IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 15, 2022, accepted March 3, 2022, date of publication March 24, 2022, date of current version April 1, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3161978

Change Detection in VHR Imagery With Severe Co-Registration Errors Using Deep Learning: A Comparative Study

VIKTORIA KRISTOLLARI AND VASSILIA KARATHANASSI

Laboratory of Remote Sensing, School of Rural, Surveying, and Geoinformatics Engineering, National Technical University of Athens, 15780 Zografou, Greece

Corresponding author: Viktoria Kristollari (vkristoll@central.ntua.gr)

This work was supported by the European Union's Framework Programme for Research and Innovation (Horizon 2020) under grant agreement no. 821054.

ABSTRACT Change detection (CD) through Earth observation techniques can offer very significant information for monitoring tasks in a time-efficient manner. Very high-resolution (VHR) images can display objects in fine detail, thus making it possible to rapidly perceive isolated changes. However, this is a challenging task because of the increased within-class variance and geometric registration errors caused by different satellite view directions and angles. Lately, deep learning (DL) CD methods have proven very appealing for the CD problem because of their flexibility to combine and process different types of information along with the increased availability of higher processing power systems. Even though previous research has developed several notable DL methodologies, it has mostly focused on images with minor co-registration errors. Based on that, the goal of this study is to evaluate the performance of five state-of-the-art DL CD methods, two unsupervised and three supervised, on VHR images with severe co-registration errors. The methods are implemented on four urban European areas of versatile morphology. In addition, before applying the CD process, four popular automatic co-registration methods were evaluated because of the importance of this pre-processing step for the successful output of the CD problem. It was shown that phase correlation used on the Fourier-Mellin Transform produced the most satisfactory co-registration results and STANet detected building-related changes most successfully. Its success can be attributed to its particular attention mechanism and its training dataset. The rest of the co-registration and CD methods showed low performance.

INDEX TERMS Change detection algorithms, artificial neural networks, very high-resolution imagery, image registration, land cover monitoring, buildings.

I. INTRODUCTION

Change detection (CD) is an important Earth observation task that aims at monitoring land cover transitions through time for a given area. In the recent past, attention has been drawn towards very high-resolution (VHR) images because smaller objects (e.g. buildings) can be displayed in detail. However, moving to VHR increases significantly the complexity of the problem, since these images present increased within-class variance and geometric registration errors [1]–[4]. The successful completion of the task becomes even more challenging when the data are collected from different sensors, since their heterogeneity is heightened [5]–[7].

Among the well-known traditional pixel-based CD methods are algebra methods such as change vector analysis, (CVA) [8]–[10] and transformation methods such as principal component analysis (PCA) [11] and multivariate alteration detection (MAD) [12]. CVA computes the spectral difference and provides change intensity and direction [13]. PCA implies the assumption of a linear relation between no-change pixels belonging to the two acquisitions [14], and selects a part of the principal components for the CD [15]. MAD, based on canonical correlation analysis (CCA) [16], also exploits unchanged pixels and aims at identifying changes from the canonical difference of multivariate images [14].

The associate editor coordinating the review of this manuscript and approving it for publication was Jiachen Yang.

Since exploring spatial information is significant, object-based CD methods (OBCD) have been developed, where the basic unit consists of pixels with similar spectral signatures [17], [18]. Although OBCD is less sensitive to co-registration noise [19], the performance of these methods highly depends on the accuracy of the segmentation process which generally alters the geometry of the objects [17].

Several techniques have been used in pixel-based CD to improve robustness to residual misregistration. The majority of the techniques have been applied on medium resolution images (spatial resolution:∼30 m). In [20], the authors proposed image smoothing via an average or median filter and alternatively adaptive grey-scale mapping which calculates total excess and deficit with respect to the image mean in a pixel window. In [21], an approach was proposed which utilizes bands where the investigated changes are not detectable, since co-registration noise is generally visible in all spectral bands. In [22], residual misregistration was detected by introducing a modeling approach that makes use of spatial brightness gradients, assuming that misregistration effects are locally uniform. In [8], a method was presented which detects co-registration noise by representing spectral change vectors in the polar domain and exploiting the direction distribution information. The same approach was followed in [10] for VHR images with the difference that the pixels in the adjacent neighborhood were also considered. In [23], the symmetric local co-registration adjustment (SLCRA) scheme was developed for HR imagery (∼5 m). The method chooses corresponding pixels by calculating the minimum dissimilarity in a window. Finally, in [14], the same approach was followed to reduce minor misregistration errors in statistically similar entities.

Recently, convolutional deep learning (DL) CD methods have drawn very high attention because of their innate ability to detect spatial context from raw data and their flexibility in the combined processing of different types of information. Another reason is the technological progress that has increased access to higher processing power systems. Hence, both unsupervised and supervised approaches have been proposed. Unsupervised methods are generally based on the comparison of feature maps produced from the bitemporal images. In [24], convolutional neural network (CNN) feature maps of the pre-trained CaffeNet on Imagenet [25] were concatenated and the change map was computed using pixel-wise Euclidean distance. The same authors in a different study [26] compared features extracted from different zooming levels of the pre-trained VGG-16 [27] on the same dataset to produce the final change map. As a pre-processing step, they applied PCA and segmented the three higher uncorrelated channels into superpixels. A similar approach was followed in [28] with the difference that the pre-trained VGG-16 deep change features were refined by a variance ranking-based method to retain only the relevant features. In [29], low-rank-based saliency computation and deep feature representation were combined. VGG-16 was fine-tuned on the AID dataset [30] and after extracting multilevel CNN features from superpixels, saliency maps that indicate pixel change probabilities were generated. In [31], the authors proposed the creation of a difference image of the feature maps produced by U-Net [32] pre-trained for semantic segmentation on the Vaihingen dataset [33]. By using networks pre-trained on the same dataset, transfer learning on U-Net was applied in [34] and an unsupervised context-sensitive deep CVA framework was proposed in [35]. Automatically selected features were combined into hypervectors that were compared pixel-wise to obtain deep change vectors for multiclass CD based on the direction of change. Finally, in [36], an unsupervised deep Siamese kernel PCA convolutional mapping network for binary and multiclass CD was designed. The multiclass CD was accomplished by a 2-D polar mapping.

Other studies have focused on approaches that avoid the costly annotation of samples. In [37], a Siamese version of VGG-16 pre-trained on AID was extended by adding a deep feature difference CNN and then transfer learning was applied combined with training on a small sample of VHR images with annotated changes. The final change map was created by a threshold. In [13], the authors applied an automatic pre-detection method of the training data and proposed a deep Siamese convolutional multiple-layers recurrent NN (RNN), which can be used both for homogeneous and heterogeneous images. Finally, in [38], pre-disaster OpenStreetMap building data were used to automatically generate training samples for a modified version of U-Net, where residual connections were added.

The increase in the availability of annotated CD datasets has greatly accelerated the research of supervised methods, which usually produce more accurate results. The SZTAKI AirChange Benchmark Set (1,5 m/px) [39] was the first VHR CD dataset that was made publicly available. This dataset has been used in many studies. In [40], the authors used it to train a Siamese CNN by the weighted contrastive loss. The changes of the image pair were detected by the distance of the feature vectors and the final output was produced by a threshold and a k-NN approach. In [41], three fully CNNs were trained on the SZTAKI dataset and instead of concatenating both connections of the encoding streams of the Siamese versions, the absolute value of their difference was concatenated. The same dataset was used in [42] to train the DeepLabV2 [43] network by an improved triplet loss function. The network was pre-trained on the PASCAL VOC 2012 dataset [44]. In addition, the SZTAKI and a building CD dataset were used in [45] to train a deep NN architecture based on the combination of an attention mechanism with information transmission by the use of bidirectional LSTMs. Finally, in [46], a modified version of U-Net was trained on the SZTAKI dataset by using a depth-wise separable convolution making the network lighter and more efficient.

Lately, more datasets have been created to promote research in the field. In [47], the first large-scale VHR semantic CD dataset was presented, and several fully CNNs for semantic CD were proposed. In [48], another dataset was used which is composed of multisource VHR images with

annotated multitype changes [49]. In this study, a multiscale convolution module was incorporated in a fully convolutional network (FCN). The authors also proposed a combination of the weighted binary cross-entropy loss (WBCE) and the dice coefficient loss to improve the training of imbalanced samples. Finally, in [50], focus was put on semantic CD and a Siamese framework with a global hierarchical (G-H) sampling mechanism was trained on three datasets with semantic annotated changes [51], [52]. The purpose of the G-H sampling mechanism is the mitigation of the imbalance problem. The authors also used the binary change mask to constrain the semantic CD results.

It is noted that since DL methods capture spatial information, it logically follows that they perform better in misregistration scenarios than pixel-based methods that exploit only spectral information. Recently, to further enhance their spatial context perception, many studies have adopted spatial attention mechanisms because they capture long-range spatial dependencies which leads to the reduction of pseudochanges [53]. Spatial attention highlights meaningful spatial relationships through reweighting of the feature maps [4]. The authors in [53] implemented dual attentive fully convolutional Siamese networks to examine spatial and spectral long-range dependencies. They also addressed the imbalance sample problem by use of the weighted double-margin contrastive loss. The network was trained and evaluated on two datasets, the multisource VHR dataset proposed in [49] and two VHR image scenes with annotated changes of buildings (WHU building dataset) [54]. In [55], a Siamese-based spatial-temporal attention CNN was introduced, along with one of the largest CD datasets of the field (changes related to buildings). In [4], an end-to-end network, called the pyramid feature-based attention-guided Siamese network was proposed. The authors introduced a co-attention mechanism and trained the network on two different building CD datasets: WHU (orthoimagery) and a challenging dataset of satellite images (with displacement). In [56], a dual-task constrained deep Siamese CNN, which contains a CD network and two semantic segmentation networks, was presented along with a dual attention module. It was trained on the WHU building dataset. In [57], deep features were extracted from a fully convolutional two-stream architecture and were fed into a deeply supervised difference discrimination network. Deep features of the raw images were fused with image difference features by attention modules and change map losses were also introduced in the intermediate layers. The CNN was trained on the dataset created in [49] and on a multisource Google Earth dataset. Finally, in [58] a scheme was proposed that contains an efficient convolution module in combination with fusion strategies based on spatial/spectral attention. The network was trained on the dataset proposed in [49] and on a recent version of WHU with semantic changes.

Even though attention mechanisms dominate the current literature on mitigating the effects of co-registration errors on VHR CD, some other approaches have also been proposed. In [59], three encoder-decoder-structured CNNs were designed to yield change maps from RGB satellite images with small color variations and co-registration errors, and a large fully-labeled dataset of Google Earth images was constructed. The ensemble of the networks outperformed each individual CNN. In [49] a conditional adversarial network was trained and evaluated on synthetic images with a small relative shift. Finally, in [3] a framework that consists of two parts was proposed by use of the WHU dataset. It involves a building change detection network that takes bi-temporal binary building maps produced from a building extraction network. The authors simulated arbitrary building changes and various building parallaxes in the binary building map to increase robustness to co-registration errors.

Although the current scientific research concerning DL CD with co-registration noise has shown promising results, it has mostly focused on images with minor co-registration errors. Based on that, the goal of this study is to assess several state-of-the-art DL CD methods on VHR images with severe co-registration noise. The study evaluates the performance of five state-of-the-art deep DL CD methods, two unsupervised and three supervised on four urban study areas of different morphology. The VHR images are selected from various satellites and exhibit high geometric distortions and co-registration errors. The fundamental logic behind the selection of the DL CD methods was the representation of each main category. Another reason was the public availability of the code proposed by the creators of the methods, to ensure correct implementation. Thus, the first unsupervised method [24] is a pre-trained network that follows a patch-to-pixel approach, while the second unsupervised method, which was developed for the purpose of our study, has an encoder-decoder architecture and was trained on the study data. Concerning the selected supervised methods, the first (FDCNN) [37] avoids the costly annotation of samples by applying transfer learning in combination with training on a small annotated CD sample of multitype changes, while the second (DASNet) [53] and the third (STANet) [55] apply spatial attention mechanisms to capture long-range spatial dependencies. DASNet was trained on the multisource VHR dataset proposed in [49] (multitype changes) and on the WHU building dataset, and STANet on a large dataset with changes related to buildings. It is noted that the supervised networks were implemented by use of the weights provided by the creators of the methods.

