A Spectral and Spatial Attention Network for Change Detection in Hyperspectral Images

Maoguo Gong, *Senior Member, IEEE*, Fenlong Jiang, *Graduate Student Member, IEEE*, A. K. Qin, *Senior Member, IEEE*, Tongfei Liu, *Graduate Student Member, IEEE* , Tao Zhan, *Member, IEEE*, Di Lu, Hanhong Zheng, and Mingyang Zhang, *Member, IEEE*

***Abstract*— Hyperspectral images (HSIs) contain rich spectral signatures that reveal more image details and, thus, enable the detection of less noticeable changes on the ground. However, HSI-based change detection (CD) is susceptible to a large amount of irrelevant or noisy spectral and spatial information due to massive spectral bands. To address these issues, we propose a novel spectral and spatial attention network (S^2AN) for HSI-based CD, which is capable to suppress CD-irrelevant spectral and spatial information via adaptive spectral and spatial attention mechanisms. S^2AN takes as input the image patch from the difference map between two HSIs and outputs the status of change for the patch. Specifically, S^2AN is composed of several repeated attention blocks, each of which contains the spectral attention (SpeA) module for directly calculating the attention score for each input channel, the Gaussian spatial attention (GSpaA) module that first constructs an adaptive Gaussian distribution and then samples it to derive the attention scores for each spatial position, and the convolutional feature extraction (CFE) module for extracting features from the attention-weighted input. It is worth mentioning that, in addition to the advantage of the attention, GSpaA also reduces the sensitivity of patch size for patch-based methods. To effectively train S^2AN when facing insufficient labeled data, a semisupervised strategy that combines supervised and unsupervised methods to augment labeled training data is proposed. Experiments on several HSI datasets in comparison to existing methods show the superiority of S^2AN.**

***Index Terms*— Attention, change detection (CD), convolutional neural network, hyperspectral images (HSIs), semisupervised learning.**

I. Introduction

OVER the past few decades, it has become increasingly important to capture and understand the changes occurring on the Earth's surface [1]. With the rapid advance of remote sensing (RS) technology, numerous multitemporal RS images have been acquired from various platforms such as satellites, aircraft, and unmanned aerial vehicles [2]. As a crucial technique for detecting changes on the ground from these images covering the same geographic area but taken at different times [3], [4], image-based change detection (CD) has been widely used and plays a vital role in environmental monitoring [5], disaster assessment [6], urban sprawl analysis [7], resource management [8], and so on.

With the rapid development of spectrometry imaging techniques, hyperspectral images (HSIs) become more often utilized in earth observation applications. Unlike multispectral and synthetic aperture radar images, HSIs typically consist of hundreds of spectral bands, including the whole visible, the near-infrared, the short infrared, and the thermal infrared spectra, which can capture richer and finer details about ground objects [9]. Therefore, HSIs are more suitable for CD tasks that contain less noticeable changes. However, HSI-based CD has the following major challenges.

1) From the spectral perspective, HSIs can capture rich spectral information via many narrow spectral bands [7], [10]. However, this inevitably introduces a lot of information that is irrelevant (and thus harmful) to CD [11], [12].
2) CD often uses neighborhood information, but not all pixels within the neighborhood contribute equally and positively to CD. Especially, each pixel in HSIs corresponds to many spectral channels, and thus, those pixels irrelevant to CD may more severely influence the performance of CD due to some accumulation operations, such as convolution [13]–[15].
3) It is very typical that there may lack sufficient labeled training data in HSI-based CD. Therefore, it is important to make full use of both labeled and unlabeled data for model training [16].

Manuscript received December 12, 2021; accepted December 24, 2021. Date of publication December 28, 2021; date of current version March 8, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62036006 and Grant 61906147, in part by the Australian Research Council (ARC) under Grant LP180100114 and Grant DP200102611, in part by the Natural Science Basic Research Program of Shaanxi Province under Grant 2021JQ-195, and in part by the Fundamental Research Funds for the Central Universities and the Innovation Fund of Xidian University. *(Corresponding author: Maoguo Gong.)*

Maoguo Gong, Fenlong Jiang, Tongfei Liu, Di Lu, Hanhong Zheng, and Mingyang Zhang are with the School of Electronic Engineering, Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University, Xi'an, Shaanxi 710071, China (e-mail: gong@ieee.org; jiangfenlong@outlook.com; liutongfei_home@hotmail.com; di_lu@stu.xidian.edu.cn; hanhong_zheng@163.com; omegazhangmy@gmail.com).

A. K. Qin is with the Department of Computing Technologies, Swinburne University of Technology, Hawthorn, VIC 3122, Australia (e-mail: kqin@swin.edu.au).

Tao Zhan is with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Computer Science and Technology, Xidian University, Xi'an 710071, China (e-mail: omegazhant@gmail.com).

Digital Object Identifier 10.1109/TGRS.2021.3139077

To address these challenges, we propose a novel spectral and spatial attention network (S^2AN) for CD in HSIs, which can suppress irrelevant and noisy information and accordingly enhance the information more relevant to CD via adaptive spectral and spatial attention mechanisms. Specifically, S^2AN takes as input the patches of the difference map derived from the two HSIs capturing the same area at different times and utilizes multiple repeated attention blocks to gradually enhance CD-relevant features, followed by several fully connected layers and a softmax layer to output the change status of the patch, reflecting whether the central pixel of the patch is changed or not.

The attention block that plays a key role in S^2AN contains three modules, i.e., spectral attention (SpeA), Gaussian spatial attention (GSpaA), and convolutional feature extraction (CFE). Here, SpeA and GSpaA are two attention modules specifically designed for deriving the spectral and spatial attention scores, respectively. Given the input, SpeA first uses a global convolutional layer to aggregate the spatial information in each channel of the input. Then, it uses several fully connected layers to calculate an attention score for each input channel. GSpaA first uses a 1×1 convolution to aggregate spectral information in each spatial location of the input and then calculates the standard deviation of a Gaussian distribution by propagating the aggregated spatial features through some fully connected layers, and the attention score for each pixel position can be obtained via sampling the 2-D Gaussian distribution with the calculated standard deviation. The attention scores derived from SpeA and GSpaA are applied on each channel and spatial position of the input to weigh the spectral and spatial information for the CD task. However, the original feature level of data is not changed. The attention-weighted input is then fed into the CFE that is composed of convolutional layers for further feature extraction. It is worth mentioning that we introduce the Gaussian sampling in GSpaA because the area closer to the central pixel can better depict the image features around the central pixel. This also brings an additional benefit that it can reduce the sensitivity of the patch-based method for the selection of the patch size and avoid fine-tuning the patch size [17], [18].

In order to effectively train the proposed model in the absence of sufficient labeled data, which is very common in HSI-based CD, we propose a simple but effective semisupervised strategy, which combines the traditional unsupervised and supervised algorithms to augment the labeled data for training. The main contributions of our work are summarized as follows.

1) We propose a novel spectral and spatial attention network, called S^2AN for CD in HSIs, which can extract CD-relevant information while suppressing CD-relevant information via the three key modules, i.e., spectral attention, Gaussian spatial attention, and convolutional feature extraction.
2) We design spectral and spatial attention modules to derive attention scores from the input feature maps. The derived spectral and spatial attention scores are imposed via weighting on the channels and pixels of the input, respectively, to enhance CD-relevant information. For spatial attention, a Gaussian sampling-based scheme is adopted to reduce the sensitivity of the patch-based method to the patch size, so as to avoid multiple attempts to select the appropriate patch size.
3) We devise a semisupervised training strategy, where the supervised and unsupervised algorithms are combined to augment limited labeled data, to handle the issue of insufficient labeled data.

The rest of this article is organized as follows. Section II introduces some related work. Section III describes the proposed approach in detail. In Section IV, experimental results are reported and discussed. The conclusion and future work are in Section VI.

II. Related Work

A. Deep Learning Based CD in HSIs

With the rapid development and widespread application in image processing recently [19]–[21], deep neural networks (DNNs), particularly convolutional neural networks (CNNs), are increasingly favored by the CD community. Some methods are extensions of traditional methods. For example, the methods in [22] and [23] extend the classic change vector analysis [24], [25] and slow feature analysis [26] methods into the field of deep learning and propose deep change vector analysis (DCVA) and deep slow feature analysis (DSFANet), respectively. Others prefer to design a direct mapping from images to changes, such as the methods in [27]–[31].

However, most of the methods mentioned above are for multispectral images. In recent years, due to the unique characteristics of HSIs, many deep learning-based CD methods for HSIs have emerged. In terms of improving representation ability, due to the 3-D characteristics of HSIs, some methods combine 3-D convolutional layers to avoid information loss caused by 2-D convolution, so as to improve the performance [32]–[35]. In addition, other methods also treat high-dimensional hyperspectral data as tensors and use tensor-based support tensor machine and restricted Boltzmann machine [36], and tensor network [37] to directly process the data. In order to reduce the interference in HSIs, Li *et al.* [8] proposed an unsupervised deep noise modeling method, which assumes that the noise obeys Gaussian distribution and accurately fits the specific distribution through KL divergence. By integrating the noise reduction results of various traditional methods, the final performance is improved. Moreover, drawing on the idea of ensemble learning, a supervised method was proposed in [38], whose innovation lies in that it combines traditional binary segmentation techniques to filter out unchanged areas, avoiding the adverse impact of its variance.

