SSA-SiamNet: Spectral-Spatial-Wise Attention-based Siamese Network for Hyperspectral Image Change Detection

Lifeng Wang, Liguo Wang, Qunming Wang, and Peter M. Atkinson

***Abstract*—Deep learning methods, especially convolutional neural network (CNN)-based methods, have shown promising performance for hyperspectral image (HSI) change detection (CD). It is acknowledged widely that different spectral channels and spatial locations in input image patches may contribute differently to CD. However, they are treated equally in existing CNN-based approaches. To increase the accuracy of HSI CD, we propose an end-to-end siamese CNN (SiamNet) with a spectral-spatial-wise attention mechanism (SSA-SiamNet). The proposed SSA-SiamNet method can emphasize informative channels and locations and suppress less informative ones to refine the spectral-spatial features adaptively. Moreover, in the network training phase, the weighted contrastive loss function is used for more reliable separation of changed and unchanged pixels and to accelerate the convergence of the network. SSA-SiamNet was validated using four groups of bi-temporal HSIs. The accuracy of CD using the SSA-SiamNet was found to be consistently greater than for ten benchmark methods.**

***Index Terms*—Attention mechanism, convolutional block attention module (CBAM), siamese network, hyperspectral images, change detection.**

I. Introduction

Remote sensing images are a common data source for global monitoring of the Earth's surface [1], [2]. Change detection (CD) can recognize the differences between multi-temporal remote sensing images and has been used widely in various applications, such as forestry and agricultural monitoring [3], [4], natural disaster assessment [5], [6] and land surface dynamic analysis [7] -[9]. With the successful launch of satellites, such as the NASA Earth Observing-1 (EO-1) and Chinese Gaofen-5, the availability of hyperspectral images (HSIs) at a global scale has increased greatly. HSIs contain rich spectral information and have inherent advantages over multispectral images in detecting land-cover changes [10]. However, the main challenges of CD using HSIs lie in the high dimensionality of the images, the redundancy of spectral information and large computational cost.

In general, conventional CD methods for HSIs can be categorized as algebra-based methods, transformation-based methods and post-classification comparison methods. The performance of algebra-based methods can be compromised due to the problem of information redundancy. Although transformation-based methods can deal with the high dimensionality, they have difficulty in selecting an appropriate threshold to detect land-cover changes. For post-classification comparison methods which compare two independent classified images pixel-by-pixel, the CD accuracy is affected directly by the propagation in classification errors of both images.

Recently, deep learning methods have shown great potential performance for HSI CD, which can solve the problem of high dimensionality to a greater extent and exploit features that are more effective than hand-crafted ones. In [11], a stacked autoencoder (SAE) was adopted to extract features from the difference image of bi-temporal HSIs to detect changes, but this method considered only the spectral features of pixels. A deep belief network (DBN) consisting of a restricted Boltzmann machine (RBM) and support tensor machine (STM) was also developed to identify changes for HSIs [12]. Similar to the SAE method, however, the DBN failed to take into account spatial features. Spectral-spatial features, if properly extracted, can be more discriminative than spectral features for HSI processing tasks [13], [14], [15]. Based on this, a noise modeling-based unsupervised fully convolutional network (FCN) framework for HSI CD was developed to learn powerful spectral-spatial features [16]. A general end-to-end two-dimensional (2D)-convolutional neural network (CNN) framework (GETNET) using a mixed-affinity matrix that integrated a subpixel representation as the input was proposed to detect changes from bi-temporal HSIs [17]. Song *et al*. [18] proposed the recurrent three-dimensional (3D) fully convolutional network (Re3FCN), which merged the advantages of a 3D fully convolutional network (FCN) and a convolutional long short-term memory (ConvLSTM) to extract joint spectral-spatial-temporal features. Nonetheless, the above methods increase computational costs and fail to consider sufficiently the information redundancy in the spectral and spatial domains.

The human vision system can selectively focus on conspicuous parts and ignore inconspicuous parts for an entire scene of interest. The attention mechanism, which is inspired by the human vision system, can be regarded as a tool biasing the allocation of available processing resources towards the most informative components of an input signal [19] . Moreover, the attention mechanism has the property of attending to important

Manuscript received XXX. This work was supported by the National Natural Science Foundation of China under Grants 41971297 and 62071084, and Tongji University under Grant 02502350047. *(Corresponding author: Qunming Wang.)*

L. Wang and L. Wang are with the College of Information and Communication Engineering, Harbin Engineering University, 145 Nantong Avenue, Harbin 150000, China (e-mail: wanglifeng_2016 @163.com).

Q. Wang is with the College of Surveying and Geo-Informatics, Tongji University, 1239 Siping Road, Shanghai 200092, China (e-mail: wqm11111@126.com).

P.M. Atkinson is with the Faculty of Science and Technology, Lancaster University, Lancaster LA1 4YR, UK; and also with Geography and Environment, University of Southampton, Highfield, Southampton SO17 1BJ, UK.

components and suppressing irrelevant ones by setting the appropriate weights to an input signal [20]. The attention mechanism was applied extensively for image captioning [21], visual question answering [22], [23] and image classification [24], [25]. Recently, several attention mechanisms were proposed to enhance the representation ability of CNNs. Specifically, a residual attention network with an encoder-decoder type attention module that refines the feature maps was proposed to enhance image classification performance [26]. However, the generation of the 3D attention map in the residual attention network requires more parameters, which leads to increased computational cost. Hu *et al.* [27] proposed the squeeze-and-excitation module that utilized global average-pooled features to achieve spectral-wise attention. The squeeze-and-excitation module, however, only refines spectral features and ignores spatial attention which also has a crucial effect on the recognition of important spatial features in an image [21]. The 3D attention map with spectral and spatial intelligence capabilities was realized by the bottleneck attention module (BAM) [28]. However, the BAM needs to be placed at each bottleneck of the network, which makes the structure of the basic network more demanding. To address the above deficiencies, a convolutional block attention module (CBAM) was developed to use global average-pooled and max-pooled features to achieve spectral-spatial-wise attention [19]. CBAM showed more satisfactory performance than other attention modules and can be plugged into each convolutional block as a plug-and-play module.

For HSIs, different spectral channels and spatial locations in the image patch contribute differently to the final CD predictions in theory. Attention mechanisms can focus on more discriminative channels and locations and have been adopted to HSI classification [29]-[31], super-resolution [32], and band selection tasks [33]. For CD using bi-temporal remote sensing images, the Pyramid feature-based attention-guided siamese network (PGA-SiamNet), integrating a pyramid-based CNN with various attention mechanisms, was developed to detect building changes in orthoimagery [34]. Lin *et al.* [35] proposed a faster R-CNN with a squeeze-and-excitation mechanism to detect ships in SAR images. For CD in bi-temporal HSIs, to the best of our knowledge, attention mechanisms have seldom been considered.

Based on the abovementioned issues, a spectral-spatial-wise attention-based siamese network, abbreviated as SSA-SiamNet, is proposed for the HSI CD task in this paper. The main contributions of this paper are as follows.

1) An end-to-end SSA-SiamNet framework is proposed to extract spectral-spatial-wise features, which can be trained from scratch. Accordingly, the learned deep features are suitable for the CD task and the method shows more competitive performance than other methods in the case of small training samples.
2) As a baseline, the siamese CNN (i.e., SiamNet) with two weight-sharing branches extracts the feature tensors mapped to the same space, which makes the calculation of subsequent distances simpler and reduces the computational complexity of the model.
3) Both the spectral-wise and spatial-wise attention modules, implemented by the CBAM, are embedded into the siamese network. The spectral-wise attention module is designed to reduce redundant information by emphasizing informative channels and suppressing less informative ones. Moreover, the spatial-wise attention module aims at focusing on the most informative locations in the adjacent pixels and ignoring less informative ones.
4) To extract more robust features and reduce the impact of imbalanced class samples, the weighted contrastive loss function is used to train the network, which makes the learned feature vectors of the changed pixel pair far away from each other and the vectors of the unchanged pixel pair close. Furthermore, learning the properties of features accelerates the convergence of the network and reduces computing time.

The rest of this paper is organized as follows. The details of the proposed SSA-SiamNet method are introduced in Section II. Section III evaluates the performances of SSA-SiamNet and shows its advantages over other CD approaches based on experimental results. Further issues about SSA-SiamNet and open questions for future research are discussed in Section IV. Finally, this paper is concluded in Section V.

II. Methods

An overview of the proposed SSA-SiamNet method is shown in Fig. 1. First, the SiamNet as a baseline is applied to simultaneously extract spectral and spatial features from the input patch pair. Second, the CBAM is embedded into the SiamNet to obtain adaptive spectral-spatial-wise features and, further, refine the features. Third, the weighted contrastive loss function is used in model training to separate the feature tensors of the changed pixel pair to be distant from each other and push those of the unchanged pixel pair to be close, which can accelerate model convergence. Finally, the Euclidean distance of the features tensors is fed into a fully connected (FC) layer and the change pixels are identified.

The four aspects in the proposed method, that is, CNN, SiamNet, CBAM and weighted contrastive loss function, are introduced in Sections II-A–II-D, respectively.

A. CNN

CNNs have been applied widely to a range of HSI tasks, such as classification, target detection and CD [36], [37]. In general, CNNs include convolutional layers, pooling layers and fully connected layers [38]. The input of a CNN is always an image patch. In each convolutional layer, the outputs of local filters are activated by a non-linear activation function, such as ReLU, Sigmoid, Tanh, etc.. The inputs of the FC layers are one-dimensional (1D) vectors.

To make full use of the spatial context information, the patches in HSI were constructed by combining the center pixel and surrounding pixels, producing patches with size $S \times S \times b$, where b represents the number of spectral bands and S is the length and width. Bi-temporal patches with the center pixel at the same location are called a patch pair.

Let $\mathbf{I} = \{\mathbf{x}(i, j) \mid 1 \le i \le h, 1 \le j \le w\}$ be an HSI, with a size of $h \times w \times b$, where h and w represent the spatial dimensions. $\mathbf{x}(i, j)$ represents the spectral bands of the pixel at the location (i, j) in HSI.

$$\mathbf{X} = \{\mathbf{x}(i,j) \mid m - \frac{S-1}{2} \le i \le m + \frac{S-1}{2}, n - \frac{S-1}{2} \le j \le n + \frac{S-1}{2}\}$$

represents the patch centered at $\mathbf{x}(m,n)$. When $\mathbf{X}$ is operated on by the l-th convolutional layer and pooling layer, the output feature map $\mathbf{H}_l(\mathbf{X})$ can be calculated as

$$\mathbf{H}_l(\mathbf{X}) = Pool(g(\mathbf{H}_{l-1}(\mathbf{X}) * \mathbf{W}_l + \mathbf{B}_l)) \qquad (1)$$

where $Pool(\bullet)$ and "$*$" represent the pooling operation and the convolution operation, respectively. $\mathbf{W}_l$ and $\mathbf{B}_l$ denote the filters and the biases of the l-th layer, respectively. The activation function is represented by $g(\bullet)$. Following the sequence of convolution and pooling operations, $\mathbf{H}_l(\mathbf{X})$ needs to be flattened to a 1D vector, and then is fed into the FC layer.

By tying weights in convolutional layers and local connections, CNNs can make full use of the spatial structure of an image patch. The pooling operation can reduce the patch size and translate invariant features. Moreover, the FC network can classify pixels according to the extracted features. Based on these advantages, CNNs have been applied extensively to acquire spectral-spatial features for HSI classification tasks.