Before applying the CD process, four popular automatic co-registration methods were evaluated since this pre-processing step is extremely important for the success of the CD problem. The selected methods cover a wide range of the existing literature approaches. The first two are Scale Invariant Feature Transform (SIFT) [60] and the Oriented FAST and Rotated BRIEF (ORB) [61] which detect local features and assign descriptors. The third is a CNN approach [62] and the fourth is the Fourier-Mellin Transform (FMT) [63] which is a global method.

In the following sections, at first a brief theoretical background of the evaluated co-registration and DL CD methods

is stated. Then, the procured images and the study areas are presented followed by the description and the discussion of the results.

II. THEORETICAL BACKGROUND

A. CO-REGISTRATION METHODS

Four popular methods were tested for the automatic co-registration of the images. These methods were SIFT, ORB, a CNN feature-based approach, and the FMT. The selected methods cover a wide range of the existing literature approaches.

1) SIFT

SIFT locates local features known as ''keypoints'' that are scale and rotation invariant. The keypoints are detected by creating different scales of the images (application of Gaussian blur) and locating local maxima and minima. Then, their orientation and magnitude are defined by calculating gradients. Thus, a unique fingerprint is created for each point called ''descriptor.'' The method consists of four parts: Scale-space extrema detection, accurate keypoint localization, orientation assignment, and keypoint descriptor generation.

2) ORB

ORB is a fusion of FAST (Features from accelerated segment test) [64] keypoint detector and BRIEF (Binary Robust Independent Elementary Features) [65] descriptor with modifications to enhance the performance. FAST is a corner detection method and BRIEF assigns descriptors by selecting a random pair of pixels in the neighborhood of a keypoint from a Gaussian distribution and comparing their brightness. The FAST modifications refer to the use of a multiscale image pyramid and the assignment of orientation, whereas the BRIEF modifications to the inclusion of orientation invariance.

3) FOURIER-MELLIN TRANSFORM

FMT-based image registration is a global method since it uses all the image pixels of both images to define the transformation parameters [66]. In this method, at first the Fast Fourier transformation (FFT) of the input images is calculated followed by the calculation of the magnitudes.

Then, the magnitudes are transformed to log-polar coordinates. Taking the Fourier transformation of a log-polar map is equivalent to the computation of the Fourier-Mellin Transform (Equation 1) [67].

$$F_M(k_1, k_2) = \int_{-\infty}^{+\infty} \int_{0}^{2\pi} f(e^r cos\varphi, e^r sin\varphi) e^{j(k_1 r + k_2 r)} d\varphi dr \quad (1)$$

where: r, φ: log-polar coordinates and k: scale.

By applying phase correlation, the angle and the scale can be retrieved. After applying rotation and scale, phase correlation can be applied again and the translation can be calculated as the final step of the 2-D image registration.

4) CO-REGISTRATION - CNN

The CNN feature-based approach uses a CNN to generate multiscale feature descriptors and then the Expectation Maximization method (EM) [68] is applied to gradually increase the selection of inliers. After detecting a feature point set X from the referenced image and a feature point set Y from the sensed image, the transformed locations of Y (Z) are obtained. The multiscale feature descriptors are generated using three pooling layers ($D_1(x), D_2(x), D_3(x)$) from a pre-trained VGG-16 network on Imagenet dataset. After defining a grid, the feature point is determined as the center of each grid cell. Features x and y are matched according to Equation 2.

$$d(x, y) = \sqrt{(2)} d_1(x, y) + d_2(x, y) + d_3(x, y) \quad (2)$$

where: $d_i(x, y)$: Euclidean distance of $D_i(x), D_i(y)$.

Inlier selection produces a $M \times N$ prior probability matrix using both convolutional feature and structural information which is then taken by a Gaussian mixture model (GMM) based transformation solver. In order to compute the matrix, at first an integrated cost matrix is computed using an element-wise Hadamard product. Then, the Jonker-Volgelant algorithm [69] is applied to solve the linear assignment on the cost matrix. Assigned point pairs are regarded as putatively corresponding.

Points in set Y are considered as GMM centroids and EM is then applied to find the optimal transformation parameters. The objective of the approach is to minimize the negative log-likelihood function. EM iteratively solves the non-rigid transformation (Equation 3) and the selection of inliers is updated in every k iterations. The process consists of the E-step where the posterior probability matrix is computed from the last iteration, and the M-step where the derivatives are solved and the parameters are updated. As a final step, the transformed image is calculated using thin plate spline interpolation.

$$Z = Y + GW \quad (3)$$

where: G: the matrix generated by a Gaussian radial basis function (GRBF) and W contains the transformation parameters.

B. CHANGE DETECTION METHODS

Five land cover DL CD methods were implemented: two unsupervised and three supervised.

1) UNSUPERVISED METHODS

The first unsupervised method was the patch-to-pixel CNN proposed in [24]. For its implementation [70], Tensorflow [71] and Keras [72] functions were applied. The method uses the VGG-19 architecture pre-trained on the Imagenet database. The size of the input image patches was 224 × 224 px and the output size was 112 × 112 px. Firstly, the feature maps are extracted from five convolutional layers ($Conv_1, Conv_2, \ldots, Conv_n$) to exploit both the

spatial (lower level features) and the semantic information (higher-level features). Since these features are not of the same size due to downsampling (pooling) operations, multilevel maps of the same size are concatenated after being resized to the same size (resampling operations), resulting in a higher-dimensional feature map.

The CD is performed using pixel-wise Euclidean distance in a feature space of $k-$ dimension (Equation 4). For the production of the final change map, the optimum threshold is defined by applying the Otsu [73] segmentation method, which detects the minimal intra-class variance of two classes. For the implementation of the first unsupervised method in our study, Otsu segmentation was applied on images of size 1120 × 1120 px, produced by joining 25 output patches (112 × 112 px) after resampling to the input size (224 × 224 px). It is noted that in the original implementation Otsu segmentation was applied on the output patches (size 112 × 112 px).

$$d_{ij} = \sum_{k=1}^{k} ((\mu_i^k)^2 - (\mu_j^k)^2)^2 \tag{4}$$

where k: feature dimension and μ_i^k and μ_j^k: features values at dimension k_{th} of the positions i and j.

In the second unsupervised method, which was developed for the purpose of our study, an encoder - decoder CNN with three convolutional layers in the encoder part (64, 32, 16 feature maps) and three convolutional layers in the decoder part (32, 64, 4 feature maps) was implemented by use of Tensorflow and Keras functions. The network was trained on patches of size: 224 × 224 of the images of the first date (four images in total (one per each study area)) and the visible and near-infrared (NIR) bands were used. The input patches were fed to the CNN by a generator function which randomly selected a study area and then a random batch of eight input patches. The model was trained for 400 epochs with 407 train steps on an NVIDIA 1070 Ti Graphical Processing Unit (GPU) for approximately six hours.

Then, similar steps to the first unsupervised method were followed. First, multilevel maps of the same size (128 × 128 px) were created via resampling for the first two and last two convolutional layers, and then the feature maps were combined to create the change map using pixel-wise Euclidean distance and manually applying an Otsu threshold for images of size 1120 × 1120 px.

2) FDCNN

The first supervised method was the feature difference CNN (FDCNN) [37], which uses transfer learning on a CNN (VGG-16) pre-trained on the AID dataset [30] (30 aerial scene types), combined with training on a small sample of VHR images with annotated changes. For its implementation [74], the Caffe framework [75] was used.

The network consists of three main parts. The first part is a two-channel Sub-VGG-16 with shared weights, composed of the first three scales of VGG-16 with input size 224 × 224 px. The second part is the FD-Net where feature difference maps of three scales are created and normalized (Equation 5). Before computing the feature difference maps, resampling is applied to generate maps of the same size. In addition, the second-period image (X_2) is differentiated from the first period (X_1) image to obtain accurate boundary information of the changes. The third part is the FF-net where the back-propagation of the network is realized by a simple CNN with few training points, which produces the final change magnitude map (CMM).

$$\mathrm{FD}(i) = \frac{|F_1^i - F_2^i|}{\max(|F_1^i - F_2^i|)},\ i = 1, \ldots, N \tag{5}$$

where FD: the feature difference map, F_1^i, F_2^i: the feature maps with inputs X_1, X_2, and N: the total number of feature maps.

The network implements an improved cross-entropy loss that uses the change magnitude of each pixel as prior knowledge for learning and a weight loss function to alleviate the tendency of the network to no-change miss-detection due to unbalanced training data. CMMs are generated by applying CVA on X_1, X_2. After obtaining the CMM, the final change map is obtained by a threshold.

3) DASNET

The second supervised method was the dual attentive fully convolutional Siamese network (DASNet) [53], which aims at capturing long-range dependencies. The network was trained on two CD datasets. One composed of multisource remote sensing images with multitype annotated changes (spatial resolution of 3 to 100 cm/px) [49] and one composed of two VHR image scenes with annotated changes of buildings (WHU building dataset) [54]. For its implementation [53], the Pytorch library [77] was used.

First, the Siam-Conv module is used to generate local features: $F_{t0}, F_{t1} \in \mathbb{R}^{C\times H\times W}$, and then the dual mechanism is applied to establish the connections between them. The feature F is fed into three convolutional layers to obtain three new features: $Fa, Fb, Fc \in \mathbb{R}^{C\times H\times W}$.

For the spatial attention, Fa, Fb, Fc are reshaped to $\mathbb{R}^{C\times N}$. Then, matrix multiplication is conducted between Fb^T and Fa and a spatial attention map is obtained through a softmax layer (Equation 6), which measures the connection between a feature at position i and a feature at position j. Fc is reshaped to $\mathbb{R}^{CxN}$ and matrix multiplication with Fs is conducted. Finally, the result is reshaped to $\mathbb{R}^{C\times H\times W}$ and added to the original feature to obtain the final output (Equation 7).

$$Fs_{ji} = \frac{e^{Fa_i \cdot Fb_j}}{\sum_{i=1}^{N} e^{Fa_i \cdot Fb_j}} \tag{6}$$

$$Fsa_j = \eta \sum_{i=1}^{N} (Fs_{ji} Fc_j) + F_j \tag{7}$$

where Fa, Fb, Fc : features succeeding Siam-Conv, F: original feature, η: scale parameter, and $N = H \times W$.

For the channel attention, F is reshaped to $\mathbb{R}^{CxN}$ and then matrix multiplication is performed between F^T and F to obtain the channel attention map. Then, similar steps as in spatial attention are followed. Equations 6, 7 are used by substituting N with the spectral dimension since it captures long-range context in the channel dimension.

The features obtained through the dual attention mechanism are aggregated.

The weighted double-margin contrastive loss was proposed to address the imbalanced sample problem. It is calculated for the spatial and channel attention modules: L_{sa}, L_{ca}, as well as the final output feature pairs L_e (Equation 8).

$$\text{Loss} = \lambda_1 L_{sa} + \lambda_2 L_{ca} + \lambda_3 L_e \tag{8}$$

where λ_i: weight of each loss

The output of DASNet is an RGB image patch of size 256 × 256 px. High red values show a high probability of change. Thus, for the implementation of DASNet in our study, the final binary change map was produced by applying an Otsu threshold in the Red band for images of size 1120 × 1120 px. These images were produced by joining output patches (256 × 256 px) after resampling to the input size, which was 224 × 224 px in the case of our study.

4) STANET

The third supervised method was the spatial-temporal attention-based network (STANet) [55]. The authors trained the network on a dataset that they proposed (LEVIR-CD), which contains professionally annotated changes related to buildings (soil/grass/hardened ground building). It was created from 637 VHR Google Earth image pairs (size: 1024 × 1024 px) from Texas, US and represents various types of buildings. For its implementation [78] the Pytorch library was used.

