IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 29, 2020, accepted December 15, 2020, date of publication December 18, 2020,
date of current version December 31, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3045740

Land Cover Change Detection With VHR Satellite Imagery Based on Multi-Scale SLIC-CNN and SCAE Features

RAN JING[1], ZHAONING GONG[2], AND HONGLIANG GUAN[3]
[1]School of Geosciences, Yangtze University, Wuhan 430100, China
[2]College of Resource Environment and Tourism, Capital Normal University, Beijing 100048, China
[3]College of Geospatial Information Science and Technology, Capital Normal University, Beijing 100048, China

Corresponding author: Zhaoning Gong (gongzhn@163.com)

This work was supported in part by the National Key R&D Program of China under Grant 2017YFC0505903.

ABSTRACT Change detection with very high resolution (VHR) satellite images is of great application values when evaluating and monitoring land use changes. However, intrinsic complexity of satellite images will introduce more difficulties to change detection tasks. In this study, a new change detection method is proposed by combining multi-scale simple linear iterative clustering-convolutional neural network (SLIC-CNN) with stacked convolutional auto-encoder (SCAE) features to improve change detection capabilities with VHR satellite images. First, the multi-scale SLIC-based image segmentation is performed on multi-temporal images to generate segment objects while keeping their edge information as much as possible. Second, the convolutional layers in a CNN architecture are used to generate change map, then, an SCAE feature-based classification procedure is performed to generate "from-to" change information. Finally, a Bayesian information criterion is used to optimize the results of change detection. In this study, the experiments carried out reveal that the multi-scale SLIC image segmentation algorithm affects the integrity of change regions; CNN features have an effect on the consistency of change regions; and SCAE features influence the performance of support vector machine (SVM) classifiers. And, features extracted from the architectures enhance the ability of information extraction from ground objects. Comparison results also show the superiority to other change detection methods.

INDEX TERMS Change detection, image segmentation, multi-scale simple linear iterative clustering-convolutional neural network (SLIC-CNN), stacked convolutional auto-encoder (SCAE), VHR satellite imagery.

I. INTRODUCTION

Change detection is the progress of determining how the land cover of a particular area changes over time. It is widely used to analyze changes in land cover, assess natural disasters, evaluate urban expansion, and detect different environmental factors [1], [2]. With the improvement of spatial resolution on imaging sensors, complex image details introduce significant challenges to change detection. According to different basic units used in VHR change detection, the algorithms are generally divided into two categories: One is pixel-based and another one is object-oriented [3]. Pixel-based change detection algorithms use pixels as basic units, and each pixel is independent of the others, regardless of their spatial or semantic relationships. Pixel-based change detection algorithms include pixel-by-pixel calculation procedure, change maps are created by performing algebraic operations, such as difference and ratio operations. The final results of change detection are typically obtained by thresholding and clustering methods. While thresholding methods are sensitive to noise, making it difficult to obtain ideal change detection results. Clustering methods have the ability to suppress noise via an effective combination of local information. However, when applying pixel-based change detection algorithms to VHR satellite images, only single pixels are considered while ignoring the relationships between neighboring pixels. Thus, the pixel-based change detection algorithms often yield poor results for the VHR satellite images.

Compared with pixel-based change detection algorithms, the object-oriented change detection algorithms make a full

The associate editor coordinating the review of this manuscript and approving it for publication was K.C. Santosh.

use of spectra, textures, shapes, and adjacency relationships of pixels. Steps of object-oriented change detection algorithms include image segmentation, object feature extraction, change map generation, and extraction of ''from-to'' change information. However, segmentation algorithms in commercial softwares are not flexible enough to obtain consistent boundaries for ground objects [4]. In many computer vision tasks, superpixel segmentation algorithms perform well, they divide an image into large amounts of boundary preserved non-overlapping image objects. And there are many research projects that adopted this kind of approach to detect changes in remote sensing images (e.g., [4]–[6]).

