Deep Siamese Multi-scale Convolutional Network for Change Detection in Multi-temporal VHR Images

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

Abstract—Very high resolution (VHR) images provide abundant ground details and spatial distribution 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, we propose a powerful multi-scale feature convolution unit (MFCU) for change detection in VHR images. The proposed unit is able to extract multi-scale features in the same layer. Based on the proposed unit, two novel deep Siamese convolutional networks, 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 in multi-temporal VHR images. For unsupervised change detection, we implement automatic pre-classification to obtain training patch samples, and the DSMS-CN fits the statistical distribution of changed and unchanged area from patch samples through multi-scale feature extraction module 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 output 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, multi-scale feature convolution, deep Siamese convolutional network, fully connected conditional random field, multi-temporal images, very high resolution images

I. INTRODUCTION

Change detection (CD) is the process of identifying differences in the state of an object or phenomenon by observing it at different time [1]. And change detection is playing a vital role in land-use and land-cover (LULC) change analysis, forest or vegetation change monitoring, ecosystem monitoring, urban expansion research, resource management and damage assessment [2]–[7].

Multi-spectral images are the most commonly used data source for change detection, and a number of methods were proposed based on it. Change vector analysis (CVA) generates a difference image (DI) and clusters DI to obtain the change result [8]. Principal component analysis (PCA) [9], as a dimension reduction methods, transforms the images into a new feature space and select a part of new bands for change detection. In [10] and [11], multivariate alteration detection (MAD) and its iterative version IRMAD are proposed, which extract the changed objects by maximizing the difference between the projection features. Based on slow feature analysis (SFA) algorithm, Wu et al. [12], [13] proposed a SFA change detection method. The method aims to find the most invariant component in multi-temporal images, in which the changed object would be highlighted.

Nowadays, with the development of satellite sensors, very-high-resolution (VHR) images are more available by various types of sensors, such as QuickBird, IKONOS, SPOT and Worldview. And VHR images can provide abundant ground details and spatial distribution information, which have crucial effects on the research of urban change analysis and building detection. Nevertheless, due to internal complexity caused by the extremely high resolution of VHR images, conventional multi-spectral change detection methods may not be applicable. Recently, deep learning (DL) has achieved significant performances in many domains [14], as well as remote sensing image interpretation [15]. And convolutional neural network (CNN) [16], as a classical and powerful DL architecture, is expert in extracting multi-level information, which is suitable for extracting spectral information and spatial context information in VHR images. In [17], a deep symmetric network is proposed for change detection of VHR heterogeneous images, and a convolutional layer plays an important role in feature extraction. Yang et al. [18] introduce a deep Siamese convolutional network for aerial image change detection, which extract features by two weight sharing convolutional branches, and generate binary change map based on the feature difference of the last layer. Three fully convolutional Siamese networks are first proposed in [19], which are trained end-to-end on change detection data set and have achieved good performances. Except CNN architecture, in [20], a change detection method based on conditional generative adversarial network (cGAN) is designed. In this method, convolutional layer is responsible for extracting features from VHR image patches. All of these
methods adopt 3x3 convolution kernel as feature extraction module. Though 3x3 convolution kernel could extract spectral features and spatial context features in some extents, it still has some powerlessness in the complex ground situations of VHR images. No research has attempted to use other sizes of convolution kernels or even multiple kernels for change detection in VHR image.

Inspired by network in network structure [21] and Inception network [22], we propose a multi-scale feature convolution unit extracting multi-scale spectral features and spatial context features in the same layer, which is suitable for VHR images. Adopting the MFCU as the basic feature extraction module, two powerful deep Siamese convolutional network are designed for unsupervised and supervised change detection. For supervised change detection, a probability graph model, fully connected conditional random field (FC-CRF) [23] is adopted to refine the change detection results. Further, based on the two deep Siamese multi-scale feature convolutional network architectures and FC-CRF, the specific supervised and unsupervised algorithms are developed.

The contributions of this paper can be summarized as follows:

1) This paper presents a multi-scale feature convolution unit for coping with internal complexity caused by the extremely high resolution of VHR images, which is able to extract multi-scale features from 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) Based on multi-scale feature convolution unit, two novel deep Siamese convolutional network are designed 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 do not require sliding patch-window, therefore the accuracy and speed of inference could be significantly improved.