The network has a Siamese structure. First, an FCN (Resnet-18 [79]) is employed to extract the bitemporal image feature maps ($X^{(1)}, X^{(2)} \in \mathbb{R}^{C\times H\times W}$). Then, $X^{(1)}, X^{(2)}$ are stacked into a feature tensor $X \in \mathbb{R}^{C\times H\times W\times 2}$ and fed to the attention module to create two attention feature maps ($Z^{(1)}, Z^{(2)}$) (Equation 9). The self-attention mechanism models attention weights between any two pixels.

$$Z = F(X) + X \tag{9}$$

where $Y = F(X)$ is a residual mapping of X to be learned.

Three tensors are introduced to illustrate the basic idea of the self-attention mechanism: query, key and value, which are obtained from the input feature tensor through three different convolutional layers. The input feature tensor is the concatenation of the bitemporal image feature maps in the temporal dimension. X is firstly transformed into three feature tensors $Q, K, V \in \mathbb{R}^{C\times H\times W\times 2}$ and then Q, K, V are reshaped to the matrices $\bar{K}, \bar{Q} \in \mathbb{R}^{C'\times N}$ and $\bar{V} \in \mathbb{R}^{C\times N}$ where $N = H \times W \times 2$ and C' is the feature dimension. Q, K are used in the computation of the attention layer. Then, the spatial-temporal attention map $A \in \mathbb{R}^{N\times N}$ is defined as the similarity matrix (Equation 10). Finally, the output matrix $\bar{Y} \in \mathbb{R}^{C\times N}$ is computed by multiplying $\bar{V}$ and A and then reshaping to $Y \in \mathbb{R}^{C\times H\times W\times 2}$.

$$A = softmax(\frac{\bar{K}^T \bar{Q}}{\sqrt{C'}}) \tag{10}$$

To capture spatial-temporal dependencies in multiple scales and alleviate misregistration issues a pyramid version is implemented, which has four branches of different scale. In each branch, the attention mechanism is applied in subregions and then aggregation is performed. The residual tensor Y and the original tensor X are then added to produce the updated tensor $Z \in \mathbb{R}^{C\times H\times W\times 2}$.

Finally, a distance map D is generated by calculating the distance between each pixel pair in the two feature maps by a residual function. During training, the model is optimized by minimizing the loss calculated by the distance map and the label map. In the testing phase, the label map is calculated by thresholding.

The training is performed by a batch-balanced contrastive loss (BCL).

III. DATA

A. DESCRIPTION OF STUDY AREAS

The satellite images used in this study were collected from four European areas: Tønsberg (Norway), Granada (Spain), Rhodes (Greece), and Venice (Italy). Tønsberg presents mostly buildings of low height with tiled roofs (gray or red tones). The urban structures are spread among large areas of forests and crops and a river also crosses the region. Granada is characterized by a very dense urban fabric, which contains very high buildings with tiled roofs of red tones. The city is also enclosed by steep mountains and a few crops. The city of Rhodes is located on an island and shows a dense urban fabric of medium-height buildings with terraces. The relief is generally flat and there is a moderate quantity of crops. A substantial percentage of the Rhodes images is covered by seawater. Finally, Venice presents very homogeneous buildings with red-tiled roofs in very close distances. As in Rhodes, the Venice images are also surrounded by a high water percentage. The presence of a high amount of ships is also noticeable. The locations and thumbnails for all four study areas are shown in Fig. 1.

B. DETAILED INFORMATION OF PROCURED IMAGES

For the detection of the land cover changes, VHR pan-sharpened images collected from Geoeye-1 (GE01) and Worldview-2/3 (WV-2/3) satellites were used. The images were globally co-registered and contained spectral information in the visual and near-infrared (VNIR) part of the light spectrum. Their time difference varied between five and six years and the area size between 17 and 33 km^2. The spatial resolution for GE01 and WV-2 images was 0.5 m, whereas for WV-3 was 0.3 m. Details about the images are shown in Table 1.

FIGURE 1. Locations and thumbnails of the four study areas.

TABLE 1. Detailed information of VHR satellite images used for the land cover CD.

Area	Collection date	Satellite	Resolution (m)	Size (km^2)
Tønsberg	20/9/2013 12/7/2019	WV-2 GE01	0.5 0.5	25
Granada	19/7/2013 2/7/2018	GE01 WV-3	0.5 0.3	21
Rhodes	23/4/2013 5/6/2019	WV-2 WV-3	0.5 0.3	33
Venice	4/5/2013 13/5/2018	GE01 WV-3	0.5 0.5	17

FIGURE 2. Example of visible/non-visible facades in Venice because of the different satellite view angles. (a) Image collected on 13/5/2018 by WV-2. (b) Image collected on 4/5/2013 by GE01.

IV. RESULTS AND DISCUSSION

A. PRE-PROCESSING STEPS

Before implementing the CD methodology, the pre-processing steps were applied. These steps included: a) creation of mosaics from the WV-3 images since the area of interest was depicted in multiple tiles, b) resampling of the WV-3 images from 0.3 m to 0.5 m spatial resolution (same as GE01, WV2), and c) co-registration.

B. CO-REGISTRATION

It is important to note that the procured images were not orthorectified, thus the co-registration process was applied locally and not globally. In more detail, SIFT, ORB, and the Fourier-Mellin transformation were tested on samples of size 1120 × 1120 px, whereas the CNN feature-based approach was tested on patches of size 224 × 224 px. The local approach is necessary because of the perspective view geometry that causes non-uniform scale according to the relief. It should be also noted that because of the different satellite view directions and angles, the images cannot be co-registered with high accuracy (e.g. visible/non-visible facades) (Fig. 2).

For both SIFT and ORB, the descriptor of one feature in the first set is matched with all other features in the second set using some distance calculation. During the matching process, outliers are excluded by the RANSAC (Random Sample Consensus) [80]. In our case, for both methods, many points were incorrectly matched. An example area in Venice showing incorrectly matched points detected by the SIFT method is shown in Fig. 3. The image shows that the algorithm fails to generate point descriptors with the adequate information needed to produce correct matches.

Concerning the CNN co-registration method, it was observed that the results were inconsistent because they were

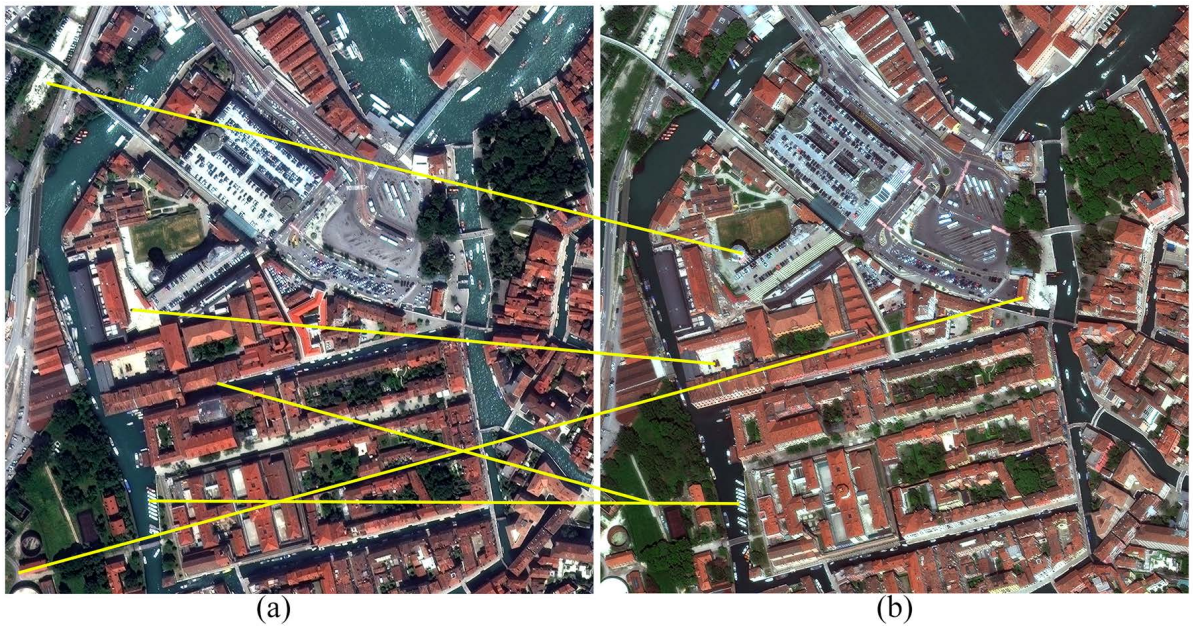

FIGURE 3. Example area in Venice showing incorrectly matched points detected by SIFT. (a) Image collected on 13/5/2018 by WV-2. (b) Image collected on 4/5/2013 by GE01.

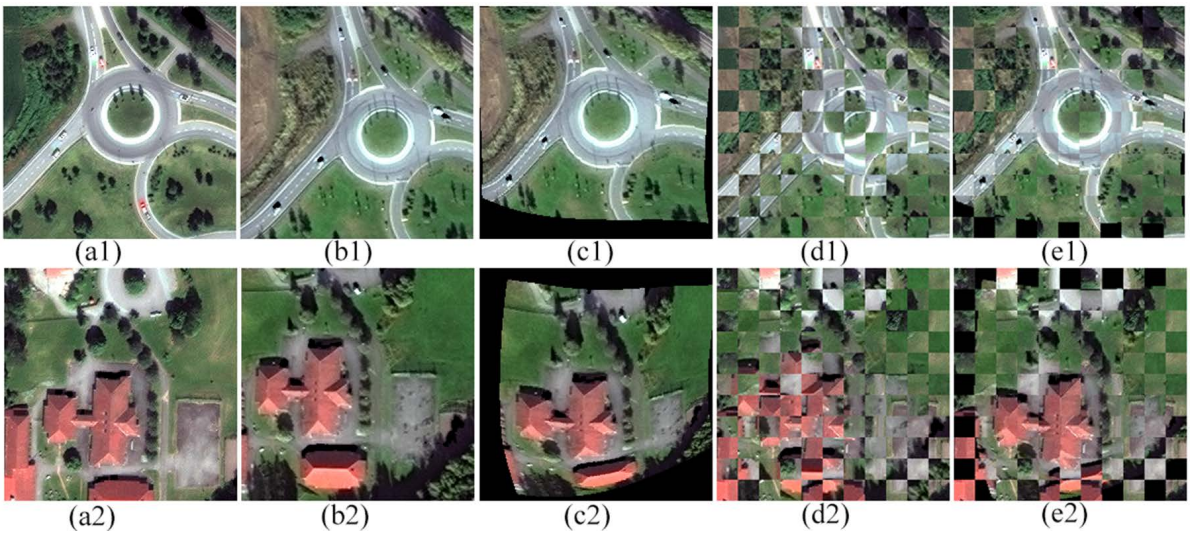

FIGURE 4. Example outputs of the CNN feature-based co-registration (Tonberg). (a1, a2) Image collected on 12/7/2019 by GE01. (b1, b2) Image collected on 20/9/2013 by WV-2. (c1, c2) Co-registered output. (d1, d2) Checkboard display of a1 & b1/ a2 & b2. (e1, e2) Checkboard display of a1 & c1/ a2 & c2.

closely reliant on the objects depicted in the tile. In more detail, the method performed well when a) urban structures with clearly defined edges (e.g. buildings, roads) were present in the patch and b) the structures were situated in the center of the tile. However, when the patch presented fuzzy objects (e.g. crops), or the pixels were situated close to the borders of the patch, distorted outputs were produced. Fig. 4 shows two examples of outputs for this method. A checkerboard display is also presented to make the results more easily perceptible.

FMT performed better than the other automatic co-registration methods, but still not as well as the manual approach where matching points are manually collected. Fig. 5 shows a comparison of two examples of co-registered outputs produced by the Fourier-Mellin Transform and the manual approach. It is shown that the Fourier-Melin Transform shows lower accuracy in areas of variable relief.