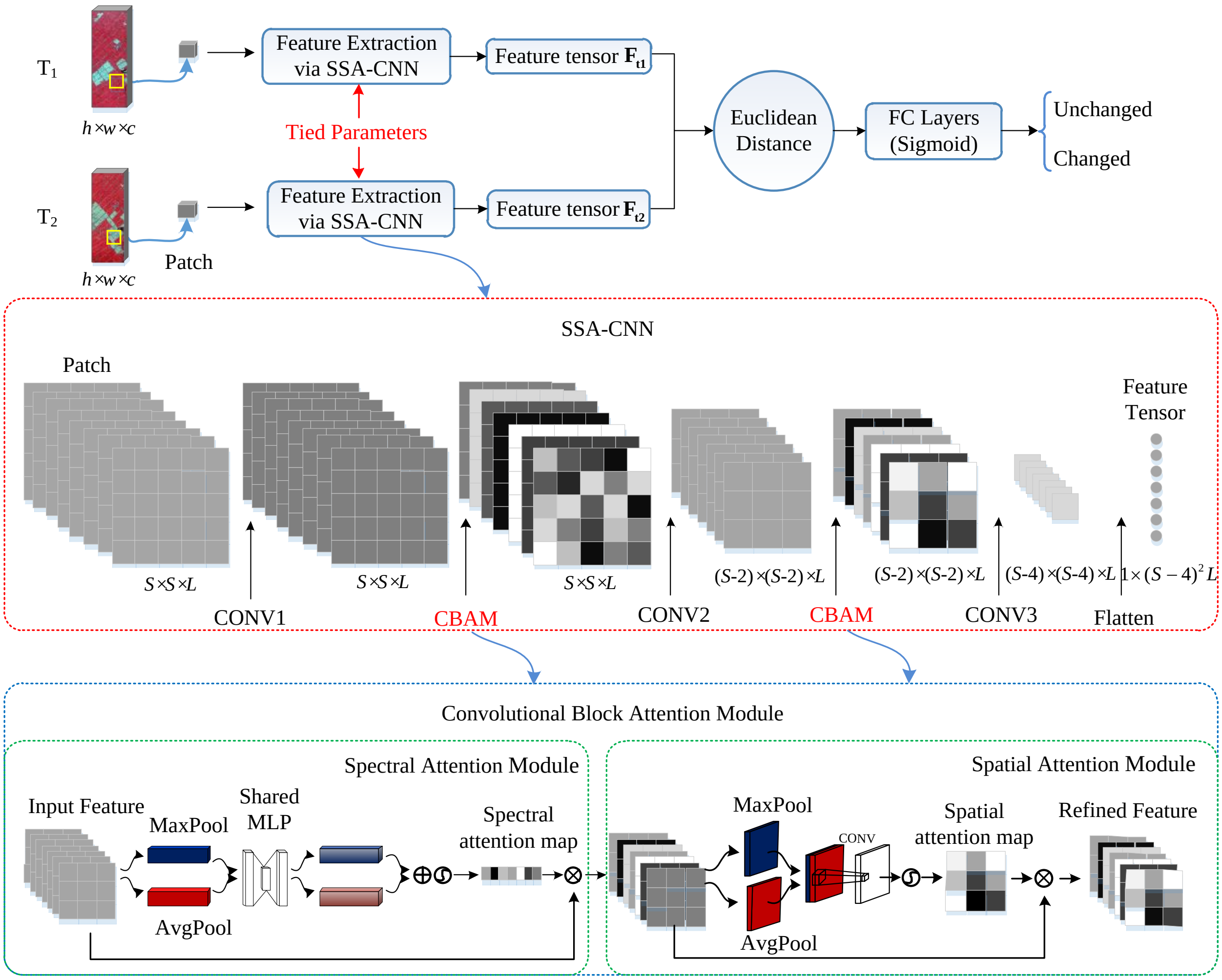


Fig. 1. Overview of the proposed SSA-SiamNet for HSI CD.

B. SiamNet

The input of the traditional CNN is the difference or concatenated image patch from the bi-temporal image patch pair, while the SiamNet extracts feature by inputting directly the image patch pairs. To compare patch pairs, three versions of CNN architectures were proposed in [39]: siamese, pseudo-siamese, and 2-channel. In a 2-channel network with only one branch, a 2-channel image combined with the input image pair is fed straight into the network. In the siamese and pseudo-siamese networks, there are two branches with the same architectures, and each of the two patches is input to separate branches. The difference between the branches of the siamese and pseudo-siamese networks is that the former shares the same weights, while the latter does not.

All of the abovementioned three versions of the network are suitable for the feature extraction task in CD. In this paper, the SiamNet is selected as the feature extraction method for the following reasons: 1) The two weight-sharing branches can extract the feature tensors mapped to the same space, which facilitates the calculation of subsequent distances; 2) Due to weight-sharing, the number of the network parameters is reduced, thus, reducing the computational complexity of the model.

As shown in Fig. 1, in SiamNet, each branch extracts a feature tensor. The Euclidean distance of the two feature tensors from the two branches is fed into the FC layer to predict the CD result. Nonetheless, the CNN structure adopted in this paper is different from the conventional CNN. In the conventional CNN architecture, the convolutional layer is used to generate a feature tensor from the input image patch. The pooling layer can enlarge the receptive field and reduce the size of the output feature map. The FC layer is similar to a classifier, which can predict the class label according to the input features. Because our objective is to extract the spectral-spatial-wise features pixel-by-pixel based on the input patch and the suitable patch size (i.e., $S\times S$) is small, the pooling layers are not adopted in the proposed architecture.

The architecture details of the designed SiamNet model are shown in Table 1. As acknowledged widely, three convolutional layers are a suitable choice for HSI classification [37]. In this paper, the CD problem is considered as the binary classification task of identifying changed and unchanged pixels. Therefore, three convolutional layers with a kernel size of 3×3 are adopted in the designed network.

In each convolutional layer, the number of kernels is set to N and ReLU is selected as the activation function. Also, batch normalization (BN) is adopted to avoid the phenomenon of gradient disappearance [40]. Moreover, L2 regularization is used to solve the over-fitting problem. Furthermore, the output of the FC layer, with two filters and a Sigmoid activation function, is a binary label indicating whether the pixel has changed.

Table 1 The architecture details of the designed SiamNet

Layers	Type	Kernel number	Kernel size	Padding
CONV1	Conv2D + BN + Activation (ReLU) + L2 (0.001)	*N*	3×3	same
CONV2	Conv2D + BN + Activation (ReLU) + L2 (0.001)	*N*	3×3	valid
CONV3	Conv2D + BN + Activation (ReLU) + L2 (0.001)	*N*	3×3	valid
FC	Fully Connected + Activation (Sigmoid)	2	-	-

C. CBAM

Our goal is that the two branches of SiamNet can learn adaptively the refined features for the CD task by using the spectral-spatial-wise attention mechanism. To this end, CBAM containing both spectral-wise and spatial-wise attention modules is adopted to obtain the refined features in this paper. The operation of CBAM is summarized as follows:

$$\begin{gathered} \mathbf{F}' = \mathbf{M}_{\text{se}}(\mathbf{F}) \otimes \mathbf{F} \\ \mathbf{F}'' = \mathbf{M}_{\text{sa}}(\mathbf{F}') \otimes \mathbf{F}' \end{gathered} \tag{2}$$

where $\mathbf{F} \in \mathbb{R}^{H\times W\times C}$ is the input feature map, $\mathbf{M}_{\text{se}} \in \mathbb{R}^{1\times1\times C}$ denotes a 1D spectral attention map, $\mathbf{M}_{\text{sa}} \in \mathbb{R}^{H\times W\times 1}$ presents a 2D spatial attention map, and $\otimes$ denotes the element-wise multiplication. In the multiplication, spectral attention values are broadcast along with the spatial dimensionality, and *vice versa*. $\mathbf{F}'$ and $\mathbf{F}''$ denote the spectral-wise-refined and spectral-spatial-wise-refined feature tensors, respectively. The details of each attention module are described in the following.

1) Spectral-wise attention module

The spectral-wise attention module refines the weights for the spectral feature maps and, thus, can emphasize meaningful channels and suppress less useful ones. This is analogous to the phenomenon that the human eye can focus on “what” is crucial in an input image. The spectral attention map is produced by squeezing the spatial dimensionality of the input features. The average-pooling and max-pooling operations were shown to be effective for generating the spectral attention map [19]. In this paper, the two operations are employed. Specifically, the average-pooling and max-pooling layers generate the average-pooled feature descriptor $\mathbf{F}_{\text{avg}}^{\text{se}}$ and max-pooled feature descriptor $\mathbf{F}_{\max}^{\text{se}}$, respectively. To generate the spectral attention map $\mathbf{M}_{\text{se}}$, both features are fed into a shared network that consists of a multi-layer perceptron (MLP) with one hidden layer. There are C/R units in the hidden layer to reduce the number of parameters, where C denotes the kernel number and R denotes the reduction ratio. Besides, the element-wise sum of the output feature vectors is activated by the Sigmoid function. Therefore, the spectral attention map can be produced by the following calculation:

$$\begin{aligned} \mathbf{M}_{\text{se}}(\mathbf{F}) &= \sigma(MLP(AvgPool(\mathbf{F})) + MLP(MaxPool(\mathbf{F}))) \\ &= \sigma(\mathbf{W}_1(\mathbf{W}_0(\mathbf{F}_{\text{avg}}^{\text{se}})) + \mathbf{W}_1(\mathbf{W}_0(\mathbf{F}_{\max}^{\text{se}}))) \end{aligned} \tag{3}$$

where σ denotes the Sigmoid function and the MLP weights $\mathbf{W}_0 \in \mathbb{R}^{C/R\times C}$ and $\mathbf{W}_1 \in \mathbb{R}^{C\times C/R}$ are shared for inputs of both average-pooled and max-pooled features.

2) Spatial-wise attention module

Different from the spectral-wise attention module, for the spatial-wise attention module, two types of pooling operations along the spectral axis are adopted to produce two feature descriptors $\mathbf{F}_{\text{avg}}^{\text{sa}} \in \mathbb{R}^{H\times W\times 1}$ and $\mathbf{F}_{\max}^{\text{sa}} \in \mathbb{R}^{H\times W\times 1}$. Then, the concatenated feature descriptor is convolved by a convolution layer. The output of the convolution layer is activated by the Sigmoid function to obtain the spatial attention map. Furthermore, the spatial-refined feature map highlighting the informative regions and suppressing the less useful ones is acquired by multiplying the input feature and the spatial attention map. In summary, the spatial attention map $\mathbf{M}_{\text{sa}}$ is produced by:

$$\begin{aligned} \mathbf{M}_{\text{sa}}(\mathbf{F}) &= \sigma(f^{N\times N}([AvgPool(\mathbf{F}); MaxPool(\mathbf{F})])) \\ &= \sigma(f^{N\times N}([\mathbf{F}_{\text{avg}}^{\text{sa}}; \mathbf{F}_{\max}^{\text{sa}}])) \end{aligned} \tag{4}$$

where $f^{N\times N}$ represents a convolution operation with a kernel size of $N\times N$.

D. The weighted contrastive loss function

An appropriate loss function can optimize the designed network in model training to extract more effective features. The weighted contrastive loss function is adopted to train the proposed network, based on the appealing property that the feature tensors of the unchanged pixel pairs are close to each other and those of changed pixel pairs are far away [41]. As mentioned earlier, $\mathbf{X}_1$ and $\mathbf{X}_2$ denote the bi-temporal patch pairs covering the same center pixel $\mathbf{x}(i, j)$. The output feature tensor of the network for the center pixel at (i, j) is represented as $\mathbf{G}(\mathbf{X})_{i,j} \in \mathbb{R}^{1\times N}$.

The Euclidean distance map between the feature vector $\mathbf{G}(\mathbf{X}_1)_{i,j}$ and $\mathbf{G}(\mathbf{X}_2)_{i,j}$ is denoted as $D(\mathbf{X}_1,\mathbf{X}_2)_{i,j}$, which is calculated as follows:

$$D(\mathbf{X}_1,\mathbf{X}_2)_{i,j} = \| \mathbf{G}(\mathbf{X}_1)_{i,j} - \mathbf{G}(\mathbf{X}_2)_{i,j} \|_2 . \tag{5}$$

Then, the loss function is expressed as follows:

$$\begin{aligned} Loss(\mathbf{W}) &= \sum_{k=1}^{p} \ell(\mathbf{W},(y,\mathbf{X}_1,\mathbf{X}_2)^k) \\ &= \sum_{k=1}^{p} (1-y_{i,j}^k)\ell_U(D_{i,j}^k) + y_{i,j}^k \ell_C(D_{i,j}^k) \end{aligned} \tag{6}$$

where y is the label for the patch pair $\mathbf{X}_1$ and $\mathbf{X}_2$, and the labels of the unchanged and changed pixel pairs are $y(i, j) = 0$ and $y(i, j) = 1$, respectively. $(y,\mathbf{X}_1,\mathbf{X}_2)^k$ is the *k*-th labeled training sample pair, and *p* is the number of training sample pairs. Moreover, ℓ_U and ℓ_C are the sectional loss functions for unchanged and changed pixel pairs, respectively, which are defined as follows [42]:

$$\ell_U(D_{i,j}^k) = \frac{1}{2}(\sigma(D_{i,j}^k))^2 \tag{7}$$

$$\ell_C(D_{i,j}^k) = \frac{1}{2}\{\max(0, q-\sigma(D_{i,j}^k))^2\} \tag{8}$$

where $q > 0$ is a margin and is set to 1 [41]. It pushes the value of the sigmoid function of changed pixel pairs, that is, $\sigma(D_{i,j}^k)$, closer to 1.

An imbalance in the numbers of class samples is a common problem in the CD task. Normally, the number of unchanged pixel pairs is much larger than that of changed pairs. To balance the class losses, it is necessary to weight each class loss. Therefore, the weighted contrastive loss function is considered in the proposed method, which is characterized as [42]:

$$\ell(\mathbf{W},(y,\mathbf{X}_1,\mathbf{X}_2)^k) = (1-y_{i,j}^k)\ell_U(D_{i,j}^k)w_U + y_{i,j}^k \ell_C(D_{i,j}^k)w_C \tag{9}$$

where w_U and w_C are the weights for unchanged and changed pixel pairs, respectively. The average frequency balancing is adopted in the loss function, and the weights w_U and w_C are calculated as follows:

$$w_U = \frac{f_{\text{avg}}}{f_U} \tag{10}$$

$$w_C = \frac{f_{\text{avg}}}{f_C} \tag{11}$$

where f_U and f_C denote the frequencies of unchanged and changed pixel pairs, respectively. Also, f_{avg} represents the average class frequency. Since there are two categories of classes in the HSI CD task identified in this paper, that is, changed and unchanged, f_{avg} is simply determined as 0.5. Thus, when the number of changed pixel pairs is less than for unchanged pairs, the resulting weight w_C is larger than 1, which can balance the contributions of the two parts in the loss function.