3) For the purpose of solving the problem of inaccurate localization caused by deep convolution architecture, fully connected conditional random field is adopted to refine the results obtained by deep Siamese fully convolutional network.

The rest of this paper is organized as follows. Section II describes the multi-scale feature convolution unit and two change detection network architectures in detail. Based on Section 2, two change detection algorithms based on the proposed network architectures are addressed in section III for unsupervised and supervised change detection. To evaluate our methods, Section IV and Section V contain quantitative and qualitative comparisons with the state-of-the-art change detection methods. In the end, Section VI draws the conclusion of our work in this paper.

II. METHODOLOGY

In this section, we propose a multi-scale feature convolution unit, which is suitable for VHR images and can extract features at multiple scales, to replace conventional single convolution unit. Based on the proposed unit, two deep Siamese convolutional network architectures are designed for unsupervised and supervised change detection of VHR images. In order to solve the problem of inaccurate localization of DCNN, we adopt the fully connected conditional random field to refine the results.

A. Multi-scale Feature Convolution Unit

The VHR images provide abundant ground details and spatial distribution information. And 3x3 convolution kernel, as a commonly used convolution unit, can extract spectral features and spatial context information of VHR images. However, the 3x3 convolution kernel has two obvious drawbacks in processing VHR images. First, the 3x3 convolution kernel has a limited receptive field and could only extracts single scale features in a same layer. But in VHR images, except normal scale features, there simultaneously exist some large-scale continuous features and a few very small salient features. 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 3x3 convolution kernel, or say, the conventional single convolution unit is somewhat incompetent in dealing with complex multi-scale ground conditions in VHR images.

For the purpose of better extracting multi-scale features from VHR images, the multi-scale feature convolution unit (MFCU) is proposed. As shown in Figure 1, the MFCU is a network in network structure [21], and it extracts multi-scale features in parallel by four ways, namely 1x1 convolution kernel, 3x3 convolution kernel, 5x5 convolution kernel and 3x3 max pooling. The 1x1 convolution kernel focuses on extracting the features of a pixel itself. The 3x3 convolution kernel extracts the features in a neighborhood. The 5x5 convolution kernel extracts features in a larger range, which is suitable for some large-scale continuous objects in VHR images. And max pooling is responsible for extracting the most salient features and efficiently avoids the smoothing effect of convolution
operation. At last, the four type features are fused to obtain the higher dimensional multi-scale features. It should be noted that the 1x1 convolution kernel before 3x3 convolution kernel and 5x5 convolution kernel is a bottleneck design [24], which can efficiently reduce the parameters of network and make network easier to train.

Compared with conventional single convolution unit, the MFCU extracts multi-scale features, which improves the feature abstraction ability of network and does not significantly increase parameters of network to some extent. Using MFCU as multi-scale feature extraction module, we design two novel deep Siamese multi-scale convolutional network architectures for unsupervised and supervised change detection in multi-temporal VHR images, respectively.

B. Deep Siamese Multi-scale Convolutional Network

The first proposed network architecture is deep Siamese multi-scale convolutional network (DSMS-CN). The DSMS-CN (Figure 2) has two components: feature extraction network (FEN) and change judging network (CJN). The FEN is a Siamese network [25] and its two branches extract features from two multi-temporal image patches using exactly the same way because of weight sharing. The former two conventional convolutional modules in each branch transform the spatial and spectral information into relative high dimensional features, and the two latter MFCU modules extract abundant multi-scale features from high dimensional features. Through calculating the absolute differences of the features from two branches, the change information is highlighted and unchanged information are eliminated. Then, the absolute differences of features from multiple-layer are fused and input into the CJN. In the CJN, a MFCU layer extracts multi-scale difference feature, and a global average pooling (GAP) layer replace fully connected layer to generate feature vector, which can make network more robust and reduce overfitting [21]. Finally, the probability of change in the center pixel of image patches is obtained by a fully connected layer.

In DSMS-CN, both conventional and multi-scale convolutional layers adopt rectified linear unit (ReLU) [26] as activation function, the last fully connected layer adopts sigmoid function to predict probability of change. Because of extracting features from multi-temporal image patches, DSMS-CN does not contain any max-pooling layer to reduce dimension.

