Change Detection in Multi-temporal VHR Images Based on Deep Siamese Multi-scale Convolutional Neural Networks

Hongruixuan Chen, *Student Member, IEEE,* Chen Wu, *Member, IEEE,* Bo Du, *Senior Member, IEEE,* and Liangpei Zhang, *Fellow, IEEE*

***Abstract*—Very-high-resolution (VHR) images can provide abundant ground details and spatial geometric information. Change detection in multi-temporal VHR images plays a significant role in urban expansion and area internal change analysis. Nevertheless, traditional change detection methods can neither take full advantage of spatial context information nor cope with the complex internal heterogeneity of VHR images. In this paper, a powerful feature extraction model entitled multi-scale feature convolution unit (MFCU) is adopted for change detection in multi-temporal VHR images. MFCU can extract multi-scale spatial-spectral features in the same layer. Based on the unit two novel deep siamese convolutional neural networks, called as deep siamese multi-scale convolutional network (DSMS-CN) and deep siamese multi-scale fully convolutional network (DSMS-FCN), are designed for unsupervised and supervised change detection, respectively. For unsupervised change detection, an automatic pre-classification is implemented to obtain reliable training samples, then DSMS-CN fits the statistical distribution of changed and unchanged areas from selected training samples through MFCU modules and deep siamese architecture. For supervised change detection, the end-to-end deep fully convolutional network DSMS-FCN is trained in any size of multi-temporal VHR images, and directly outputs the binary change map. In addition, for the purpose of solving the inaccurate localization problem, the fully connected conditional random field (FC-CRF) is combined with DSMS-FCN to refine the results. The experimental results with challenging data sets confirm that the two proposed architectures perform better than the state-of-the-art methods.**

***Index Terms*—Change detection, very-high-resolution images (VHR images), multi-temporal images, multi-scale feature convolution, deep siamese convolutional neural network, fully connected conditional random field (FC-CRF)**

I. Introduction

CHANGE detection is one of the major and hot topics in the remote sensing field. The most commonly used definition of change detection is given by Singh [1]: change detection is the process of identifying differences in the state of an object or phenomenon by observing it at different times. Now, change detection is widely applied to ecosystem monitoring, resource management, land-use/land-cover change analysis, urban expansion research, and damage assessment [2]–[7].

Multi-spectral images with medium- and low- spatial resolution are the most commonly used data source for change detection, thus numerous change detection methods have been developed and newer methods are still emerging. Change vector analysis (CVA) is one of the most classic change detection methods that generates a difference image (DI) and performs clustering algorithms in DI to obtain the change detection result [8]. CVA is also the backbone for some of the more advanced change detection methods [9]–[11]. As a dimension reduction method, principal component analysis (PCA) uses an orthogonal transformation to convert original images into a new orthogonal feature space and chooses a part of principal components for change detection [12]. Multivariate alteration detection (MAD) and its iterative version iteratively re-weighted MAD (IRMAD) extract changed pixels by maximizing the difference between the transformed variables [13], [14]. A novel change detection method is proposed by Wu et al. in [15], [16]. This method based on slow feature analysis (SFA) theory tries to seek the most invariant components in multi-temporal images for change detection.

Nowadays, with the development of Earth observation technology, VHR images are more available by lots of satellite sensors, such as IKONOS, Worldview, SPOT, GaoFen (GF), and QuickBird. VHR images have abundant detailed geometric information, which has crucial effects on the research of urban change analysis and building detection [4], [16]–[18]. Therefore, VHR images change detection has caught more and more attention in the remote sensing field [19]–[35].

Nevertheless, as spatial resolution increases, the internal heterogeneity of the same class also increases. The aforementioned change detection methods for low- and medium-spatial resolution images only explore spectral features and neglect spatial context information, thus they may not apply for change detection in VHR images. To achieve better change detection performance, a lot of methods are designed to utilize spatial-spectral features. In [19], a method based on texton forest is developed to capture spatial context information. Tan et al explore texture and morphological profiles in VHR images for change detection [23]. Based on the assumption that the spectral vectors of the pixels belonging to the same

Manuscript submitted July 8, 2020.

H. Chen is with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, P.R. China (e-mail:Qschrx@whu.edu.cn).

C. Wu is with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, P.R. China (e-mail: chen.wu@whu.edu.cn, corresponding author).

B. Du is with the School of Computer Science, and Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan, P.R. China (email: gunspace@163.com).

L. Zhang is with the Remote Sensing Group, State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, P.R. China (e-mail: zlp62@whu.edu.cn).

type object are similar, some object-based change detection (OBCD) methods are designed and achieve relatively good results [24]–[26]. Besides, some probabilistic graph models, such as conditional random field (CRF) [36] and Markov random field (MRF) [37], are introduced to utilize spatial context information for change detection [27]–[31]. However, these methods only extract low-level features from VHR images for change detection, which are not robust and insufficient for representing the key information of original images.

Recently, deep learning (DL) has achieved significant performance in many domains [38], including remote sensing image interpretation [39], [40]. Convolutional neural network (CNN), as a classical and powerful DL architecture has the capacity to capture multi-level features in an automatic manner [41], which is suitable for processing VHR images. Consequently, a variety of methods based on CNN have been proposed for change detection in multi-temporal VHR images. Saha et al. [11] propose an unsupervised method called deep change vector analysis (DCVA) for binary and multi-class change detection. In DCVA, a pre-trained deep CNN is adopted to extract deep spatial-spectral features from multi-temporal images. In [42], a deep convolutional neural network entitled symmetric convolutional coupling network (SCCN) is introduced for change detection in heterogeneous images. In SCCN, the convolutional layer is responsible for feature extraction. For multi-temporal aerial image change detection, Zhan et al. [32] design a deep siamese convolutional network, which extracts features through two weight-shared branches. In [43], the fully convolutional siamese network is first introduced into change detection and three networks are designed. The three networks are trained in an end-to-end manner and achieve good performance. For the purpose of extracting high-level spatial-spectral features from multi-source VHR images, a deep siamese convolutional neural network is proposed in [44]. After high-level spatial-spectral features are extracted, a multiple-layers recurrent neural network is designed to mine change information. Except for CNN architecture, a change detection method based on conditional generative adversarial network (cGAN) is designed and a convolutional layer aims to extract spatial-spectral features from VHR images [45]. In this method, the convolutional layer is responsible for extracting features from VHR image patches.

All of these above methods adopt a single scale convolution kernel as the feature extraction module. Although the single scale convolution kernel could extract spatial-spectral features to a certain degree, it still has some powerlessness in coping with the complex ground situations of VHR images. Few research has attempted to explore other sizes of convolution kernels or even multiple convolution kernels for change detection in VHR image. Whats more, unsupervised and supervised change detection often face different scenarios, it is hard to find a general architecture that is suitable for both tasks simultaneously. On the one hand, in supervised change detection, these network structures proposed for unsupervised change detection, such as SCCN, have limited fitting ability, which makes it difficult to obtain accurate change detection result. On the other hand, the architectures developed for supervised change detection cannot be adequately trained in unsupervised tasks, also resulting in unsatisfactory results. Therefore, it is necessary to design different network architectures for unsupervised and supervised change detection, respectively.

Considering the above issues comprehensively and inspired by Inception network [46], the multi-scale feature convolution unit proposed in [46] is adopted to extract multi-scale spatial-spectral features in the same layer, which is suitable for processing VHR images. Adopting MFCU as the basic feature extraction module, two powerful deep siamese convolutional neural networks are designed for unsupervised and supervised change detection, respectively. For supervised change detection, a probabilistic graph model, FC-CRF [47] is adopted to refine the change detection results. Based on the two networks and FC-CRF, the specific supervised and unsupervised algorithms are developed.

The contributions of this paper [1] are summarized as follows:

1) This paper introduces a multi-scale feature convolution unit into change detection, which has the capacity to extract multi-scale spatial-spectral features from VHR images. Whats more, the unit is a flexible module and can be used in any deep neural networks designed for the tasks involving VHR images. To the authors best knowledge, this is the first time that such multi-scale feature convolution unit is exploited for change detection.
2) For the fact that unsupervised and supervised change detection often face different scenarios, based on MFCU, two novel deep siamese convolutional neural networks are developed for unsupervised and supervised change detection, respectively. Among them, the network used for supervised change detection is able to process images of any size and does not require sliding patch-window, thus the accuracy and inference speed could be significantly improved.
3) In the proposed supervised change detection algorithm, for the purpose of solving the problem of inaccurate localization caused by deep convolution architecture, FC-CRF is adopted to refine the results obtained by DSMS-FCN.

The rest of this paper is organized as follows. Section II introduces the MFCU, DSMS-CN, and specific unsupervised change detection algorithm in detail. In section III, the DSMS-FCN, FC-CRF, and corresponding supervised change detection algorithm are described. To evaluate the proposed methods, the experiments of unsupervised and supervised change detection are carried in section IV and section V, respectively. In the end, Section VI draws the conclusion of our work in this paper.

II. Unsupervised Change Detection

In this section, MFCU, the network designed for unsupervised change detection and specific change detection algorithm are elaborated. Though MFCU is introduced in this section, it is also the basic module of the network proposed for supervised change detection.

[1]This paper is an improved version of the original conference paper: https://ieeexplore.ieee.org/document/8866947

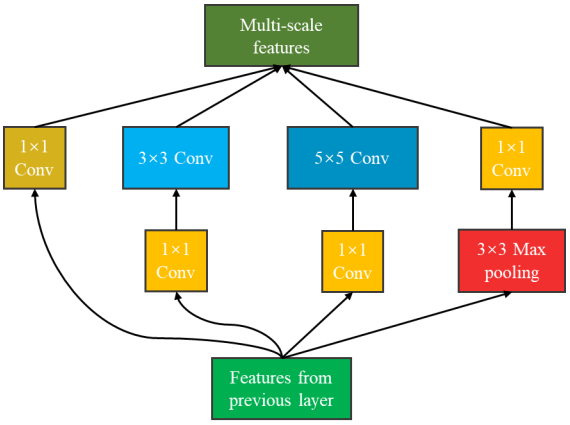


Fig. 1: Illustration of MFCU. Different from the conventional single-scale convolution unit, MFCU can extract multi-scale spatial-spectral features in parallel by four ways in the same layer.