III. Experiments

A. Datasets and parameter setting

To evaluate the effectiveness of the proposed SSA-SiamNet method, four HSI datasets were used for test in the experiments.

1) Datasets

In this paper, all HSI datasets were acquired by the Hyperion sensor onboard the EO-1 satellite. The EO-1 Hyperion sensor provides HSIs with a spectral resolution of about 10 nm and spatial resolution of about 30 m. Moreover, it covers the 0.4-2.5μm spectral range with 242 spectral bands. In the experiments, spectral bands with a low signal-to-noise ratio (SNR) were eliminated.

The first dataset is designated “Farmland” [17] and the three-channel false-color composites (bands 33, 22 and 11 as RGB) of the bi-temporal images are shown in Fig. 2(a) and Fig. 2(e). The bi-temporal images were acquired on May 3, 2006 and April 23, 2007, and cover farmland in Yancheng, Jiangsu province, China. The spatial size is 450×140 pixels and 155 spectral bands were selected for CD after noisy band removal. The changes in the image are caused mainly by crop rotation.

The second dataset, named “River” [17], covers an area in Jiangsu province, China. The two HSIs were acquired on May 3, 2013 and December 31, 2013. The false-color composites (bands 33, 22 and 11 as RGB) of the two images are shown in Fig. 2(b) and Fig. 2(f). This dataset has a spatial size of 463×241 pixels and contains 198 bands after removing noisy bands. The changes in the dataset are due mainly to the removal of sediment in the river.

The third dataset “Santa Barbara”, covers an agricultural area in Santa Barbara, California, USA. The bi-temporal images are shown in Fig. 2(c) and Fig. 2(g) (bands 33, 22 and 11 as RGB). The two images were acquired in 2013 and 2014. This dataset has a spatial size of 984×740 pixels and contains 224 bands after noisy band removal.

The fourth dataset “Bay Area”, covers an area in the San Francisco Bay Area, California, USA. The bi-temporal images were acquired in 2013 and 2015 as shown in Fig. 2(d) and Fig. 2(h) (bands 33, 22 and 11 as RGB). The spatial extent is 600×500 pixels and 224 bands were considered.

Ground-reference maps for the four datasets are shown in Fig. 4(h)-Fig. 7(h), where the white, black and gray parts represent changed, unchanged and unknown pixels, respectively.

2) Data preprocessing

For testing, all patch pairs were divided into training and testing sets according to a pre-defined proportion. To increase the learning ability, more training samples were simulated by flipping and rotating each training patch pair by 90°, 180°, and 270°.

3) Quantitative evaluation metrics

The metrics used for quantitative assessment are the overall accuracy (OA), Kappa coefficient (Kappa), Precision (Pr), Recall (Re), and F1-score (F1), which are calculated as follows

$$OA = \frac{TP+TN}{TP+FP+TN+FN} \tag{12}$$

$$P_C = \frac{(TP+FP)(TP+FN)+(FN+TN)(FP+TN)}{(TP+FP+TN+FN)^2} \tag{13}$$

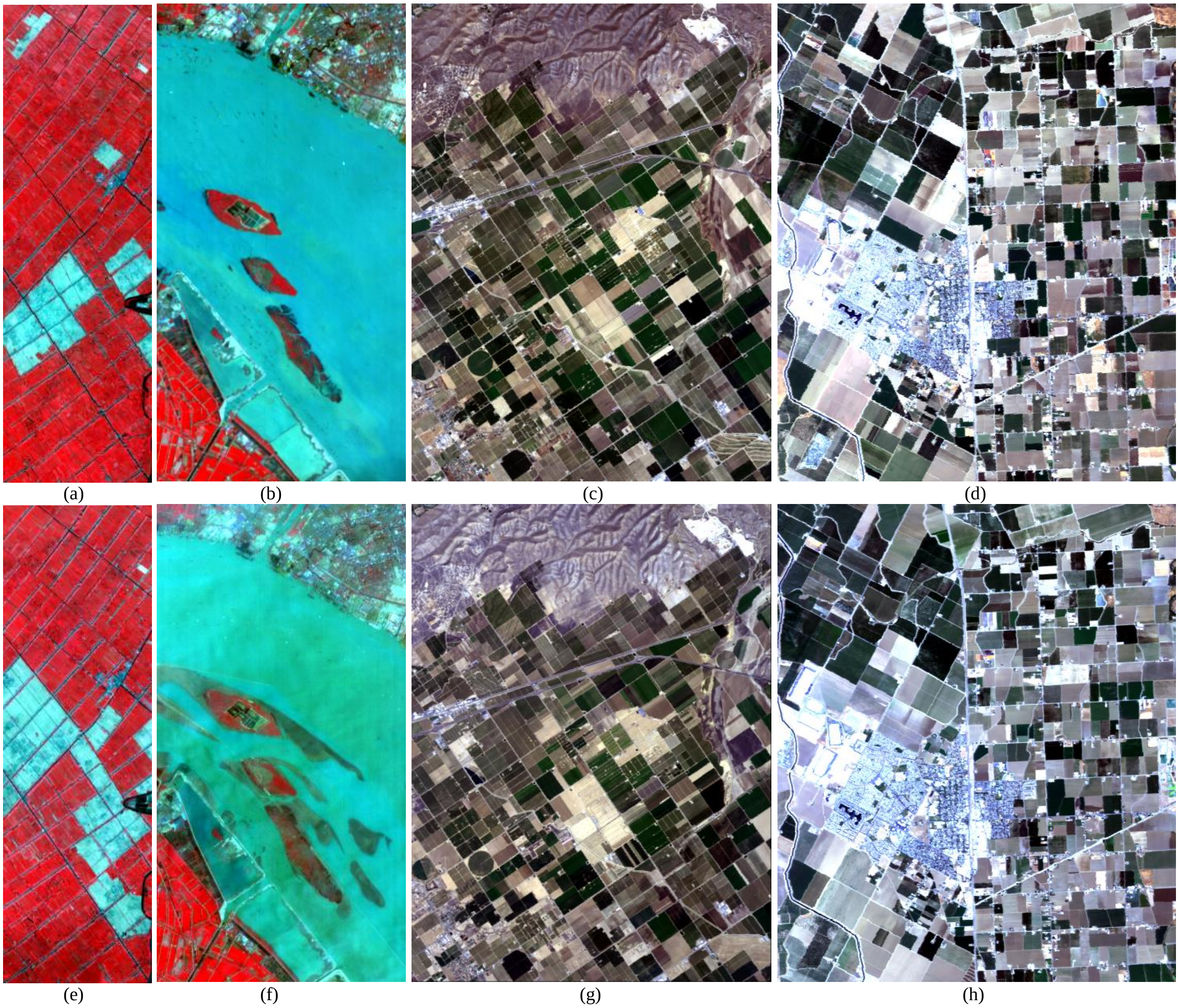


Fig. 2. The four HSI datasets used in the experiments (bands 33, 22 and 11 as RGB). (a) and (e) are Farmland images acquired on May 3, 2006 and April 23, 2007. (b) and (f) are River images acquried on May 3, 2013 and December 31, 2013. (c) and (g) are Santa Barbara images in 2013 and 2014. (d) and (h) are Bay Area images in 2013 and 2015.

$$Kappa = \frac{OA - P_C}{1 - P_C} \tag{14}$$

$$Pr = \frac{TP}{TP + FP} \tag{15}$$

$$Re = \frac{TP}{TP+FN} \tag{16}$$

$$F1 = 2 \times \frac{Pr \times Re}{Pr + Re}. \tag{17}$$

In Eqs. (12)-(17), there are four intermediate indices: 1) true positives (*TP*), that is, the number of correctly detected changed pixels; 2) true negatives (*TN*), which represents the number of correctly detected unchanged pixels; 3) false positives (*FP*), that is, the number of false-alarm pixels; and 4) false negatives (*FN*), that is, the number of missed changed pixels.

4) *Parameter setting*

The SSA-SiamNet was trained from scratch. The optimizer was the Root Mean Square prop (RMSprop) with $\rho = 0.9$, the learning rate was set to 0.001 in the first 100 epochs and 0.0001 in the second 100 epochs, and the number of total epochs was 200. The kernel numbers for the Farmland, River, Santa Barbara, and Bay Area datasets were set to 24, 24, 32 and 32, respectively, which will be discussed in Section III-D. The batch size was set to 32 for the Farmland dataset and 64 for the other three datasets. The patch size and reduction ratio for all four datasets were determined as 5×5 and 8, respectively. To avoid biased estimation, 10 independent tests were carried out using Tensorflow and Keras in a single NVIDIA GTX 1080Ti GPU with 64G memory.

B. *Comparison with other methods*

To demonstrate the performance of the proposed SSA-SiamNet method, we compared it with ten benchmark methods, including CVA [43], SVM [44], GETNET (without unmixing) [17], 2D-CNN [45], 3D-CNN [46], Diff-ResNet [47], Con-ResNet [47], Diff-RSSAN [48], Con-RSSAN [48], and SiamNet. For Diff-ResNet or Diff-RSSAN, the difference image of bi-temporal HSIs was classified by ResNet or RSSAN. For SVM, 2D-CNN, 3D-CNN, Con-ResNet and Con-RSSAN, the concatenate bi-temporal HSIs were used as input for CD.

In SVM, the radial basis function (RBF) kernel was used for all four datasets. In the experiments, the deep learning-based benchmark methods developed for HSI classification with multiple land cover classes were employed. Since the change detection task is considered as binary classification in this paper, if the above deep learning-based classification methods are applied directly to detect changes, it is easy to cause the problem of overfitting. Therefore, a dropout layer was added between convolution layers for 2D-CNN and 3D-CNN and residual blocks for ResNet and RSSAN. The dropout parameters were set to 0.2 for 2D-CNN and 3D-CNN and 0.4 for Diff-ResNet, Con-ResNet, Diff-RSSAN and Con-RSSAN. For all deep learning-based benchmark methods, the L2 regularization coefficient, kernel numbers and batch size were set to 0.001, 32 and 64, respectively. The loss function was binary cross-entropy for all deep learning-based methods except for SiamNet and SSA-SiamNet.

For a fair comparison, 5% of the samples from the Farmland and River datasets and 1% of the samples from the ground-references of the Santa Barbara and Bay Area datasets were selected as the training sets. The details are shown in Table 2.

Table 2 The numbers of pixel pairs in training and testing sets for the four datasets

Dataset	Training Set			Testing Set		
	Unchanged	Changed	Total	Unchanged	Changed	Total
Farmland	914	2236	3150	17363	42487	59850
River	5094	485	5579	96791	9213	106004
Santa Barbara	804	521	1325	79614	51613	131227
Bay Area	342	392	734	33869	38878	72747

1) Quantitative comparison

The 10-time average CD results of the methods on four datasets are reported in Table 3. Generally, the following observations can be made. First, CVA produce the smallest accuracy for four datasets, especially for the Santa Barbara and Bay Area datasets. This is because CVA are unsupervised CD methods that do not use any training information, which may not separate various changes satisfactorily, especially when their features are very close to those of unchanged pixels.

Second, deep learning-based methods (i.e., GETNET, 2D-CNN, 3D-CNN, Diff-ResNet, Con-ResNet, Diff-RSSAN, Con-RSSAN, SiamNet, and SSA-SiamNet) generally perform more satisfactorily than the SVM method in all metrics for all datasets except for the Farmland dataset. Compared with conventional machine learning-based methods, CNN-based methods make fuller use of spatial information through multi-layer convolution, which facilitates the CD task.

Third, 3D-CNN enhances the performance of 2D-CNN for all datasets except for the Santa Barbara dataset, as the former considers additionally the convolution in the spectral dimension and can acquire more discriminating features. However, the increase in accuracy is still not obvious. More precisely, the increases in OAs for the Farmland, River, Santa Barbara and Bay Area datasets are 0.09%, 0.09%, -0.10% and 0.47%, respectively. Furthermore, in most cases, the results of the SiamNet are superior to those of 3D-CNN, and the former has fewer parameters and lower computational costs, which also shows that the SiamNet method is more advantageous in dealing with the CD problem. In addition, the accuracy of Con-ResNet (or Con-RSSAN) is greater than Diff-ResNet (or Diff-RSSAN).