C. Deep Siamese Multi-scale Fully Convolutional Network

The second proposed network architecture is deep Siamese multi-scale fully convolutional network (DSMS-FCN). Same as the general FCN [27], the DSMS-FCN (Figure 3) consists of two parts: the left is an encoder network and another side is a decoder network.

The encoder network has two equal weight sharing branches, and the features of multi-temporal images are extracted in a same approach. Each branch has four subsampling layers, where each subsampling layer consists of a convolution module and a 2x2 max-pooling layer. The former two convolution modules consist of 3x3 conventional single convolution units and the latter two modules consist of MFCUs. Based on the concept of skip-connection in the U-Net [28], the features of subsampling layer and upsampling layer at the same scale are concatenated during upsampling phase, which can recover localization information and produce more accurate binary changed map with precise boundaries. The motivation for concatenating the absolute value of the difference between two branches with the features of upsampling layer is that change detection is trying to detect the differences between multi-temporal images. Furthermore, it is worth pointing out that only one branches features enter into decoder network. This is because the two branches share weights and changes in multi-temporal VHR image are only in the minority, thus most of features extracted from both sides are same. Meanwhile, difference of features is delivered to decoder network through the skip-connection structure. Consequently, using features of only...
Fig. 3: Illustration of the proposed deep Siamese multi-scale fully convolutional network. It has an encoder network and a decoder network. The encoder network extracts high dimensional robust features from whole multi-temporal images and the decoder network generates change maps based on feature difference between multi-temporal VHR images and high dimensional features of one branch. Color legend: green means conventional 3x3 convolution module, blue means multi-scale feature convolution module, red means 2x2 max-pooling layer and purple means transpose convolution module.

one branch could avoid feature redundancy and overfitting.

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

D. Fully Connected Conditional Random Field

Though DSMS-FCN adopts skip-connection structure to deliver localization information, it still cannot solve the problem of inaccurate localization caused by invariance of features and large receptive field [29]. And fully connected conditional random field (FC-CRF), as a variation of conditional random field (CRF), is remarkably successful in processing the localization problem. Compared with CRF, FC-CRF considers short-range and long-range information simultaneously, thus it can better recover the local structure. Therefore, FC-CRF is adopted to refine the localization information of results obtained by DSMS-FCN.

FC-CRF is a conditional probability distribution model that outputs another set of random variables given a set of input random variables. According to [23], the FC-CRF is designed as follows:

$$P(Y|X) = \frac{1}{Z(X)} \exp\left( \sum_{c \in C_G} \Phi_c(Y_c|X) \right)$$

(1)

It is characterized by a Gibbs distribution, where $G$ is a complete graph on $Y$ and $C_G$ is the set of all unary and pairwise cliques, which induces a potential $\Phi_c$. The corresponding Gibbs energy of FC-CRF is

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

(2)

where $i$ and $j$ range from 1 to $N$. 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

$$\phi_u(y_i) = -\log P(y_i)$$

(3)

where $P(y_i)$ is probability of change at pixel $i$, which is computed by DSMS-FCN. And in FC-CRF model, each pixel...
Fig. 4: Flowchart of our unsupervised change detection architecture. 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, the DSMS-CN is trained on image patches and predicts the patches that need to be classified to obtain whole binary change map.

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:

\[
\begin{align*}
\phi_{p}(y_i, y_j) &= \mu(y_i, y_j)k(f_i, f_j) \\
k(f_i, f_j) &= \sum_{m=1}^{K} w^{(m)}k^{(m)}(f_i, f_j)
\end{align*}
\]

where \( \mu \) 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 \( w^{(m)} \).

\( f_i \) and \( f_j \) are feature vectors for pixels i and j in an feature space.

In change detection problem, the kernels are

\[
k(f_i, f_j) = w_1 \exp\left(-\frac{||c_i - c_j||^2}{2\sigma^2_{\alpha}} - \frac{||X_i - X_j||^2}{2\sigma^2_{\beta}}\right) + w_2 \exp\left(-\frac{||c_i - c_j||^2}{2\sigma^2_{\alpha}}\right)
\]

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

As proposed in [23], the inference of FC-CRF adopts mean field approximation algorithm (MFA) which approximates \( P \) through the product of a set of marginal distribution \( \prod_{i} G_i \). And the message passing step of MFA can be significantly accelerated by high dimensional filter algorithm which reduces the complexity of message passing from quadratic to linear, resulting in an efficient approximate inference algorithm for FC-CRF. The specific combination of FC-CRF and DSMS-FCN is presented in section 3-B.