A. Multi-scale Feature Convolution Unit

The VHR images can provide abundant ground details, texture information, and spatial distribution information [44]. As a commonly used convolution unit, 3×3 convolution kernel could extract spatial-spectral features from VHR images. However, the 3×3 convolution kernel has two obvious disadvantages in feature extraction of VHR images. First, the 3×3 convolution kernel could only extract single scale spatial-spectral features in the same layer. But in VHR images, there exist different features, varying from small scales to large scales. Besides, as a weighted summation operation, convolution has a smoothing effect, thus some changes existing in multi-temporal images could be erased. Therefore, it is undeniable that the 3×3 convolution kernel, or say, the conventional single-scale convolution unit is somewhat incompetent in dealing with complex multi-scale ground conditions in VHR images.

Therefore, to extract multi-scale spatial-spectral features, the Inception module first proposed in [46] is adopted, and we rename it as MFCU in this paper. As illustrated in Fig. 1, MFCU has a network in network structure [48] and can extract multi-scale spatial-spectral features by 1×1 convolution kernel, 3×3 convolution kernel, 5×5 convolution kernel, and 3×3 max pooling, respectively. In the above four ways, the 1×1 convolution kernel concentrates on extracting the features of a pixel itself. The 3×3 convolution kernel is able to extract the spatial-spectral features. The 5×5 convolution kernel extracts spatial-spectral features over a larger range, which applies for a few large-scale continuous objects in VHR images. And the 3×3 max pooling could extract the most salient features and efficiently avoids the smoothing effect of the convolution operation. At last, the four type features are fused to obtain the higher dimensional multi-scale features. In addition, a bottleneck design [49] is adopted in MFCU, which uses the 1×1 convolution kernel to reduce the feature dimensions. This structure can efficiently reduce the number of parameters and make the training process of network easier.

Compared with the conventional single-scale convolution unit, MFCU could extract multi-scale features, which makes the feature extraction ability of network more powerful and does not increase parameters of network.

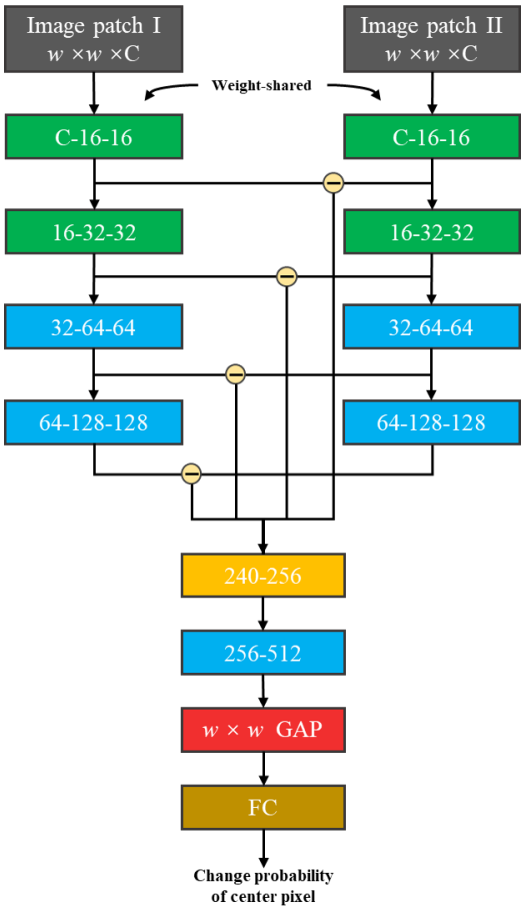


Fig. 2: Architecture of the proposed DSMS-CN. DSMS-CN extracts features from multi-temporal VHR image patches with a fixed size and outputs change probability of center pixels. Legend: green block indicates a conventional 3×3 convolution module, blue block means the MFCU module, yellow block is an 1×1 convolution module, red block is a w×w global average pooling layer and brown block means a fully connected layer. The numbers in each block represent the change of the number of feature channels in each module.

B. Deep Siamese Multi-scale Convolutional Network

Using MFCU as a basic feature extraction module, the network entitled DSMS-CN is illustrated in Fig. 2, which consists of two sub-networks: feature extraction network and change judging network. DSMS-CN processes image patches with a fixed size and outputs change probabilities of the center pixels.

The feature extraction network of DSMS-CN is a siamese convolutional network [50]. Its two branches extract spatial-spectral features from two multi-temporal image patches with a fixed size using exactly the same way. In each branch, the former two conventional convolutional modules transform the original image patches into relative high dimensional representation, then abundant multi-scale features are extracted by

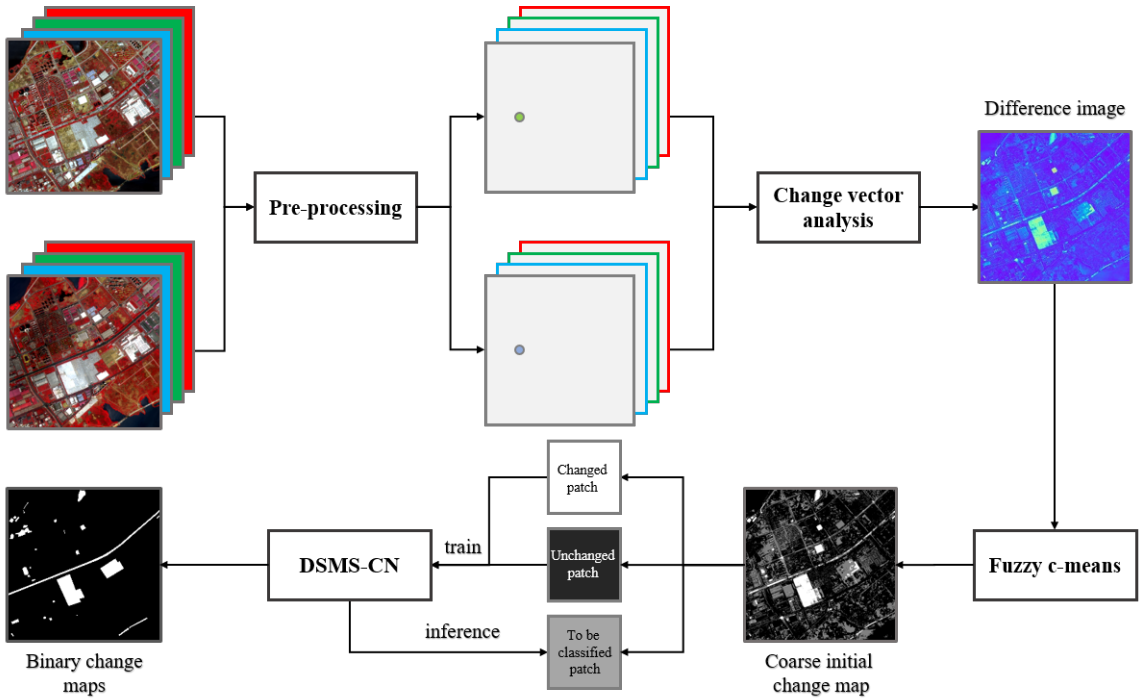


Fig. 3: Flowchart of the unsupervised change detection alogrithm. The co-registration and radiometric correction are first implemented to make the geometry and radiometric conditions of multi-temporal VHR images consistent. Then suitable training image patches are chosen by CVA and FCM. Finally, training DSMS-CN with the selected image patches, and then DSMS-CN classifies the center pixels of the remaining image patches to obtain full change detection result. Note that the change detection decision made by DSMS-CN is at the pixel level instead of the patch level.

the two latter MFCU modules. To highlight change information, the absolute values of feature differences are calculated [11]. Then the feature differences with different level are concatenated and a 1×1 convolution layer is used to fuse these features.

In the change judging network, multi-scale features are further extracted by an MFCU first. For the purpose of making the network more robust and mitigating overfitting [48], a global average pooling layer (GAP) replaces the fully connected layer to generate feature vector. Finally, the change probability of the center pixel in each patch is obtained by a fully connected layer.

In DSMS-CN, the activation functions of both conventional and multi-scale convolutional layers are rectified linear units (ReLU), the sigmoid function is adopted in the last fully connected layer to predict the change probability. For the purpose of preserving the information to the largest degree, DSMS-CN does not adopt the max-pooling layer to reduce dimension.

C. Unsupervised Change Detection Algorithm

Adopting DSMS-CN to obtain change detection results, the specific unsupervised change detection algorithm is presented. The pipeline of the algorithm is illustrated in Fig. 3. The first step is image pre-processing, including co-registration and radiometric correction. Image registration is defined as the process of aligning two or more images of the same scene acquired at diverse times [51]. Through collecting matched point-pairs, constructing transform models and transforming images, the multi-temporal VHR images are geometrically aligned. Then radiometric relative normalization is implemented to reduce the radiometric difference between multi-temporal VHR images caused by distinct imaging conditions, such as light intensity, sun zenith angle, and atmospheric conditions. The specific implementation is normalizing the multi-temporal VHR images with zero mean and unit variance, respectively.

After pre-processing images, the suitable training samples are chosen by automatic pre-classification. The main purpose of this step is to choose reliable samples for training the proposed network. CVA is first adopted to calculate DI of multi-temporal VHR images. Then, fuzzy c-means clustering algorithm (FCM), based on the memberships of pixels, is performed to partition DI into three clusters: w_c, w_{uc} and w_{tbc}. The pixels belonging to w_c and w_{uc} are reliable pixels that have high change and non-change probabilities, respectively. And the pixels in w_{tbc} are unreliable and need to be classified. The $w \times w$ neighborhood area of pixels in w_c and w_{uc} are chosen as training samples. This automatic sample selection method based on pre-classification has been utilized in many change detection models [52]–[60]. Besides, some advanced change detection models, such as DCVA [11], based on CVA are developed.