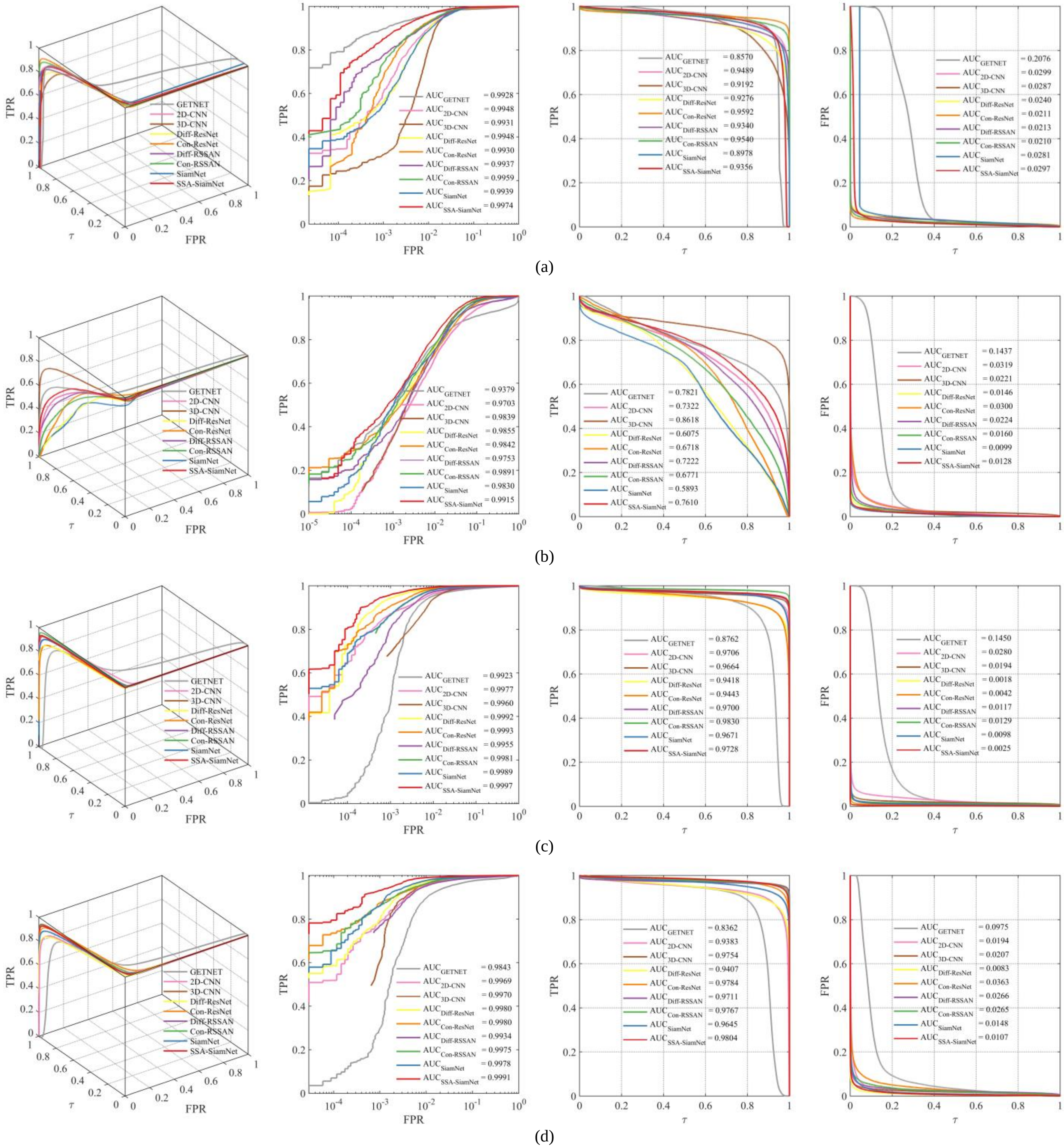
Last but not least, it is seen clearly that the most accurate results are produced by the proposed SSA-SiamNet method. For the Farmland dataset, the OA, Kappa, F1-score, Precision, and Recall metrics of SSA-SiamNet are 0.93%, 0.023, 1.49%, 1.75%, and 1.90% larger than those of SiamNet.

Table 3 Accuracy of different CD methods for the four datasets (the bold value indicates the most accurate result in each term)

Data set	Method	Metrics				
		OA (%)	Kappa	F1 (%)	Pr (%)	Re (%)
Farmland	CVA	96.08	0.9063	93.42	91.01	95.97
	SVM	**98.17**	**0.9555**	**96.84**	97.05	96.62
	GETNET	97.96	0.9507	96.51	95.64	**97.40**
	2D-CNN	96.98	0.9273	94.87	93.49	96.29
	3D-CNN	97.07	0.9294	95.02	93.48	96.62
	Diff-ResNet	97.24	0.9328	95.22	95.63	94.82
	Con-ResNet	97.36	0.9361	95.47	95.14	95.80
	Diff-RSSAN	97.25	0.9328	95.21	95.98	94.46
	Con-RSSAN	97.46	0.9385	95.63	95.50	95.77
	SiamNet	96.94	0.9253	94.67	95.71	93.66
	SSA-SiamNet	97.87	0.9481	96.30	**97.20**	95.41
River	CVA	94.22	0.6972	72.84	61.56	89.16
	SVM	96.71	0.7693	78.68	90.17	69.79
	GETNET	97.04	**0.8057**	**82.18**	85.64	78.98
	2D-CNN	96.77	0.7677	78.46	93.39	67.65
	3D-CNN	96.86	0.8052	82.24	80.94	**83.58**
	Diff-ResNet	96.79	0.7762	79.34	89.85	71.03
	Con-ResNet	96.82	0.7864	80.36	86.70	74.88
	Diff-RSSAN	96.99	0.8025	81.89	85.81	78.31
	Con-RSSAN	97.04	0.7964	81.23	90.24	73.86
	SiamNet	96.93	0.7959	81.25	86.67	76.47
	SSA-SiamNet	**97.18**	**0.8053**	82.04	**91.89**	74.10
Santa Barbara	CVA	88.51	0.7574	85.07	87.02	83.21
	SVM	96.11	0.9177	94.92	97.55	92.43
	GETNET	97.63	0.9504	96.98	97.29	96.68
	2D-CNN	97.69	0.9517	97.08	96.53	97.64
	3D-CNN	97.59	0.9495	96.93	97.35	96.51
	Diff-ResNet	98.02	0.9582	97.43	**99.78**	95.18
	Con-ResNet	98.19	0.9619	97.67	99.01	96.36
	Diff-RSSAN	98.28	0.9640	97.81	98.44	97.18
	Con-RSSAN	98.65	0.9716	98.28	98.18	98.38
	SiamNet	98.51	0.9688	98.10	98.78	97.42
	SSA-SiamNet	**98.86**	**0.9760**	**98.53**	99.65	**97.44**
Bay Area	CVA	86.83	0.7378	86.83	93.20	81.28
	SVM	94.64	0.8926	94.93	96.11	93.77
	GETNET	95.87	0.9171	96.09	97.07	95.14
	2D-CNN	97.41	0.9479	97.59	96.94	98.26
	3D-CNN	97.88	0.9575	98.01	98.42	97.60
	Diff-ResNet	97.54	0.9507	97.71	97.51	97.90
	Con-ResNet	98.07	0.9612	98.20	97.79	**98.62**
	Diff-RSSAN	98.20	0.9639	98.32	98.19	98.46
	Con-RSSAN	98.36	0.9671	98.46	98.88	98.04

	SiamNet	98.17	0.9633	98.27	**99.28**	97.29
	SSA-SiamNet	**98.77**	**0.9753**	**98.85**	99.10	98.59

To further demonstrate the effectiveness of the proposed algorithm, the 3D receiver operating characteristic (ROC) curves along with their three 2D ROC curves, including (true positive rate (TPR), false positive rate (FPR)), (TPR, τ (i.e., detector threshold)), and (FPR, τ) [49], [50] for the nine deep learning-based methods are shown in Fig. 3. The area under the curve (AUC) values for the three 2D ROC curves were also calculated and shown in the legends in Fig. 3. The ROC curves were produced based on the results before binarization (i.e., the sigmoid operation in the final layer of the network). Overall, the proposed SSA-SiamNet method is more accurate than the benchmark methods in terms of the AUC values for all three types of ROC curves in most cases. This demonstrates that the 3D ROC provides a very useful evaluation tool to evaluate the change detection performance.



(a)

(b)

(c)

(d)

Fig. 3. The 3D ROC curves of different methods for the four datasets. (a) Farmland dataset. (b) River dataset. (c) Santa Barbara dataset. (d) Bay Area dataset.

2) *Qualitative comparison*

In addition to the quantitative comparison in Table 3 and Fig.3, the CD maps of the 11 methods are compared qualitatively, as shown in Figs. 4-7.

Different from the other methods, the phenomenon of “salt-and-pepper” noise is apparent for the CVA method as it does not use training information in CD. Moreover, compared with the SVM method, the CD maps of the deep learning-based methods are more similar to ground-reference, which is in line with the quantitative results in Table 3. Furthermore, the CD map of the proposed SSA-SiamNet method is close to the ground-reference map for each dataset.

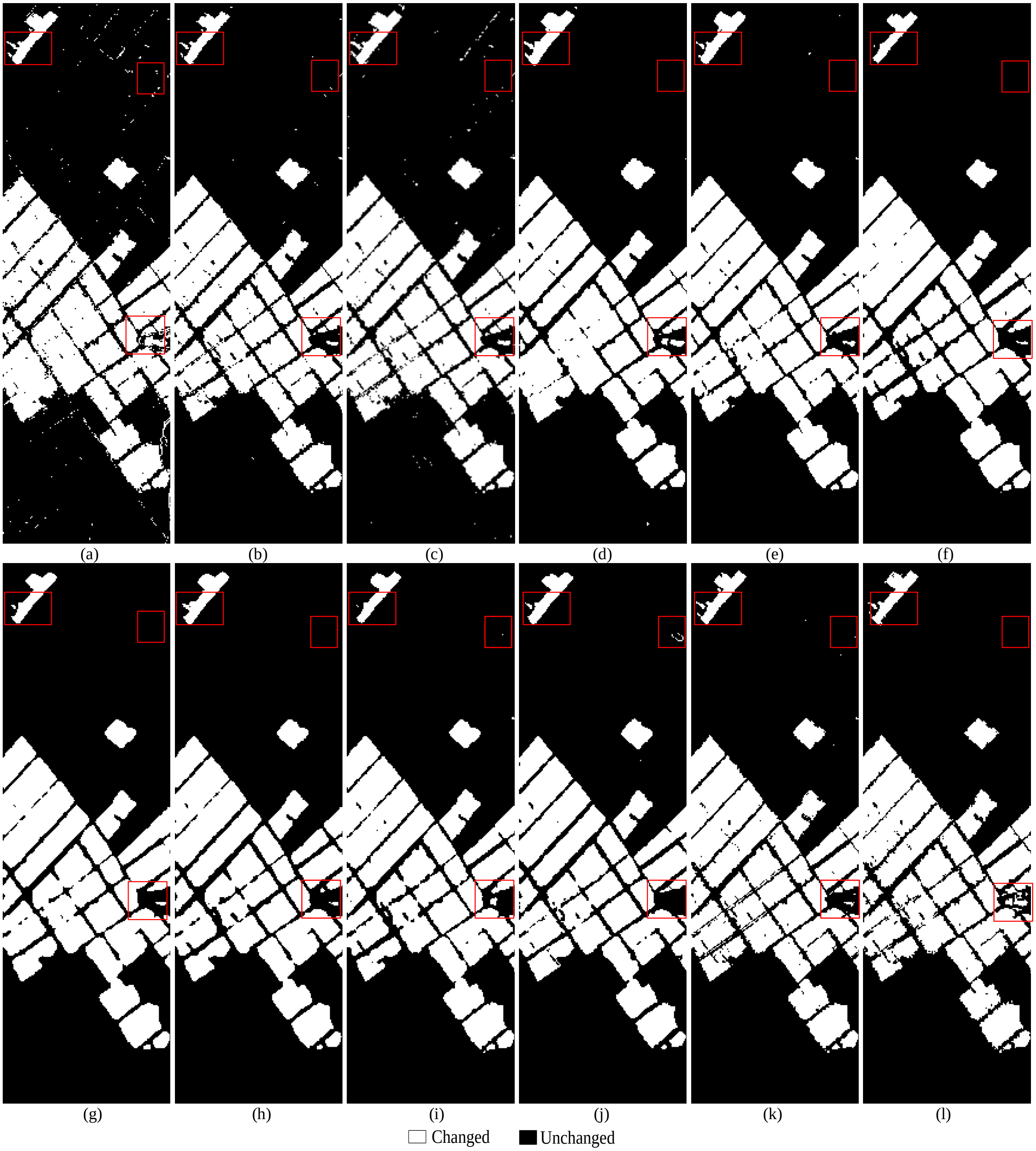


Fig. 4. CD maps of the different methods for the Farmland dataset. (a) CVA. (b) SVM. (c) GETNET. (d) 2D-CNN. (e) 3D-CNN. (f) Diff-ResNet. (g) Con-ResNet. (h) Diff-RSSAN. (i) Con-RSSAN. (j) SiamNet. (k) SSA-SiamNet. (l) Ground-reference map.