III. EMPLOYMENT OF CHANGE DETECTION ALGORITHMS

Based on the two aforementioned deep Siamese multi-scale convolutional network architectures and FC-CRF, in this section, we formally propose the detailed employments of unsupervised and supervised change detection algorithms.

A. Unsupervised Change Detection

In our proposed unsupervised change detection method, the DSMS-CN is adopted to predict change result of multi-temporal VHR images. And the flowchart of our unsupervised change detection algorithm is shown in Figure 4.

The first step of our architecture is image pre-processing, consisting of co-registration and radiometric correction. Image registration is the process of aligning two or more images
of the same scene obtained at different times [30]. Through collecting matched point-pairs, establishing transform model and transforming images, the two given multi-temporal VHR images are geometrically aligned. Then radiometric relative normalization is adopted to eliminate the radiometric difference between multi-temporal VHR images caused by different sun angle, light intensity and atmospheric conditions. The specific implementation of radiometric relative normalization is normalizing the two 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 find pixels which have extremely high changed or unchanged probabilities. CVA is first adopted to generate difference map (DM) of multi-temporal images. Then, fuzzy c-means clustering (FCM), based on the memberships of pixels, is implemented to partition DM into three clusters: \( w_c \), \( w_{uc} \) and \( w_{tbc} \). The pixels belonging to \( w_c \) and \( w_{uc} \) are reliable pixels which have high probabilities to be changed or unchanged. And the pixels in \( w_{tbc} \) are uncertain and need to be classified. Finally, the \( w \times w \) neighborhood area of pixels in \( w_c \) and \( w_{uc} \) are chosen as training image patches.

Eventually, the DSMS-CN is trained on chosen training image patches. After training process is completed, the pixels in \( w_{tbc} \) are categorized by the DSMS-CN and the whole changed map is generated. Owing to sigmoid function of fully connected layer of DSMS-CN, the threshold segmentation step could be simplified and just adopt 0.5 as threshold to get binary change map. We could see that the whole process is completely unsupervised.

### B. Supervised Change Detection

The flowchart of proposed supervised change detection algorithm is shown in Figure 5. In supervised algorithm, the fully convolutional architecture, DSMS-FCN, is adopted to predict change maps of multi-temporal VHR images. 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, the DSMS-FCN is directly trained end-to-end on change detection data set without any pre-training or transfer learning. The inputs of the DSMS-FCN are two whole multi-temporal VHR images, and the output is a change map. Unlike our unsupervised architecture and majority of recent patch-based approach [18], the DSMS-FCN is able to process images of any size and do not require sliding patch-window, therefore the accuracy and speed of inference could be significantly improved.

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 DM is computed by CVA and the change probability map of multi-temporal VHR images are inferred by DSMS-FCN. Then the unary potential of FC-CRF is generated by change probability map and the pairwise potential of FC-CRF is computed by DM. Based on its fully connected structure, the FC-CRF can efficiently extract accurate localization information by considering short-range and long-range information simultaneously. Thus 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 EXPERIMENTS

A. Data set

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

The second VHR data set was also acquired by GF-2 with the resolution of 4m, denoted as HY. The images cover the Hanyang area of Wuhan city. The size of multi-temporal images are 1000 x 1000 pixels with four bands. Figure 7 shows the pseudo-color images and ground truth of change and unchanged. (a) and (b) are the pseudo-color images acquired on April 4, 2016 (denoted as HY-1) and September 1, 2016 (denoted as HY-2), respectively. (c) is the ground truth. The changed area (red) contains 20026 pixels, and unchanged area (green) contains 484143 pixels. The remaining pixels are undefined.