Taking into consideration the performance of the four automatic co-registration methods analyzed above, it was decided to co-register the images manually, so that the co-registration errors are minimized as much as possible, given the case studies. For the implementation of the manual co-registration process, at first a grid with cells of size 1120 × 1120 px was created for each image and matching points were selected manually for 261 grid cells in total (Tønsberg: 84,

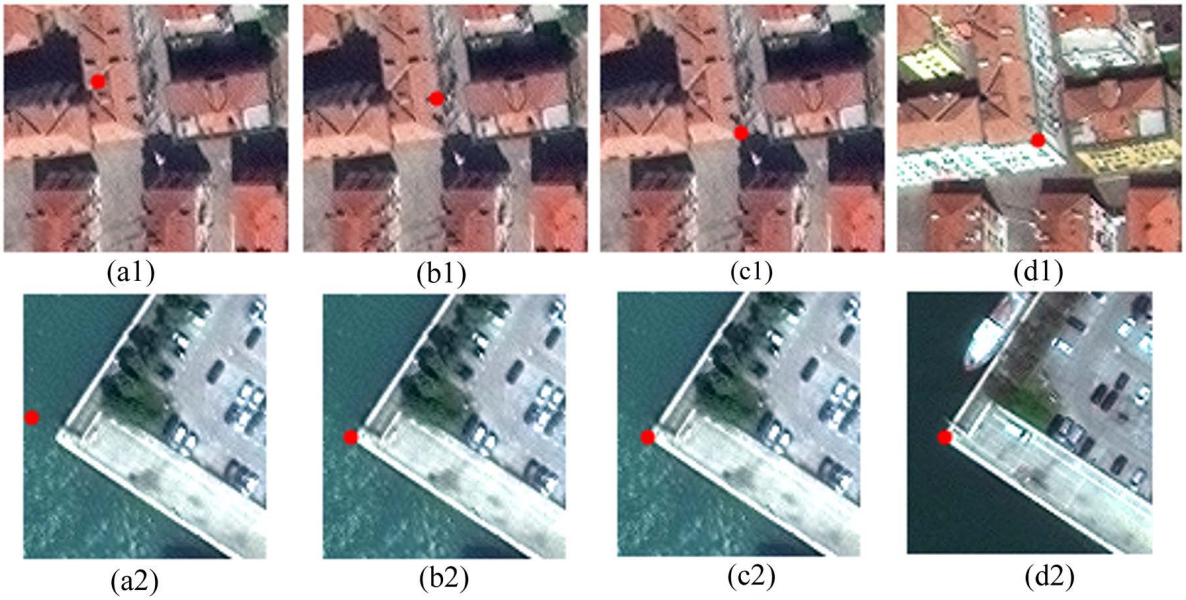

FIGURE 5. Comparison of Fourier-Mellin Transform and manual co-registration (Example outputs in Venice). (a1, a2) Image collected on 13/5/2018 by WV-2. (b1, b2) Co-registered output of Fourier-Mellin Transform. (c1, c2) Manually co-registered output. (d1, d2) Image collected on 4/5/2013 by GE01. The red bullet shows the position for a point.

Granada: 70, Rhodes: 59, Venice: 48). At least four points were selected for each grid cell and then the affine transformation was applied. The selection of the number of points was based on a visual evaluation of the scene height variance and the magnitude of geometric distortions. Thus, the number and height variance of the points increased according to the difficulty of each case.

Fig. 6 shows the box plots of the Root Mean Squared Error (RMSE) for the four areas of interest. The RMSE was calculated by use of the points that had been selected for the manual co-registration. It can be seen that Granada showed the highest mean RMSE (∼ 4m) followed by Tønsberg (∼ 3m), Venice (∼ 2m), and Rhodes (∼ 1.5m). Granada also showed the highest variance as it can be seen from the higher distance between the first (Q1) and third quartile (Q3) (∼ 4.5 m) and the values of the outliers (isolated incidents) reaching RMSE values of ∼15 m. Lower Q3-Q1 values are presented for Tønsberg (∼2.5 m), Venice, and Rhodes (<2 m). Low variance for Rhodes could be explained by similar view directions of WV-2 and WV-3.

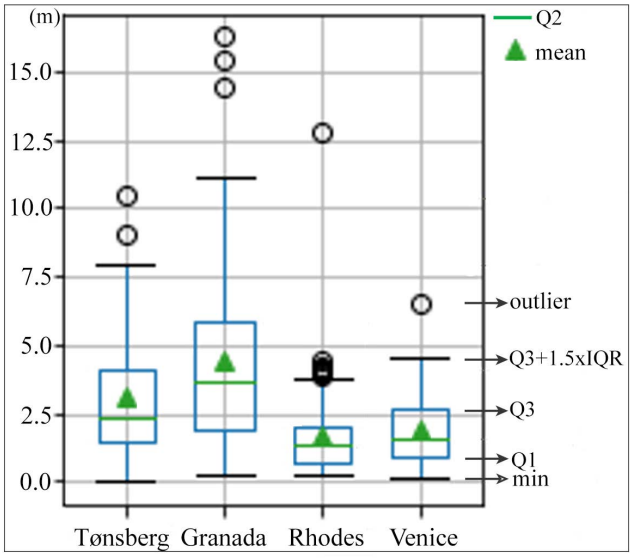

FIGURE 6. Box plots showing the distribution of the co-registration RMSE for the four areas of interest.

C. CHANGE DETECTION METHODS

The first unsupervised CD method and the three supervised applied in this study made use of the publicly available code proposed by the creators of each method, to ensure the correct implementation. It is noted that in this study we refer to the DASNet network trained on multitype changes as "DASNetCDD" and to the DASNet network trained on changes of buildings as "DASNetBCDD."

The methods were evaluated both qualitatively by visually observing the outputs of the methods and quantitatively by calculating evaluation metrics.

1) QUALITATIVE EVALUATION

For the qualitative evaluation, several samples of outputs were observed for all the algorithms. Fig. 7 shows the results for example areas in Tønsberg (Figs 7 (a1-h1)) and Granada (Figs 7 (a2-h2)) produced by the unsupervised and the supervised methods. Similarly, Fig. 8 shows the respective results for example areas in Rhodes (Figs 8 (a1-h1)) and Venice (Figs 8 (a2-h2)). The red square shows the significant changes.

The results of the first unsupervised method (Figs 7 (c1, c2), 8 (c1, c2)) show a high commission error caused by different satellite view directions and angles

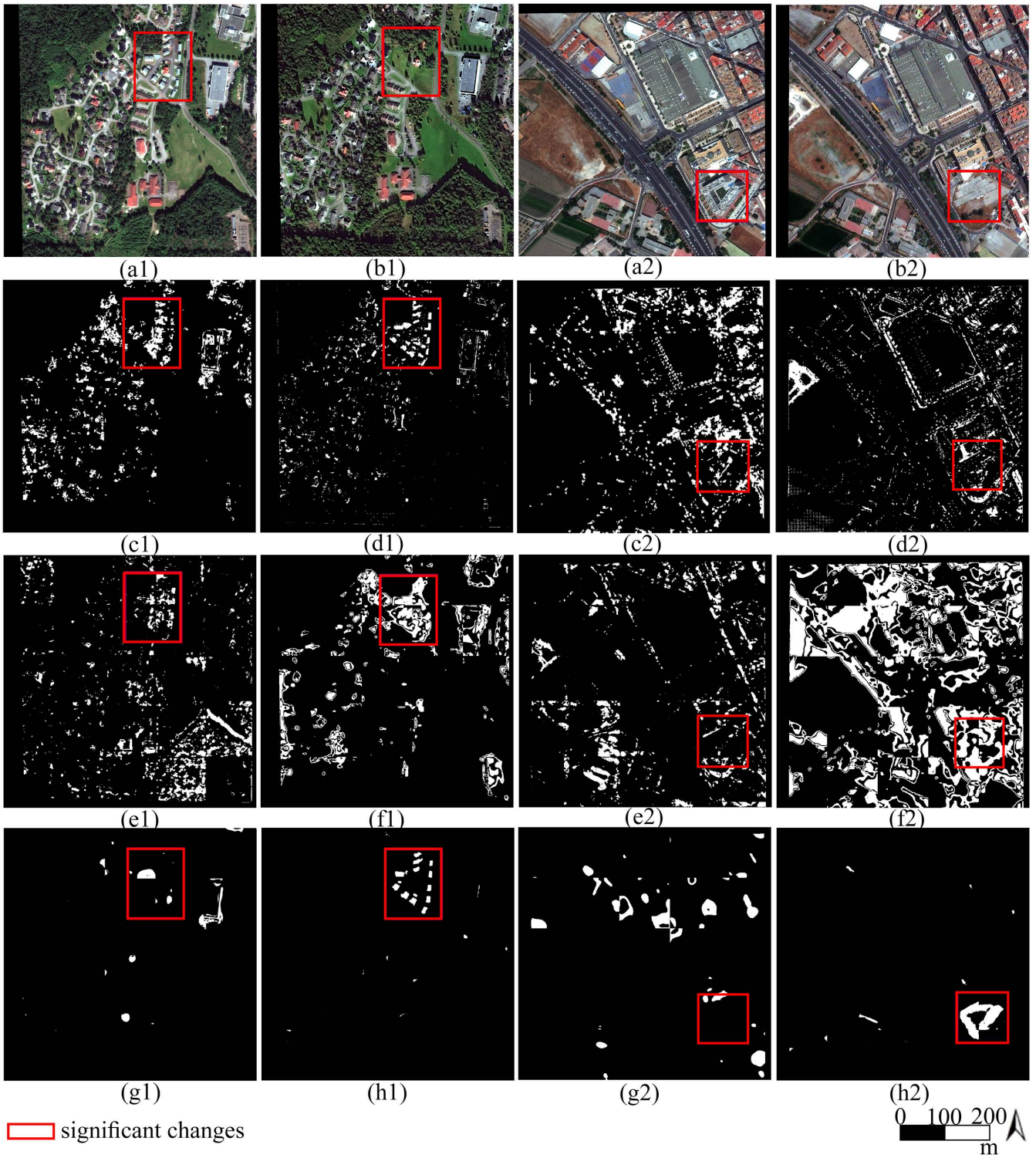

FIGURE 7. Example areas in Tønsberg (1st & 2nd column) and Granada (3st & 4th column) showing results of the unsupervised and supervised methods. (a1, a2) Image of the latest date. (b1, b2) Image of the earliest date. (c1, c2) 1st Unsupervised method. (d1, d2) 2nd Unsupervised method. (e1, e2) FDCNN. (f1, f2) DASNetCDD. (g1, g2) DASNetBCDD. (h1, h2) STANet.

(e.g visible/non-visible facades), radiometric differences, and insufficient co-registration. It is noted that radiometric differences cause diverse spectral information for the same object and geometric distortions cause object shifts. Similarly to the results of the first unsupervised method, the results of the second unsupervised method show a high commission error caused by the same issues (Figs 7 (d1, d2), 8 (d1, d2))). The second unsupervised method also showed high sensitivity to seasonal changes (e.g. crops).