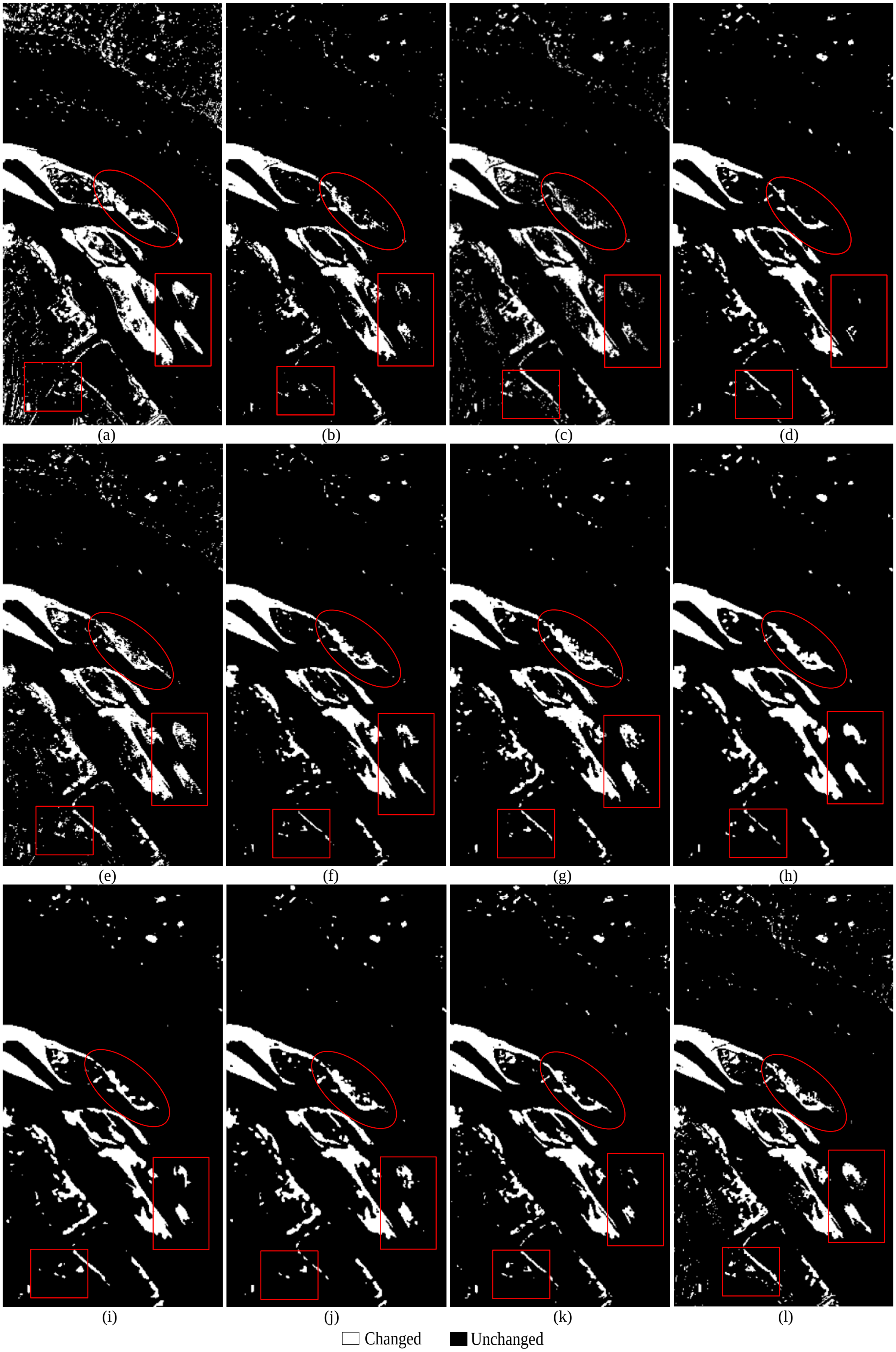
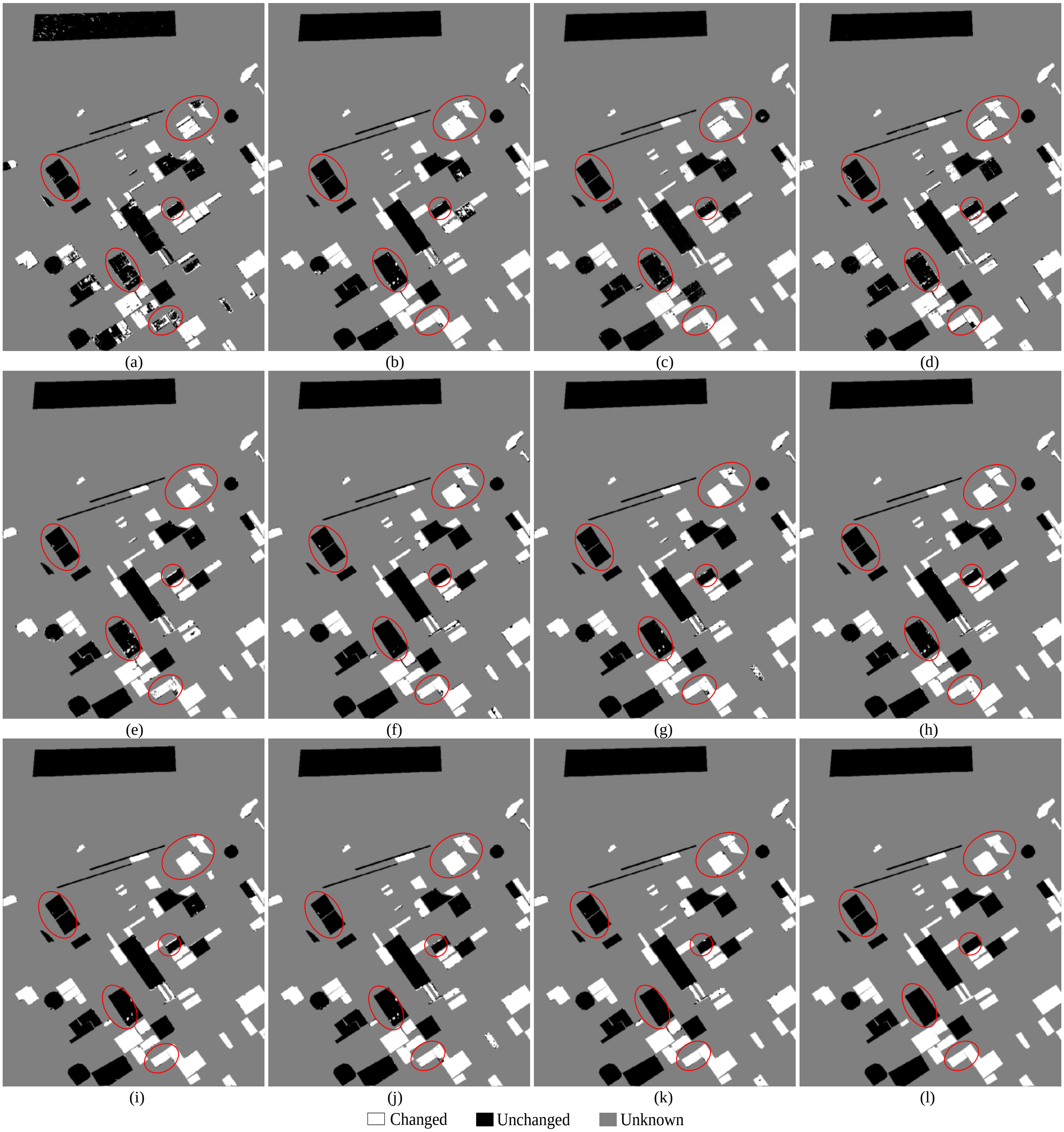


Fig. 5. CD maps of the different methods for the River dataset. (a) CVA. (b) SVM. (c) GETNET. (d) 2D-CNN. (e) 3D-CNN. (f) Diff-ResNet. (g) Con-ResNet. (h) Diff-RSSAN. (i) Con-RSSAN. (j) SiamNet. (k) SSA-SiamNet. (l) Ground-reference map.



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

(e) (f) (g) (h)

(i) (j) (k) (l)

Fig. 6. CD maps of the different methods for the Santa Barbara dataset. (a) CVA. (b) SVM. (c) GETNET. (d) 2D-CNN. (e) 3D-CNN. (f) Diff-ResNet. (g) Con-ResNet. (h) Diff-RSSAN. (i) Con-RSSAN. (j) SiamNet. (k) SSA-SiamNet. (l) Ground-reference map.

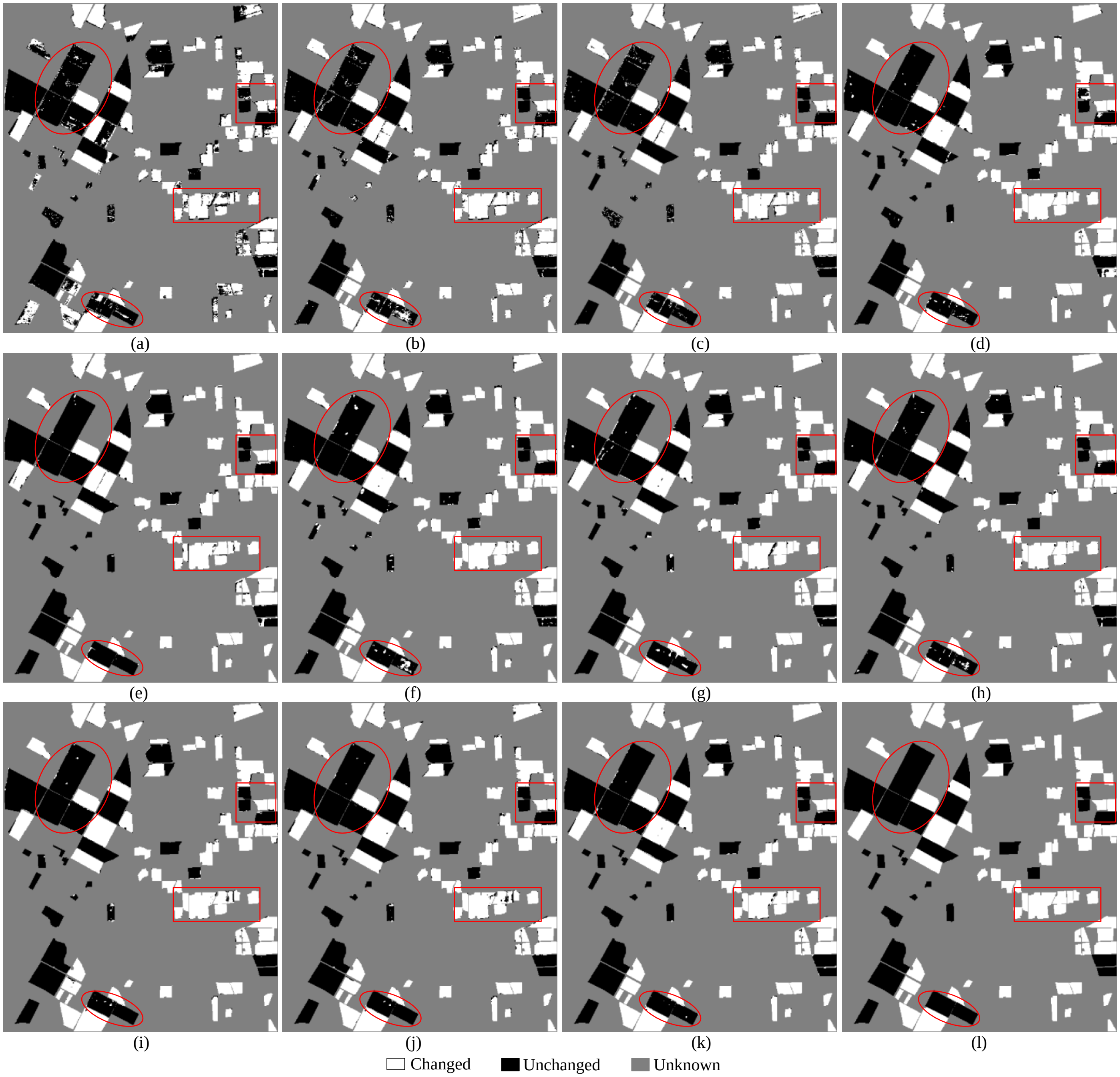


Fig. 7. CD maps of the different methods for the Bay Area dataset. (a) CVA. (b) SVM. (c) GETNET. (d) 2D-CNN. (e) 3D-CNN. (f) Diff-ResNet. (g) Con-ResNet. (h) Diff-RSSAN. (i) Con-RSSAN. (j) SiamNet. (k) SSA-SiamNet. (l) Ground-reference map.

3) *Computational cost*

Table 4 lists the computational costs and parameters for the ten supervised methods when the proportions of training and testing samples were set according to Table 2. The training and testing time increase with the number of pixel samples in the datasets. In the training phase, the deep learning-based methods required more time than the SVM. The reason is that the sample input is a patch of size $5\times5\times b$ in the deep learning-based methods except for GETNET, while the input is a vector of size $1\times b$ in the SVM. The GETNET method took the longest time amongst all the training-based methods, due to the relatively large patch of size $b\times b\times1$ and large number of parameters to be determined. In the testing phase, the testing time of the SVM method for the River and Santa Barbara datasets is longer than for the other two datasets, as the number of test samples for these two datasets is relatively larger. In addition, the training and testing time of Con-ResNet (or Con-RSSAN) is longer than Diff-ResNet (or Diff-RSSAN). The reason is that the concatenate bi-temporal HSIs with 2*b* bands were used in the former, while the difference images with *b* bands were used in the latter.