These abovementioned approaches all achieve "state of the art" on their respective target problems. However, they only consider the macroscopic spatial–temporal expression while do not analyzing the detailed information inside the HSIs. Therefore, some studies have examined this more detailedly. Marinelli *et al.* [39] paid attention to exploiting the change information present in each band. With a lot of human intervention and manual statistical information retrieval, each spectral band can detect useful change information adaptively. Another approach called GETNET presented in [16] takes a more automated way. It first performs hyperspectral unmixing and then represents each pixel via a feature adjacency matrix

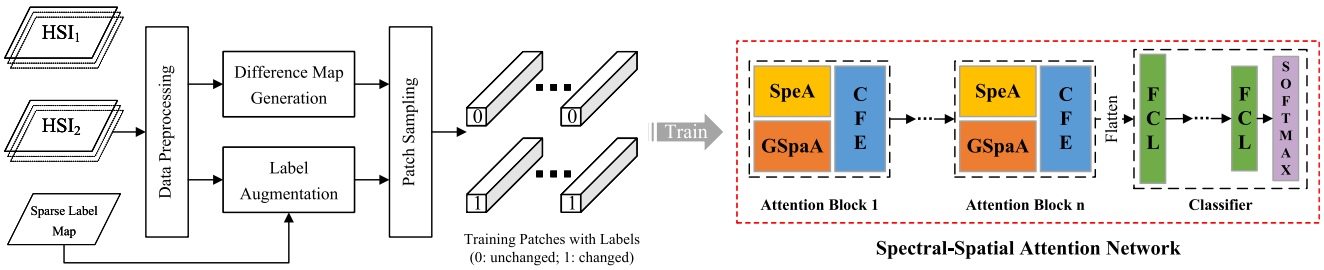


Fig. 1. Flowchart of the proposed method for CD in HSIs. First, two original HSIs need to be preprocessed. Then, the processed HSIs are compared to generate a difference map. At the same time, they combine the real sparse label map to augment the labels. After that, patch sampling is carried out according to the augmented labels to form the training set, where 0 is the unchanged class and 1 is the changed class. Finally, the constructed training set is input into the designed S^2AN for training.

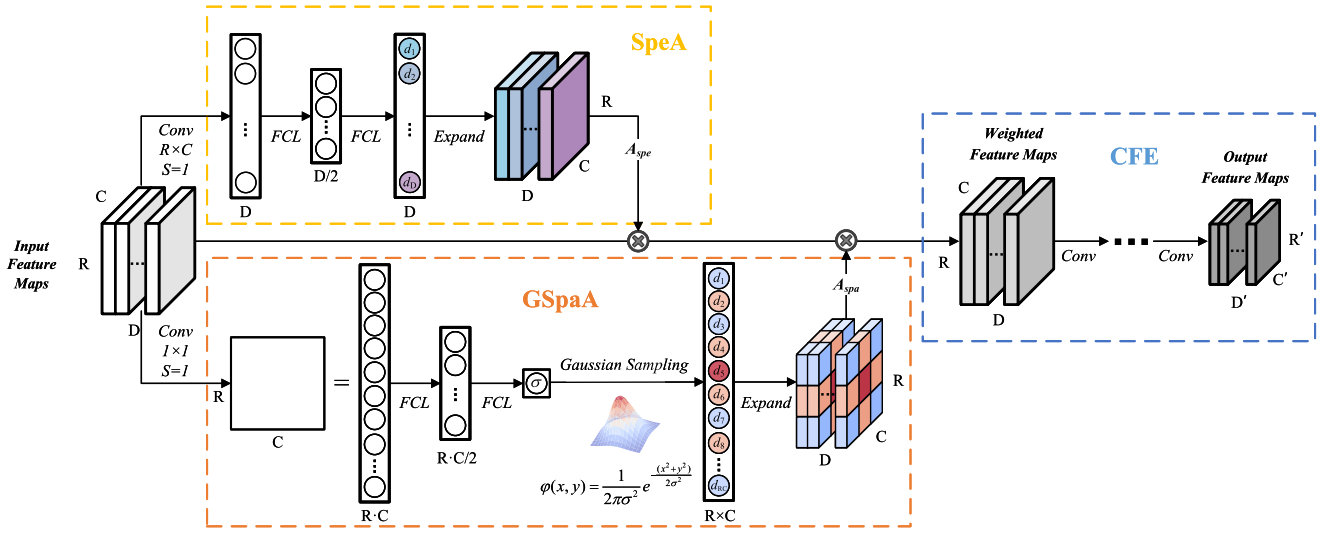


Fig. 2. Diagram of the spectral-spatial attention block, which is composed of three modules, i.e., SpeA, GSpaA, and CFE. Given an input to the block, e.g., feature maps, SpeA and GSpaA first aggregate spectral and spatial features via the convolutional layers. Then, SpeA calculates the attention score of each input channel through several fully connected layers, while GSpaA calculates the standard deviation of a 2-D Gaussian distribution and then samples this Gaussian distribution to obtain the attention score of each spatial position. Finally, the calculated attention scores are applied to each channel and spatial position of the input. The attention-weighted inputs are further extracted by CFE via several convolutional layers.

on the spectrum. A CNN is employed to further extract the features of this matrix and determine whether the corresponding pixel is changed. However, the above two methods only consider the spectral domain and do not take the spatial perspective into account. Similar to GETNET, the method proposed in [40] utilizes unmixing to analyze the change of spectral information in subtle granularity but also incorporates spatial structure information obtained by the principal component analysis. In addition, a morphological attribute profile-based method was proposed to extract meaningful morphological features, and the final changes were analyzed by combining spectral information [41]. Zhan *et al.* [35] used the designed 3-D convolutional network to extract the features of the difference image from two spatial directions and one spectral direction respectively, taking into full consideration the data characteristics of hyperspectral 3-D cubes.

B. Attention in CD

By simulating the behavior and mechanism of the human visual system, attention has been widely applied in many fields [42], [43], including, but not limited to, computer vision [44], [45], natural language processing, and presentation learning [46]. At the specific application level, an attention mechanism can be used to improve the sensitivity to features containing important information, so as to highlight useful information, suppress noise, and reduce redundancy.

In recent years, attention-based models have also become increasingly popular in RS applications, such as scene classification [47]–[49], object detection [50]–[52], and building extraction [53]–[55]. At the same time, some methods, albeit few, have emerged for CD. Liu *et al.* [56] considered the information correlation of two temporal images and used the attention mechanism to weigh different temporal features. In [57], a coattention module was designed to emphasize the parts of the image pairs that are significantly different and reduce sensitivity to similar parts. In addition, Long *et al.* [58] devised an attention module for the difference feature maps of different levels, and each element in the feature maps is weighted from both the space and the spectrum. Similarly, in the upsampling stage of the end-to-end U-Net designed by Peng *et al.* [59], they also applied an attention mechanism

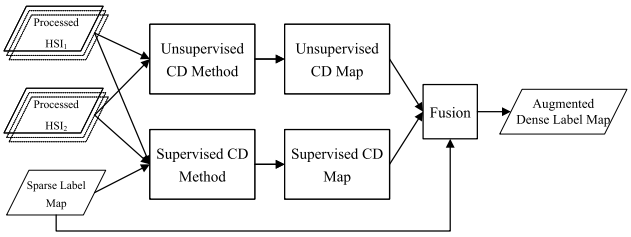


Fig. 3. Diagram of the label augmentation module. First, an unsupervised CD method is used to analyze the two processed HSIs and generate the corresponding unsupervised CD map. Then, combined with a real sparse label map, a supervised CD map is generated by a supervised method. Finally, the augmented dense label map can be obtained by fusing the two result maps and the sparse label map.

to the elements in the multilevel feature maps. In [31], apart from the spatial and spectral attention mechanisms, the proposed method also designs a double-margin contrastive loss to improve attention to the changed region. In brief, the attention mechanism is an effective tool for further screening and weighting features. It is necessary to make a comprehensive analysis from both spatial and spectral perspectives.

III. Proposed Method

To address the issues of irrelevant information interference and effective training with limited labeled HSI data, we propose the S^2AN for CD in HSIs, which can suppress irrelevant information via adaptive spectral and spatial mechanisms. As shown in Fig. 1, S^2AN is composed of one or more attention blocks with the same architecture, followed by several fully connected layers and a softmax layer to determine whether the central pixel of the input patch is changed. The detailed structure of the attention block featured by S^2AN is illustrated in Fig. 2. It can be seen that the block is composed of three modules, i.e., SpeA, GSpaA, and CFE. Given an input to the block, e.g., feature maps, SpeA and GSpaA first aggregate spectral and spatial features, respectively. Then, SpeA calculates the attention score of each input channel through several fully connected layers, while GSpaA calculates the standard deviation of a 2-D Gaussian distribution and then samples this Gaussian distribution to obtain the attention score of each spatial position. Finally, the calculated attention scores are applied to each channel and spatial position of the input. The attention-weighted inputs are further extracted by CFE via several convolutional layers. In the following, we will introduce the process of our whole method in detail.

A. Data Preprocessing

1) Preprocessing of HSI Data: Before comparing and analyzing the two HSIs, it is essential to perform some preprocessing for them [60]. Taking Earth Observing-1 Hyperion images as examples, they have to go through fixing bad and outlier pixels, local destriping, atmospheric correction, minimum noise fraction smoothing, and so on before being used for other tasks [61]. Moreover, for the CD task, precise geometric coregistration is required, which is instrumental in avoiding false detection caused by image misalignment. In the end, the coregistered images need to be normalized to meet the subsequent computing requirements.

2) Sparse Label Map Generation: Since efficient CD in complex scenarios cannot be separated from the guidance of supervised information, limited manual labeling that only consumes a small amount of labor and time cost is necessary. In order to simulate this kind of labeling in the experimental environment, we randomly sampled a few real samples from the reference map of the HSI dataset to generate a real sparse label map.