IV. UNSUPERVISED CHANGE DETECTION EXPERIMENTS

A. Data set

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

The second VHR data set was also acquired by GF-2 with the resolution of 4m, denoted as HY. The images cover the Hanyang area of Wuhan city. The size of multi-temporal images are 1000 x 1000 pixels with four bands. Figure 7 shows the pseudo-color images and ground truth of change and unchanged. (a) and (b) are the pseudo-color images acquired on April 4, 2016 (denoted as HY-1) and September 1, 2016 (denoted as HY-2), respectively. (c) is the ground truth. The changed area (red) contains 20026 pixels, and unchanged area (green) contains 484143 pixels. The remaining pixels are undefined.
Fig. 9: Visualized results obtained by proposed method and comparison methods on WH data set. (a) MAD. (b) IRMAD. (c) SFA. (d) ISFA. (e) CVA. (f) PCA-Kmeans. (g) DSCN. (h) DSMS-CN. The DSMS-CN achieves the best visual results, followed by CVA and DSCN.

Table I: Accuracy Assessment On Change Detection Results Of Different Methods On WH Data Set

| Method     | Rec. | Pre. | OA   | F1   | Kappa |
|------------|------|------|------|------|-------|
| MAD        | 0.8968 | 0.1905 | 0.8446 | 0.3143 | 0.2662 |
| IRMAD      | 0.5708 | 0.3651 | 0.9212 | 0.3651 | 0.3289 |
| SFA        | 0.8866 | 0.1695 | 0.8230 | 0.2846 | 0.2335 |
| ISFA       | 0.6408 | 0.2790 | 0.9200 | 0.3888 | 0.3530 |
| CVA        | 0.5679 | 0.7682 | 0.9711 | 0.6530 | 0.6434 |
| PCA-Kmeans | 0.7303 | 0.5836 | 0.9686 | 0.6488 | 0.6325 |
| DSCN       | 0.4484 | 0.7852 | 0.9729 | 0.5708 | 0.5564 |
| DSMS-CN    | 0.6202 | 0.8611 | 0.9809 | 0.7210 | 0.7114 |

Fig. 10: Results obtained by pre-classification on HY data set, which are (a) difference image (DM) by CVA, (b) coarse init change map (CICM) by FCM.

B. Experiment settings

In our unsupervised architecture, the weights of DSMS-CN are initialized by he-normal way [31]. In order to overcome skew-class problem, the loss function of DSMS-CN apply weighted binary cross-entropy function, its definition is

\[ L = w_p \hat{y} \log y + (1 - \hat{y}) \log (1 - y) \]  

where \( w_p \) is the reciprocal of the proportion of changed and unchanged classes in training samples. By means of setting a larger weight for change class, the change samples can play a more important role at training phase. Adam optimizer [32] is adopted to optimize loss function (learning rate is set to...
Fig. 11: Visualized results obtained by proposed method and comparison methods on HY data set. (a) MAD. (b) IRMAD. (c) SFA. (d) ISFA. (e) CVA. (f) PCA-Kmeans. (g) DSCN. (h) DSMS-CN. The DSMS-CN achieves the best visual results, followed by CVA and DSCN.

1e-4). Dropout [33] and weight decay [34] are used to avoid overfitting during training phase. As discussed in [35], [36], the generally best choice of VHR image patch size is 5, in our experiments we also use 5 as image patch size.

To evaluate our method, seven state-of-the art methods are used for comparison, namely CVA, MAD [10], IRMAD [11], SFA [12], ISFA [12], PCA-Kmeans [37], and DSCN [18]. The IRMAD and ISFA are iterative versions of MAD and SFA. Except DSCN, other methods are unsupervised methods. The DSCN is a deep siamese convolutional network and perform well in aerial images. We train it with exactly the same way as DSMS-CN. In our experiments, the hyper-parameters of all methods use the optimal values recommended in original references. OTSU, K-means and FCM are adopted as segmentation methods, the best results of comparison methods are chosen as final results. As evaluation criteria, recall rate, precision rate, overall accuracy (OA), F1 score and kappa coefficient are adopted for quantitative analysis. The following part discusses the experimental results for data sets.

C. Experiments 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 DM. As shown in Fig. 8(a), the areas of warmer tones have the greater probability of change. Based on DM, FCM is implemented to generate coarse init change map (CICM). In Fig. 8 (b), the white pixels belong to changed class, the dark pixels belong to unchanged class, and gray pixels are candidates to be classified by the proposed method. All the pixels in change class are chosen as training samples. However, since spectral difference between multi-temporal images of the pixels in unchanged class is relatively constant, we only randomly select a part of black area as training samples, and the specific number is four times the number of change pixel.