Table 4 The computational costs of different methods for the four datasets

Dataset	Method	Training Time (s)	Testing Time (s)	Parameters
Farmland	SVM	0.8	4.32	2
	GETNET	357.29	66.03	93.42M

	2D-CNN	163.68	8.51	73.13K
	3D-CNN	397.57	12.72	155.07K
	Diff-ResNet	198.43	7.11	48.19K
	Con-ResNet	320.37	11.26	53.15K
	Diff-RSSAN	181.28	6.54	49.25K
	Con-RSSAN	221.21	9.66	103.05K
	SiamNet	214.92	8.50	87.84K
	SSA-SiamNet	318.65	10.35	88.60K
River	SVM	4.42	32.55	2
	GETNET	1003.05	169.98	154.18M
	2D-CNN	297.47	99.57	91.71K
	3D-CNN	674.88	91.26	159.19K
	Diff-ResNet	484.79	14.69	49.57K
	Con-ResNet	546.79	144.79	55.91K
	Diff-RSSAN	318.86	11.54	63.65K
	Con-RSSAN	394.13	85.36	135.14K
	SiamNet	431.48	105.71	106.42K
	SSA-SiamNet	583.15	128.56	107.18K
Santa Barbara	SVM	0.74	16.96	2
	GETNET	310.1	588.96	198.27M
	2D-CNN	70.7	193.26	139.81K
	3D-CNN	111.4	190.66	277.03K
	Diff-ResNet	124.91	102.79	50.40K
	Con-ResNet	130.85	200.15	57.57K
	Diff-RSSAN	83.12	100.84	72.68K
	Con-RSSAN	91.21	197.58	156.25K
	SiamNet	87.76	193.46	166.08K
	SSA-SiamNet	126.53	202.17	167.32K
Bay Area	SVM	0.31	9.78	2
	GETNET	169.34	160.69	198.27M
	2D-CNN	50.49	87.10	139.81K
	3D-CNN	106.06	96.52	277.03K
	Diff-ResNet	80.29	82.37	50.40K
	Con-ResNet	84.64	88.27	57.57K
	Diff-RSSAN	51.91	81.51	72.68K
	Con-RSSAN	62.32	88.34	156.25K
	SiamNet	71.68	93.39	166.08K
	SSA-SiamNet	104.47	95.82	167.32K

Compared with 3D-CNN, the SSA-SiamNet method generally required less training and testing time, because more parameters need to be determined in 3D-CNN, and the weighted contrast loss function in SSA-SiamNet can accelerate the network convergence. Moreover, the proposed SSA-SiamNet method required more training and testing time than the SiamNet for all datasets, as SSA-SiamNet has a more complex architecture (both the spectral-wise and spatial-wise attention modules are considered in addition to SiamNet). However, the proposed method produced more accurate CD results than the other methods. Therefore, the slight increase in training and testing time is generally acceptable.

C. Model analysis

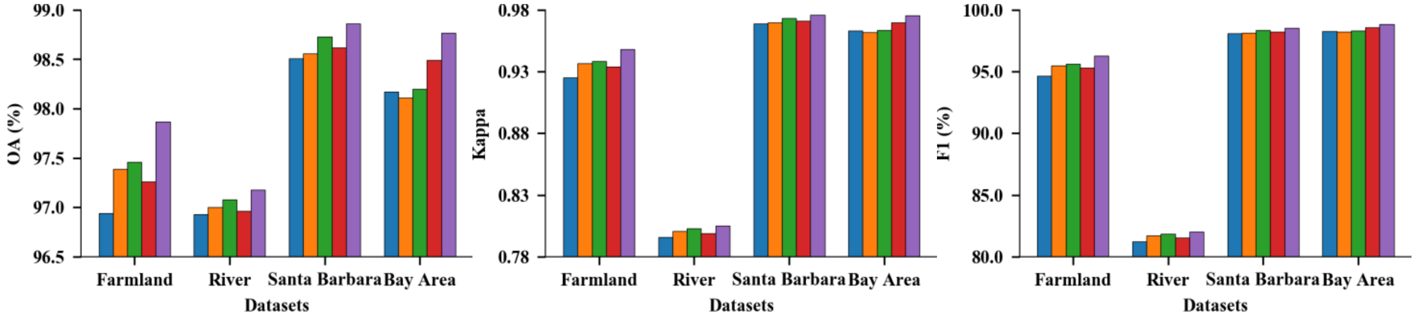
1) Application of the attention module

To validate the effectiveness of the attention mechanism, we compared the SiamNet, SiamNet only with spectral attention module (SE-SiamNet), SiamNet only with spatial attention module (SA-SiamNet), SiamNet with spatial-spectral attention module (SASE-SiamNet) that first considers spatial attention then spectral attention, and SiamNet with spectral-spatial attention module (i.e., the proposed SSA-SiamNet) that first considers spectral attention then spatial attention. The accuracy of the different models in terms of OA, Kappa and F1-score metrics is shown in Fig. 8. From Fig. 8, we can observe that both SE-SiamNet and SA-SiamNet are more accurate than SiamNet for the four datasets, which demonstrates the effectiveness of the attention mechanisms. This is consistent with the conclusion in [48]. Moreover, SA-SiamNet produces larger OA than SE-SiamNet. For the Farmland, River, Santa Barbara and Bay Area datasets, the increases in OA of SA-SiamNet over SE-SiamNet are 0.07%, 0.08%, 0.17% and 0.09%, respectively. Furthermore, the order of spectral and spatial attention can affect the results, and SSA-SiamNet methods can further enhance the CD performance compared with SASE-SiamNet. The increases in OA of SSA-SiamNet over SiamNet for the Farmland, River, Santa Barbara and Bay Area datasets are 0.93%, 0.25%, 0.35% and 0.60%, respectively. Most importantly, the proposed SSA-SiamNet achieves the most accurate result amongst the five versions, suggesting it is the most appropriate version in integrating the complementary information of the two attention mechanisms.

2) Analysis of the weighted contrastive loss function

The numbers of changed and unchanged pixel pairs of the four datasets are shown in Table 5, where f_U and f_C represent the percentage of unchanged and changed pixel pairs, respectively. Obviously, there is a large difference between the numbers of unchanged and changed samples for the Farmland and River datasets. To overcome the sample imbalances, the weighted contrastive loss function was adopted to train the network in SSA-SiamNet, and the unchanged weight w_U and changed weight w_C are also shown in Table 5.

To demonstrate the validity of the weighted contrastive loss function, loss functions with and without weights were used to train the network, and the results are shown in Fig. 9. Overall, the performance of using the weighted contrastive loss function is superior to that of the unweighted function. Moreover, the increase in accuracy is particularly obvious for the Farmland and River datasets, which is consistent with the degree of sample imbalance in Table 5. Also, amongst the three metrics of F1, Pr, and Re, the increase in Re is the largest, revealing that the weighted contrastive loss function can help to detect the changed samples more accurately.

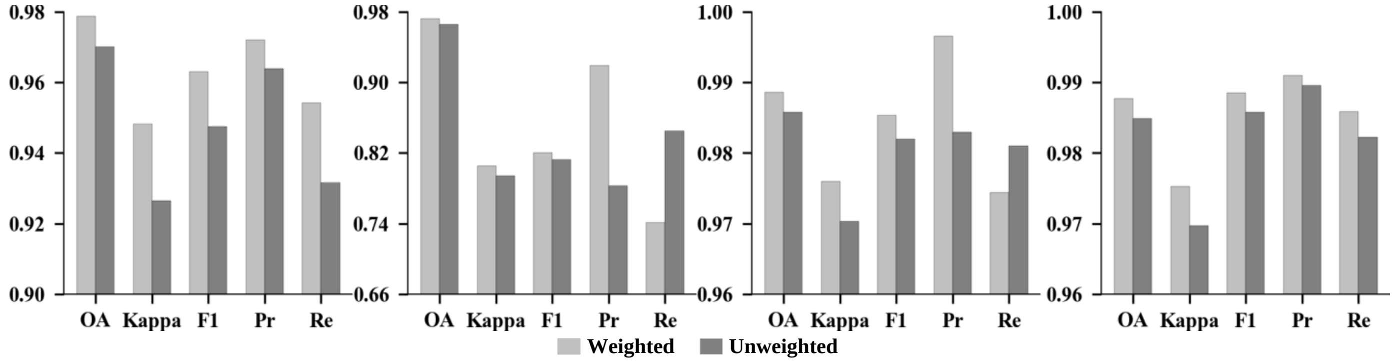


(a) (b) (c)

Fig. 8. The accuracy of the SiamNet with different attention modules for all four datasets. (a) OA. (b) Kappa. (c) F1-Score.

Table 5 The numbers of pixel pairs for the four datasets

Dataset	Pixel Pairs				f_U	f_C	w_U	w_C
	Unchanged	Changed	Unknown	Total				
Farmland	18277	44723	0	63000	29.01%	70.99%	1.7235	0.7043
River	101885	9698	0	111583	91.31%	8.69%	0.5476	5.7529
Santa Barbara	80418	52134	595608	728160	11.04%	7.16%	4.5290	6.9835
Bay Area	34211	39270	226519	300000	11.40%	13.09%	4.3860	3.8197

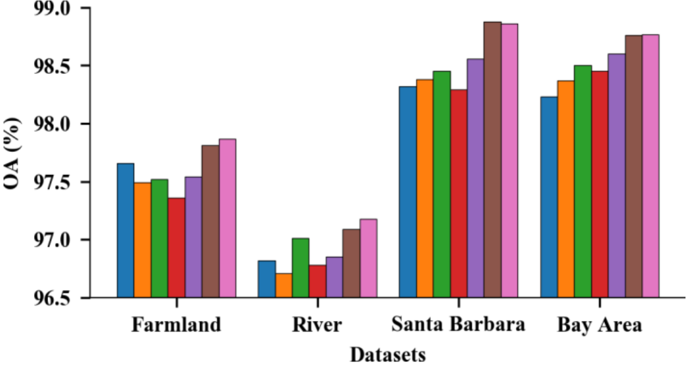


(a) (b) (c) (d)

Fig. 9. The accuracy of the loss function and weighted loss function for the four datasets. (a) Farmland dataset. (b) River dataset. (c) Santa Barbara dataset. (d) Bay Area dataset.

3) *Comparison with other attention mechanisms*

To demonstrate the advantage of CBAM, we compared it with some state-of-the-art attention mechanisms, including the squeeze-and-excitation network (SENet) [27], non-local network (NLNet) [51], global context network (GCNet) [52], position attention module (PAM) [53], channel attention module (CAM) [53] and dual attention network (DANet) [53]. For a fair comparison, the mechanisms were incorporated into the proposed method by replacing the CBAM part. The reduction ratio was set to 8 in GCNet, SENet, DANet, and PAM for all datasets. The kernel numbers in DANet, CAM and NLNet were consistent with the proposed method. The kernel size was set to 3 in DANet and PAM for all datasets. From Fig. 10, it is seen DANet and CBAM have similar performances, and they are obviously more accurate than the other attention mechanisms. This is because SENet, NLNet, GCNet and CAM only consider spectral attention while PAM only considers spatial attention. However, DANet and CBAM take both into consideration. Moreover, CBAM tends to be more accurate than DANet. Thus, CBAM is considered to be the most suitable choice for the network structure proposed in this paper.



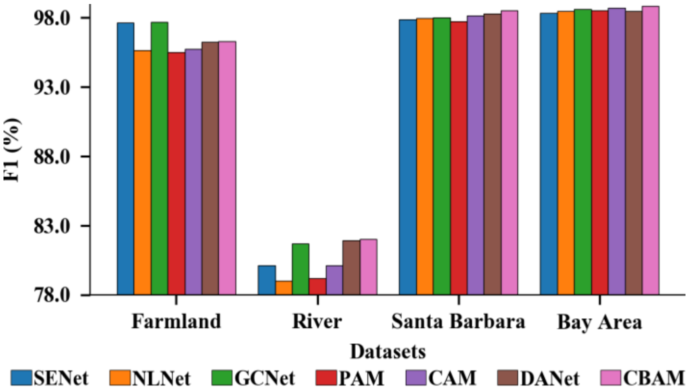


Fig. 10. The accuracy (in terms of OA and F1) of different attention mechanisms.

D. *Impact of parameters*

In the proposed method, several hyperparameters, such as the kernel numbers of convolutional filters, batch size, patch size and the proportion of training samples, can affect the model training process and, further, the detection results. Thus, the influence of these hyperparameters is investigated in this section. When analyzing the impact of a certain parameter, other parameters were fixed in the experiment.