B. Training Data Generation

1) Difference Map Generation: In order to make the model more targeted for hyperspectral differences, here, the difference map I_D is utilized as the processing object of the subsequent network, which can be calculated by the following formula:

$$I_D = |I_1 - I_2| \tag{1}$$

where I_1 and I_2 represent two processed images to be detected.

Then, considering the small scale of data, CD can be regarded as the pixel-by-pixel classification of images, that is, patch-based paradigm, which usually analyzes the patch image to judge the category of the central pixel. Therefore, the neighborhood of each pixel in I_D is cut into multiple patches as input data.

2) Label Augmentation: The method of supervised learning uses labeled training data to train without considering the inherent characteristics of the data itself, so its performance is often restricted by the amount of training data. However, unsupervised methods do not introduce supervised information and only analyze data by mining inherent characteristics, which saves labor costs, but often fails to fully fit complex data scenes. Therefore, we adopt a semisupervised learning strategy in this article, which not only utilizes the limited annotations but also combines the characteristics of the data itself. Specifically, traditional unsupervised and supervised approaches are combined to augment the labeled data. As shown in Fig. 3, a classical unsupervised algorithm, for example, CVA, is employed to generate pseudolabels for all pixels, which is represented by R_u. Then, with the real sparse label map L_t mentioned in Section III-A2, a traditional supervised algorithm is used to generate pseudolabels for unlabeled pixels, and the whole map including L_t is represented by R_s. Finally, joint analysis of the two results can screen out more reliable pseudolabels. It is obvious that only the part with the same detection category is trustworthy. That is, for a two-class classification problem, only samples with either 0 or 1 in both of R_u and R_s can be selected for training. In this way, the pseudolabels L_p obtained can be represented by the XNOR operation [$\odot$] of R_u and R_s, that is,

$$L_p = R_u \odot R_s. \tag{2}$$

However, it is worth noting that this process may incorrectly filter out the real labels in L_t. Therefore, the final augmented dense label map, i.e., the training labels $\mathbf{y}$, needs to fuse these real labels, which can be expressed as the union set of L_p and L_t, that is,

$$\mathbf{y} = L_p \cup L_t. \tag{3}$$

3) Training Patch Sampling: According to the augmented training labels $\mathbf{y}$ generated above, the corresponding patches are selected to construct a training set for training the designed network.

C. S^2AN

The data prepared in the previous section can be fed into the designed S^2AN for training and testing, as shown in Fig. 1. It can be seen that S^2AN is mainly composed of several repeated spectral–spatial attention blocks, each of which contains the SpeA module, GSpaA module, and CFE module. These modules are introduced in detail in the following.

1) SpeA: The coupled spatial and spectral information in the patch affects each other. In order to make the spectral attention more targeted, we first extract the aggregated feature of spatial information. To achieve this, a global convolution operation is employed. Formally, given the feature maps $\mathbf{x} \in \mathbb{R}^{R\times C\times D}$, the SpeA first utilizes D convolution kernels $\tilde{\kappa} = \{\tilde{\kappa}_1, \tilde{\kappa}_2, \ldots, \tilde{\kappa}_d, \ldots, \tilde{\kappa}_D | \tilde{\kappa}_d \in \mathbb{R}^{R\times C\times D}\}$ to extract and aggregate the spatial information of each channel, where R, C, and D represent the height, width, and depth of the feature maps or kernels, respectively. This will result in the feature maps $\tilde{\mathbf{v}}$ of size D, whose the dth channel value $\tilde{v}_d$ can be calculated as

$$\tilde{v}_d = \mathbf{x} * \tilde{\kappa}_d \tag{4}$$

where $*$ defines the convolution operation.

Then, two sequential fully connected layers are employed to calculate the attention score for each channel. It is worth noting that the attention score should be bounded, and the weighted feature should maintain the same magnitude as the original one. The *Sigmoid* (ς) activation function exactly meets these requirements. Therefore, the spectral attention scores can be expressed as

$$\mathbf{A_{spe}} = \varsigma(\tilde{\mathbf{w}}^\top \tilde{\mathbf{v}} + \tilde{b}) = \frac{1}{1 + e^{\tilde{\mathbf{w}}^\top \tilde{\mathbf{v}} + \tilde{b}}} \tag{5}$$

where $\tilde{\mathbf{w}}$ and $\tilde{b}$ are the weight and bias of the last fully connected layer.

Before being applied to $\mathbf{x}$, the score of each channel needs to be expanded to each spatial position of the channel, as depicted in the top half of Fig. 2. Finally, the weighted feature maps can be obtained by multiplying the corresponding elements of $\mathbf{A_{spe}}$ and $\mathbf{x}$, as shown in the following equation:

$$\tilde{\mathbf{x}} = \mathbf{x} \circ \mathbf{A_{spe}} \tag{6}$$

where $\circ$ denotes the *Hadamard Product* and $\tilde{\mathbf{x}}$ is the spectrally weighted feature maps.

2) GSpaA: In view of the lack of labeled HSI CD data, the patch-based method is often considered, which establishes a mapping from the neighborhood around a pixel to the category of the pixel, which takes the neighborhood information into account and, thus, improves the robustness of detection. However, it is coarse and inappropriate to treat the pixels at each position in the patch equally. In fact, The pixels closer to the central pixel can better depict the image features around the central pixel. We expect our spatial attention module to be able to adaptively control the size of the region most relevant to detecting the change of the central pixel.

To tackle these problems, the 2-D Gaussian kernel function is considered to be introduced into the spatial attention mechanism because it satisfies the property that the value decreases from the center toward the periphery. Formally, we use the 2-D Gaussian function $\varphi(p, q)$ to represent the correlation between the pixel and the change of the central pixel, which is expressed as

$$\varphi(p, q) = \frac{1}{2\pi\sigma^2} e^{-\frac{(p^2+q^2)}{2\sigma^2}}. \tag{7}$$

where (p, q) represents the spatial coordinates of a pixel, and σ is the standard deviation. At this time, σ becomes the only parameter that determines the shape of the function, thereby controlling the size of the relevant region. Therefore, the neural network can be used to learn a corresponding σ for each patch, so as to realize adaptive control.

Back to the spatial attention module, such as SpeA, GSpaA needs to be more specific to spatial information, so it also needs to condense spectral information for each spatial position first. This process can be implemented by 1 convolution kernel $\hat{\kappa}$ of size $1 \times 1 \times D$, outputting a feature map $\hat{\mathbf{m}}$ of size $R \times C$, which can be calculated by the following equation:

$$\hat{\mathbf{m}} = \mathbf{x} * \hat{\kappa}. \tag{8}$$

Similar to SpeA, two fully connected layers are used to calculate attention scores, except that, in order to satisfy the form of the input, $\hat{\mathbf{m}}$ needs to be first drawn into a 1-D vector $\hat{\mathbf{v}}$. Moreover, considering that the nonnegative property of the standard deviation and the negative output values should also be activated efficiently, the Softplus activation function (ζ) is adopted, that is,

$$\hat{\sigma} = \zeta(\hat{\mathbf{w}}^\top \hat{\mathbf{v}} + \hat{b}) = \log\left(1 + e^{\hat{\mathbf{w}}^\top \hat{\mathbf{v}} + \hat{b}}\right) \tag{9}$$

where $\hat{\mathbf{w}}$ and $\hat{b}$ are the weight and bias of the last fully connected layer.

Next, the Gaussian function $\hat{\varphi}(p, q)$ based on the obtained $\hat{\sigma}$ can be directly generated. By sampling $\hat{\varphi}(p, q)$, the contribution of each position pixel to CD can be obtained. However, it should be noted that, in the actual calculation, the attention scores should range from 0 to 1. Thus, the further normalization is implemented. Specifically, the normalized spatial attention score $\mathbf{A_{spa}}(p, q)$ at position (p, q) can be calculated as

$$\mathbf{A_{spa}}(p, q) = \frac{\hat{\varphi}(p, q)}{\hat{\varphi}(0, 0)} = \frac{\frac{1}{2\pi\hat{\sigma}^2} e^{-\frac{(p^2+q^2)}{2\hat{\sigma}^2}}}{\frac{1}{2\pi\hat{\sigma}^2}} = e^{-\frac{(p^2+q^2)}{2\hat{\sigma}^2}} \tag{10}$$

where $\hat{\varphi}(0, 0)$ is the max of $\hat{\varphi}$ because $\hat{\varphi}$ is monotonically decreasing.

The resulting attention score $\mathbf{A}_{\mathbf{spa}}(p, q)$ for each position first needs to be expanded to all spectral channels. Then, the spatially weighted feature maps $\hat{\mathbf{x}}$ can be calculated by multiplying the corresponding elements of $\mathbf{x}$ and $\mathbf{A}_{\mathbf{spa}}$, that is,

$$\hat{\mathbf{x}} = \mathbf{x} \circ \mathbf{A}_{\mathbf{spa}}. \tag{11}$$

3) Convolution Feature Extraction: So far, we have recalibrated the weight of the feature maps from the spectral and spatial perspectives. However, the level of the feature maps remains unchanged. Therefore, it is necessary to further extract deep abstract features via convolution operations. The calculation process is similar to that of (4) and (8), but the used convolution kernels are in different numbers and sizes.

In this way, multiple spatial and spectral attention modules can be stacked to weight and extract features at different scales. Finally, a fully connected classifier is employed to analyze the features to categorize the input patch.