Figure 9 shows the binary change maps obtained by proposed method and comparison methods. The results obtained by MAD and SFA are unsatisfactory, a large number of unchanged pixels are classified as change. Through iterative weighting, the results obtained by IRMAD and ISFA are better in visual. Nevertheless, it is obvious that a lot of building roofs are misclassified into change class, which means ISFA and IRMAD cannot solve overexposed problem. Besides, some slightly changed pixels are classified as unchanged class. Compared with IRMAD and ISFA, CVA directly calculating a

| Method     | Rec.    | Pre.    | OA      | F1      | Kappa   |
|------------|---------|---------|---------|---------|---------|
| MAD        | 0.7709  | 0.5989  | 0.9074  | 0.6741  | 0.6211  |
| IRMAD      | 0.7110  | 0.4356  | 0.8497  | 0.5402  | 0.4565  |
| SFA        | 0.7876  | 0.6313  | 0.9165  | 0.7008  | 0.6530  |
| ISFA       | 0.7120  | 0.6481  | 0.9162  | 0.6786  | 0.6305  |
| CVA        | 0.6631  | 0.8579  | 0.9445  | 0.7480  | 0.7174  |
| PCA-Kmeans | 0.5899  | 0.8948  | 0.9405  | 0.7110  | 0.6794  |
| DSCN       | 0.6428  | 0.7865  | 0.9340  | 0.7074  | 0.6706  |
| DSMS-CN    | 0.6769  | 0.8982  | 0.9503  | 0.7720  | 0.7448  |
difference image and performing clustering achieve a relatively good result. Nevertheless, changes in a few buildings and small objects, such as roads, are not recognized and plenty of margins of buildings are misclassified as changed class. By means of transforming spatial blocks based on PCA, PCA-Kmeans fuses spatial context information to correctly classify small objects. But some margins of buildings are still misclassified as changed class. Through convolution operation, in the results obtained by DSCN, margins of buildings are correctly classified and DSCN can ignore the effects of overexposed. But many changes of building are not recognized. Same as DSCN, the DSMS-CN is completely immune to overexposed problem. And despite existing a few omissions and residual errors, the binary change map obtained by DSMS-CN achieve the best qualitative result.

Table I reports the quality analysis results based on five evaluation criteria as described in Section 4-A. As observed in Figure 9, IRMAD has a low accuracy, and its F1 and kappa are only 0.3651 and 0.3289. ISFA performs slightly better than IRMAD, where its F1 and kappa are 0.3888 and 0.3530. However, ISFA still has similar shortcomings with IRMAD. This is because both MAD and SFA are based on the central limit theorem, whereas the Gaussianity of VHR images is not obvious. Therefore, they are not suitable for change detection of the VHR images covering a small region, even though they can achieve outstanding results in low- and medium-resolution images. Through the training samples are selected by CVA and FCM, the DSMS-CN achieves the best result with OA of 0.9809, F1 of 0.7210, and kappa coefficient of 0.7114. It means that our proposed DSMS-CN can effectively fit the distributions of ground changes of VHR images from pre-classification samples based on powerful extraction ability of MFCU and deep Siamese convolutional structure, and achieve the best performance. By contrast, the DSCN directly processing features cannot compete with our approach.

D. Experiments on HY data set

The DM and CICM are shown in Figure 10. Same as the WH data set, all the pixels in change class are chosen as training samples and for unchanged pixels, four times the number of the changed pixels are selected as samples. The qualitative results obtained by proposed method and comparison methods are shown in Figure 11. The residual errors and noisy of the results obtained by MAD and SFA are obvious, where a lot of unchanged pixels are misclassified as changes. Because of the complexity of VHR images, through iterative weighting, the results obtained by IRMAD and ISFA are worse and have more misclassifications. The binary change maps obtained by CVA and PCA-Kmeans are better in visual. However, the CVA misclassifies a part of margins of buildings as changed class and PCA-Kmeans misses some obvious changes. The result obtained by DSCN is similar to that obtained by CVA, but more building margins are misclassified as changes. The change detection result by applying DSMS-CN on HY is shown in Fig. 11 (h). Though a few details in changed regions are missed, most unchanged regions are well preserved. It can be observed that DSNS-CN outperforms the other methods in comparison and best qualitative result.