Eventually, DSMS-CN is trained on the chosen training image patches. After the training process is completed, the pixels in w_{tbc} are detected by DSMS-CN and the full change

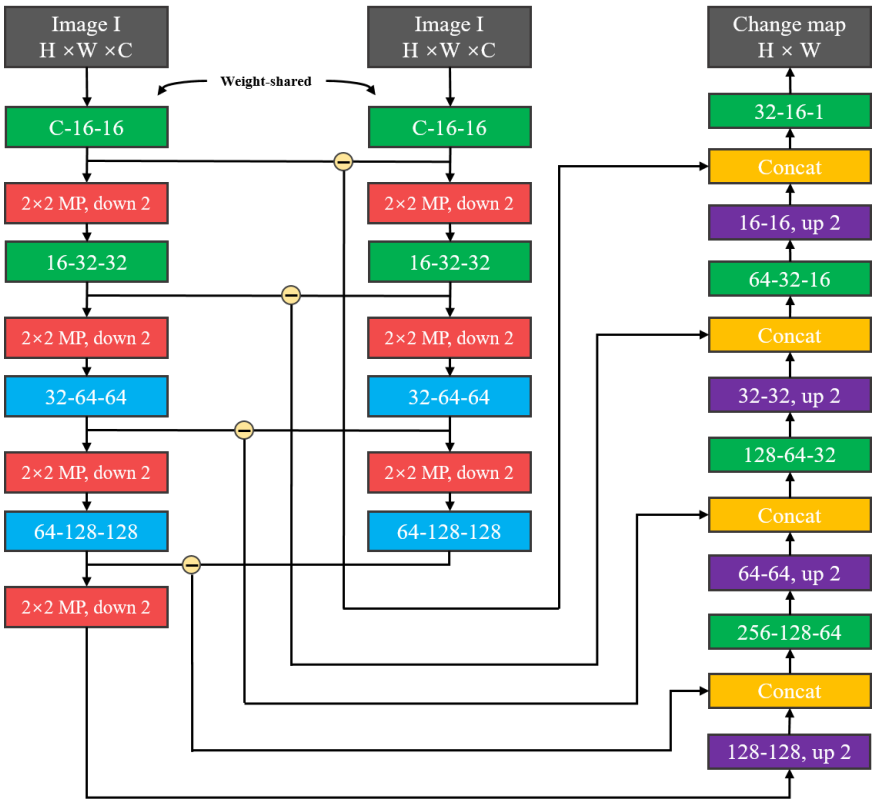


Fig. 4: Architecture of the proposed DSMS-FCN. The encoder network directly processes multi-temporal VHR images to get high dimensional multi-scale feature maps and the decoder network generates change maps based on feature difference from multiple layers and high dimensional multi-scale feature map of one branch. Legend: green block indicates a conventional 3×3 convolution module, blue block means the MFCU module, red block module is a 2×2 max-pooling layer and purple block module is a transpose convolution module. The numbers in each block represents the change of the number of feature channels.

detection result is generated. Owing to sigmoid function of fully connected layer of DSMS-CN, the threshold segmentation step could be simplified and just sets threshold as 0.5 to get binary change map. Though training the proposed DSMS-CN is a supervised learning process, the entire process of the algorithm is an unsupervised fashion without any priori knowledge.

III. Supervised Change Detection

In this section, the network proposed for supervised change detection, FC-CRF, and specific supervised change detection algorithm are introduced.

A. Deep Siamese Multi-scale Fully Convolutional Network

Based on MFCU (for details, please refer to Section II-A), the proposed network for supervised change detection is a fully convolutional network entitled DSMS-FCN. DSMS-FCN consists of an encoder network and a decoder network. The specific architecture of DSMS-FCN is shown in Fig. 4.

The encoder network of DSMS-FCN has two weight-shared branches. Each branch has four subsampling layers and each subsampling layer includes a convolution module and a 2×2 max-pooling layer. The former two convolution modules consist of 3×3 convolution kernels and the latter two modules consist of MFCUs. Based on the skip-connection structure proposed in the U-Net [61], the features in subsampling layer and upsampling layer at the same scale are concatenated during the upsampling phase, which could recover localization information to a certain degree and generate more accurate binary change maps with precise boundaries. The motivation for concatenating the absolute values of feature differences with the features in the upsampling layer is that change detection aims at detecting the differences between multi-temporal images. Furthermore, it is worth pointing out that only one branchs features enter into decoder network. This is because the two branches share weights and changes in multi-temporal VHR images are only in the minority, thus most of features extracted from both sides are same. Meanwhile, feature differences are delivered to decoder network through the skip-connection structure. Consequently, using features of only one branch could avoid feature redundancy and require a smaller number of parameters.

In DSMS-FCN, all convolutional layers and transpose convolutional layers adopt ReLU as activation function, except the last convolutional layer, which adopts sigmoid function to predict change probability of the whole image. Owing to the fully convolutional architecture of DSMS-FCN, it can process multi-temporal VHR images of any size.

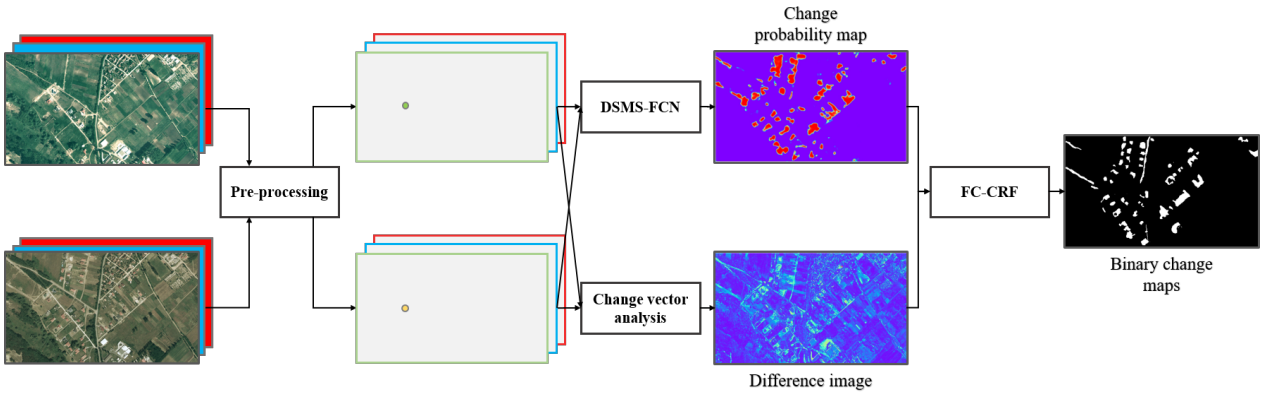


Fig. 5: Flowchart of the supervised change detection alogrithm. The pre-processing is first performed on the data set. DSMS-FCN and FC-CRF are trained on the training set. After the training step is completed, DSMS-FCN infers change probability map of given multi-temporal VHR images. Then the FC-CRF refine the results obtained by DSMS-FCN depending on the change probability map and DI acquired by CVA. At last, the FC-CRF generates a more precise binary change map.

B. Fully Connected Conditional Random Field

Though DSMS-FCN adopts the skip-connection structure to deliver localization information, it still suffers from the problem of inaccurate localization caused by invariance of features and large receptive field [62]. To tackle this problem, FC-CRF [47] is adopted to refine the localization information of the results obtained by DSMS-FCN. Compared with CRF, FC-CRF considers short-range and long-range information simultaneously, thus it can better recover the local structure.

FC-CRF is a conditional probability distribution model that outputs another set of random variables given a set of input random variables. According to [47], the energy function of FC-CRF is defined as follows:

$$E(Y|X) = \sum_{i} \phi_u(y_i) + \sum_{i<j} \phi_p(y_i, y_j) \tag{1}$$

where i and j range from 1 to N, ϕ_u indicates unary potential, and ϕ_p indicates pair-wise potential. In change detection, $X = \{x_1, x_2, \cdots, x_N\}$ is the observed image acquired by the difference of multi-temporal images and $Y = \{y_1, y_2, \cdots, y_N\}$ is the binary change map.

The domain of each y_i is $L = \{0, 1\}$. The unary potential

$$\phi_u(y_i) = -logP(y_i) \tag{2}$$

where $P(y_i)$ is change probability of pixel i, which is computed by DSMS-FCN. In addition, each pixel pair has a corresponding pairwise term no matter how far apart they are. The pairwise $\phi_p(y_i, y_j)$ potential has the form:

$$\begin{cases} \phi_p(y_i, y_j) = \mu(y_i, y_j) k(f_i, f_j) \\ k(f_i, f_j) = \sum\limits_{m=1}^{n} w^{(m)} k^{(m)}(f_i, f_j) \end{cases} \tag{3}$$

Here, μ is a penalty, $\mu(y_i, y_j) = 1 \; if \; y_i \neq y_j$ and zero otherwise.$k^{(m)}$ is a Gaussian kernel and is weighted by $w^{(m)}$ and n is the number of kernels. f_i and f_j are feature vectors for pixels i and j in a feature space.