D. Training and Testing Process

1) Training S^{2}AN: In the case of limited labeled samples, a semisupervised training strategy is used for our proposed network, which utilizes both labeled and unlabeled data. Specifically, label augmentation is used to generate pseudolabeled data, and the network is trained with both real labeled and pseudolabeled data, as described in detail in Section III-B.

Given the high confidence augmented training dataset $\mathcal{D}_{\text{train}}$ containing input–target pairs $(\mathbf{x}, \mathbf{y})$, training the S^{2}AN can be seen as minimizing a loss $\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}})$ between the model output $\hat{\mathbf{y}}$ and the targets. For general classification problems, the cross entropy is often considered, that is,

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{i} \mathbf{y} \log(\hat{\mathbf{y}}). \tag{12}$$

Here, the stochastic gradient descent (SGD) optimizer is utilized to update network parameters and reduce the loss.

2) Testing S^{2}AN: After several training iterations, the loss will converge, and the parameters will reach the optimal, revealing that S^{2}AN is well trained. During the test, the difference map of HSI pairs is generated first, and then, patches are fed into the well-trained S^{2}AN for prediction. Finally, the change state of each pixel can be determined, thus forming the final change map.

IV. Experiments

In this section, we will conduct an experimental analysis of our algorithm on several public real datasets. First, the datasets and experimental setup are introduced. Then, for verifying the advantages of the proposed algorithm, we use several representative CD algorithms for comparison experiments. After that, to prove that the proposed algorithm is robust to the patch size, we analyzed its sensitivity under several different patch sizes. Finally, in order to prove the role of the semisupervised training strategy and the spatial and spectral attention modules, we conducted a rigorous ablation experiment and interpretability analysis on them.

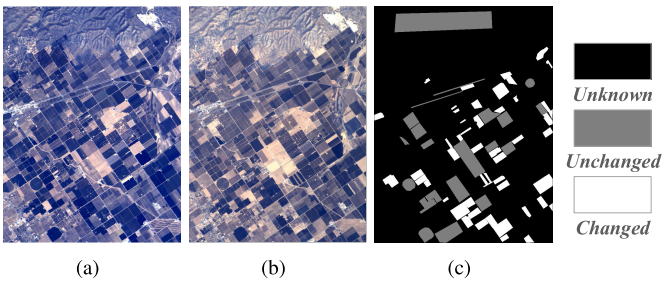


Fig. 4. Santa Barbara dataset. (a) Image acquired in 2013. (b) Image acquired in 2014. (c) Reference map.

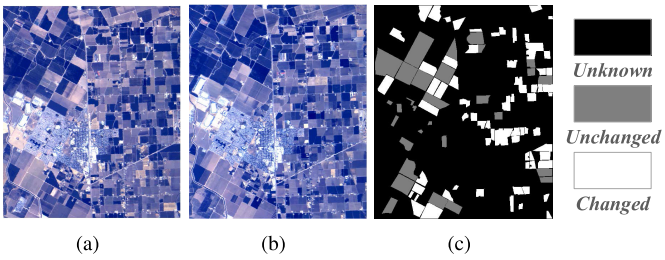


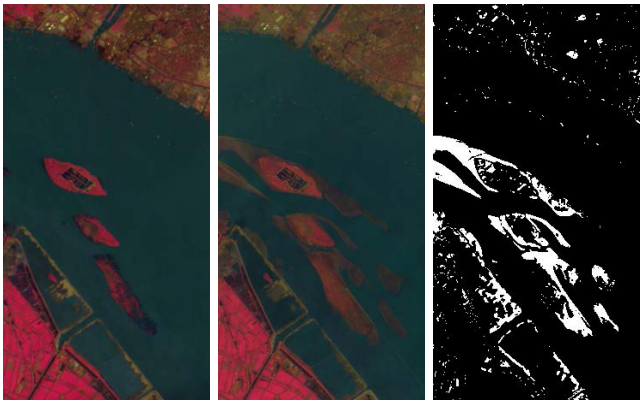
Fig. 5. Bay Area dataset. (a) Image acquired on 2013. (b) Image acquired on 2015. (c) Reference map.

A. Datasets

In order to verify the performance of the proposed method in a real environment, three publicly available HSI CD datasets are used in the experiments, including the *Santa Barbara* dataset, the *Bay Area* dataset, and the *River* dataset. Among them, the first two datasets were published by Lopez-Fandino *et al.* [38], and the last dataset was released in [16]. The details of these datasets are given as follows.

1) Santa Barbara and Bay Area: As shown in Fig. 4, the two HSIs in the Santa Barbara dataset were taken in 2013 and 2014, respectively, via the AVIRIS sensor over the Santa Barbara region in California. The dimensions of the two images are both 984 × 740 pixels with 224 spectral bands. Similarly, with the same sensor hovering above the city of Patterson, California, two scenes of the Bay Area were captured in 2013 and 2015, respectively, as shown in Fig. 5(a) and (b), whose sizes are both 600 × 500 × 224. The reference ground truth of the two datasets is separately shown in Figs. 4(c) and 5(c), where the black areas are unknown and not of interest, and the gray and white parts correspond to the changed and unchanged areas, respectively. These two datasets both reflect the land surface changes caused by buildings and crops, which is of great significance to the study of urban expansion and arable land use.

2) River: The third dataset "River" is shown in Fig. 6, where Fig. 6(a) and (b) shows the pseudocolor images of the two HSIs captured on May 3, 2013, and December 31, 2013, respectively, in Jiangsu province, China, while Fig. 6(c) shows the ground-truth image for reference. The sizes of the two HSIs are 463 × 241 with 198 spectral bands after removing the noise bands. This dataset mainly reflects the changes in river water, which can help decision support in applications such as flood monitoring and water regulation. In addition, this dataset contains changes with multiple scales at the same time,



(a) (b) (c)

Fig. 6. River dataset. (a) Image acquired on May 3, 2013. (b) Image acquired on December 31, 2013. (c) Reference map.

which can test the adaptability and robustness of the model to changing scales.

B. Experimental Setup

1) Evaluation Criteria: In order to effectively evaluate the detection performance of the proposed and several compared baseline algorithms, some classical evaluation criteria are employed in our experiments. According to the reference image, we first calculate the confusion matrix of the results, where four values need to be calculated, i.e., true positive (TP), true negative (TN), false positive (FP), and false negative (FN), which separately defines the number of pixels that are correctly classified as a changed class, correctly classified as an unchanged class, misclassified as a changed class, and misclassified as a changed class.

According to the confusion matrix, we can, thus, calculate the following evaluation criteria.

1) The overall accuracy (OA) is the proportion of all correctly classified positive and negative samples to all samples

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}. \tag{13}$$

2) The overall error (OE) refers to the number of all samples that are misclassified

$$\mathrm{OE} = \mathrm{FP} + \mathrm{FN}. \tag{14}$$

3) Due to the imbalance of some categories, the accuracy rate cannot accurately reflect the consistency of the actual prediction results. In contrast, the kappa coefficient (κ) [62] can better reflect the global consistency, and it can be calculated by the following formulas:

$$\kappa = \frac{\mathrm{OA} - \mathrm{PRE}}{1 - \mathrm{PRE}} \tag{15}$$

$$\mathrm{PRE} = \frac{(\mathrm{TP} + \mathrm{FP}) \cdot \mathrm{RC} + (\mathrm{TN} + \mathrm{FN}) \cdot \mathrm{RU}}{(\mathrm{FP} + \mathrm{FN} + \mathrm{TP} + \mathrm{TN})^2} \tag{16}$$

where RC and RU indicate the number of real changed and unchanged pixels in the reference image, respectively. The larger the value of kappa is, the better detection consistency one method obtains.

TABLE I
IMPLEMENTATION DETAILS OF SPEA, GSPAA, AND CFE MODULES

Module	Layer	Size
SpeA	Conv	$R \times C \times D$
	BatchNorm	D
	ReLU	-
	Linear	$D/2$
	BatchNorm	$D/2$
	ReLU	-
	Linear	D
	Sigmoid	-
GSpaA	Conv	$1 \times 1 \times 1$
	BatchNorm	1
	ReLU	-
	Linear	$D/2$
	BatchNorm	$D/2$
	ReLU	-
	Linear	1
	BatchNorm	1
	Softplus	-
CFE	Conv	$3 \times 3 \times c$, $c = \{512, 256, 128, 64, 32\}$
	BatchNorm	c
	ReLU	-

4) In statistics, in order to effectively measure the accuracy of the binary classification model, F_1 score is often considered, which takes into account the precision and recall of the model at the same time and can be regarded as the harmonic mean of them. It can be calculated as

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{17}$$

where $P = (\mathrm{TP}/(\mathrm{TP} + \mathrm{FP}))$ and $R = (\mathrm{TP}/(\mathrm{TP} + \mathrm{FN}))$ represent the precision and recall rate, respectively. Note that the value of F_1 ranges from 0 to 1; the higher the value is, the better detection result one method achieves.

2) Implementation Details: In the experiments, the proposed S^{2}AN was trained on an Nvidia RTX 3090 GPU with PyTorch [63] implementation. In the implementation process, the sizes of some layers of the proposed attention module need to be changed according to the size of the input feature maps. Therefore, assuming that the input feature size is $R \times C \times D$. At this point, the detailed structures of SpeA, GSpaA, and CFE corresponding to this input can be shown in Table I. These modules are stacked five times to extract deep features. The final classification is then performed using the fully connected layers of sizes 256 and 2 to obtain the final change map. Specifically, the computational cost of S^{2}AN is approximately 1.2442G multiply–accumulate operations (MACs), which is calculated by the Pytorch-based THOR library.