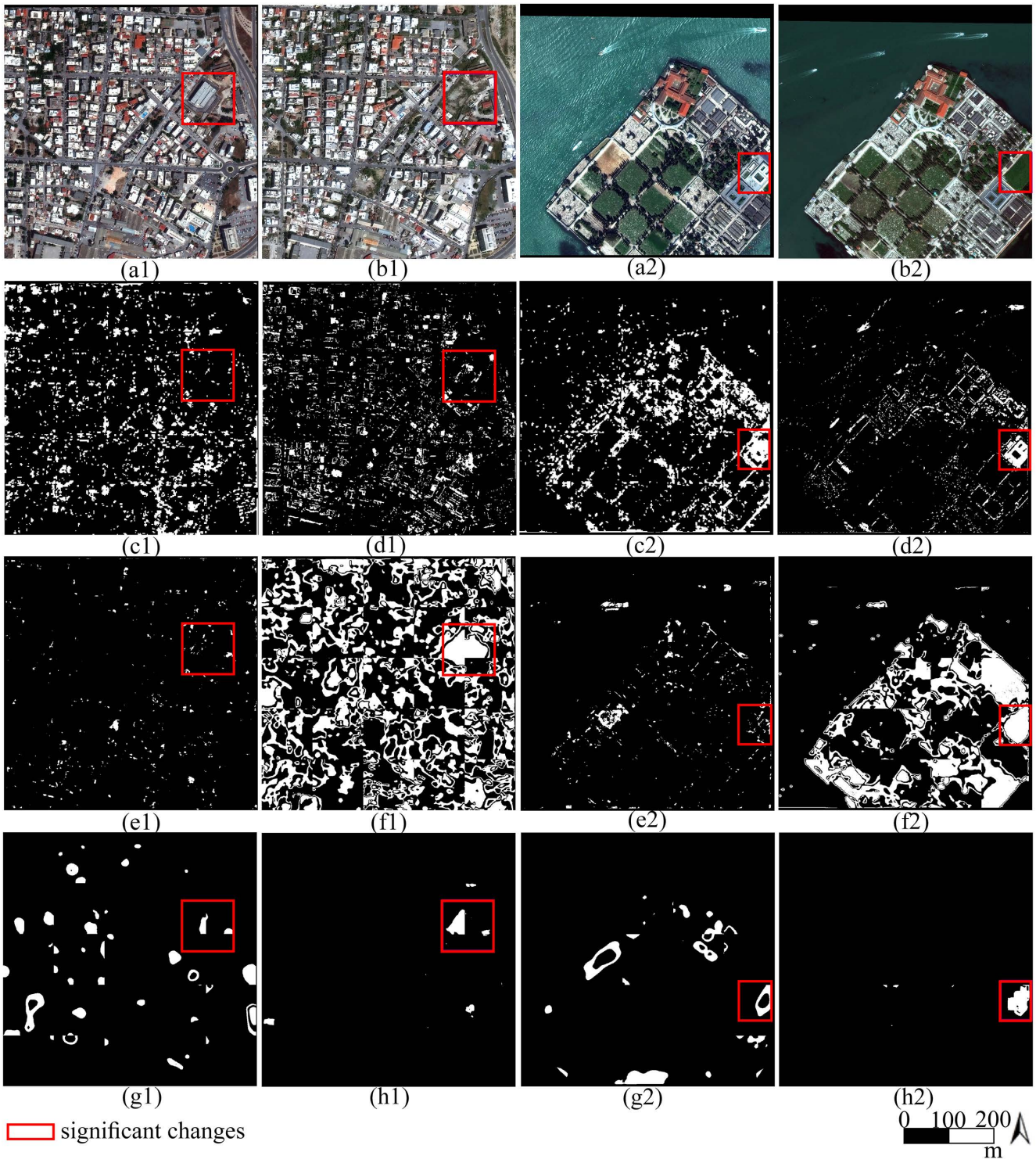

FIGURE 8. Example areas in Rhodes (1st & 2nd columns) and Venice (3st & 4th columns) showing results of the unsupervised and supervised methods. (a1, a2) Image of the latest date. (b1, b2) Image of the earliest date. (c1, c2) 1st Unsupervised method. (d1, d2) 2nd Unsupervised method. (e1, e2) FDCNN. (f1, f2) DASNetCDD. (g1, g2) DASNetBCDD. (h1, h2) STANet.

Since the unsupervised methods are based on comparing the distance of feature maps, it reasonably follows that a large number of pseudochanges will occur in the final result. It should be noted however, that feature maps display the object in various detail levels, thus the output is expected to show lower commission error than directly comparing the original bitemporal images.

Concerning the supervised CD methods, the outputs of FDCNN (Figs 7 (e1, e2), 8 (e1, e2)) show a large commission error and it can be observed that even insignificant

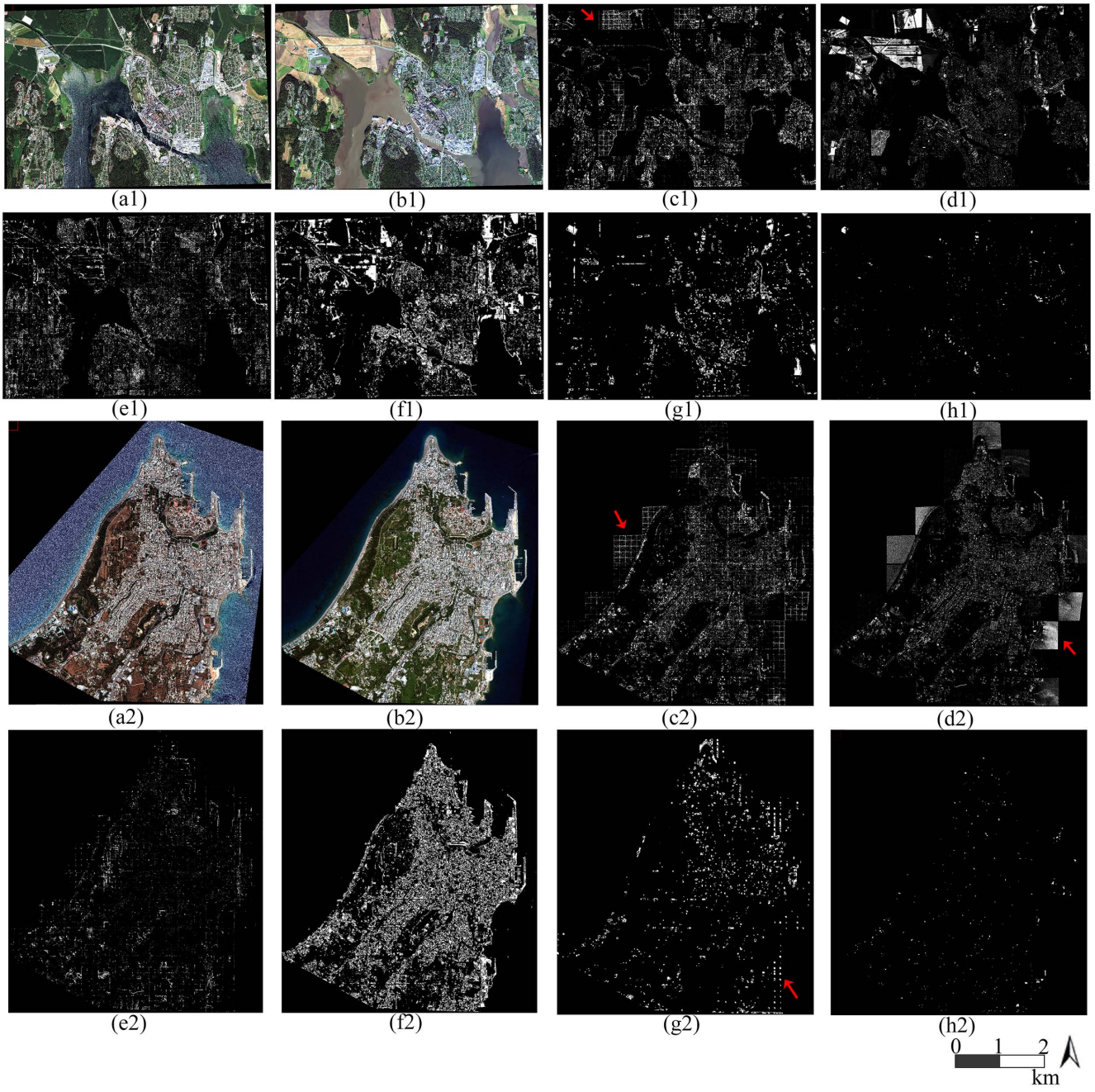

FIGURE 9. Results of the supervised and unsupervised methods for the whole area of Tønsberg (1st & 2nd rows) and Rhodes (3rd & 4th rows). (a1, a2) Image of the latest date. (b1, b2) Image of the earliest date. (c1, c2) 1st Unsupervised method. (d1, d2) 2nd Unsupervised method. (e1, e2) FDCNN. (f1, f2) DASNetCDD. (g1, g2) DASNetBCDD. (h1, h2) STANet. The red arrows show edge noise or water pseudochanges.

changes in vegetation scenes are incorrectly detected (mostly pseudochanges in the forest). Similarly, large commission error is produced by DASNetCDD (Figs 7 (f1, f2), 8 (f1, f2)), where high sensitivity for radiometric differences is presented. It can be also observed that there is distortion in the shapes of the objects. It should be noted that the training set of DASNetCDD was dissimilar to our study areas (e.g. it contained images with snow). DASNetBCDD (Figs 7 (g1, g2), 8 (g1, g2)) also incorrectly detects non-existent changes of buildings while simultaneously showing high omission error. Finally, better results are shown by STANet (Figs 7 (h1, h2), 8 (h1, h2)) as it can be seen that changes related to buildings are detected more successfully than in all previously applied unsupervised and supervised methods. It can also be easily seen that the commission error is lower. The good performance of this method can be attributed to the proposed attention mechanism in combination with the large professionally annotated dataset.

Figs 9, 10 show the results produced by the unsupervised and the supervised methods for the whole study area of

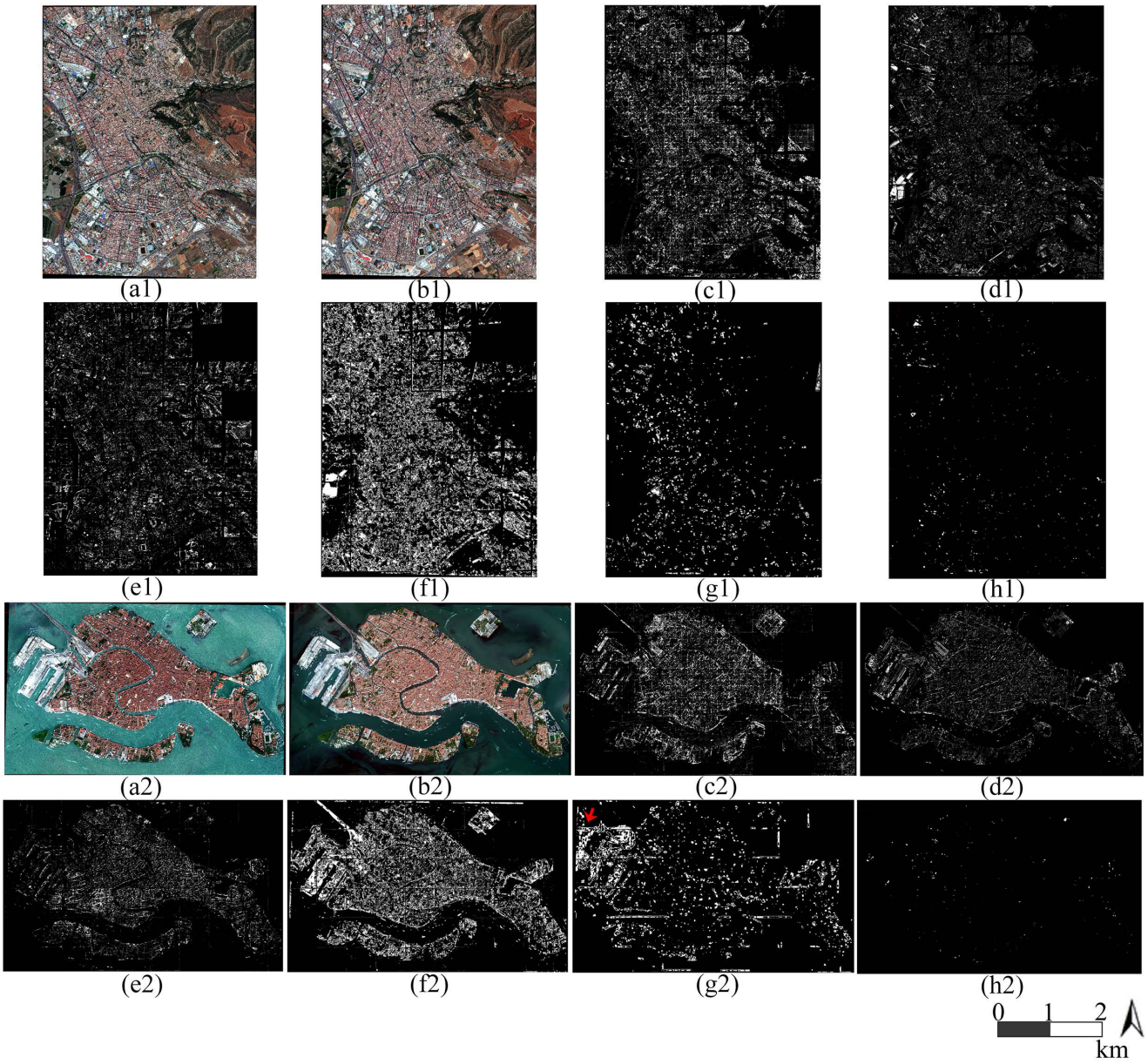

FIGURE 10. Results of the supervised and unsupervised methods for the whole area of Granada (1st & 2nd rows) and Venice (3rd & 4th rows). (a1, a2) Image of the latest date. (b1, b2) Image of the earliest date. (c1, c2) 1st Unsupervised method. (d1, d2) 2nd Unsupervised method. (e1, e2) FDCNN. (f1, f2) DASNetCDD. (g1, g2) DASNetBCDD. (h1, h2) STANet. The red arrow shows water pseudochanges.