In change detection problem, the kernels are

$$\begin{aligned} k(f_i, f_j) = w_1 exp(-\frac{\| c_i - c_j \|_2^2}{2\sigma_\alpha^2} - \frac{\| d_i - d_j \|_2^2}{2\sigma_\beta^2}) \\ + w_2 exp(-\frac{\| c_i - c_j \|_2^2}{2\sigma_\gamma^2}) \end{aligned} \tag{4}$$

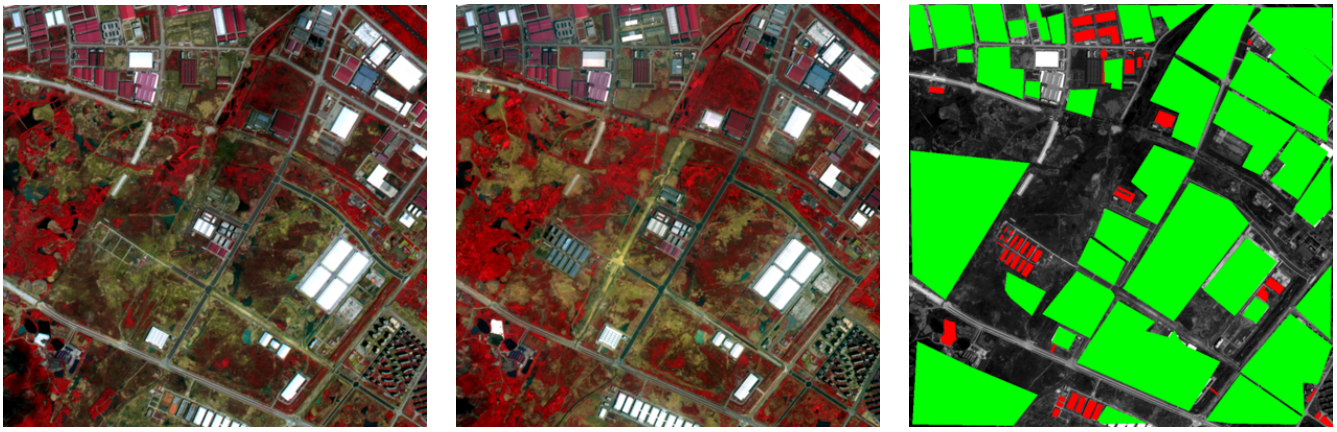
where the first kernel depends on both pixel co-ordinates (denoted as c) and spectral difference intensities (denoted as d). The second smoothness kernel only depends on pixel co-ordinates.

The inference of FC-CRF adopts mean field approximation algorithm (MFA), , which can be significantly accelerated by high dimensional filter algorithm [47].

C. Supervised Change Detection Algorithms

Based on the proposed DSMS-FCN and FC-CRF, the pipeline of the supervised change detection algorithm is shown in Fig. 5. Same as unsupervised architecture, the first step is pre-processing, which can eliminate the difference of geometry and radiometric conditions of multi-temporal VHR images. Then, DSMS-FCN is trained on change detection data sets in an end-to-end manner. The inputs of the DSMS-FCN are a pair of multi-temporal VHR images, and the output is a corresponding change probability map. Different from the proposed DSMS-CN and majority of patch-based models [32], [42], [44], DSMS-FCN can process images of any size and does not require sliding patch-window. Therefore, the change detection accuracy and inference speed could get significantly improved [63].

As mentioned before, skip-connection structure of DSMS-FCN cannot solve the problem of inaccurate localization, thus we make use of FC-CRF to refine the results obtained by DSMS-FCN. The parameters of FC-CRF is trained on training set or validation set and the training process between FC-CRF and DSMS-FCN are decoupled. Firstly, the DI is computed with CVA and the change probability map of multi-temporal VHR images is inferred by DSMS-FCN. Then the unary potential of FC-CRF is generated by change probability map



(a) (b) (c)

Fig. 6: WH data set. (a) Pre-change. (b) Post-change. (c) is ground truth where red means change and green indicates non-change.

(a) (b) (c)

Fig. 7: HY data set. (a) Pre-change. (b) Post-change. (c) is ground truth where red means change and green indicates non-change.

and the pairwise potential of FC-CRF is computed by DI. Based on its fully connected structure, the FC-CRF can efficiently extract accurate localization information by considering short-range and long-range information simultaneously. Consequently, the FC-CRF can further refine the results obtained by DSMS-FCN and eventually obtain a better binary change map with more accurate boundaries. Because of adopting high dimension filter algorithm to accelerate MFA, the inference of FC-CRF is very fast in practice. Therefore, the entire algorithm may take some time during training phase, but the speed of inference is still fast.

IV. Unsupervised Change Detection Experiment

A. Data Set

In the unsupervised change detection experiment, the first VHR data set called as WH was captured by GaoFen-2 (GF-2) sensor on April 4, 2016 and September 1, 2016, covering the city of Wuhan, China. The image size is 1000×1000 with four bands consisting of red, green, blue and near-infrared. Its spatial resolution is 4 m. Fig. 6 shows the pseudo-color images and ground truth of change and non-change. (a) and (b) are the pseudo-color images acquired on April 4, 2016 and September 1, 2016, respectively. (c) is the ground truth. The changed area (red) contains 20026 pixels, and the unchanged area (green) contains 484143 pixels. The remaining pixels are undefined.

The second data set is HY data set with image size of 1000×1000, the two multi-temporal VHR images in this data set were also acquired by GF-2. The images cover the Hanyang city. Fig. 7 shows the pseudo-color images and ground truth. The changed area (red) contains 59051 pixels, and the unchanged area (green) contains 416404 pixels.

It could be observed that in both data sets, the changed area only occupies a small part, thus there exists a heavy skewed-class problem between changed and non-change classes, which brings greater challenge to change detection. In addition, there exists the over-exposed problems on some buildings in VHR images, which break the linear relationship of radiometric intensity of unchanged regions between multi-temporal images and cannot be eliminated by radiometric normalization [4],

[64]. Hence, the over-exposed problem makes accurate change detection more difficult.

B. Experiment Settings

Firstly, the weights and bias of DSMS-CN are initialized by he-normal way [65]. In order to overcome the skew-class problem, weighted binary cross-entropy (WBCE) function is applied as the loss function of DSMS-CN:

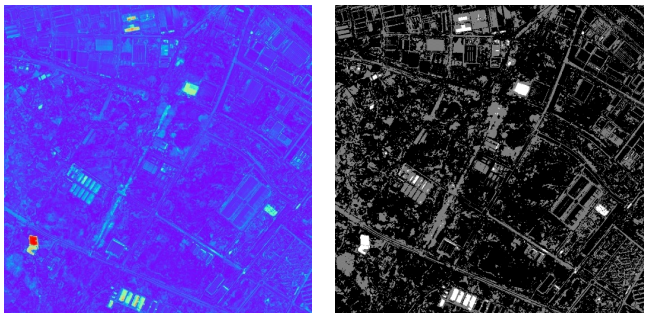
$$L = w_p \hat{y} log y + (1 - \hat{y}) log(1 - y), \tag{5}$$

where w_p is the reciprocal of the proportion of non-change and change classes in training samples. Through setting a larger weight for change class, the change samples would play a more important role at training phase. Adam optimizer [66] is chosen to train the network (learning rate is set to 1e-4). Dropout [67] and weight decay [68] are used to avoid overfitting during training phase. The image patch size for DSMS-CN is set as 13 for both data sets. The specific influence at different values is discussed in section IV-E.

To evaluate our method, nine widely used change detection methods are adopted for comparison, they are summarized as follows.

1) IRMAD [14], which is an iteratively weighted extension of MAD and has shown good performance in multi-temporal image change detection.
2) ISFA [15], which is an unsupervised change detection approach based on slow feature analysis theory.
3) CVA [8], which is one of the most classic unsupervised change detection methods.
4) OBCD [69], an unsupervised change detection method for VHR images, which adopts the object as the change detection basic unit.
5) LSTM [70], a deep learning-based method, has recently shown promising performance in change detection.
6) PCANet [52], which utilizes Gabor wavelets and FCM as the pre-classification method to select training samples, and then trains a PCANet [48] model with the selected image patches.
7) DSFA [58], a deep learning-based change detection algorithm, which works by extracting nonlinear features from multi-temporal images with two-stream DNN and detecting changes with SFA.
8) DCVA [11], an effective unsupervised method for binary and multi-class change detection in VHR images, which adopts a pre-trained CNN extract deep spatial-spectral features from multi-temporal VHR images.
9) DSCN [32], a deep learning-based change detection method, which uses a deep siamese convolutional network to detect changes and performs well in aerial images.

Among these methods, IRMAD, ISFA, CVA, OBCD, and DCVA are unsupervised models without training samples. PCANet and DSFA are pre-classification-based methods. LSTM and DSCN are supervised models, we train them on the same samples as the DSMS-CN. Precision rate, recall rate, overall accuracy (OA), F1 score, and kappa coefficient (KC) are used for accuracy assessment.



(a) (b)

Fig. 8: Pre-classification results on the WH data set. (a) DI. (b) CICM.

C. Experimental Result and Analysis on WH Data Set

According to our unsupervised architecture, the WH data set is first pre-processed. Then, CVA fuses information of four bands and computes DI. As shown in Fig. 8-(a), the areas of warmer tones have a greater change probability. Based on DI, FCM is implemented to generate coarse initial change map (CICM). In Fig. 8-(b), the white pixels belong to change class, the dark pixels belong to non-change class, and gray pixels are candidates to be classified by the proposed method.

All the pixels in the change class are chosen as training samples. However, since spectral difference between multi-temporal images of the pixels in non-change class is relatively constant, we only randomly select a part of pixels in the non-change class as training samples. On the WH data set, the number of selected unchanged pixels is four times to the changed ones. Section IV-E further discusses the impact of the proportion of change and non-change classes in the training sample.