During the training, the SGD optimizer with the momentum of 0.5 and the weight decay of 0.001 was employed to update the parameters. In addition, we also applied a multiplicative learning rate reduction strategy to ensure the stable convergence of the model at the later stage, where the

initial learning rate was set to 0.001. We used a minibatch training fashion, and the batch size was set to 32. Besides, each forward propagation process of our method takes about 10.5 ms, in which GSpaA, SpeA, CFE, and others account for 77.5%, 15.5%, and 7%, respectively. It can be seen that the attention mechanism compared with the pure feature extraction part consumes more time; especially, the GSpaA module due to the sampling process requires more computation. However, from the view of the final performance improvement, this additional time is worthwhile. Because the overall speed is very fast, in actual use, the attention mechanism will not increase the consuming time.

C. Comparison Results and Analysis

In order to verify the superiority and effectiveness of the proposed method, some classical methods are selected as compared algorithms, which are summarized as follows.

1) CVA, which compares spectral vector differences and utilizes a threshold segmentation to get the change map [25].
2) RCVA, which introduces neighborhood information on the basis of CVA to improve the robustness of detection [64].
3) KNN, a very classic supervised learning method, the principle of which is that, when predicting the category of a new sample, it is judged according to the category of the nearest K samples. In our experiment, it was used to introduce supervised information into the pseudolabel method.
4) SVM, a support vector machine model that uses a small number of samples for supervised training.
5) DCVA, which introduces the idea of transfer learning, employs the model pretrained by other datasets and provides the deep representations for each pixel by upsampling and concatenating the features of multiple scales. Then, the change map can be obtained by thresholding the differences between these representations [22]. It should be noted that, in order to fit the number of input channels, the dimensionality reduction transformation is first performed on the two HSIs via the principal component analysis (PCA).
6) DSFANet, which is a representative slow feature analysis method based on deep learning. It first transforms the images into a nonlinear space through a neural network and then uses the transformed features to detect slow feature changes [23].
7) GETNET, which designs a better matrix representation for each pixel in an HSI combined with unmixing and processes it through the convolutional neural network to improve detection performance [16].
8) TDSSC, which takes spatial and spectral information into full consideration by processing data from one spectral direction and two spatial directions [35].
9) CNN, which is a simplified version of the proposed model without the attentional mechanism. It is used in ablation experiments to analyze the specific role of the attention modules.

TABLE II
SEMISUPERVISED TRAINING SAMPLE SELECTION IN THE EXPERIMENT

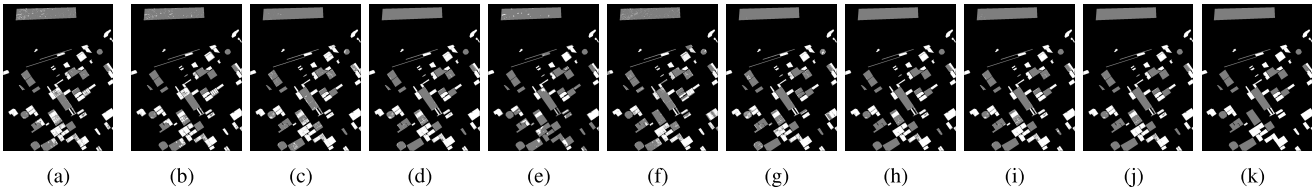
Dataset	Class	Ground Truth	Training	Ratio(%)
Barbara	Changed	52134	50	0.0959
	Unchanged	80418	50	0.0622
Bay	Changed	38425	50	0.1301
	Unchanged	34211	50	0.1462
River	Changed	9698	50	0.5156
	Unchanged	101885	50	0.0491

TABLE III
QUANTITATIVE EVALUATION OF EXPERIMENTAL RESULTS OBTAINED BY DIFFERENT METHODS ON THE SANTA BARBARA DATASET

Algorithm	OA	OE	κ	F_1
CVA	0.8712	17078	0.7320	0.8396
RCVA	0.8674	17576	0.7226	0.8322
KNN	0.9102	11898	0.8122	0.8864
SVM	0.9321	8999	0.8568	0.9120
DCVA	0.7921	27564	0.5313	0.6696
DSFANet	0.8676	17549	0.7174	0.8224
GETNET	0.8967	13688	0.7802	0.8627
TDSSC	0.9329	8891	0.8570	0.9104
KNN+CVA+S^{2}AN	**0.9346**	**8664**	**0.8621**	**0.9152**
KNN+RCVA+S^{2}AN	0.9326	8931	0.8566	0.9105

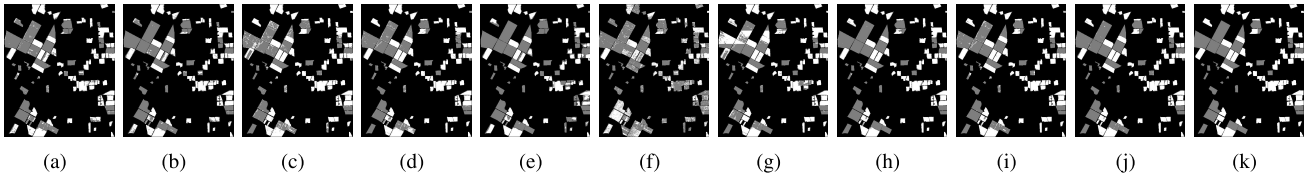
In the comparative experiment, we only used 50 changed and 50 unchanged samples randomly selected from the reference map for all methods. The specific selection ratio of each dataset is shown in Table II. For the sake of fairness, several deep learning algorithms, including the proposed S2AN, GETNET, and TDSSC, have adopted the proposed sample selection strategy. In addition, since the proposed algorithm is highly robust to the patch size, we choose an empirical rough value of 15.

1) Results on Santa Barbara and Bay Area Datasets: Because of the same sensor and similar coverage scenarios, the results on the Santa Barbara and Bay Area datasets, as separately shown in Figs. 7 and 8, will be discussed together, and the quantitative evaluations are displayed in Tables III and IV, respectively. As we expected, many spots are observed in the results of CVA and RCVA, which indicates that simple numerical calculations based on pixel values are not feasible to detect changes in HSIs, even though the latter has taken the neighborhood information into account. In contrast, due to the introduction of a little supervised information, KNN and SVM achieve better performance, and their OA on both datasets exceeds 0.9. This is mainly because these two datasets tend to have relatively regular boundaries, and the covered scene is relatively simple, including buildings and farmland only. Next, let us take a look at the performance of several deep learning-based approaches. Interestingly, it can be observed that DCVA and DSFANet have the worst performance on the



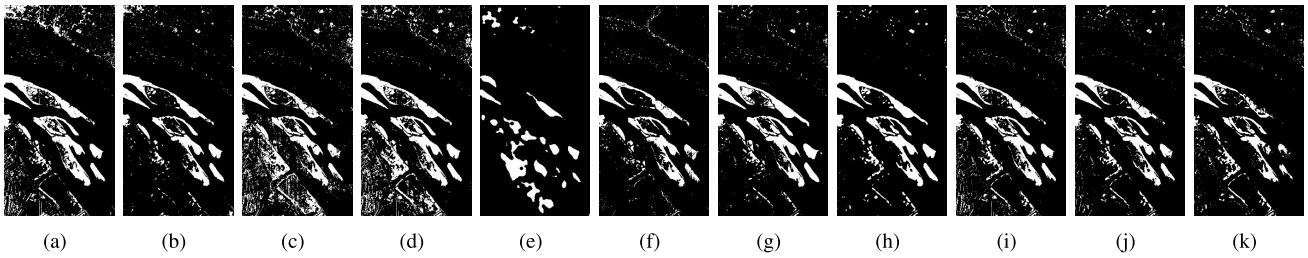
(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k)

Fig. 7. Santa Barbara dataset. (a) CVA. (b) RCVA. (c) KNN. (d) SVM. (e) DCVA. (f) DSFANet. (g) GETNET. (h) TDSSC. (i) KNN+CVA+S^{2}AN. (j) KNN+RCVA+S^{2}AN. (k) Reference map.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k)

Fig. 8. Bay Area dataset. (a) CVA. (b) RCVA. (c) KNN. (d) SVM. (e) DCVA. (f) DSFANet. (g) GETNET. (h) TDSSC. (i) KNN+CVA+S^{2}AN. (j) KNN+RCVA+S^{2}AN. (k) Reference map.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k)

Fig. 9. River dataset. (a) CVA. (b) RCVA. (c) KNN. (d) SVM. (e) DCVA. (f) DSFANet. (g) GETNET. (h) TDSSC. (i) KNN+CVA+S^{2}AN. (j) KNN+RCVA+S^{2}AN. (k) Reference map.