1) *Impact of kernel numbers*

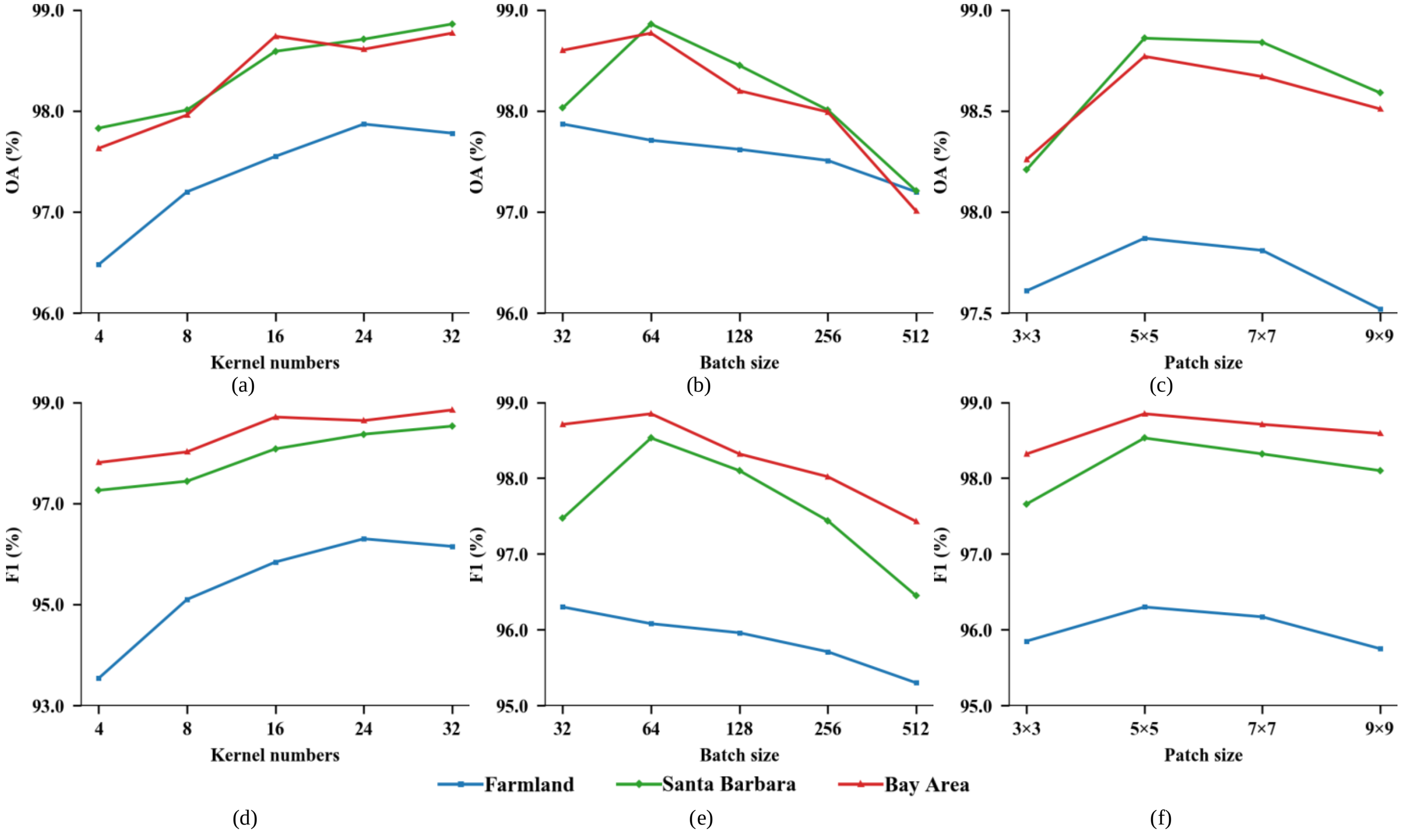
The kernel numbers of convolutional filters affect the representation capability and computational burden of SSA-SiamNet. As mentioned in Table 1, each convolutional layer has the same kernel number. Different kernel numbers ($N \in \{4, 8, 16, 24, 32\}$) were examined in this experiment. As shown in Fig. 11(a) and Fig. 11(d), as the kernel numbers increase, the OA and F1 values first increase and then decrease for the Farmland and River datasets. The optimal kernel numbers for the Farmland, River, Santa Barbara, and Bay Area datasets are 24, 24, 32 and 32, respectively.

2) Impact of batch size

To evaluate the effect of the batch size on the performance of SSA-SiamNet, a set of batch sizes {32, 64, 128, 256, 512} were considered. As shown in Fig. 11(b) and Fig. 11(e), as the batch size increases from 32 to 512, the OA and F1 values first increase and then decrease for all datasets except for the Farmland dataset, and the optimal batch size is 32 for the Farmland dataset and 64 for the other three datasets.

3) Impact of patch size

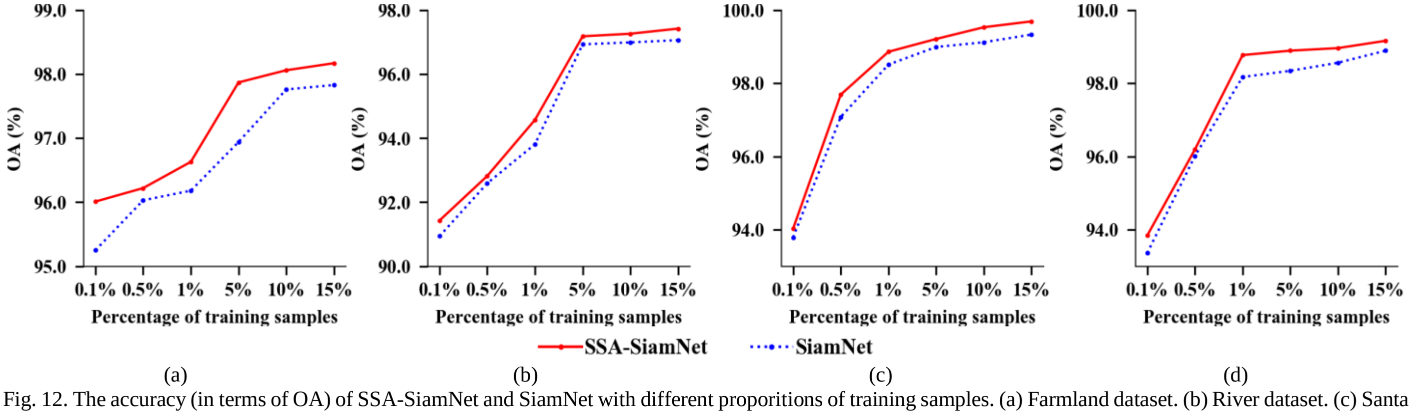
The size of the input patch reflects the amount of data used around the center pixel. To explore the effect of different patch sizes on the proposed method, we examined the set of patch sizes: 3×3, 5×5, 7×7 and 9×9. Amongst them, when the patch size is 3×3, the padding is set to be the same in the three convolutional layers. From Fig. 11(c) and Fig. 11(f), it is seen that with the increase in the patch size, the accuracy of CD does not change obviously, suggesting that the proposed method is robust to patch size.



(a) (b) (c)

(d) (e) (f)

Fig. 11. The accuracy (in terms of OA and F1) of SSA-SiamNet with different hyperparameters for the four datasets.



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

Fig. 12. The accuracy (in terms of OA) of SSA-SiamNet and SiamNet with different proporitions of training samples. (a) Farmland dataset. (b) River dataset. (c) Santa Barbara dataset. (d) Bay Area dataset.

4) Impact of the proportion of training samples

In this experiment, we examined the influence of the proportion of training samples for SiamNet and SSA-SiamNet. The set of proportions {0.1%, 0.5%, 1%, 5%, 10%, 15%} was considered.

As shown in Fig. 12, the OA curves for all four datasets show obvious increasing trends when the proportion *r* increases from 0.1% to 5%. However, the increases in OA are much smaller when the proportion further increases from 5% to 15%. Overall, the OA of the proposed SSA-SiamNet method is much larger than for SiamNet, and the advantages are more noticeable when r = 5% for both the Farmland and Bay Area datasets, respectively, and when r = 1% and r = 0.5% for the River and Santa Barbara datasets, respectively.

IV. Discussion

The advantage in terms of CD accuracy of the proposed SSA-SiamNet method over the benchmark methods is due mainly to the application of CBAM that contains the spectral-spatial-wise attention module. The spectral-spatial-wise attention mechanism can enhance informative channels and suppress less informative ones in the spectral domain, and meanwhile, emphasize informative neighborhood pixels and suppress uncorrelated ones in the spatial domain. Thus, the incorporation of CBAM into the SiamNet can refine the spectral-spatial features adaptively. Moreover, the proposed method achieves a good balance between CD accuracy and computational cost.

In this paper, 2D-CNN provides a simple solution to extract features directly from the center pixel and surrounding pixels. It took less computational cost than the other deep learning-based methods. According to Table 3, however, the accuracy of 2D-CNN is lower on the contrary. 3D-CNN with 3D kernels for the 3D convolution operation can extract spatial and spectral features simultaneously from HSI cubes, and achieve greater change detection performance than 2D-CNN. However, it requires the longest training and testing time in all deep learning-based methods. ResNet can be regarded as an extension of CNN with skip connections, which can promote the propagation of gradients and perform robustly with deep architecture. With respect to RSSAN, a spectral-spatial attention module is embedded into the residual block, which can avoid overfitting and accelerate the training of ResNet. Overall, the concatenate-based methods (i.e., Con-ResNet and Con-RSSAN) are more accurate than the difference-based ones (i.e., Diff-ResNet and Diff-RSSAN). This is probably because the concatenate patch contains richer spectral information. Compared with other CNN-based algorithms, the computational cost of SSA-SiamNet is reduced by using the weighted contrastive loss function. More specifically, the SiamNet using two weight-sharing branches to acquire spectral-spatial features reduces the computational complexity of the model correspondingly. Then, the weighted contrastive loss function can accelerate the convergence of the network. In addition, the SSA-SiamNet method is fairly robust to parameters such as batch size, patch size, etc.. This property ensures robust predictions under various conditions and helps to promote applicability in practice.

The proposed method is also applicable to data acquired by other platforms. This paper demonstrated the effectiveness of SSA-SiamNet through experiments based on HSIs from the EO-1 satellite. Similarly, the method can be extended simply to the HSIs acquired by other satellites and even UAVs. For example, the Gaofen-5 satellite launched by China in 2018 provides HSI at the global scale with a spatial resolution of 30 m and a spectral resolution of 5 nm for 150 visual bands and 10 nm for 180 short wave infrared bands (330 bands in all). It is believed that the SSA-SiamNet method will have great potential for CD based on multi-temporal Gaofen-5 HSIs. Additionally, the proposed algorithm is also potentially suitable for the CD task using other remote sensing images, such as multispectral images (MSI), synthetic aperture radar images (SAR), very high resolution (VHR) images, etc.. If the proposed model is applied to other data sources, the hyperparameters should be determined reasonably, such as the kernel number, batch size, patch size, and reduction ratio. For example, different from HSIs, multispectral images contain much fewer bands, so it is rational to reduce the kernel numbers and reduction ratio. For VHR images with much more spatial information, the patch size may need to be increased to fully characaterize the spatial texture. As the spatial resolution of remote sensing images increases, the central pixel is likely to be more closely related to the neighboring pixels. As a result, the effect of spatial attention could be more obvious, and the CD accuracy could be hopefully increased. It would be interesting in future research to investigate the relation between increases in CD accuracy and spatial resolution. On the other hand, the proposed CD method could also be applied to more application domains, such as ecological and environmental change monitoring, tracking urban development, natural disasters assessment, mapping coastline changes, forest and farmland monitoring, and so on.

The SSA-SiamNet method was demonstrated to be appropriate for CD between images acquired by the same platform with the same spatial and spectral resolution. In reality, however, timely CD may be required in cases where only bi-temporal images from different sensors with different spatial and spectral resolutions are available for use. To address this issue, it is worthwhile to further extend current the SSA-SiamNet for CD between multi-resolution images. The reliable geometric registration between the images is an important premise in this case. Moreover, how to match the spatial and spectral resolution of both images would be a very interesting issue. The key would be to make full use of the complementary information in both images and retain as much spatial and spectral information as possible for more reliable CD.

The continuous monitoring of land cover changes can be realized through analyzing time-series remote sensing images [54], [55], [56]. One of the main challenges in time-series analysis is to identify the exact breakpoints, that is, when changes occur along the timeline for a given location. The proposed method is developed for bi-temporal image CD and is expected to be extended to multi-temporal image CD. For example, the SiamNet can be extended to multiple parallel networks. Suppose the time-series contains five remote sensing images, $\mathbf{T}_1$, $\mathbf{T}_2$, $\mathbf{T}_3$, $\mathbf{T}_4$ and $\mathbf{T}_5$. Each image can be inputted into the parallel network separately, and the output is one of the four potential changes (i.e., change occurs between $\mathbf{T}_1$ and $\mathbf{T}_2$, $\mathbf{T}_2$ and $\mathbf{T}_3$, $\mathbf{T}_3$ and $\mathbf{T}_4$, or $\mathbf{T}_4$ and $\mathbf{T}_5$). In addition, to make full use of the information in the time-series, the CNN structure in the proposed method can be potentially replaced by a recurrent neural network (RNN), and the siamese long short-term memory (LSTM) RNN can be further considered [57]. All these provide interesting avenues for future research.