Compared with above two types of change detection methods, deep learning technique shows a potential advantage for feature extraction and semantic segmentation in change detection tasks [7]–[11]. Earlier architectures, such as the Deep Belief Network [12] and Multilayer Perceptron [13], only consist of fully connected layers, requiring an update of large numbers of parameters when using images as input, which consumes additional computing resources. Simultaneously, such architectures only take 1-D vectors as input, hard to use neighborhood relationships between pixels. In order to better handle image problems, convolutional neural network (CNN) is proposed. CNN has the ability to learn hierarchical features, from low level features to high level features, which is beneficial to a comprehensive representation on VHR satellite images. While compared with other application fields (e.g., visual object recognition and object detection in [14], [15]), there are fewer open source remote sensing datasets available. So, natural scene datasets are often used to pre-train the parameters of CNN to mitigate the lack of datasets [16], [17]. With the features generated from trained CNN architectures, most methods take change detection as a classification problem, i.e., a binary classification. Hou *et al.* [18] conducted a binary change detection experiment based on low-rank saliency detection and feature extraction using a CNN architecture. Zhang *et al.* [19] adopted a change detection framework including a denoising autoencoder and mapping network to perform the binary change detection. To extract change information from multi-sensor remote sensing images, Wang *et al.* [20] proposed an OB-DSCNH module, deep change features were extracted effectively and the depth and width of the network were partly increased. For multi-label change detection, Khan *et al.* [21] applied a deep neural network (DNN) architecture to conduct change detection in forests with long-term serial images, dividing the change results into unchanged, harvest events, or fire events. With UAV images, Sun *et al.* [22] proposed MTL-CD method for detecting find-grained ''from-to'' building changes. However, many deep learning-based change detection methods (e.g., [23]) directly input multi-temporal images into the architectures, which could result in rough detection results in a certain degree.

In this study, we propose a deep learning-based change detection method, in which a multi-scale simple linear iterative clustering (SLIC) segmentation algorithm is applied to generate segmented objects that retain edge information. Fine-tuned layers of a CNN architecture are used to generate change map. A stacked convolutional autoencoder (SCAE)-based classifier is trained to generate ''from-to'' change information of ground objects. And the change map and ''from-to'' change information are combined via the Bayesian information criterion to obtain optimized change detection results, which improve the accuracy of change detection with VHR satellite images. The results of proposed method are evaluated with reference map and compared with other change detection methods.

II. METHODOLOGY

In this study, the proposed method consists of two parts: (1) Change map generation. Multi-temporal images are first segmented by the proposed multi-scale SLIC segmentation algorithm. And then, a fusion rule is used to merge the segmented objects. The fused image objects are subsequently input into a feature extractor based on CNN architecture. The output feature maps are used to calculate a change map with the structural similarity index measure (SSIM) algorithm. (2) ''from-to'' change information extraction. A mask operation is conducted on the image pairs to screen unchanged parts. The ''from-to'' change information is obtained with an SCAE feature classifier. And the change information is optimized by the Bayesian criterion optimization algorithm. Figure 1 shows the diagram of the proposed method.

A. MULTI-SCALE SLIC-CNN CHANGE MAP GENERATION

Change map reflects the variation of land cover types in multi-temporal remote sensing images. And the ''from-to'' change information of land cover types is extracted based on the obtained change map. For change map generation, we first use multi-scale SLIC segmentation algorithm to segment the images into segmented objects. Then a CNN architecture is used to generate the change map.

1) MULTI-SCALE SLIC IMAGE SEGMENTATION

VHR satellite images have a large number of pixels. Compared with pixels in an image, segmented objects fuse spectra, texture, and contextual characteristics of the pixels, which is more reasonable as the basic unit to analyze the VHR satellite images. In this paper, we introduce the SLIC algorithm [24] to generate segmented objects. SLIC is a simple and effective segmentation algorithm, which uses k-means as the basic algorithm to generate segmented objects that are compact and internally homogeneous. However, scenic information and pixel relationships vary in VHR satellite images of different spatial resolutions [25]. When SLIC algorithm is performed under single scale image, object characteristics at other scales are lost. Therefore, it is difficult to ensure accurate object boundaries when performing image segmentation at a single scale. To fully utilize information in different image scales, multi-scale transformation is used, as shown in Figure 2.