TABLE IV

QUANTITATIVE EVALUATION OF EXPERIMENTAL RESULTS OBTAINED BY DIFFERENT METHODS ON THE BAY AREA DATASET

Algorithm	OA	OE	κ	F_1
CVA	0.8761	9000	0.7534	0.8745
RCVA	0.8790	8789	0.7598	0.8746
KNN	0.9137	6268	0.8268	0.9187
SVM	0.9258	5388	0.8516	0.9280
DCVA	0.8248	12726	0.6546	0.8062
DSFANet	0.6337	26604	0.2800	0.5834
GETNET	0.8442	11319	0.6862	0.8567
TDSSC	0.9437	4093	0.8876	0.9439
KNN+CVA+S^{2}AN	0.9463	3898	0.8925	0.9484
KNN+RCVA+S^{2}AN	**0.9505**	**3594**	**0.9012**	**0.9514**

TABLE V

QUANTITATIVE EVALUATION OF EXPERIMENTAL RESULTS OBTAINED BY DIFFERENT METHODS ON THE RIVER DATASET

Algorithm	OA	OE	κ	F_1
CVA	0.9216	8744	0.6272	0.6681
RCVA	0.9465	5972	0.6760	0.7054
KNN	0.9258	8274	0.6532	0.6917
SVM	0.9242	8463	0.6504	0.6896
DCVA	0.8847	12866	0.2466	0.3094
DSFANet	0.9461	6018	0.6645	0.6941
GETNET	0.9472	5887	0.7073	0.7361
TDSSC	0.9593	4537	0.7598	0.7821
KNN+CVA+S^{2}AN	0.9550	5023	0.7563	0.7807
KNN+RCVA+S^{2}AN	**0.9667**	**3711**	**0.8010**	**0.8192**

Santa Barbara and Bay Area datasets, respectively, indicating that they are less robust to different scenarios. This is because DCVA uses features transferred from other datasets, while DSFANet uses deep slow features, which may not be universal enough. Compared with them, due to the introduction of a small number of true annotations, GETNET, TDSSC, and our proposed method achieve better performance. Among them, since GETNET only considers spectral information, it is not advantageous in detecting the regular changed and unchanged regions of interest. TDSSC, however, obtains results second only to our method because it takes more space and spectrum information into account. The reason for the performance gap may be that our method can search the region most relevant to the change more carefully.

2) Results on the River Dataset: Fig. 9 shows the results of the compared and the proposed algorithms on the River dataset, and the related quantitative evaluation values are listed in Table V. We first take a look at the results of several traditional methods. CVA correctly detects most of the variation areas and obtains a high OA, but it could not

Fig. 10. Sensitivity analysis about the PS on the river dataset. (a)–(e) CNN (PS = 13, 15, 27, 31, 39). (f)–(j) S^2AN (PS = 13, 15, 27, 31, 39).

effectively filter some isolated noise points. Compared with CVA, RCVA effectively inhibits the interference of isolated noise, and the result is relatively clean. KNN and SVM adopt the supervised learning fashion, and they get similar performance, but limited training samples lead to relatively high false detection. Then, because of the ability of deep feature extraction, most of the deep learning approaches achieve better performance with the exception of DCVA. In our analysis, there may be several reasons for the poor results of DCVA. First, the pretrained model used for DCVA does not cover the water scene, so the scene transferability is poor. Then, the deep features used may lead to relatively rough results. For DSFANet, GETNET, and TDSSC, there are no big gaps in their performance. On the River dataset, DSFANet proves the validity of the deep slow feature. Because the River dataset has the change regions of different scales, GETNET does not have disadvantages similar to those in the Barbara and Bay datasets. However, TDSSC still achieves better results due to a more comprehensive analysis of space spectrum information. The proposed algorithm achieves optimal performance under the support of both the attention mechanism and the sample selection strategy. In particular, KNN+RCVA+S^2AN achieves the highest OA and the lowest OE, while κ and F_1 also reach more than 0.8.

D. Stability Analysis on the Patch Size

In the patch-based methods, the patch size is always a very important parameter, which usually affects the fineness of the result and requires multiple attempts to find the appropriate value. The GSpaA module proposed in this article can effectively solve this problem and adaptively determine the relevant area of each patch. In order to verify the sensitivity and stability of the proposed module to patches of different sizes, we conducted experiments on the River dataset, which has the change regions of different scales, and thus, the adaptability of the model to regions of different sizes can be well verified.

Specifically, we use an ordinary CNN model, actual an S^2AN without attention modules, as a reference. We analyze the performance of the two models when the patch size is 13, 15, 23, 31, and 39. In addition, in order to avoid the interference brought by inaccurate training samples and maximize the performance of the models, we select 200 sample points separately from the changed and the unchanged areas of the reference map as the standard training set. The experimental results are shown in Fig. 10. It can be intuitively seen that, with the increase in the patch size, the results

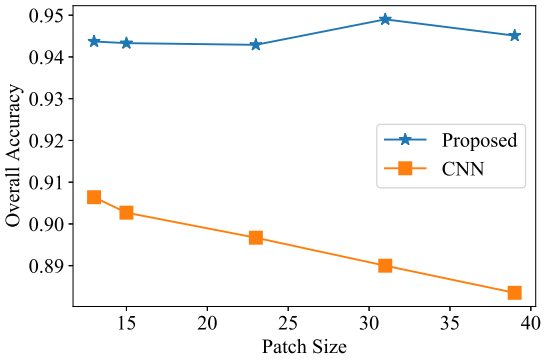


Fig. 11. OA curves with respect to the patch size.

of CNN get rougher and rougher, mainly because the large patch size usually leads to the decrease of the distinction between different patches, the fine-grained changes will be erased, and the edges will become rough so that only relatively large changed areas can be retained. On the contrary, S^2AN can maintain similar fineness under different patch sizes, and the edges of surface objects are sharp and distinguishable. These intuitive phenomena are also objectively reflected in the curve shown in Fig. 11, which represents the change of OA with different patch sizes. Specifically, S^2AN's OA has been stabilized around 0.94, but CNNs attenuate from 0.91 to less than 0.88. Moreover, it can be observed that S^2AN's results are consistently higher than CNN's, which fully demonstrates the effect of the proposed attention mechanism on improving performance and stability.

Observing the experimental phenomenon, we believe that, in the patch, the area most relevant to the change of the central pixel can be found stably. When the boundary of the patch gets larger and larger, the GSpaA can still limit the core region within that range, rather than considering the whole area, such as CNN, thereby reducing the sensitivity of the entire model to the patch size. In short, this part fully demonstrates the effectiveness of the proposed Gaussian spatial attention module and its low sensitivity to patches of different sizes. With the module, multiple attempts are no longer necessary to determine the most appropriate patch size; merely, a rough size is enough.

E. Ablation Study and Interpretability Analysis

In this section, we provide a more detailed analysis of the role of each module in the proposed method. First, the advantages of the proposed semisupervised sample selection

TABLE VI

QUANTITATIVE EVALUATION OF EXPERIMENTAL RESULTS OBTAINED BY METHODS WITH DIFFERENT TRAINING SAMPLES ON THE RIVER DATASET

Algorithm	OA	OE	KC	F1
Supervised S^2AN	0.9439	6260	0.7083	0.7385
CVA+S^2AN	0.9216	8744	0.6272	0.6681
RCVA+S^2AN	0.9258	8274	0.6532	0.6917
KNN+CVA+S^2AN	0.9550	5023	0.7563	0.7807
KNN+RCVA+S^2AN	**0.9667**	**3711**	**0.8010**	**0.8192**

TABLE VII

QUANTITATIVE EVALUATION OF EXPERIMENTAL RESULTS OBTAINED BY THE MODELS WITH DIFFERENT ATTENTION ON THE RIVER DATASET

Algorithm	OA	OE	KC	F1
KNN+RCVA+CNN	0.9590	4575	0.7654	0.7879
KNN+RCVA+GSpaA	0.9649	3914	0.7974	0.8166
KNN+RCVA+SpeA	0.9623	4208	0.7777	0.7984
KNN+RCVA+S^2AN	**0.9667**	**3711**	**0.8010**	**0.8192**

strategy are discussed. Then, we decouple the spatial and spectral attention modules, as well as analyzing their respective roles, while performing an intuitive interpretability analysis.

1) Semisupervised Strategy: By combining the unsupervised data mining and the guidance of supervised information, the semisupervised sample selection strategy can provide enough samples with high confidence, which reflects the idea of ensemble learning. Here, the effect of the strategy continues to be validated on the River dataset. The comparison strategies include CVA, RCVA, supervision, KNN+CVA, and KNN+RCVA.

The quantitative evaluation results of the methods using different training samples are displayed in Table VI. From the table, it can be seen that KNN+CVA- and KNN+RCVA-based instances have obvious improvement compared with CVA, RCVA, and supervised-based ones. Moreover, due to the high accuracy of RCVA itself and the introduction of supervised information, the instance based on KNN+RCVA achieves the best performance with almost no additional computing costs, which fully proves that the proposed strategy is simple, lightweight, and effective. Of course, this strategy can only make up for the defects of the two methods to a certain extent, and when the scene is too complex, the training samples screened may still be dissatisfactory. However, in most cases, our strategy is a better option than traditional ones indeed.

2) Attention Modules: In the proposed model, there are two kinds of attention modules that are SpeA module for channels and the GSpaA module for spaces, respectively. Here, a set of ablation experiments on the River dataset are designed to discuss the role played by each module in detail. Specifically, we compare the performance of the model under four conditions, i.e., not using any attention module, i.e., CNN,

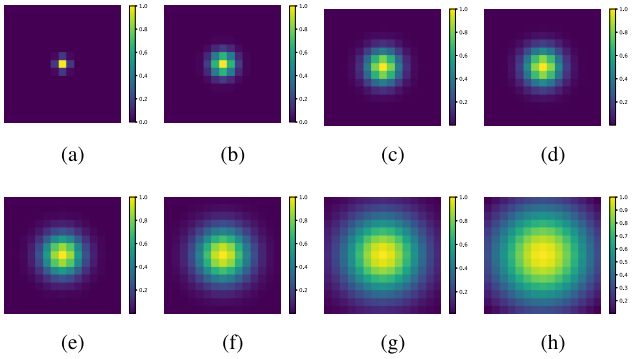


Fig. 12. Normalized spatial attention score heat maps corresponding to eight representative patches, where $\hat{\sigma}$ is the standard deviation of the corresponding Gaussian function. (a) $\hat{\sigma} = 0.5239$. (b) $\hat{\sigma} = 1.0691$. (c) $\hat{\sigma} = 1.5088$. (d) $\hat{\sigma} = 1.6300$. (e) $\hat{\sigma} = 1.8107$. (f) $\hat{\sigma} = 2.3487$. (g) $\hat{\sigma} = 3.5042$. (h) $\hat{\sigma} = 4.6379$.