Table II reports the quality analysis results. As observed in qualitative results, due to the complexity of VHR images, the accuracy of IRMAD and ISFA is unsatisfactory. And MAD and SFA is slightly better than IRMAD and ISFA. The results obtained by CVA and PCA-K-means is relatively better, with kappa coefficient of 0.7174 and 0.6794, respectively. Because of misclassifying more margins of buildings, the accuracy of DSCN is lower than CVA with 0.7074 of F1 score and 0.6796 of kappa coefficient. By contrast, the DSMS-CN achieves the best result with precision rate of 0.8982, OA of 0.9503, F1
Fig. 13: Visualized results obtained by proposed methods on the Szada-1 of ACD. (a) is MSFC-EF. (b) is MSFC-EF-FCCRF. (c) is DSMS-FCN. (d) is DSMS-FCN-FCCRF.

Fig. 14: Visualized results obtained by proposed 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.

score of 0.7720 and kappa coefficient of 0.7448, which proves the powerful extraction ability of MFCU and deep Siamese convolutional structure again.

V. SUPERVISED CHANGE DETECTION EXPERIMENTS

A. ACD data set

In order to train the proposed network and evaluate our method, the SZTAKI AirChange Benchmark set (ACD) is employed [38], [39]. 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 x 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 [18], [19], [38], [39].

B. Experiment settings

First, to augment available training data, all possible flips and rotations of multiple of 90 degree are used. For the data set split, we adopt the way was proposed in [18] and [19]: 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 Figure 12. 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 our supervised algorithm, the weights of DSMS-FCN are initialized by he-normal way. In order to overcome skew-class problem, the loss function of DSMS-FCN apply weighted binary cross-entropy function. Adam optimizer is adopted to optimize loss function (learning rate is set to 2e-4). Dropout is used to avoid overfitting during training phase. As proposed in [23], 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.

The methods used for comparison are DSCN [18], CXM [38], SCCN [17] and three fully convolutional networks proposed in [19], using the values in [18] and [19]. The three comparative full-convolution networks are FC-EF, FC-Siam-Conc, FC-Siam-Diff. For the purpose of better reflecting the role of FC-CRF, we separately evaluate the results obtained by DSMS-FCN and combination of DSMS-FCN and FC-CRF. In addition, in order to verify the superiority of the MFCU compared to conventional single convolution unit, we modify the fully convolution architecture FC-EF proposed in [19] and replace the 3x3 convolution kernel with our MFCU. The modified network is denoted as MSFC-EF. Same as the unsupervised experiment, recall rate, recall rate, precision rate, overall accuracy, F1 score and kappa coefficient are adopted as evaluation criteria.

| Method             | Rec.   | Pre.   | OA     | F1     | Kappa |
|--------------------|--------|--------|--------|--------|-------|
| 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           | 0.5133 | 0.5975 | 0.9435 | 0.5522 | 0.5223|
| DSMS-FCN-FCCRF     | 0.5494 | 0.5736 | 0.9476 | 0.5612 | 0.5334|
TABLE IV: Accuracy assessment on change detection results of different methods on the Tiszadob-3 of ACD.

| Method       | Rec.   | Pre.   | OA     | F1    | Kappa  |
|--------------|--------|--------|--------|-------|--------|
| DSCN         | 0.883  | 0.851  | NA     | 0.867 | NA     |
| CxM          | 0.617  | 0.934  | 0.743  | NA    | 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 |

C. Experiment on ACD data set

Figure 13 and Figure 14 are illustrations of our results on ACD. We could see that both our DSMS-FCN and improved FC-EF network, namely MSFC-EF, can achieve satisfactory qualitative results. And 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 DSMS-FCN, which obviously outperforms all the other comparison methods in recall metric of 0.5133, overall accuracy of 0.9435 and F1 score of 0.5522. Utilizing the MFCU, each metric of MSFC-EF is better than FC-EF. And combined with FC-CRF, overall accuracy, F1 score and kappa coefficient of MSFC-EF and DSMS-FCN are further improved. The DSMS-FCN-FCCRF achieves the best recall metric, overall accuracy, F1 score and kappa coefficient.

On the Tiszadob-3, though the DSMS-FCN is not the best, it still superior than DSCN, CXM, SCCN and other two fully convolutional networks. The FC-EF network achieves a very high F1 score of 0.9340, but the MSFC-EF utilizing the MFCU still obtain improvement of performance. All metric of MSFC-EF is the better than FC-EF. Adopting FC-CRF, overall accuracy, F1 score and kappa coefficient of MSFC-EF and DSMS-FCN are further improved. The MSFC-EF-FCCRF achieves the best performance.