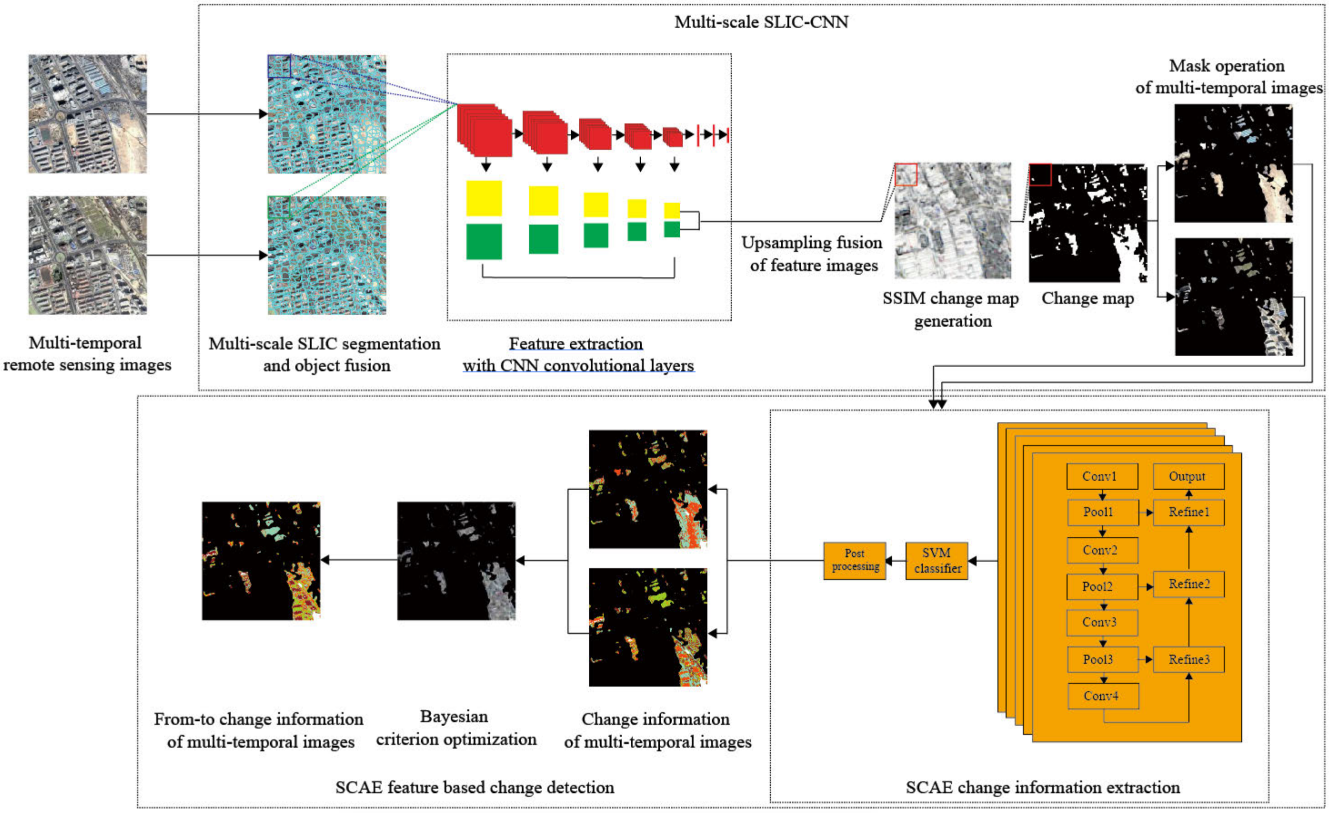
Multi-scale SLIC-CNN
Multi-temporal remote sensing images
Multi-scale SLIC segmentation and object fusion
Feature extraction with CNN convolutional layers
Upsampling fusion of feature images
SSIM change map generation
Change map
Mask operation of multi-temporal images
Conv1
Pool1
Conv2
Pool2
Conv3
Pool3
Conv4
Output
Refine1
Refine2
Refine3
SVM classifier
Post processing
Change information of multi-temporal images
Bayesian criterion optimization
From-to change information of multi-temporal images
SCAE feature based change detection
SCAE change information extraction

FIGURE 1. Diagram of the Methodology. Red, yellow, and green blocks represent the feature maps generated by CNN architecture of time phase 1 and time phase 2, respectively. After image segmentation, image objects are used as input for the bottom of the feature extractor, where convolutional layers are used to automatically extract features from the image objects. On the top of feature extractor, up-sampled features are merged to obtain the change map mask via the SSIM algorithm. Masked image patches are fed into SCAE architecture and a Bayesian optimization algorithm is used to obtain optimized "from-to" change information.

Figure 2 presents that segmentation result from a higher spatial resolution image will contain shattered objects, where inner pixels of the segmented objects mainly originate from single ground object. With a lower spatial resolution image, segmentation algorithm will generate more sparse objects, and in some image areas with uniform characteristics, segmented objects show a rounded shape, which is due to principles of the SLIC algorithm. The algorithm selects cluster centers and groups the pixels around them in a certain distance that result in a regular object shape. The inner pixels within the objects are more likely from different ground objects. In order to fuse information from multi-scale images, segmented objects obtained have to be merged together. We make segmented objects from higher-scale images as initial segmentation while lower-scale segmentation results are merged layer-by-layer according to their upper level layers (Figure 3).