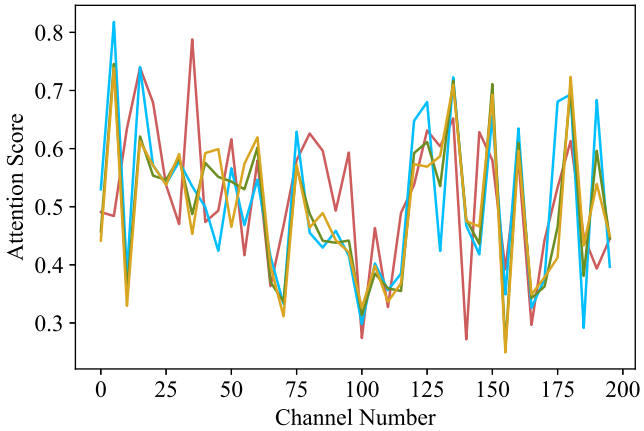


Fig. 13. Spectral attention score curves of four colors corresponding to four random patches, in which, 40 data points were obtained by sampling 198 channels with equal spacing of five steps.

using GSpaA only, using SpeA only, and using GSpaA and SpeA at the same time. It is worth noting that the patch size is uniformly set to 15, and the KNN+RCVA is utilized to obtain training samples.

The relevant evaluation results are presented in Table VII. It can be observed that, compared with CNN without attention mechanism, the model with SpeA or GSpaA attention module can actually improve the detection performance to a certain extent, which proves the positive effects of these two modules. When the two attention modules are both employed, the detection performance gets further improvement. In order to analyze the function of the SpeA and GSpaA modules more intuitively, some spatial and spectral attention scores corresponding to several random patches were captured during model inference, and the corresponding visualization results, including spatial score heat map and spectral score curve, are displayed in Figs. 12 and 13, respectively. It is obvious that it can be observed that each patch can get its own spatial and spectral attention score via our model. Especially, for GSpaA, it realizes adaptive core region division, thus reducing the sensitivity of the patch size for the whole model.

V. CONCLUSION

In this article, a novel spatial and spectral attention network is proposed to detect changes in HSIs. The whole network

is composed of several blocks, each of which includes three modules, i.e., SpeA, GSpaA, and CFE. These modules ensure that the proposed model can enhance relevant information and suppress irrelevant information from both spectral and spatial perspectives. SpeA calculates the attention score for each channel of the input, while GSpaA calculates the standard deviation of a Gaussian distribution, and the attention score for each spatial position of the input can be sampled from this Gaussian distribution. It brings an additional benefit, that is, it reduces the sensitivity of the model to the patch size, avoiding multiple attempts to determine the most appropriate patch size. In addition, a simple but effective label augmentation strategy is proposed by organically combining unsupervised and supervised methods, which further improves the feasibility of the HSI CD when labeled samples are limited.

In future work, we will continue to use effective methods, such as self-supervised learning and few-shot learning, to deal with CD tasks under complex conditions.

ACKNOWLEDGMENT

The authors would like to thank the Researchers for the open-source datasets and codes. They would also like to thank the Editor-in-Chief, Associate Editors, and Reviewers for the insightful comments and suggestions.

REFERENCES

[1] L. Bruzzone and F. Bovolo, "A novel framework for the design of change-detection systems for very-high-resolution remote sensing images," *Proc. IEEE*, vol. 101, no. 3, pp. 609–630, Mar. 2013.

[2] A. Arias-Montano, C. A. C. Coello, and E. Mezura-Montes, "Multiobjective evolutionary algorithms in aeronautical and aerospace engineering," *IEEE Trans. Evol. Comput.*, vol. 16, no. 5, pp. 662–694, Oct. 2012.

[3] L. Khelifi and M. Mignotte, "Deep learning for change detection in remote sensing images: Comprehensive review and meta-analysis," *IEEE Access*, vol. 8, pp. 126385–126400, 2020.

[4] W. Shi, M. Zhang, R. Zhang, S. Chen, and Z. Zhan, "Change detection based on artificial intelligence: State-of-the-art and challenges," *Remote Sens.*, vol. 12, no. 10, p. 1688, May 2020.

[5] C. Manzo, A. Mei, E. Zampetti, C. Bassani, L. Paciucci, and P. Manetti, "Top-down approach from satellite to terrestrial rover application for environmental monitoring of landfills," *Sci. Total Environ.*, vols. 584–585, pp. 1333–1348, Apr. 2017.

[6] P. Washaya, T. Balz, and B. Mohamadi, "Coherence change-detection with Sentinel-1 for natural and anthropogenic disaster monitoring in urban areas," *Remote Sens.*, vol. 10, no. 7, p. 1026, Jun. 2018.

[7] S. Liu, D. Marinelli, L. Bruzzone, and F. Bovolo, "A review of change detection in multitemporal hyperspectral images: Current techniques, applications, and challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 140–158, Jun. 2019.

[8] X. Li, Z. Yuan, and Q. Wang, "Unsupervised deep noise modeling for hyperspectral image change detection," *Remote Sens.*, vol. 11, no. 3, p. 258, Jan. 2019.

[9] P. Ghamisi *et al.*, "Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 37–78, Dec. 2017.

[10] D. Hong, N. Yokoya, and X. X. Zhu, "Learning a robust local manifold representation for hyperspectral dimensionality reduction," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 2960–2975, Jun. 2017.

[11] D. Letexier and S. Bourennane, "Noise removal from hyperspectral images by multidimensional filtering," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 7, pp. 2061–2069, Jul. 2008.

[12] Q. Yuan, L. Zhang, and H. Shen, "Hyperspectral image denoising employing a spectral–spatial adaptive total variation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 10, pp. 3660–3677, Mar. 2012.

[13] J. M. Bioucas-Dias *et al.*, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 2, pp. 354–379, May 2012.

[14] R. Heylen, M. Parente, and P. Gader, "A review of nonlinear hyperspectral unmixing methods," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 1844–1868, May 2014.

[15] S. Liu, L. Bruzzone, F. Bovolo, M. Zanetti, and P. Du, "Sequential spectral change vector analysis for iteratively discovering and detecting multiple changes in hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4363–4378, Feb. 2015.

[16] 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 Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 3–13, Jul. 2018.

[17] M. Gong, T. Zhan, P. Zhang, and Q. Miao, "Superpixel-based difference representation learning for change detection in multispectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2658–2673, Feb. 2017.

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

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, Jun. 2016, pp. 770–778.

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, Jun. 2016, pp. 779–788.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[22] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised deep change vector analysis for multiple-change detection in VHR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3677–3693, Jan. 2019.

[23] B. Du, L. Ru, C. Wu, and L. Zhang, "Unsupervised deep slow feature analysis for change detection in multi-temporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9976–9992, Sep. 2019.

[24] R. D. Johnson and E. S. Kasischke, "Change vector analysis: A technique for the multispectral monitoring of land cover and condition," *Int. J. Remote Sens.*, vol. 19, no. 3, pp. 411–426, 1998.

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

[26] C. Wu, B. Du, and L. Zhang, "Slow feature analysis for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2858–2874, May 2014.

[27] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, H. Ding, and X. Huang, "SemiCDNet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5891–5906, Jul. 2021.

[28] C. Wu, H. Chen, B. Du, and L. Zhang, "Unsupervised change detection in multitemporal VHR images based on deep kernel PCA convolutional mapping network," *IEEE Trans. Cybern.*, early access, Jul. 8, 2021, doi: 10.1109/TCYB.2021.3086884.

[29] C. Zhang *et al.*, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 183–200, Aug. 2020.

[30] H. Chen, C. Wu, B. Du, and L. Zhang, "Deep Siamese multi-scale convolutional network for change detection in multi-temporal VHR images," in *Proc. 10th Int. Workshop Anal. Multitemporal Remote Sens. Images (MultiTemp)*, Aug. 2019, pp. 1–4.

[31] J. Chen *et al.*, "DASNet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1194–1206, 2020.

[32] A. Song, J. Choi, Y. Han, and Y. Kim, "Change detection in hyperspectral images using recurrent 3D fully convolutional networks," *Remote Sens.*, vol. 10, no. 11, p. 1827, Nov. 2018.

[33] T. Zhan *et al.*, "SSCNN-S: A spectral-spatial convolution neural network with Siamese architecture for change detection," *Remote Sens.*, vol. 13, no. 5, p. 895, Feb. 2021.

[34] S. T. Seydi, M. Hasanlou, and M. Amani, "A new end-to-end multi-dimensional CNN framework for land cover/land use change detection in multi-source remote sensing datasets," *Remote Sens.*, vol. 12, no. 12, p. 2010, Jun. 2020.

[35] T. Zhan *et al.*, "TDSSC: A three-directions spectral–spatial convolution neural network for hyperspectral image change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 377–388, Nov. 2021.

