[57] P. Fischer, A. Dosovitskiy, and T. Brox, "Descriptor matching with convolutional neural networks: a comparison to SIFT," *arXiv preprint arXiv:1405.5769*, 2014.

[58] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015, pp. 4353–4361.

[59] A. Vetrivel, N. Kerle, M. Gerke, F. Nex, and G. Vosselman, "Towards automated satellite image segmentation and classification for assessing disaster damage using data-specific features with incremental learning," September 2016. [Online]. Available: http://proceedings.utwente.nl/369/

[60] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[61] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 447–456.

[62] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[63] F. Bovolo and L. Bruzzone, "A split-based approach to unsupervised change detection in large-size multitemporal images: application to tsunami-damage assessment," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 6, pp. 1658–1670, 2007.

[64] F. G. Hall, D. E. Strebel, J. E. Nickeson, and S. J. Goetz, "Radiometric rectification: toward a common radiometric response among multidate, multisensor images," *Remote Sensing of Environment*, vol. 35, no. 1, pp. 11–27, 1991.

[65] A. Kieri, "Context dependent thresholding and filter selection for optical character recognition," 2012.

[66] P. D. Wellner, "Adaptive thresholding for the digitaldesk," *Xerox, EPC1993-110*, pp. 1–19, 1993.

[67] J. A. Lee and M. Verleysen, *Nonlinear dimensionality reduction*. Springer Science & Business Media, 2007.

[68] D. Marinelli, F. Bovolo, and L. Bruzzone, "A novel method for unsupervised multiple change detection in hyperspectral images based on binary spectral change vectors," in *Analysis of Multitemporal Remote Sensing Images (MultiTemp), 2017 9th International Workshop on the*. IEEE, 2017, pp. 1–4.

[69] C. M. Bishop, "Pattern recognition," *Machine Learning*, 2006.

[70] B. Auffarth, M. López, and J. Cerquides, "Comparison of redundancy and relevance measures for feature selection in tissue classification of ct images," in *Industrial Conference on Data Mining*. Springer, 2010, pp. 248–262.

[71] T. Fawcett, "ROC graphs: Notes and practical considerations for researchers," *Machine learning*, vol. 31, no. 1, pp. 1–38, 2004.