As shown in Figure 3, the segmented objects generated from higher-scale images are generally rough and present a phenomenon of under-segmentation. As the scale decreased, the segmentation results are gradually refined, and the segmented objects obtained from higher-scale images are used as basis to establish multi-level adjacency relationships with other lower-scale objects. The centroid coordinate s_1 of the upper-scale object is calculated and mapped to lower-scale objects s_2–s_n layer-by-layer. Region adjacent graph (RAG) between segmented objects at various image scales is also established to calculate the Euclidean distance of s_2–s_n as well as 8-connected adjacent objects with their corresponding upper-scale objects in color space. Objects with minimal distance are merged to eliminate redundancy. The merging process is divided into two cases: (1) If the neighbor object has the minimal color distance, its label number is incorporated into the mapped object; and (2) if the mapped object has the minimal color distance, its corresponding label number is set in an identical manner as the label number of the second minimal distance.

Due to changes happening in ground objects, the segmentation results of multi-temporal images are inconsistent. As shown in Figure 4, changes are substantial in the multi-temporal images (2009 and 2017). Any single segmentation result is unable to represent the real object boundaries. Therefore, segmentation results of different time phases are fused to ensure consistent boundaries with all ground objects in the multi-temporal images.

Segmentation results from multi-temporal images are first simply merged by addition to obtain more segmented objects, which ensure that the segmentation boundaries are

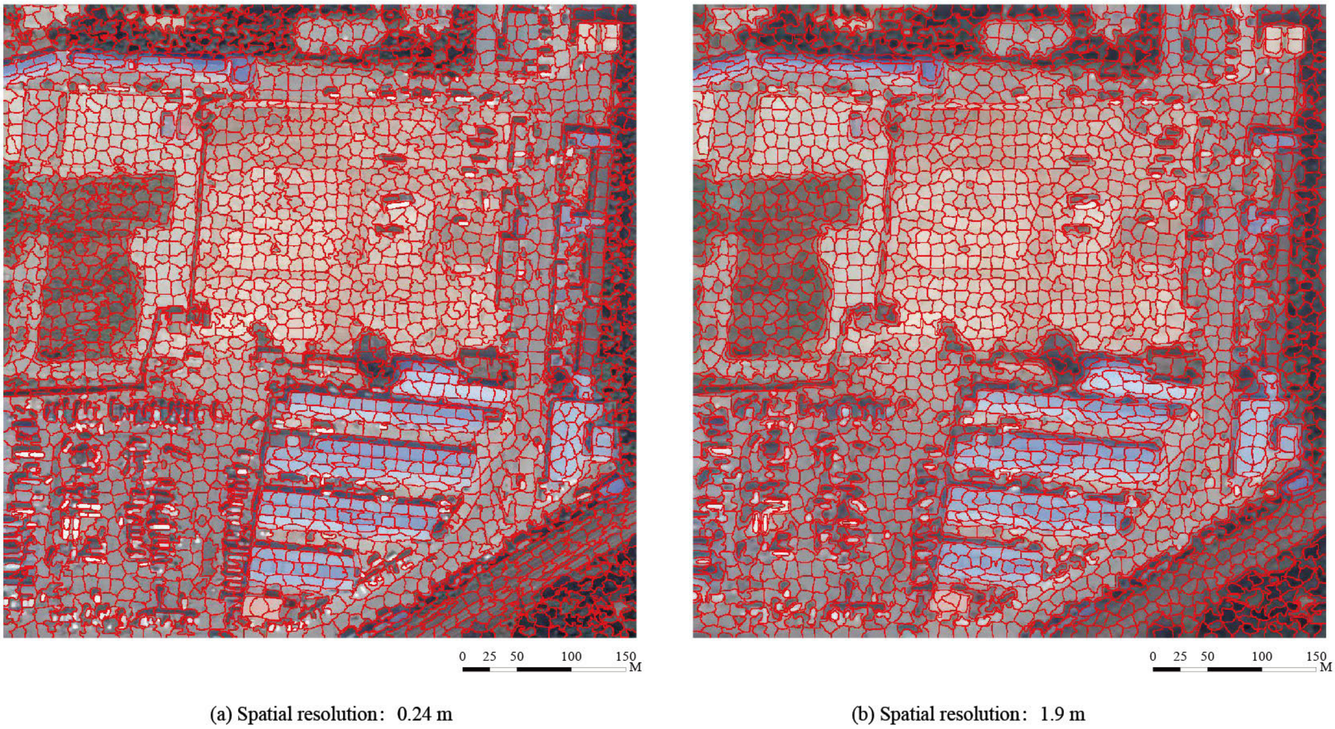

FIGURE 2. Segmented objects at different spatial resolutions with parameter 50: (a) 0.24 m and (b) 1.9 m. Images are obtained from Google Earth v7.1.8. Parameter 50 is used to show clear boundary lines, but the optimal number of divided objects is larger than 50.