Tønsberg (Figs 9 (a1-h1)), Rhodes (Figs 9 (a2-h2)), Granada (Figs 10 (a1-h1)), and (Figs 10 (a2-h2)). The observation of these figures leads to some further conclusions. In more detail, it can be seen that a) the first unsupervised method sometimes shows noise at the edges of the input CNN patch (Figs 9 (c1, c2)), and b) the second unsupervised method and DASNetBCDD exhibit sensitivity to sunglint/watercolor differences (Figs 9 (d2), 10 (g2)). The above mentioned issues are indicated by red arrows.

2) QUANTITATIVE EVALUATION

Quantitative evaluation was performed by the calculation of metrics. These metrics were recall (Equation 11), which corresponds to omission error, precision (Equation 12), which corresponds to commission error, F_1 score (Equation 13) which combines recall and precision metrics.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{12}$$

$$\text{F}_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{13}$$

where: TP: True positive, FN: False negative and FP: False positive

The quantitative evaluation was performed a) for the whole area of the four study areas for the results of STANet

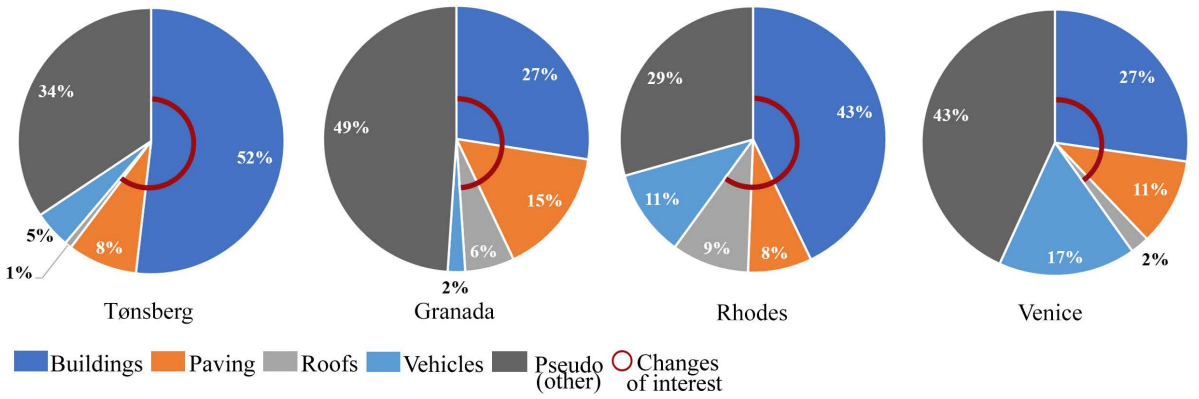

FIGURE 11. Percentages of the types of changes detected by STANet for the whole study area. The "pseudo (other)" category refers to changes caused mostly by co-registration errors and radiometric differences.

(261 images of size: 1120 × 1120 px) and b) for a representative sample which contained ∼ 20% of the results of all algorithms (59 images of size: 1120 × 1120 px). Correctly and incorrectly detected objects were defined by carefully observing the results in a laborious and time-consuming process (required ∼ 1.5 month to complete). We believe that the representative sample is sufficient for the evaluation of the performance since this percentage is the common practice for the test set. It is noted that instead of creating ground-truth maps (e.g. polygons) from the photo-interpretative process, the evaluation metrics were calculated by directly counting the number of objects for each category). We decided to follow this approach because it is significantly simpler and less time-consuming.

It is noted that false negatives were calculated by taking into account only the undetected buildings, whereas true positives by considering detected buildings as well as paving, roofs, and areas of dense tree growth (i.e. soil ⟶forest). False positives were considered changes that are not of interest in this study (i.e. changes related to vehicles and seasonal changes (e.g. agricultural fields)) and pseudochanges. We categorized pseudochanges to those found in forests or in the water (e.g. sunglint) and to those caused by other reasons (e.g. co-registration and radiometric differences).

i. STANet evaluation for the whole study area

The STANet evaluation metrics for all four study areas and the training set (reported by the creators of STANet) are shown in Table 2. By observing the table it can be seen that the omission error is lower than the commission error. The lowest omission error is presented in Rhodes (7%) and the highest in Venice (26%). It is mostly caused in cases not present in the training set. The commission error is higher than ∼ 40% for all study areas and can be attributed mainly to the co-registration errors caused by the different satellite view directions and angles. Radiometric differences were the second reason for the commission error. This error percentage is expected since in much better conditions (training set composed of images from the same satellite with small co-registration errors) the network showed a 16% commission error. Tønsberg and Rhodes present the lowest commission errors (∼ 40%) and the highest F_1 scores followed by Granada and Venice.

TABLE 2. Evaluation metrics for the results of STANet.

Area	Recall	Precision	F_1 - score
Tønsberg	0.88	0.61	0.72
Granada	0.90	0.49	0.63
Rhodes	0.93	0.60	0.73
Venice	0.74	0.40	0.51
Training set	0.91	0.84	0.87

The pie charts displayed in Fig. 11 show the percentages of the types of changes detected by STANet for the four study areas. The highest pseudochanges are presented in Granada and Venice because of the presence of high building blocks and the different view directions and angles of GE01 and WV. Another challenge for Granada was its mountainous terrain because geometric distortions are increased. The lower pseudochanges for Rhodes can be attributed to the similar view direction of WV-2 and WV-3, whereas for Tønsberg to the low building height and higher similarity with the training set. It should be noted that as shown in the box plots of Fig. 6, Granada presented the highest mean RMSE in the co-registration process, whereas Rhodes the lowest. It is also interesting to notice the high amount of vehicles (ships) that exist in Venice and Rhodes.

ii. Evaluation of all methods on the test set

The evaluation metrics (recall, precision, F_1-score) for all the methods for a representative sample (test set (∼ 20% of the results)) are shown in Table 3 for the unsupervised methods and FDCNN, and in Table 4 for DASNetCDD, DASNetBCDD and STANet. In addition, a new evaluation metric was defined for the needs of the study ("precisionCD" (Equation 14)) that associates the commission error with the percentage of the pixels that were classified as change. We believe this index provides a better understanding of the magnitude of the commission error because it directly corresponds to its depiction in the image. The values of precisionCD for the test set are shown in Table 5. Finally, the percentages of the types of changes detected by all algorithms

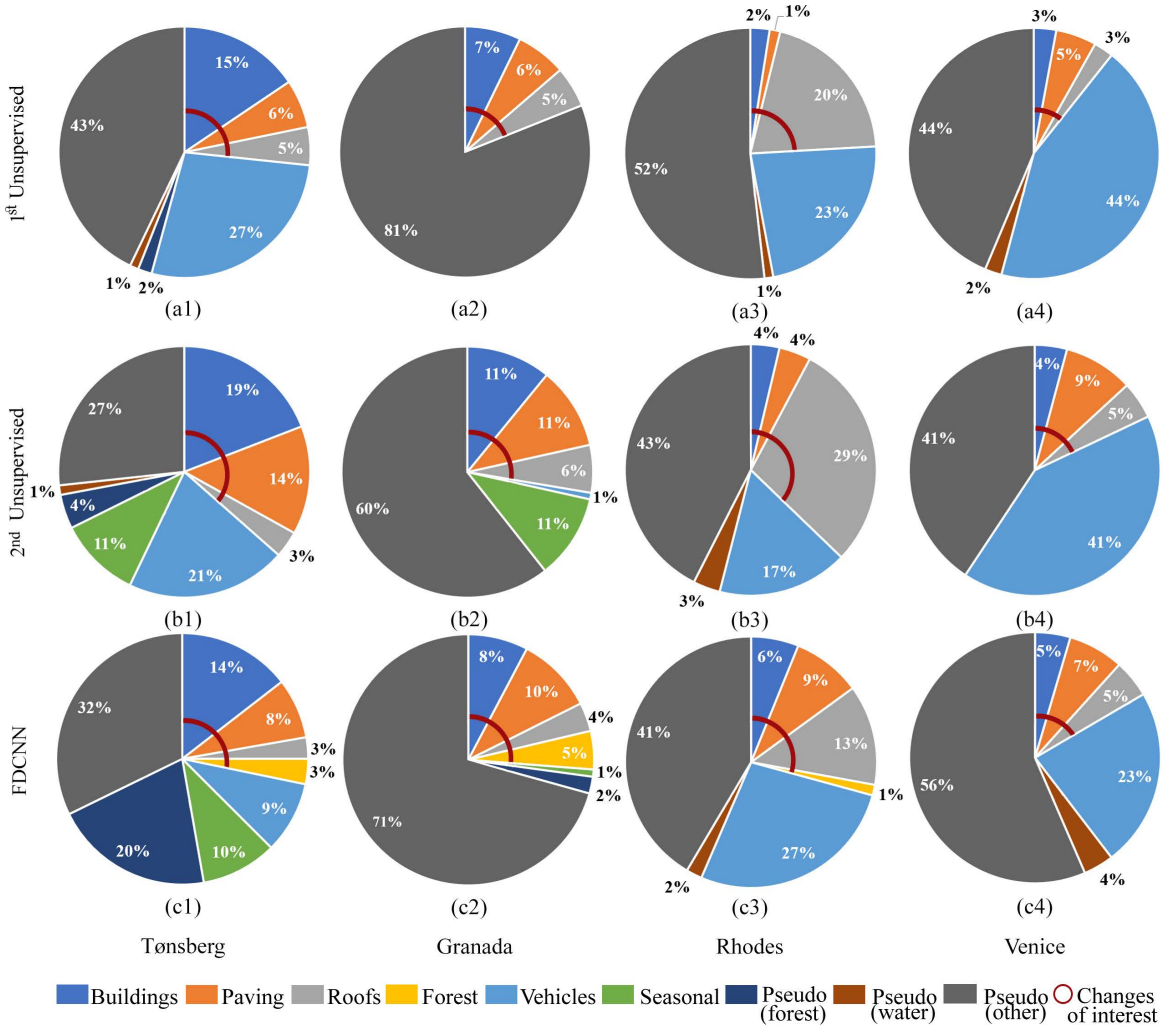

FIGURE 12. Percentages of the types of changes detected on the test set by the 1st unsupervised method (a1-a4), the 2nd Unsupervised method (b1-b4), and FDCNN (c1-c4).

TABLE 3. Evaluation metrics on the test set (1st Unsupervised, 2nd Unsupervised, FDCNN).

	1st Unsupervised			2nd Unsupervised			FDCNN		
Area	**Recall**	**Precision**	**F_1 - score**	**Recall**	**Precision**	**F_1 - score**	**Recall**	**Precision**	**F_1 - score**
Tønsberg	0.86	0.27	0.41	0.93	0.37	0.53	0.78	0.28	0.41
Granada	0.76	0.19	0.30	0.86	0.28	0.42	0.74	0.26	0.39
Rhodes	0.88	0.24	0.38	0.96	0.37	0.54	0.73	0.29	0.42
Venice	0.82	0.11	0.19	0.89	0.18	0.30	0.77	0.17	0.27
mean	0.83	0.20	0.32	0.91	0.30	0.45	0.75	0.25	0.37

on the test set are displayed via pie charts on Figs. 12, 13.

$$\text{precisionCD} = (1 - \text{precision}) \cdot \%\text{CP} \tag{14}$$

where: CP: pixels detected as change.