The binary change maps obtained by DSMS-CN and comparison methods are presented in Fig. 9. The change detection results acquired by IRMAD and ISFA are unsatisfactory, it is obvious that a lot of unchanged buildings are detected as nto changes, which indicates ISFA and IRMAD suffer from the over-exposed problem. Compared with IRMAD and ISFA, CVA directly calculating a DI and performing clustering achieves a relatively good result. However, changes of a few buildings roads are not recognized and many pixels in building margins are falsely detected as changed ones. Due to adopting object as the basic unit, there is almost no noise in the result of OBCD. But limited to only exploring low-level spatial-spectral features, lots of building changes are not detected. Fig. 9-(e) shows the result of LSTM, there exist some noises caused by insufficient utilization of spatial context information. The result of PCANet is shown in Fig. 9-(f), through cascade PCA filters, the noises are well suppressed. Extracting nonlinear spatial-spectral features by DNN, most of the changes are well preserved by DSFA as shown in Fig. 9-(g). However, because of detecting changes through SFA, DSFA is inevitably affected by the over-exposed problem. Through comparing the deep spatial-spectral features extracted from input images, DCVA shows a good change detection performance though a part of

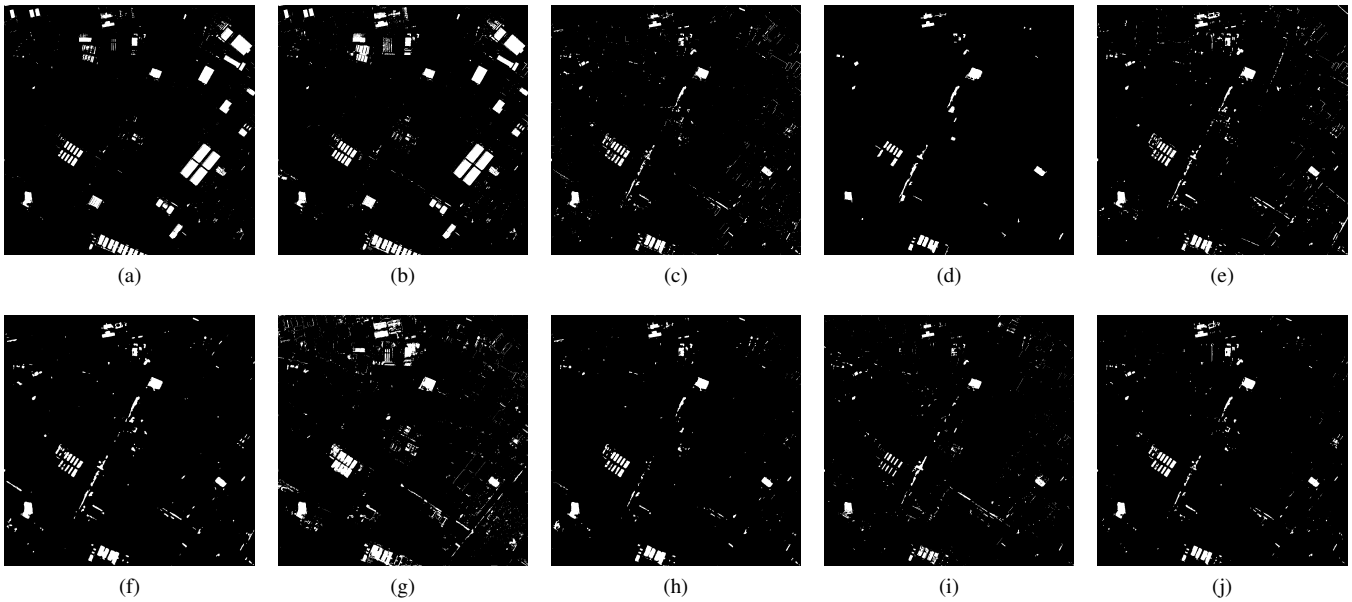


Fig. 9: Change detection results obtained by different methods on the WH data set. (a) IRMAD. (b) ISFA. (c) CVA. (d) OBCD. (e) LSTM. (f) PCANet. (g) DSFA. (h) DCVA. (i) DSCN. (j) DSMS-CN.

TABLE I: ACCURACY ASSESSMENT ON THE BINARY CHANGE MAPS ACQUIRED BY DIFFERENT METHODS ON THE WH DATA SET

Method	Pre.	Rec.	OA	F1	KC
IRMAD	0.3651	0.5708	0.9212	0.3651	0.3289
ISFA	0.2790	0.6408	0.9200	0.3888	0.3530
CVA	0.7682	0.5679	0.9711	0.6530	0.6434
OBCD	**0.9409**	0.4906	0.9785	0.6449	0.6350
LSTM	0.7761	0.6125	0.9782	0.6892	0.6780
PCANet	0.8386	0.6282	0.9805	0.7183	0.7084
DSFA	0.6936	**0.7770**	0.9776	0.7329	0.7212
DCVA	0.8665	0.6196	0.9812	0.7225	0.7131
DSCN	0.7852	0.4484	0.9729	0.5708	0.5564
DSMS-CN	0.8120	0.6844	**0.9813**	**0.7428**	**0.7331**

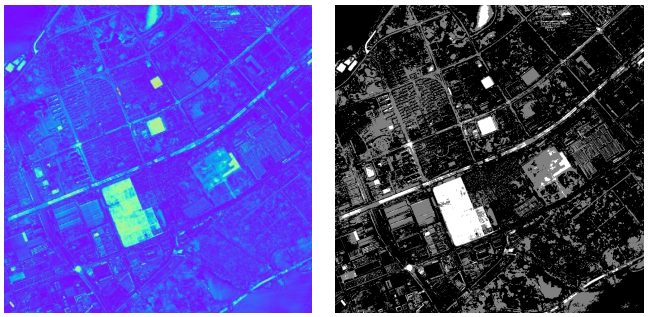


Fig. 10: Pre-classification results on the HY data set. (a) DI. (b) CICM.

changes is misclassified. From Fig. 9-(i), it can be observed that plenty of building changes are not detected by DSCN. Compared with these methods, the binary change map of the proposed DSMS-CN is better in visual, as shown in Fig. 9-(j).

Table I reports the quantitative analysis results based on five evaluation criteria as described in Section IV-B. DSMS-CN achieves the best result with OA of 0.9812, F1 of 0.7428, and KC of 0.7331. It indicates that DSMS-CN can effectively fit the distributions of ground changes in VHR images from pre-classification samples based on the powerful extraction ability of MFCU and deep siamese convolutional structure, and achieve the best performance. By contrast, the performance of SVM and DSCN cannot compete with our approach.

D. Experimental Result and Analysis on HY Data Set

The DI and CICM of the HY data set are shown in Fig. 10. Same as the WH data set, all the pixels in change class are chosen as training samples. But the number of selected unchanged pixels is the same as the changed ones. Section IV-E further discusses the impact of the proportion of change and non-change classes in the training sample.

The qualitative results obtained by the proposed method and comparison methods are shown in Fig. 11. Fig. 11-(a) and -(b) are the results obtained by IRMAD and ISFA. Because of the complex ground situations of VHR images, the results have many falsely detected pixels. Many unchanged buildings are misclassified as changes and the road changes are almost not detected. By contrast, the binary change maps obtained by CVA and OBCD are better in visual. However, CVA misclassifies a part of building margins as change class and OBCD ignores some obvious building and road changes. The result of LSTM is shown in Fig. 11-(e). Through three gates and cell state to capture the temporal dependency, the main changes regions are detected successfully. But it still has plenty of noise. In Fig. 11-(f), compared to LSTM, the binary change map of PCANet is better with less noise. As shown in Fig.

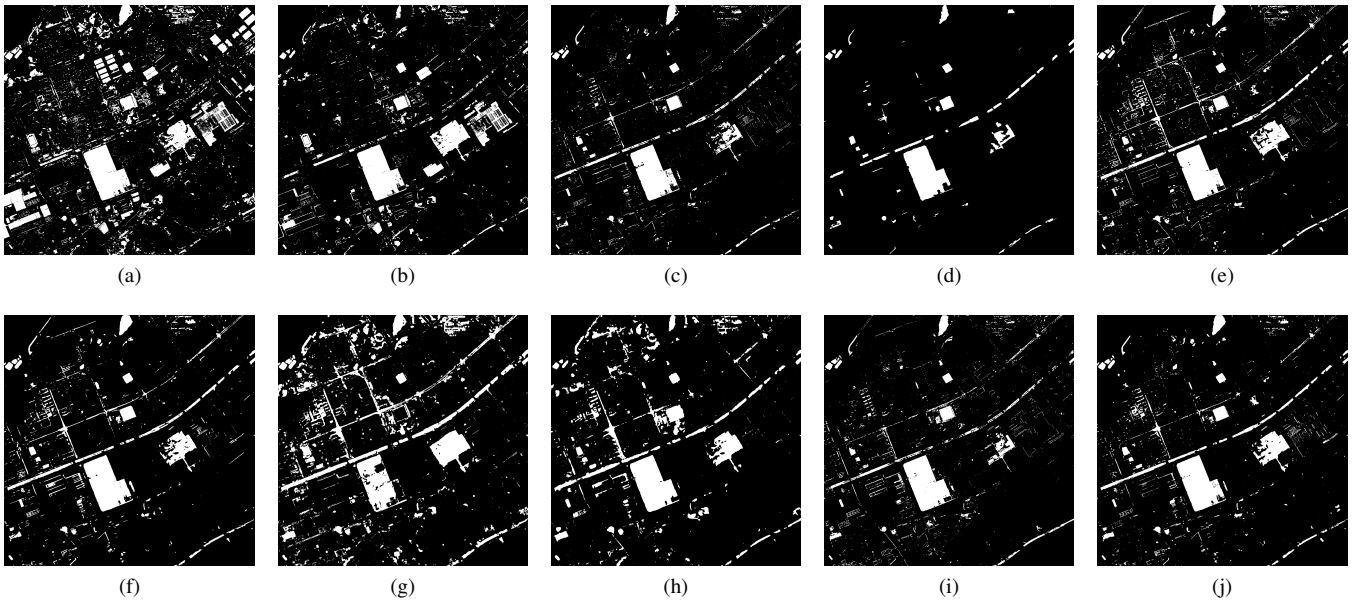


Fig. 11: Change detection results obtained by different methods on the HY data set. (a) IRMAD. (b) ISFA. (c) CVA. (d) OBCD. (e) LSTM. (f) PCANet. (g) DSFA. (h) DCVA. (i) DSCN. (j) DSMS-CN.