[36] F. Huang, Y. Yu, and T. Feng, "Hyperspectral remote sensing image change detection based on tensor and deep learning," *J. Vis. Commun. Image Represent.*, vol. 58, pp. 233–244, Jan. 2019.
[37] F. Zhou and Z. Chen, "Hyperspectral image change detection by self-supervised tensor network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Sep. 2020, pp. 2527–2530.
[38] J. López-Fandiño, A. S. Garea, D. B. Heras, and F. Argüello, "Stacked autoencoders for multiclass change detection in hyperspectral images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Nov. 2018, pp. 1906–1909.
[39] D. Marinelli, F. Bovolo, and L. Bruzzone, "A novel change detection method for multitemporal hyperspectral images based on binary hyperspectral change vectors," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4913–4928, Jul. 2019.
[40] Q. Guo, J. Zhang, and Y. Zhang, "Multitemporal hyperspectral images change detection based on joint unmixing and information coguidance strategy," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9633–9645, Nov. 2021.
[41] Z. Hou, W. Li, L. Li, R. Tao, and Q. Du, "Hyperspectral change detection based on multiple morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, early access, Jul. 1, 2021, doi: 10.1109/TGRS.2021.3090802.
[42] A. Vaswani *et al.*, "Attention is all you need," *Proc. Adv. Neural Inform. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.
[43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[44] M. Jaderberg *et al.*, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 2017–2025.
[45] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7794–7803.
[46] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," 2017, *arXiv:1710.10903*.
[47] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Sep. 2018.
[48] W. Cai and Z. Wei, "Remote sensing image classification based on a cross-attention mechanism and graph convolution," *IEEE Geosci. Remote Sens. Lett.*, early access, Oct. 1, 2020, doi: 10.1109/LGRS.2020.3026587.
[49] W. Tong, W. Chen, W. Han, X. Li, and L. Wang, "Channel-attention-based densenet network for remote sensing image scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4121–4132, 2020.
[50] Q. Zhang *et al.*, "Dense attention fluid network for salient object detection in optical remote sensing images," *IEEE Trans. Image Process.*, vol. 30, pp. 1305–1317, 2021.
[51] Y. Li, Q. Huang, X. Pei, L. Jiao, and R. Shang, "RADet: Refine feature pyramid network and multi-layer attention network for arbitrary-oriented object detection of remote sensing images," *Remote Sens.*, vol. 12, no. 3, p. 389, Jan. 2020.
[52] X. Hua, X. Wang, T. Rui, H. Zhang, and D. Wang, "A fast self-attention cascaded network for object detection in large scene remote sensing images," *Appl. Soft Comput.*, vol. 94, Sep. 2020, Art. no. 106495.
[53] H. Guo, Q. Shi, B. Du, L. Zhang, and H. Ding, "Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4287–4306, May 2020.
[54] M. Guo, H. Liu, Y. Xu, and Y. Huang, "Building extraction based on U-Net with an attention block and multiple losses," *Remote Sens.*, vol. 12, no. 9, p. 1400, Apr. 2020.
[55] J. Cai and Y. Chen, "MHA-Net: Multipath hybrid attention network for building footprint extraction from high-resolution remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5807–5817, 2021.
[56] R. Liu, Z. Cheng, L. Zhang, and J. Li, "Remote sensing image change detection based on information transmission and attention mechanism," *IEEE Access*, vol. 7, pp. 156349–156359, 2019.
[57] H. Jiang, X. Hu, K. Li, J. Zhang, J. Gong, and M. Zhang, "PGA-SiamNet: Pyramid feature-based attention-guided Siamese network for remote sensing orthoimagery building change detection," *Remote Sens.*, vol. 12, no. 3, p. 484, Feb. 2020.
[58] L. Chen, D. Zhang, P. Li, and P. Lv, "Change detection of remote sensing images based on attention mechanism," *Comput. Intell. Neurosci.*, vol. 2020, pp. 1–11, Aug. 2020.
[59] X. Peng, R. Zhong, Z. Li, and Q. Li, "Optical remote sensing image change detection based on attention mechanism and image difference," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7296–7307, Sep. 2021.
[60] J. M. Amigo and C. Santos, "Preprocessing of hyperspectral and multispectral images," in *Data Handling in Science and Technology*. Amsterdam, The Netherlands: Elsevier, 2020, vol. 32, pp. 37–53.
[61] B. Datt, T. R. McVicar, T. G. Van Niel, D. L. B. Jupp, and J. S. Pearlman, "Preprocessing EO-1 Hyperion hyperspectral data to support the application of agricultural indexes," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 6, pp. 1246–1259, Aug. 2003.
[62] G. H. Rosenfield and K. Fitzpatrick-Lins, "A coefficient of agreement as a measure of thematic classification accuracy," *Photogram. Eng. Remote Sens.*, vol. 52, no. 2, pp. 223–227, 1986.
[63] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Adv. Neural Inf. Process. Syst.*, vol. 32, pp. 8026–8037, 2019.
[64] F. Thonfeld, H. Feilhauer, M. Braun, and G. Menz, "Robust change vector analysis (RCVA) for multi-sensor very high resolution optical satellite data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 50, pp. 131–140, Apr. 2016.

Maoguo Gong (Senior Member, IEEE) received the B.S. degree (Hons.) in electronic engineering and the Ph.D. degree in electronic science and technology from Xidian University, Xi'an, China, in 2003 and 2009, respectively.

Since 2006, he has been a Teacher with Xidian University. He was promoted to an Associate Professor and a Full Professor, in 2008 and 2010, respectively, with exceptive admission. He is leading or has completed over 20 projects as the Principle Investigator, funded by the National Natural Science Foundation of China, the National Key Research and Development Program of China, and others. He has authored or coauthored over 100 articles in journals and conferences, and holds over 20 granted patents as the First Inventor. His research interests are broadly in the area of computational intelligence, with applications to optimization, learning, data mining, and image understanding.

Dr. Gong is the Executive Committee Member of the Chinese Association for Artificial Intelligence and a Senior Member of the Chinese Computer Federation. He was a recipient of the Prestigious National Program for Support of the Leading Innovative Talents (selected by the Central Organization Department of China), the Leading Innovative Talent in the Science and Technology (selected by the Ministry of Science and Technology of China), the Excellent Young Scientist Foundation (selected by the National Natural Science Foundation of China), the New Century Excellent Talent in University (selected by the Ministry of Education of China), the Young Teacher Award by the Fok Ying Tung Education Foundation, and the National Natural Science Award of China. He is an Associate Editor or an Editorial Board Member for over five journals including the IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION and the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

Fenlong Jiang (Graduate Student Member, IEEE) was born in 1996. He is currently pursuing the Ph.D. degree in electronic science and technology with the School of Electronic Engineering, Xidian University, Xi'an, China.

He is with the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education of China. His research interests include deep learning, computer vision, remote sensing image interpretation, and processing.

A. K. Qin (Senior Member, IEEE) received the B.Eng. degree from Southeast University, Nanjing, China, in 2001, and the Ph.D. degree from Nanyang Technology University, Singapore, in 2007.

From 2007 to 2017, he was with the University of Waterloo, Waterloo, ON, Canada; INRIA Grenoble Rhône-Alpes, Montbonnot-Saint-Martin, France; and RMIT University, Melbourne, VIC, Australia. He joined Swinburne University of Technology, Hawthorn, VIC, Australia, in 2017, where he is currently a Professor. He is currently the Director of Swinburne Intelligent Data Analytics Laboratory, the Deputy Director of Swinburne Space Technology and Industry Institute, and the Program Lead of Swinburne Data Science Research Institute. His major research interests include machine learning, evolutionary computation, computer vision, remote sensing, services computing, and pervasive computing.

Dr. Qin was a recipient of the 2012 IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION Outstanding Paper Award. He is currently the Chair of the IEEE Computational Intelligence Society (CIS) Neural Networks Technical Committee and the Vice-Chair of the IEEE CIS Emergent Technologies Task Force on Multitask Learning and Multitask Optimization.

Tongfei Liu (Graduate Student Member, IEEE) is currently pursuing the Ph.D. degree in electronic science and technology with the School of Electronic Engineering, Xidian University, Xi'an, China, and the master's degree from the School of Computer Science and Engineering, Xi'an University of Technology, in 2020.

He is interested in deep learning, computational intelligence, and land cover change detection and classification, through VHR remote sensing images (including satellite and aerial images).

Tao Zhan (Member, IEEE) received the B.E. degree in electronic information engineering from the Henan University of Technology, Zhengzhou, China, in 2013, and the Ph.D. degree in pattern recognition and intelligent system from Xidian University, Xi'an, China, in 2019, respectively.

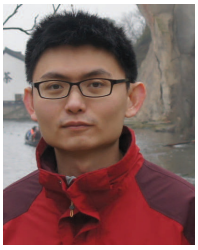
He is currently a Lecturer with the School of Computer Science and Technology, Xidian University. His research interests include computational intelligence and remote sensing image understanding.

Di Lu received the B.S. degree in communication engineering from Hohai University, Jiangsu, China. He is currently pursuing the master's degree in electronic science and technology with the School of Electronic Engineering, Xidian University, Xi'an, China.

His research interests include remote sensing image processing and machine learning.

Hanhong Zheng received the B.S. degree in communication engineering from Hohai University, Jiangsu, China. He is currently pursuing the master's degree in electronic science and technology at the School of Electronic Engineering, Xidian University, Xi'an, China.

His research interests include remote sensing image processing and machine learning.



Mingyang Zhang (Member, IEEE) received the B.S. degree in automation and the Ph.D. degree in pattern recognition and intelligent system from Xidian University, Xi'an, China, in 2012 and 2018, respectively.

Since 2018, he has been a Lecturer with the School of Electronic Engineering, Xidian University. His research interests include computational intelligence and remote sensing image understanding.