In Tables 3, 4 it can be observed that DASNetCDD displays the lowest omission error (<9%) followed by the second unsupervised method (<14%). STANet and the first unsupervised method show an average omission error of ∼15%, while the lowest performance is exhibited by FDCNN and DASNet-BCDD with an average of ∼25%. Regarding commission error, STANet shows the best performance (>37%) followed by the second unsupervised method with a minimum difference of 22%. The highest commission errors are shown by the first unsupervised method and DASNetCDD with an average of 80%. Similarly, STANet displays the highest F_1-score with an average value 0.66 followed by the second unsupervised method (0.45). The lowest F_1-scores are displayed by the first unsupervised method and DASNetCDD (∼0.33). Regarding study areas, in general, Tønsberg and Rhodes present the lowest commission errors and the highest F_1-scores.

TABLE 4. Evaluation metrics on the test set (DASNetCDD, DASNetBCDD, STANet).

	DASNetCDD			DASNetBCDD			STANet		
Area	**Recall**	**Precision**	**F_1 - score**	**Recall**	**Precision**	**F_1 - score**	**Recall**	**Precision**	**F_1 - score**
Tønsberg	0.91	0.25	0.39	0.73	0.36	0.49	0.92	0.63	0.75
Granada	0.93	0.16	0.27	0.77	0.28	0.41	0.85	0.52	0.65
Rhodes	0.97	0.25	0.40	0.77	0.26	0.39	0.94	0.59	0.73
Venice	0.95	0.15	0.25	0.77	0.12	0.21	0.69	0.42	0.52
mean	0.94	0.20	0.33	0.76	0.26	0.37	0.85	0.54	0.66

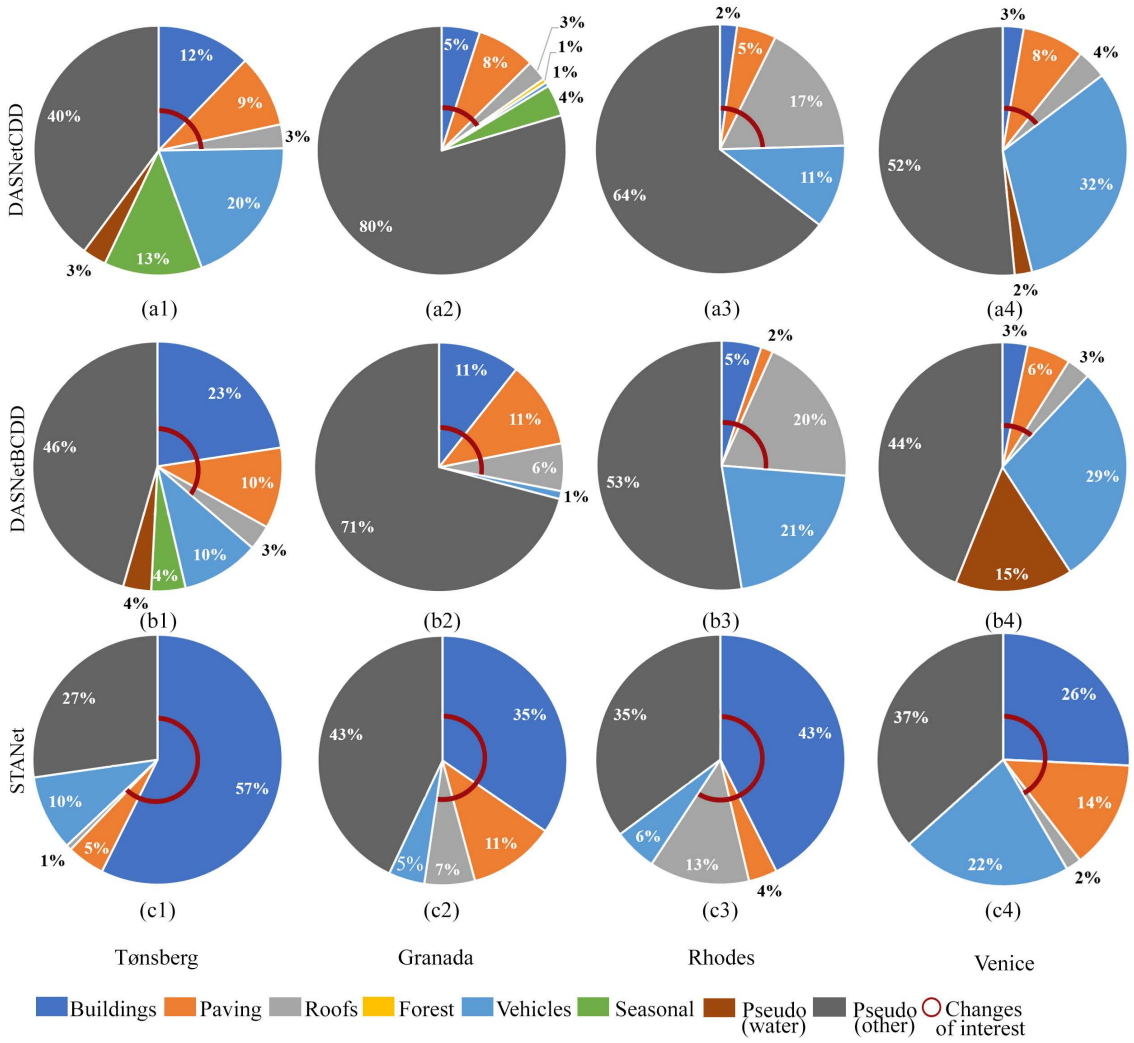

FIGURE 13. Percentages of the types of changes detected on the test set by DASNetCDD (a1-a4), DASNetBCDD (b1-b4), and STANet (c1-c4).

In Table 5, the values of the precisionCD metric show that when the commission error of STANet is translated into pixels, it is easily understandable that the pixels miss-classified by STANet as change, are 13 times less than DASNetBCCD which also focuses on changes of buildings. In addition, it can be observed that the commission error of DASNetCDD corresponds to the highest number of pixels and that the respective errors of the unsupervised methods, as well as of FDCNN and DASNetBCDD correspond to a similar amount of pixels.

From the pie charts displayed in Figs. 12, 13, it can be seen that STANet presents the highest percentages of the changes of interest for all four study areas. It is noted that this behavior is expected since the percentage of the changes of interest directly corresponds to the precision values. In addition, STANet presents the lowest percentages of pseudochanges of the ''other'' category (e.g. co-registration errors, radiometric differences). Further interesting observations are the high sensitivity shown by: a) the second unsupervised method for the detection of seasonal changes followed by DASNetCDD

TABLE 5. Calculation of precisionCD on the test set.

Area	1st Unsupervised	2nd Unsupervised	FDCNN	DASNetCDD	DASNetBCDD	STANet
Tønsberg	0.0480	0.0552	0.0520	0.0951	0.0282	0.0030
Granada	0.0802	0.0518	0.0536	0.2042	0.0313	0.0048
Rhodes	0.0563	0.0347	0.0129	0.1687	0.0188	0.0013
Venice	0.0855	0.0472	0.0378	0.1886	0.0689	0.0023
mean	0.0675	0.0472	0.0391	0.1641	0.0368	0.0028

and FDCNN, b) FDCNN for the detection of pseudochanges in the forest, and c) DASNetBCDD for the detection of pseudochanges in water (e.g. sunglint). It is noted that seasonal changes were included in the VHR images used in the training set of FDCNN. In addition, all methods are sensitive to the detection of changes in the presence of vehicles (mostly ships) and that both unsupervised methods show the highest miss-detection on this type of change. STANet shows the lowest percentage of vehicle changes. Finally, regarding study areas, Granada shows the highest percentage of pseudochanges of the ''other'' category in the results of all the algorithms while Tønsberg the lowest.

Concerning the need for a human operator, it is not required for the implementation of the unsupervised methods as well as DASNet and STANet. However, for the implementation of FDCNN in our study, a threshold was manually selected. Finally, it should be noted that inference time for the second unsupervised method and STANet was $\sim$ 0.05 sec for an image patch (size: 224 x 224) while for the rest of the methods (first unsupervised, FDCNN, DASNet) was $\sim$ 0.3 sec. The methods were implemented in a machine with i7-8700K CPU and NVIDIA 1070 Ti GPU.

V. CONCLUSION

In this study, five state-of-the-art DL CD methods were evaluated for VHR images with severe co-registration errors. In addition, before applying the CD process, four popular automatic co-registration methods were evaluated because of the importance of this pre-processing step for the successful output of the CD algorithms. The study was performed on images depicting four European areas with versatile urban patterns.

The implemented co-registration methods covered a wide range of the existing literature approaches. It was observed that SIFT and ORB, as well as a CNN-based method, displayed low performance, while results were more satisfying for the Fourier-Mellin Transform. However, given the crucial role of co-registration in the final CD result, it was decided to follow the more accurate manual approach, which produced mean RMSE between 1.5 and 4 m.

Concerning the CD methods, two unsupervised and three supervised were applied. The supervised method called STANet, produced satisfying results concerning the detection of buildings which are considered the most important indicator for the assessment of urban development. In addition, the commission error for this method was smaller than all other tested methods and was mostly attributed to the remaining co-registration issues. Its success can be attributed to the proposed attention mechanism in combination with a large professionally annotated dataset. The other methods showed a high commission error caused by different satellite view directions and angles that caused geometric distortions, co-registration errors, radiometric differences, seasonal changes and changes related to vehicles.

ACKNOWLEDGMENT

This work was implemented in the framework of the HYPERION project. The content of this publication is the sole responsibility of NTUA (Work Package 6, Task 6.3) and does not necessarily reflect the opinion of the European Union. For all figures permission has been obtained from the owners.

REFERENCES

[1] M. Volpi, D. Tuia, F. Bovolo, M. Kanevski, and L. Bruzzone, ''Supervised change detection in VHR images using contextual information and support vector machines,'' *Int. J. Appl. Earth Observ. Geoinf.*, vol. 20, pp. 77–85, Feb. 2013.

[2] D. Cerra, S. Plank, V. Lysandrou, and J. Tian, ''Cultural heritage sites in danger—Towards automatic damage detection from space,'' *Remote Sens.*, vol. 8, no. 9, p. 781, Sep. 2016.

[3] S. Ji, Y. Shen, M. Lu, and Y. Zhang, ''Building instance change detection from large-scale aerial images using convolutional neural networks and simulated samples,'' *Remote Sens.*, vol. 11, no. 11, p. 1343, Jun. 2019.

[4] H. Jiang, X. Hu, K. Li, J. Zhang, J. Gong, and M. Zhang, ''PGA-SiamNet: Pyramid feature-based attention-guided Siamese network for remote sensing orthoimagery building change detection,'' *Remote Sens.*, vol. 12, no. 3, p. 484, Feb. 2020.

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

[6] L. Su, M. Gong, P. Zhang, M. Zhang, J. Liu, and H. Yang, ''Deep learning and mapping based ternary change detection for information unbalanced images,'' *Pattern Recognit.*, vol. 66, pp. 213–228, Jun. 2017.

[7] L. T. Luppino, F. M. Bianchi, G. Moser, and S. N. Anfinsen, ''Unsupervised image regression for heterogeneous change detection,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9960–9975, Dec. 2019.

[8] F. Bovolo and L. Bruzzone, ''A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 1, pp. 218–236, Jan. 2006.

[9] J. Chen, X. Chen, X. Cui, and J. Chen, ''Change vector analysis in posterior probability space: A new method for land cover change detection,'' *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 2, pp. 317–321, Mar. 2010.

[10] F. Thonfeld, H. Feilhauer, M. Braun, and G. Menz, ''Robust change vector analysis (RCVA) for multi-sensor very high resolution optical satellite data,'' *Int. J. Appl. Earth Observ. Geoinf.*, vol. 50, pp. 131–140, Aug. 2016.

[11] K. Pearson, ''LIII. On lines and planes of closest fit to systems of points in space,'' *London, Edinburgh, Dublin Philosoph. Mag. J. Sci.*, vol. 2, no. 11, pp. 559–572, 1901.