TABLE II: ACCURACY ASSESSMENT ON THE BINARY CHANGE MAPS ACQUIRED BY DIFFERENT METHODS ON THE HY DATA SET

Method	Pre.	Rec.	OA	F1	KC
IRMAD	0.4356	0.7110	0.8497	0.5402	0.4565
ISFA	0.6481	0.7120	0.9162	0.6786	0.6305
CVA	0.8579	0.6631	0.9445	0.7480	0.7174
OBCD	**0.9802**	0.6161	0.9507	0.7566	0.7308
LSTM	0.7761	0.7680	0.9438	0.7720	0.7400
PCANet	0.7891	0.7681	0.9458	0.7785	0.7476
DSFA	0.7432	**0.8380**	0.9440	0.7877	0.7556
DCVA	0.7928	0.7708	0.9466	0.7816	0.7512
DSCN	0.7865	0.6428	0.9340	0.7074	0.6706
DSMS-CN	0.8304	0.7736	**0.9523**	**0.8011**	**0.7740**

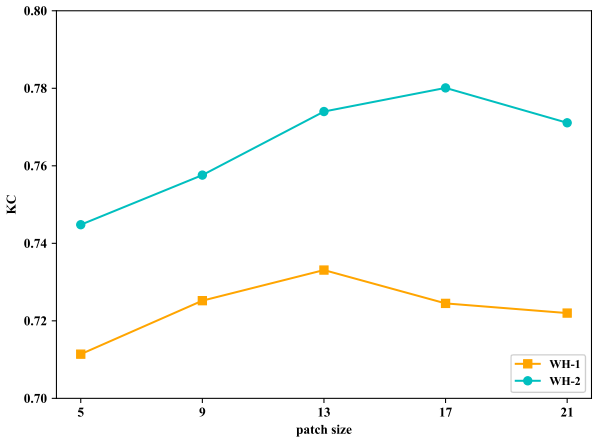


Fig. 12: Relationship between the performance of the DSMS-CN and patch size.

11-(g), the road changes are well detected by DSFA, yet there is some internal fragmentation of changed regions. Fig. 11-(h) shows that utilizing deep spatial-spectral features for change detection, the result generated by DCVA is obviously better than the result of CVA. The result obtained by DSCN is similar to the ones acquired by CVA, but more building margins are misclassified as change class. The change detection result by applying DSMS-CN on the HY data set is shown in Fig. 11-(j). Compared with other methods, DSMS-CN is successful to identify most of the land-cover changes without losing many details and well preserve most of the unchanged regions, which further proves the powerful extraction ability of MFCU and effectiveness of the proposed network. Table II reporting the quantitative analysis results demonstrate it. Clearly, DSMS-CN achieves the highest OA, F1, and KC again.

E. Discussion

For the proposed DSMS-CN, there are two hyper-parameters play an important role, namely input image patch size and proportion of non-change and change classes in training samples. The relationship between the change detection performance of the DSMS-CN and input image patch size is shown in Fig. 12. It can be observed that the performance of DSMS-CN becomes better with the increasing of patch size on both data sets. This is because the larger the patch size is, the more neighborhood information the patch contains, which is beneficial for exploring multi-scale features by DSMS-CN to achieve better performance. However, if patch size is too large, more irrelevant interference information would be considered, resulting in impaired performance. Concretely speaking, the optimal values of patch size on the WH data

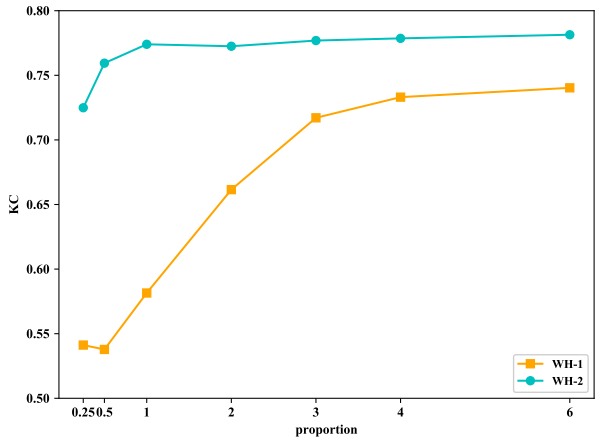


Fig. 13: Relationship between the performance of the DSMS-CN and proportion of non-change and change classes in training samples.

set and HY data set are 13 and 17, respectively. But as the patch size increases, the computational expense would increase dramatically. Therefore, as a trade-off between performance and computational expense, the patch size is set to 13 on both data sets.

Besides, the proportion of non-change and change classes in training samples also affects the performance of the proposed method. The relationship between them over two data sets is shown in Fig. 13. Specifically, on the WH data set, the number of samples in change class generated by pre-classification is not many, so if the proportion of two classes is small, the number of non-change class in the training samples are too few training samples, causing an unsatisfactory change detection result. As the proportion increases, training samples contain enough unchanged land-cover types, hence the performance is improved. When the proportion is larger than 4, the performance becomes stable. Considering the computational expense, the proportion over WH data set is set to 4. The performance tendency on the HY data sets is similar. However, a sufficient number of samples in change class can be generated by pre-classification on the HY data set, hence the performance is acceptable even if the proportion is only 0.25. When the proportion is larger than 1, the performance of the DSMS-CN becomes stable. Thus on the HY data set, the proportion is set to 1.

V. Supervised Change Detection Experiment

A. ACD Data Set

In order to train the proposed network and evaluate our method, the SZTAKI AirChange Benchmark set (ACD) is employed [27], [71]. The data set contains three sets of registered multi-temporal RGB aerial image pairs acquired in different seasonal conditions and associated ground truth. The size of each image is 952×640 and their spatial resolution is 1.5m. The main differences between image pairs are new built-up regions, fresh plough-land and groundwork before building over. As a public data set, ACD has already been used in [27], [32], [43], [71].

B. Experiment Settings

First, to augment available training data, all possible flips and rotations of multiple of 90 degrees are used. For the data set split, we adopt the way proposed in [32] and [43]: the top-left 784x448 corner of the Szada-1 and Tiszadob-3 are cropped for testing, and the rest of the images are used for training. The test image pairs is shown in Fig. 14. Szada and Tiszadob are treated separately into two different data sets, and the images named Archieve are ignored, since it contains only one image pair.

In the supervised algorithm, he-normal way is used to initialize the weights and bias of DSMS-FCN. To overcome the skew-class problem, the loss function of DSMS-FCN applies WBCE. Adam optimizer is chosen to train the network (learning rate is set to 2e-4). Dropout is used to avoid overfitting during the training phase. As proposed in [47], the weights and parameters of the Gaussian kernel of FC-CRF are determined by grid search on training set and the penalty of pairwise potential is trained by L-BGFS algorithm.

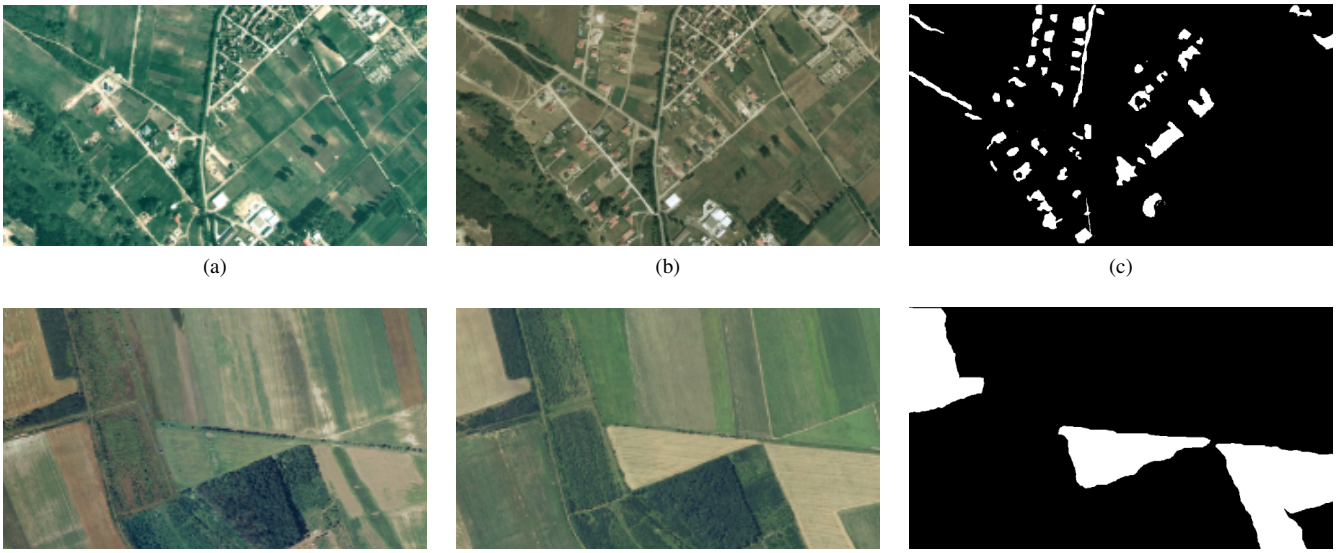
To evaluate the effectiveness of the proposed models, six state-of-the-art methods are used for comparison, including DSCN [32], CXM [27], SCCN [42] and three fully convolutional networks proposed in [43]. The three comparative fully-convolution networks are FC-EF, FC-Siam-Conc, FC-Siam-Diff. The experimental results of the comparison methods use the values obtained in the literature [32] and [43]. For the purpose of better reflecting the role of FC-CRF, we separately evaluate the results obtained by DSMS-FCN and the combination of DSMS-FCN and FC-CRF. In addition, in order to verify the superiority of MFCU compared to conventional single-scale convolution unit, we modify the fully convolution architecture FC-EF proposed in [43] and replace the 3×3 convolution kernel with our MFCU. The modified network is denoted as MSFC-EF. Same as the unsupervised experiment, precision rate, recall rate, OA, F1 score, and KC are employed as evaluation criteria.