The experimental results that the MSFC-EF can outperform the original FC-EF in the data of Szada-1 and Tiszadob-3, which illustrates that the proposed MFCU has the ability to improve the performance of change detection with deep network. And depending on FC-CRF, the results obtained by deep fully convolutional network can be further improved. It is worth noting that in two test images, the performance of fully convolutional networks is obviously better than the patch based method DSCN and other architectures.

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

In this paper, a powerful multi-scale feature convolution unit is presented, which is different from conventional single convolution unit only extracting single-scale feature in one layer, and able to extract features at multi-scale in the same layer by a network in network structure. Based on the unit, two novel deep Siamese convolutional network architectures are designed for unsupervised and supervised change detection in VHR images. The DSMS-CN, used for unsupervised change detection is trained on pre-classification samples automatically generated by CVA and FCM. The DSMS-FCN is responsible for supervised change detection. As a fully convolutional network, the DSMS-FCN is able to process images of any size and do not require sliding patch-window, thus the accuracy and speed of inference could be significantly improved. To overcome 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 challenging WH and HY data sets, the qualitative and quantitative results indicate that the DSMS-CN outperforms other seven state-of-the-art methods with better overall accuracy, F1 score and kappa coefficient, 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 fully convolutional networks, a patch-based deep network method and other state-of-the-art methods, the DSMS-FCN delivers better performance, MFCU also exhibits powerful feature extraction capability and FC-CRF dose make the results obtained by deep fully convolutional networks more accurate. Besides, 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.

Our future work includes but is not limited to 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.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 61601333, 61822113 and 41801285.

REFERENCES

[1] A. Singh, “Review article digital change detection techniques using remotely-sensed data,” *International journal of remote sensing*, vol. 10, no. 6, pp. 989–1003, 1989.
[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, “Review article 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, p. 980, 2018.
[5] D. Lu, P. Mausel, E. Brondizio, and E. Moran, “Change detection techniques,” *International journal of remote sensing*, vol. 25, no. 12, pp. 2365–2401, 2004.
[6] D. Brunner, G. Lemoine, and L. Bruzzone, “Earthquake Damage Assessment of Buildings Using VHOR 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 D. F. Prieto, “Automatic analysis of the difference image for unsupervised change detection,” IEEE Transactions on Geoscience and Remote Sensing, vol. 38, no. 3, pp. 1171–1182, 2000.

[9] J. Deng, K. Wang, Y. Deng, and G. Qi, “Pca-based land-use change detection and analysis using multitemporal and multisensor satellite data,” International Journal of Remote Sensing, vol. 29, no. 16, pp. 4823–4838, 2008.

[10] 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, no. 1, pp. 1–19, 1998.

[11] 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.

[12] 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.

[13] 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.

[14] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] R. Caye Daudt, B. Le Saux, and A. Boulch, “Fully Convolutional Siamese Networks for Change Detection,” 2018 25th IEEE International Conference on Image Processing (ICIP), 2018, pp. 4063–4067.

[20] X. Niu, M. Gong Saux, 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.

[21] M. Lin, S. Yan and Q. Chen. (Dec. 2013). “Network in network.” [Online]. Available: https://arxiv.org/abs/1312.4400

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9.

[23] P. Krheinbhl, and V. Koltn, “Efficient inference in fully connected crfs with gaussian edge potentials,” Advances in neural information processing systems 24: 25th annual conference on neural information processing systems 2011., 2011, pp. 109–117.

[24] K. Simonyan, and A. Zisserman. (Sep. 2014). “Very Deep Convolutional Networks for Large-Scale Image Recognition.” [Online]. Available: https://arxiv.org/abs/1409.1556

[25] J. Bromley, I. Guyon, Y. Lecun, E. Sckinger, and R. Shah, “Signature Verification Using a Siamese Time Delay Neural Network,” Advances in neural information processing systems, vol. 7, no. 4, pp. 737–744, 1995.

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

[27] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.

[28] A. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in International Conference on Medical image computing and computer-assisted intervention, 2015, pp. 234–241.