[12] A. A. Nielsen, K. Conradsen, and J. J. Simpson, ''Multivariate alteration detection (MAD) and MAF postprocessing in multispectral, bitemporal image data: New approaches to change detection studies,'' *Remote Sens. Environ.*, vol. 64, no. 1, pp. 1–19, Apr. 1998.

[13] H. Chen, C. Wu, B. Du, L. Zhang, and L. Wang, "Change detection in multisource VHR images via deep Siamese convolutional multiple-layers recurrent neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2848–2864, Apr. 2020.
[14] N. Falco, P. R. Marpu, and J. A. Benediktsson, "A toolbox for unsupervised change detection analysis," *Int. J. Remote Sens.*, vol. 37, no. 7, pp. 1505–1526, 2016.
[15] J. S. Deng, K. Wang, Y. H. Deng, and G. Qi, "PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data," *Int. J. Remote Sens.*, vol. 29, no. 16, pp. 4823–4838, 2008.
[16] B. Thompson, "Canonical correlation analysis," in *Reading and Understanding More Multivariate Statistics*, L. G. Grimm and P. R. Yarnold, Eds. Washington, DC, USA: American Psychological Association, 2000, pp. 285–316. [Online]. Available: https://psycnet.apa.org/record/2000-00427-009
[17] M. Hussain, D. Chen, A. Cheng, H. Wei, and D. Stanley, "Change detection from remotely sensed images: From pixel-based to object-based approaches," *ISPRS J. Photogram. Remote Sens.*, vol. 80, pp. 91–106, Jun. 1998.
[18] J. L. Gil-Yepes, L. A. Ruiz, J. A. Recio, Á. Balaguer-Beser, and T. Hermosilla, "Description and validation of a new set of object-based temporal geostatistical features for land-use/land-cover change detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 121, pp. 77–91, Nov. 2016.
[19] G. Chen, K. Zhao, and R. Powers, "Assessment of the image misregistration effects on object-based change detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 87, pp. 19–27, Jan. 2014.
[20] P. Gong, E. F. Ledrew, and J. R. Miller, "Registration-noise reduction in difference images for change detection," *Int. J. Remote Sens.*, vol. 13, no. 4, pp. 773–779, Mar. 1992.
[21] L. Bruzzone and S. B. Serpico, "Detection of changes in remotely-sensed images by the selective use of multi-spectral information," *Int. J. Remote Sens.*, vol. 18, no. 18, pp. 3883–3888, 1997.
[22] D. A. Stow, "Reducing the effects of misregistration on pixel-level change detection," *Int. J. Remote Sens.*, vol. 20, no. 12, pp. 2477–2483, Jan. 1999.
[23] J. Theiler and B. Wohlberg, "Local coregistration adjustment for anomalous change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 8, pp. 3107–3116, Aug. 2012.
[24] A. M. El Amin, Q. Liu, and Y. Wang, "Convolutional neural network features based change detection in satellite images," *Proc. SPIE*, vol. 10011, Jul. 2016, Art. no. 100110W.
[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[26] A. M. El Amin, Q. Liu, and Y. Wang, "Zoom out CNNs features for optical remote sensing change detection," in *Proc. 2nd Int. Conf. Image, Vis. Comput. (ICIVC)*, Jun. 2017, pp. 812–817.
[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[28] C. Zhang, L. He, and L. Jiang, "Refined deep features for unsupervised change detection in high resolution remote sensing images," in *Proc. 9th Int. Conf. Agro-Geoinform. (Agro-Geoinformatics)*, Jul. 2021, pp. 1–4.
[29] B. Hou, Y. Wang, and Q. Liu, "Change detection based on deep features and low rank," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2418–2422, Dec. 2017.
[30] G. S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Apr. 2017.
[31] K. L. de Jong and A. S. Bosman, "Unsupervised change detection in satellite images using convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.
[32] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241. [Online]. Available: https://link.springer.com/book/10.1007/978-3-319-24553-9#about
[33] ISPRS. (2018). *2D Semantic Labeling—Vaihingen Data*. [Online]. Available: http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-vaihingen.htm
[34] J. Liu, K. Chen, G. Xu, X. Sun, M. Yan, W. Diao, and H. Han, "Convolutional neural network-based transfer learning for optical aerial images change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 1, pp. 127–131, Jan. 2020.
[35] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised deep change vector analysis for multiple-change detection in VHR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3677–3693, Jun. 2019.
[36] C. Wu, H. Chen, B. Do, and L. Zhang, "Unsupervised change detection in multi-temporal VHR images based on deep kernel PCA convolutional mapping network," 2019, *arXiv:1912.08628*.
[37] M. Zhang and W. Shi, "A feature difference convolutional neural network-based change detection method," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7232–7246, Oct. 2020.
[38] S. Ghaffarian, N. Kerle, E. Pasolli, and J. J. Arsanjani, "Post-disaster building database updating using automated deep learning: An integration of pre-disaster OpenStreetMap and multi-temporal satellite data," *Remote Sens.*, vol. 11, no. 20, p. 2427, Oct. 2019.
[39] C. Benedek and T. Szirányi, "Change detection in optical aerial images by a multilayer conditional mixed Markov model," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 10, pp. 3416–3430, Oct. 2009.
[40] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep Siamese convolutional network for optical aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1845–1849, Oct. 2017.
[41] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional Siamese networks for change detection," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 4063–4067.
[42] M. Zhang, G. Xu, K. Chen, M. Yan, and X. Sun, "Triplet-based semantic relation learning for aerial remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 266–270, Feb. 2019.
[43] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.
[44] DeepAI. (2012). *PASCAL VOC Dataset*. [Online]. Available: https://deepai.org/dataset/pascal-voc
[45] R. Liu, Z. Cheng, L. Zhang, and J. Li, "Remote sensing image change detection based on information transmission and attention mechanism," *IEEE Access*, vol. 7, pp. 156349–156359, 2019.
[46] R. Liu, D. Jiang, L. Zhang, and Z. Zhang, "Deep depthwise separable convolutional network for change detection in optical aerial images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1109–1118, 2020.
[47] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Multitask learning for large-scale semantic change detection," *Comput. Vis. Image Understand.*, vol. 187, Oct. 2019, Art. no. 102783.
[48] X. Li, M. He, H. Li, and H. Shen, "A combined loss-based multiscale fully convolutional network for high-resolution remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[49] M. A. Lebedev, Y. V. Vizilter, O. V. Vygolov, V. A. Knyaz, and A. Y. Rubis, "Change detection in remote sensing images using conditional adversarial networks," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 42, pp. 565–571, May 2018.
[50] Q. Zhu, X. Guo, W. Deng, S. Shi, Q. Guan, Y. Zhong, L. Zhang, and D. Li, "Land-use/land-cover change detection based on a Siamese global learning framework for high spatial resolution remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 184, pp. 63–78, Feb. 2022.
[51] 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 Trans. Geosci. Remote Sens.*, vol. 56, no. 7, pp. 4002–4015, Jul. 2018.
[52] S. Shi, Y. Zhong, J. Zhao, P. Lv, Y. Liu, and L. Zhang, "Land-use/land-cover change detection based on class-prior object-oriented conditional random field framework for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.
[53] J. Chen, Z. Yuan, J. Peng, L. Chen, H. Huang, J. Zhu, Y. Liu, and H. Li, "DASNet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1194–1206, 2021. [Online]. Available: https://github.com/lehaifeng/DASNet
[54] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.
[55] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sens.*, vol. 12, no. 10, p. 1662, May 2020.
[56] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep Siamese convolutional network model," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 811–815, May 2021.
[57] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangguan, Li Huang, and G. Liu, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 183–200, Aug. 2020.

[58] A. Raza, H. Huo, and T. Fang, "EUNet-CD: Efficient UNet++ for change detection of very high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[59] K. Lim, D. Jin, and C.-S. Kim, "Change detection in high resolution satellite images using an ensemble of convolutional neural networks," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2018, pp. 509–515.
[60] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Jun. 2004.
[61] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2564–2571.
[62] Z. Yang, T. Dan, and Y. Yang, "Multi-temporal remote sensing image registration using deep convolutional features," *IEEE Access*, vol. 6, pp. 38544–38555, 2018.
[63] D. Casasent and D. Psaltis, "New optical transforms for pattern recognition," *Proc. IEEE*, vol. 65, no. 1, pp. 77–84, Jan. 1977.
[64] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2006, pp. 430–443. [Online]. Available: https://link.springer.com/book/'10.1007/11744023#about
[65] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary robust independent elementary features," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 778–792. [Online]. Available: https://link.springer.com/book/10.1007/978-3-642-15561-1#about
[66] X. Guo, Z. Xu, Y. Lu, and Y. Pang, "An application of Fourier-Mellin transform in image registration," in *Proc. 5th Int. Conf. Comput. Inf. Technol. (CIT)*, 2005, pp. 619–623.
[67] D. Gupta and M. K. Patil, "A review on image registration," *Int. J. Eng. Res. Technol.*, vol. 3, no. 2, pp. 2630–2633, 2014.
[68] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc., B, Methodol.*, vol. 39, no. 1, pp. 1–22, 1977.
[69] R. Jonker and A. Volgenant, "A shortest augmenting path algorithm for dense and sparse linear assignment problems," *Computing*, vol. 38, no. 4, pp. 325–340, Nov. 1987.
[70] A. M. E. Amin, Q. Liu, and Y. Wang. (2016). *Unstructured-Change-Detection-Using-CNN*. [Online]. Available: https://github.com/vbhavank/Unstructured-change-detection-using-CNN
[71] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Operating Syst. Design Implement.*, 2016, pp. 265–283.
[72] F. Chollet *et al.* (Jun. 2018). *Keras: The Python Deep Learning Library*. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2018ascl.soft06022C/exportcitation
[73] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.
[74] M. Zhang and W. Shi. (2020). *FDCNN*. [Online]. Available: https://github.com/MinZHANG-WHU/FDCNN
[75] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 675–678.
[76] J. Chen, Z. Yuan, J. Peng, L. Chen, H. Haozhe, J. Zhu, Y. Liu, and H. Li. (2020). *DASNet*. [Online]. Available: https://github.com/lehaifeng/DASNet
[77] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Dec. 2019, pp. 8026–8037.
[78] H. Chen and Z. Shi. (2020). *STANet for Remote Sensing Image Change Detection*. [Online]. Available: https://github.com/justchenhao/STANet
[79] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[80] M. A. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

VIKTORIA KRISTOLLARI received the M.S. degree in rural and surveying engineering from the National Technical University of Athens (NTUA), Athens, Greece, in 2016, where she is currently pursuing the Ph.D. degree in remote sensing. Since 2017, she has been a Researcher with the Laboratory of Remote Sensing, NTUA, and has participated in EU projects. She has authored publications for peer-reviewed international scientific journals and conferences. Her research interests include multispectral and hyperspectral image processing via artificial neural networks.

She was a recipient of the Excellence Award of the Limmat Stiftung Foundation, in 2016, and was granted a Three Year Scholarship by the NTUA Research Committee for her Ph.D. research, in 2017.

VASSILIA KARATHANASSI received the B.S. degree in rural and surveying engineering from the National Technical University of Athens (NTUA), Greece, in 1984, the M.S. degree in urban planning-geography from Paris V, France, in 1985, and the Ph.D. degree in remote sensing from NTUA, in 1990.

Since 2000, she has been a Professor with the School of Rural, Surveying, and Geoinformatics Engineering, NTUA, specialized in hyperspectral/multispectral remote sensing and InSAR/DInSAR processing and applications. She teaches multiple undergraduate and postgraduate courses and she has supervised more than 40 undergraduate, eight master's theses, eight Ph.D. theses (four of them completed), and one postdoctoral research. She has published research work includes more than 100 articles and one chapter in the book *Hyperspectral Remote Sensing*. Furthermore, she is involved in EU and national excellence/competitive research projects as a Coordinator, a Principal Investigator, and a Researcher toward the design, development, and validation of state-of-the-art methodologies, and cutting-edge technology in remote sensing and earth observation.