C. Experimental Result and Analysis

Fig. 15 and Fig. 16 are illustrations of our results on the ACD data set. We could see that both our DSMS-FCN and improved FC-EF network, namely MSFC-EF, can achieve satisfactory qualitative results. Combined with FC-CRF, more low-level information is considered, thus the results obtained by deep fully convolutional network are refined.

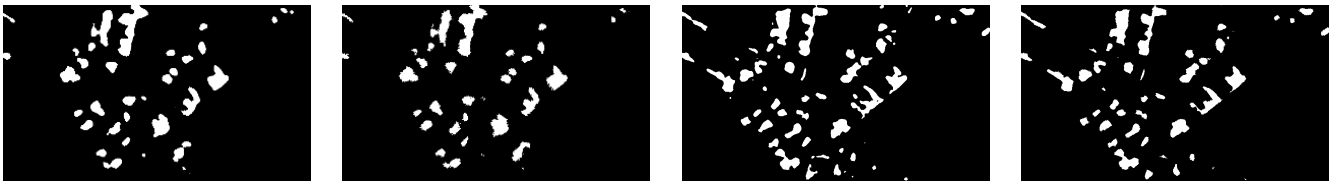
Table III and Table IV report the quality analysis results on two test image pairs. The results obtained on the Szada-1 show the superiority of the proposed network, which obviously outperforms all the other comparison methods in recall metric, F1 score, and OA. Utilizing MFCU, each metric of MSFC-EF is better than FC-EF. Combined with FC-CRF, OA, F1 score and KC of MSFC-EF and DSMS-FCN are further improved. The DSMS-FCN-FCCRF achieves the best recall metric, OA, F1 score and KC.

On the Tisza-3, though the performance of the DSMS-FCN is not the best, it is still superior to DSCN, CXM, SCCN, and other two FCNs. FC-EF achieves a very high F1 score of 0.9340, but MSFC-EF utilizing MFCU still obtain



(a) (b) (c) (d) (e) (f)

Fig. 14: The multi-temporal images selected as the test set from the ACD data set. (a) Pre-change of Szada-1. (b) Post-change of Szada-1. (c) Ground truth. (d) Pre-change of Tiszadob-3. (e) Post-change of Tiszadob-3. (f) Ground truth.



(a) (b) (c) (d)

Fig. 15: Change detection results obtained by different methods on the Szada-1 of ACD. (a) MSFC-EF. (b) MSFC-EF-FCCRF. (c) DSMS-FCN. (d) DSMS-FCN-FCCRF.

(a) (b) (c) (d)

Fig. 16: Change detection results obtained by different methods on the Tiszadob-3 of ACD. (a) is MSFC-EF. (b) is MSFC-EF-FCCRF. (c) is DSMS-FCN. (d) is DSMS-FCN-FCCRF.

improvement of performance. All metric of MSFC-EF is the better than FC-EF. Adopting FC-CRF, OA, F1 score, and KC of MSFC-EF and DSMS-FCN are further improved. The MSFC-EF-FCCRF achieves the best performance.

Depending on FC-CRF, the results obtained by deep FCN can be further improved. It is worth noting that in the two test image-pairs, the performance of training architectures is obviously better than the patch-based method DSCN and other architectures.

The numbers of total trainable parameters in the five FCN architectures are shown in Fig. 17. The number of parameters in the proposed DSMS-FCN is in the smallest quantity. Compared to the FC-Siam-Conc and FC-Siam-Diff, the total number of parameters in DSMS-FCN is reduced by about 46.9% and 38.4%, respectively. However, DSMS-FCN shows obviously better performance on both data sets, which implies that the proposed network has both better change detection ability and smaller computational cost. Besides, the model size of MSFC-EF is reduced by about 14.5% compared to the original network FC-EF. Meanwhile, the MSFC-EF can outperform the original model FC-EF in the data of Szada-1 and Tisza-3. In contrast to the conventional single-scale convolution unit, MFCU has fewer parameters and more powerful feature extraction ability, which can efficiently improve the

TABLE III: ACCURACY ASSESSMENT ON THE BINARY CHANGE MAPS OF DIFFERENT METHODS ON THE SZADA-1 OF ACD DATA SET

Method	Pre.	Rec.	OA	F1	KC
DSCN	0.412	0.574	NA	0.479	NA
CXM	0.365	0.584	NA	0.449	NA
SCCN	0.224	0.347	NA	0.287	NA
FC-EF	0.4357	0.6265	0.9308	0.5140	NA
FC-Siam-Conc	0.4093	0.6561	0.9246	0.5041	NA
FC-Siam-Diff	0.4138	**0.7238**	0.9240	0.5266	NA
MSFC-EF	0.4725	0.6237	0.9374	0.5377	0.5048
MSFC-EF-FCCRF	0.4888	0.6014	0.9400	0.5393	0.5076
DSMS-FCN	0.4835	0.6753	0.9392	0.5635	0.5317
DSMS-FCN-FCCRF	**0.5278**	0.6339	**0.9457**	**0.5772**	**0.5484**

TABLE IV: ACCURACY ASSESSMENT ON THE BINARY CHANGE MAPS OF DIFFERENT METHODS ON THE Tiszadob-3 OF ACD DATA SET

Method	Pre.	Rec.	OA	F1	KC
DSCN	0.883	0.851	NA	0.867	NA
CXM	0.617	0.934	NA	0.743	NA
SCCN	0.927	0.798	NA	0.858	NA
FC-EF	0.9028	0.9674	0.9766	0.9340	NA
FC-Siam-Conc	0.7207	0.9687	0.9304	0.8265	NA
FC-Siam-Diff	0.6561	0.8829	0.9137	0.7778	NA
MSFC-EF	0.9489	0.9763	0.9870	0.9624	0.9545
MSFC-EF-FCCRF	**0.9514**	**0.9765**	**0.9874**	**0.9638**	**0.9560**
DSMS-FCN	0.8691	0.8690	0.9552	0.8690	0.8420
DSMS-FCN-FCCRF	0.8918	0.8856	0.9620	0.8886	0.8658

performance of the deep network.

Furthermore, the inference time of DSMS-FCN is below 0.1s per image and the inference time of FC-CRF is under 1s per image, which means our method can efficiently process multi-temporal VHR images in real-time.

VI. Conclusion

This paper utilizes a powerful multi-scale feature convolution unit for change detection in VHR images. Differing from conventional single-scale convolution unit that only extracts single-scale features in one layer, MFCU is capable of extracting multi-scale spatial-spectral features in the same layer by four ways. Specifically, the 1×1 convolution kernel focus on extracting the features of a pixel itself. The 3×3 convolution kernel can extract spatial-spectral features. The 5×5 convolution kernel extracts larger-scale features. The 3×3 max pooling is able to extract the most salient features and alleviate the smoothing effect of the convolution operation.

Based on MFCU, two novel deep siamese convolutional neural networks are designed for unsupervised and supervised change detection. In unsupervised change detection, DSMS-CN is trained on the samples generated by automatic pre-classification. In supervised change detection, DSMS-FCN is capable of processing images of any size and does not require sliding patch-window, thus the accuracy and inference speed could be significantly improved. To overcome the inaccurate localization problem, the FC-CRF is adopted to refine the results obtained by DSMS-FCN. Through using the output of DSMS-FCN as unary potential, the FC-CRF is combined with DSMS-FCN.

On the unsupervised change detection experiments with the two challenging data sets, the experimental results indicate that DSMS-CN outperforms the other comparison methods with better F1 score, OA, and KC, which confirms the powerful extraction ability of MFCU and outstanding fitting capability of DSMS-CN. In the supervised change detection experiments with ACD data set, compared with three FCNs and other state-of-the-art methods, DSMS-FCN delivers a competitive performance. And MFCU can be treated as a general module and embedded into deep CNN by replacing the original single-scale convolution unit to improve the performance. What's more, FC-CRF does make the results obtained by deep FCNs more accurate. Lastly, the inference time of our supervised architecture is below 1s per image, thus it can predict change maps of multi-temporal VHR images in real-time.

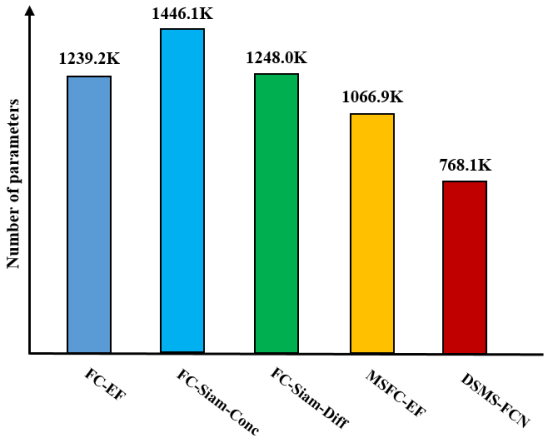


Fig. 17: Comparison of the five FCN architectures in terms of model size.

Our future work will focus on applying deep siamese architecture for change detection in heterogeneous VHR images and utilizing FC-CRF to improve the performance of traditional change detection methods in VHR images.

References

[1] A. Singh, "Review Articlel: Digital change detection techniques using remotely-sensed data," *International Journal of Remote Sensing*, vol. 10, no. 6, pp. 989–1003, 1989.

[2] G. Xian, C. Homer, and J. Fry, "Updating the 2001 National Land Cover Database land cover classification to 2006 by using Landsat imagery change detection methods," *Remote Sensing of Environment*, vol. 113, no. 6, pp. 1133–1147, 2009.

[3] P. Coppin, I. Jonckheere, K. Nackaerts, B. Muys, and E. Lambin, "Digital change detection methods in ecosystem monitoring: A review," *International Journal of Remote Sensing*, vol. 25, no. 9, pp. 1565–1596, 2004.

[4] H. Luo, C. Liu, C. Wu, and X. Guo, "Urban change detection based on Dempster-Shafer theory for multitemporal very high-resolution imagery," *Remote Sensing*, vol. 10, no. 7, pp. 20–22, 2018.