There are also some limitations of the proposed method. First, SSA-SiamNet is a supervised method, and its performance may be limited by the lack of available ground-reference labels for the changed and unchanged classes. Therefore, it would be interesting to develop solutions to explore the automatic generation of more training samples based on the available ones. On the other hand, the proposed method deals with binary CD concerned with "change or not", rather than multiple changes that solve "from-to" problems. Therefore, it is a promising avenue to explore methods with attention mechanisms for multiple CD in future research.

V. Conclusion

In this paper, an end-to-end framework named SSA-SiamNet was proposed to detect land cover changes in bi-temporal HSIs. The proposed method, which integrates CBAM with SiamNet, extracts spectral-spatial-wise features adaptively from the input patch pairs. The extracted features highlight influential information and suppress less informative channels and pixels in the spectral and spatial domains, respectively. Then, the Euclidean distance of the learned feature tensors from the two weight-sharing branches are fed into the FC layer to identify changes. Moreover, the designed network is trained using the weighted contrastive loss function, which can accelerate the convergence of the network. Experimental results on four HSIs showed that the proposed method can produce more accurate CD results than ten state-of-the-art methods.

References

[1] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang. "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 4340-4354, 2021.

[2] B. Rasti., D. Hong, R. Hang, P. Ghamisi, X. Kang, J. Chanussot, and J.A. Benediktsson, "Feature extraction for hyperspectral imagery: the evolution from shallow to deep: overview and toolbox," *IEEE Geoscience and Remote Sensing Magazine*, vol. 8, no. 4, pp.60-88, 2020.

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

[4] T. Adao et al., "Hyperspectral imaging: a review on UAV-based sensors, data processing and applications for agriculture and forestry," *Remote Sensing*, vol. 9, no. 11, p. 1110, 2017.

[5] S. Liu, M. Chi, Y. Zou, A. Samat, J. A. Benediktsson, and A. Plaza, "Oil spill detection via multitemporal optical remote sensing images: a change detection perspective," *IEEE Geoscience Remote Sensing Letters*, vol. 14, no. 3, pp. 324-328, 2017.

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

[7] X. Huang, D. Wen, J. Li, and R. Qin, "Multi-level monitoring of subtle urban changes for the megacities of China using high-resolution multi-view satellite imagery," *Remote Sensing of Environment*, vol. 196, pp. 56-75, 2017.

[8] 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, 2016.

[9] S. Liu, D. Marinelli, L. Bruzzone, and F. Bovolo, "A review of change detection in multitemporal hyperspectral images: current techniques, applications, and challenges," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 2, pp. 140-158, 2019.

[10] D. Hong, N. Yokoya, J. Chanussot, and X. Zhu. "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1923-1938, 2019.

[11] J. Lopezfandino, A. S. Garea, D. B. Heras, and F. Arguello, "Stacked autoencoders for multiclass change detection in hyperspectral images," in *International Geoscience and Remote Sensing Symposium*, 2018, pp. 1906-1909.

[12] F. Huang, Y. Yu, and T. Feng, "Hyperspectral remote sensing image change detection based on tensor and deep learning," *Journal of Visual Communication and Image Representation*, vol. 58, pp. 233-244, 2019.

[13] H. Yu *et al*., "Global spatial and local spectral similarity-based manifold learning group sparse representation for hyperspectral imagery classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3043-3056, 2020.

[14] S. Jia, Z. Lin, B. Deng, J. Zhu, and Q. Li, "Cascade superpixel regularized Gabor feature fusion for hyperspectral image classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 5, pp. 1638-1652, May 2020.

[15] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph Convolutional Networks for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1-13, 2020.

[16] X. Li, Z. Yuan, and Q. Wang, "Unsupervised deep noise modeling for hyperspectral image change detection," *Remote Sensing*, vol. 11, no. 3, 2019.

[17] Q. Wang, Z. Yuan, Q. Du, and X. Li, "GETNET: A general end-to-end 2-D CNN framework for hyperspectral image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 3-13, 2019.

[18] A. Song, J. Choi, Y. Han, and Y. Kim, "Change detection in hyperspectral images using recurrent 3D fully convolutional networks," *Remote Sensing*, vol. 10, no. 11, 2018.

[19] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," in *European Conference on Computer Vision*, 2018, pp. 3-19.

[20] A. Vaswani *et al*., "Attention is all you need," in *Neural Information Processing Systems*, 2017, pp. 5998-6008.

[21] L. Chen et al., "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Computer Vision and Pattern Recognition*, 2017, pp. 6298-6306.

[22] Y. Zhu, O. Groth, M. S. Bernstein, and L. Feifei, "Visual7W: Grounded question answering in images," in *Computer Vision and Pattern Recognition*, 2016, pp. 4995-5004.

[23] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola, "Stacked attention networks for image question answering," in *Computer Vision and Pattern Recognition*, 2016, pp. 21-29.

[24] J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and J. Li, "Visual attention-driven hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 10, pp. 8065-8080, 2019.

[25] L. Mou and X. X. Zhu, "Learning to pay attention on spectral domain: A spectral attention module-based convolutional network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 1, pp. 110-122, 2020.

[26] F. Wang *et al*., "Residual attention network for image classification," in *Computer Vision and Pattern Recognition*, 2017, pp. 6450-6458.

[27] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Computer Vision and Pattern Recognition*, 2018, pp. 7132-7141.

[28] J. Park, S. Woo, J. Lee, and I. S. Kweon, "BAM: Bottleneck attention module," in *Computer Vision Pattern Recognition*, 2018, pp. 7132-7141.

[29] H. Sun, X. Zheng, X. Lu, and S. Wu, "Spectral-spatial attention network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1-14, 2019.

[30] W. Ma, Q. Yang, Y. Wu, W. Zhao, and X. Zhang, "Double-branch multi-attention mechanism network for hyperspectral image classification," *Remote Sensing*, vol. 11, no. 11, 2019.

[31] R. Hang, Z. Li, Q. Liu, P. Ghamisi, and S. S. Bhattacharyya, "Hyperspectral image classification with attention-aided CNNs," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1-13, 2020.

[32] J. Li *et al*., "Hyperspectral image super-resolution by band attention through adversarial learning," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1-15, 2020.

[33] Y. Cai, X. Liu, and Z. Cai, "BS-Nets: An end-to-end framework for band selection of hyperspectral image," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 3, pp. 1969-1984, 2020.

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

[35] Z. Lin, K. Ji, X. Leng, and G. Kuang, "Squeeze and excitation rank faster R-CNN for ship detection in SAR images," *IEEE Geoscience Remote Sensing Letters*, vol. 16, no. 5, pp. 751-755, 2019.

[36] S. Hao, W. Wang, Y. Ye, E. Li, and L. Bruzzone, "A deep network architecture for super-resolution-aided hyperspectral image classification

with classwise loss," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 8, pp. 4650-4663, 2018.

[37] S. Hao, W. Wang, Y. Ye, T. Nie, and L. Bruzzone, "Two-Stream Deep Architecture for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2349-2361, 2018.

[38] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 6232-6251, 2016.

[39] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Computer Vision and Pattern Recognition*, 2015, pp. 4353-4361.

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

[41] R. Hadsell, S. Chopra, and Y. Lecun, "Dimensionality reduction by learning an invariant mapping," in *Computer Vision and Pattern Recognition*, 2006, vol. 2, pp. 1735-1742.

[42] 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.

[43] W. A. Malila, "Change vector analysis: An approach for detecting forest changes with Landsat," in *Proc. LARS Symp.*, 1980, pp. 326-335.

[44] H. Nemmour and Y. Chibani, "Multiple support vector machines for land cover change detection: An application for mapping urban extensions," *Isprs Journal of Photogrammetry*, vol. 61, no. 2, pp. 125-133, 2006.

[45] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 6232-6251, 2016.

[46] Y. Li, H. Zhang, and Q. Shen, "Spectral–Spatial Classification of Hyperspectral Imagery with 3D Convolutional Neural Network," *Remote Sensing*, vol. 9, no. 1, 2017.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[48] M. Zhu, L. Jiao, F. Liu, S. Yang, and J. Wang, "Residual Spectral-Spatial Attention Network for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1-14, 2020.

[49] M. Song, X. Shang, and C.-I. Chang, "3-D Receiver Operating Characteristic Analysis for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 11, pp. 8093-8115, 2020.

[50] C.-I. Chang, "An Effective Evaluation Tool for Hyperspectral Target Detection: 3D Receiver Operating Characteristic Curve Analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 6, pp. 5131-5153, 2021.

[51] X. Wang, R. Girshick, A. Gupta, and K. He. "Non-local neural networks." in *Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[52] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: non-local networks meet squeeze-excitation networks and beyond," in *International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 1971-1980.

[53] J. Fu, J. Liu, H. Tian, Y. L, Y. Bao, Z. Fang, and H. Lu. "Dual attention network for scene segmentation," in *Computer Vision and Pattern Recognition*, 2019, pp. 7794–7803.

[54] Q. Wang, Y. Tang, X. Tong, and P. M. Atkinson, "Virtual image pair-based spatio-temporal fusion," *Remote Sensing of Environment*, vol. 249, 2020.

[55] Q. Wang, X. Ding, X. Tong, and P. M. Atkinson, "Spatio-temporal spectral unmixing of time-series images," *Remote Sensing of Environment*, vol. 259, 2021.

[56] E. L. Bullock, C. E. Woodcock, and C. E. Holden, "Improved change monitoring using an ensemble of time series algorithms," *Remote Sensing of Environment*, vol. 238, 2020.

[57] Z. Sun, L. Di, and H. Fang, "Using long short-term memory recurrent neural network in land cover classification on Landsat and Cropland data layer time series," *International Journal of Remote Sensing*, vol. 40, no. 2, pp. 593-614, 2018.

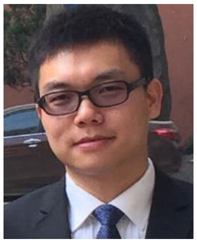
Lifeng Wang received the B.S. and M.S. degree from the Northeast Agricultural University, Harbin, China, in 2014 and 2017, respectively.

She is currently pursuing the Ph.D. degree in the College of Information and Communication Engineering Harbin Engineering University, Harbin, China.

Her research interests include hyperspectral image change detection and deep learning.

Liguo Wang received his M.A. degree in 2002 and Ph.D. degree in signal and information processing in 2005 from Harbin Institute of Technology, Harbin, China.

He held postdoctoral research position from 2006 to 2008 in the College of Information and Communications Engineering, Harbin Engineering University, where he is currently a Professor. His research interests are remote sensing image processing and machine learning. He has published three books, 25 patents, and more than 170 papers in journals and conference proceedings.



Qunming Wang received the Ph.D. degree from the Hong Kong Polytechnic University, Hong Kong, in 2015.

He is currently a Professor with the College of Surveying and Geo-Informatics, Tongji University, Shanghai, China. He was a Lecturer (Assistant Professor) with Lancaster Environment Centre, Lancaster University, Lancaster, U.K., from 2017 to 2018. His 3-year Ph.D. study was supported by the hypercompetitive Hong Kong Ph.D. Fellowship and his Ph.D. thesis was awarded as the Outstanding Thesis in the Faculty. He has authored or coauthored 60 peer-reviewed articles in international journals such as *Remote Sensing of Environment*, *IEEE Transactions on Geoscience and Remote Sensing*, and *ISPRS Journal of Photogrammetry and Remote Sensing*. His research interests include remote sensing, image processing, and geostatistics.

Dr. Wang serves as Associate Editor for *Science of Remote Sensing* (sister journal of *Remote Sensing of Environment*) and *Photogrammetric Engineering & Remote Sensing*, and was Associate Editor for *Computers and Geosciences* (2017−2020).

Peter M. Atkinson received the Ph.D. degree from the University of Sheffield (NERC CASE award with Rothamsted Experimental Station) in 1990. More recently, he received the MBA degree from the University of Southampton in 2012.

He is currently Distinguished Professor of Spatial Data Science and Dean of the Faculty of Science and Technology at Lancaster University, UK. He was previously Professor of Geography at the University Southampton, where he is currently Visiting Professor. He is also Visiting Professor at the Chinese Academy of Sciences, Beijing. He previously held the Belle van Zuylen Chair at Utrecht University, the Netherlands, is a recipient of the Peter Burrough Award of the International Spatial Accuracy Research Association and is a Fellow of the Learned Society of Wales. The main focus of his research is in remote sensing, geographical information science and spatial (and space-time) statistics applied to a range of environmental science and socio-economic problems. He has published over 300 peer-reviewed articles in international scientific journals and around 50 refereed book chapters. He has also edited nine journal special issues and eight books.

Professor Atkinson is Editor-in-Chief of Science of Remote Sensing, a sister journal of Remote Sensing of Environment. He also sits on the editorial boards of several further journals including Geographical Analysis, Spatial Statistics, International Journal of Applied Earth Observation and Geoinformation, and Environmental Informatics.