[5] D. Lu, P. Mausel, E. Brondízio, and E. Moran, "Change detection techniques," *International Journal of Remote Sensing*, vol. 25, no. 12, pp. 2365–2407, 2004.

[6] D. Brunner, G. Lemoine, and L. Bruzzone, "Earthquake damage assessment of buildings using VHR optical and SAR imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 5, pp. 2403–2420, 2010.

[7] M. E. Zelinski, J. Henderson, and M. Smith, "Use of Landsat 5 for change detection at 1998 Indian and Pakistani nuclear test sites," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 8, pp. 3453–3460, 2014.

[8] L. Bruzzone and Diego Fernàndez 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.

[9] F. Bovolo, S. Marchesi, L. Bruzzone, S. Member, 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.

[10] 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.

[11] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised deep change vector analysis for multiple-change detection in VHR Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 6, pp. 3677–3693, 2019.

[12] J. S. Deng, K. Wang, Y. H. Deng, G. J. Qi, K. Wang, Y. H. Deng, and G. J. Q. Pca, "PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data," *International Journal of Remote Sensing*, vol. 1161, 2008.

[13] 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 Sensing of Environment*, vol. 64, pp. 1–19, 1998.

[14] A. A. Nielsen, "The regularized iteratively reweighted MAD method for change detection in multi- and hyperspectral data," *IEEE Transactions on Image Processing*, vol. 16, no. 2, pp. 463–478, 2007.

[15] C. Wu, B. Du, and L. Zhang, "Slow feature analysis for change detection in multispectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 5, pp. 2858–2874, 2014.

[16] C. Wu, L. Zhang, and B. Du, "Kernel Slow Feature Analysis for Scene Change Detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 4, pp. 2367–2384, 2017.

[17] Y. Tang, L. Zhang, and X. Huang, "Object-oriented change detection based on the Kolmogorov Smirnov test using high-resolution multispectral imagery," *International Journal of Remote Sensing*, vol. 32, no. 20, pp. 5719–5740, 2011.

[18] D. Wen, X. Huang, L. Zhang, and J. A. Benediktsson, "A novel automatic change detection method for urban high-resolution remotely sensed imagery based on multiindex scene representation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 1, pp. 609–625, Jan 2016.

[19] Z. Lei, T. Fang, H. Huo, and D. Li, "Bi-Temporal texton forest for land cover transition detection on remotely sensed imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 2, pp. 1227–1237, 2014.

[20] 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," *arXiv preprint arXiv:1912.08628*, 2019.

[21] H. Chen, C. Wu, B. Du, and L. Zhang, "Dsdanet: Deep siamese domain adaptation convolutional neural network for cross-domain change detection," *arXiv preprint arXiv:2006.09225*, 2020.

[22] ——, "Deep Siamese Multi-scale Convolutional Network for Change Detection in Multi-Temporal VHR Images," in *2019 10th International Workshop on the Analysis of Multitemporal Remote Sensing Images, MultiTemp 2019*, 2019.

[23] K. Tan, X. Jin, A. Plaza, X. Wang, L. Xiao, and P. Du, "Automatic Change Detection in High-Resolution Remote Sensing Images by Using a Multiple Classifier System and Spectral-Spatial Features," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 8, pp. 3439–3451, 2016.

[24] 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 Journal of Photogrammetry and Remote Sensing*, vol. 80, pp. 91–106, 2013.

[25] 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.

[26] C. Huo, Z. Zhou, C. P. Hanqing Lu, and K. Chen, "Fast Object-Level Change Detection for VHR Images," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 1, pp. 118–122, 2010.

[27] C. Benedek and T. Sziranyi, "Change detection in optical aerial images by a multilayer conditional mixed markov model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 10, pp. 3416–3430, Oct 2009.

[28] G. Moser, E. Angiati, S. Member, and S. B. Serpico, "Multiscale Unsupervised Change Detection on Optical Images by Markov Random Fields and Wavelets," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 4, pp. 725–729, 2011.

[29] T. Hoberg, F. Rottensteiner, R. Q. Feitosa, and C. Heipke, "Conditional Random Fields for Multitemporal and Multiscale Classification of Optical Satellite Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 2, pp. 659–673, 2015.

[30] L. Zhou, G. Cao, Y. Li, and Y. Shang, "Change Detection Based on Conditional Random Field With Region Connection Constraints in High-Resolution Remote Sensing Images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 8, pp. 3478–3488, 2016.

[31] P. Lv, Y. Zhong, J. Zhao, and L. Zhang, "Unsupervised Change Detection Based on Hybrid Conditional Random Field Model for High Spatial," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 7, pp. 4002–4015, 2018.

[32] 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.

[33] S. Saha, F. Bovolo, and L. Bruzzone, "Destroyed-buildings detection from VHR SAR images using deep features," in *Image and Signal Processing for Remote Sensing XXIV*, L. Bruzzone and F. Bovolo, Eds., vol. 10789, International Society for Optics and Photonics. SPIE, 2018, pp. 336 – 344. [Online]. Available: https://doi.org/10.1117/12.2325149

[34] S. Saha, L. Mou, X. X. Zhu, F. Bovolo, and L. Bruzzone, "Semisupervised change detection using graph convolutional network," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2020.

[35] S. Saha, Y. T. Solano-Correa, F. Bovolo, and L. Bruzzone, "Unsupervised deep transfer learning-based change detection for hr multispectral images," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2020.

[36] C. Sutton, A. McCallum *et al.*, "An introduction to conditional random fields," *Foundations and Trends® in Machine Learning*, vol. 4, no. 4, pp. 267–373, 2012.

[37] S. Z. Li, "Markov random field models in computer vision," in *European conference on computer vision*. Springer, 1994, pp. 361–370.

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

[39] 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.

[40] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep Learning in Remote Sensing:A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.

[41] Y. LECUN, L. BOTTOU, Y. BENGIO, and P. HAFFNER, "Gradient-Based Learning Applied to Document Recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[42] J. Liu, M. Gong, K. Qin, and P. Zhang, "A Deep Convolutional Coupling Network for Change Detection Based on Heterogeneous Optical and Radar Images," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 3, pp. 545–559, 2018.

[43] R. Caye Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proceedings - International Conference on Image Processing, ICIP*, 2018, pp. 4063–4067.

[44] 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 Transactions on Geoscience and Remote Sensing*, vol. 58, no. 4, pp. 2848–2864, 2020.

[45] X. Niu, M. Gong, T. Zhan, and Y. Yang, "A Conditional Adversarial Network for Change Detection in Heterogeneous Images," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 1, pp. 45–49, 2019.

[46] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 1–9.

[47] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with Gaussian edge potentials," in *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011*, no. 4, 2011, pp. 1–9.

[48] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013.

[49] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," pp. 1–14, 2014.

[50] J. Bromley, I. Guyon, Y. Lecun, E. Sckinger, and R. Shah, "Signature verification using a siamese time delay neural network." vol. 7, 01 1993, pp. 737–744.
[51] B. Zitová and J. Flusser, "Image registration methods: A survey," *Image and Vision Computing*, vol. 21, no. 11, pp. 977–1000, 2003.
[52] 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.
[53] M. Gong, H. Yang, and P. Zhang, "Feature learning and change feature classification based on deep learning for ternary change detection in SAR images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 129, pp. 212–225, 2017.
[54] M. Gong, T. Zhan, P. Zhang, and Q. Miao, "Superpixel-based difference representation learning for change detection in multispectral remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 5, pp. 2658–2673, 2017.
[55] M. Li, M. Li, P. Zhang, Y. Wu, W. Song, and L. An, "SAR Image Change Detection Using PCANet Guided by Saliency Detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 3, pp. 402–406, 2019.
[56] J. Geng, X. Ma, X. Zhou, and H. Wang, "Saliency-Guided Deep Neural Networks for SAR Image Change Detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 10, pp. 1–13, 2019.
[57] Y. Li, C. Peng, Y. Chen, L. Jiao, L. Zhou, and R. Shang, "A Deep Learning Method for Change Detection in Synthetic Aperture Radar Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 5751–5763, 2019.
[58] B. Du, L. Ru, C. Wu, and L. Zhang, "Unsupervised Deep Slow Feature Analysis for Change Detection in Multi-Temporal Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 12, pp. 9976–9992, 2019.
[59] F. Liu, L. Jiao, X. Tang, S. Yang, W. Ma, and B. Hou, "Local Restricted Convolutional Neural Network for Change Detection in Polarimetric SAR Images," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 3, pp. 818–833, 2019.
[60] A. Song and Y. Kim, "Transfer Change Rules from Recurrent Fully Convolutional Networks for Hyperspectral Unmanned Aerial Vehicle Images without Ground Truth Data," *Remote Sensing*, vol. 12, no. 1099, pp. 1–20, 2020.
[61] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9351, pp. 234–241, 2015.
[62] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.
[63] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017.
[64] N. Dhingra, A. Nandal, M. Manchanda, and D. Gambhir, "Fusion of fuzzy enhanced overexposed and underexposed images," *Procedia Computer Science*, vol. 54, pp. 738 – 745, 2015.
[65] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034.
[66] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," 2014, pp. 1–15.
[67] S. Nitish, H. Geoffrey, K. Alex, S. Ilya, and S. Ruslan, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.
[68] J. Moody, S. Hanson, A. Krogh, and J. Hertz, "A simple weight decay can improve generalization," *Adv. Neural Inf. Process. Syst*, vol. 4, pp. 950–957, 01 1995.
[69] B. Desclée, P. Bogaert, and P. Defourny, "Forest change detection by statistical object-based method," *Remote Sensing of Environment*, vol. 102, no. 1-2, pp. 1–11, 2006.
[70] 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, pp. 1–22, 2016.
[71] C. Benedek and T. Sziranyi, "A mixed markov model for change detection in aerial photos with large time differences," in *2008 19th International Conference on Pattern Recognition*, Dec 2008, pp. 1–4.