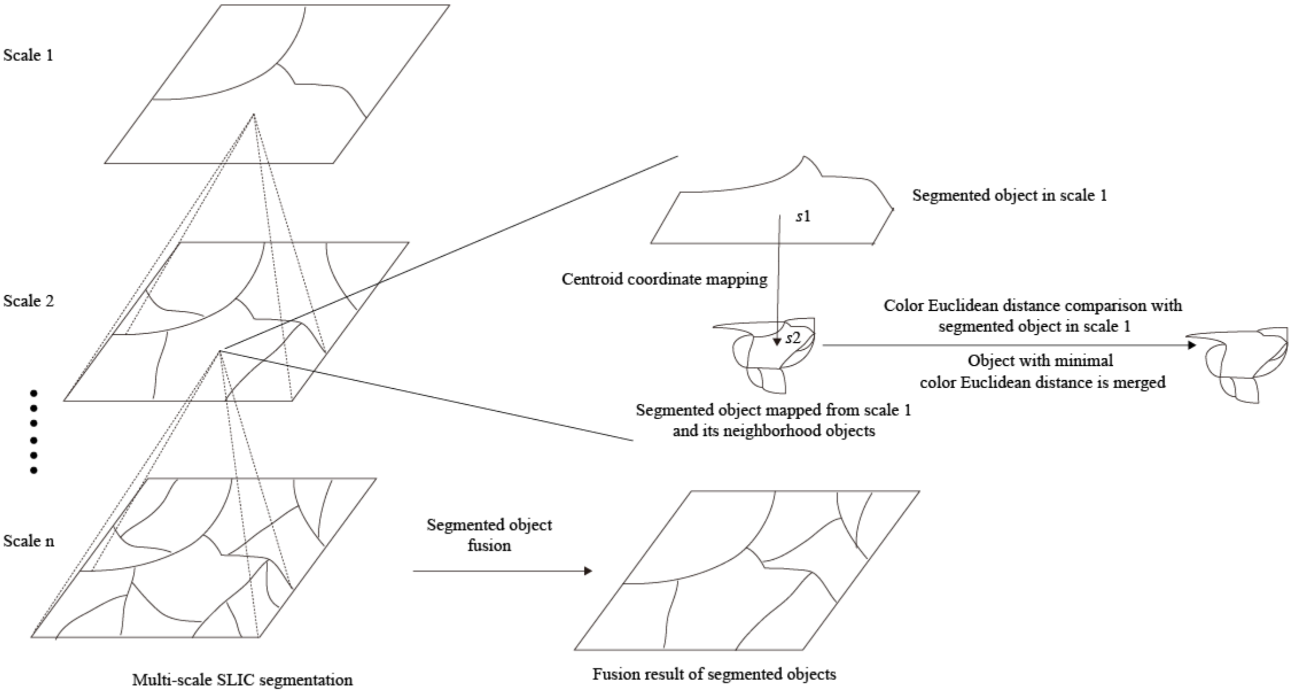
Scale 1
Scale 2
Scale n
Segmented object in scale 1
s1
Centroid coordinate mapping
s2
Color Euclidean distance comparison with segmented object in scale 1
Object with minimal color Euclidean distance is merged
Segmented object mapped from scale 1 and its neighborhood objects
Segmented object fusion
Multi-scale SLIC segmentation
Fusion result of segmented objects

FIGURE 3. Multi-scale SLIC image segmentation. From scale 1-n, the spatial resolution gradually increases. Fusion processes, based on the upper-level results, are performed on the segmentation results at higher spatial resolutions, and object merge at scale 2 is expanded to show the process of fusion.

consistent with ground objects of each time phase. A fusion process is then conducted, the rule is judging the overlap area ratio of objects *P1* and *P2*, which have the same object label *R*. The object label mentioned here is not the same concept with subsequent procedure, it is automatically numbered by the segmentation algorithm, where each object has a unique label number. The overlap area is calculated based on the statistics of overlapping region

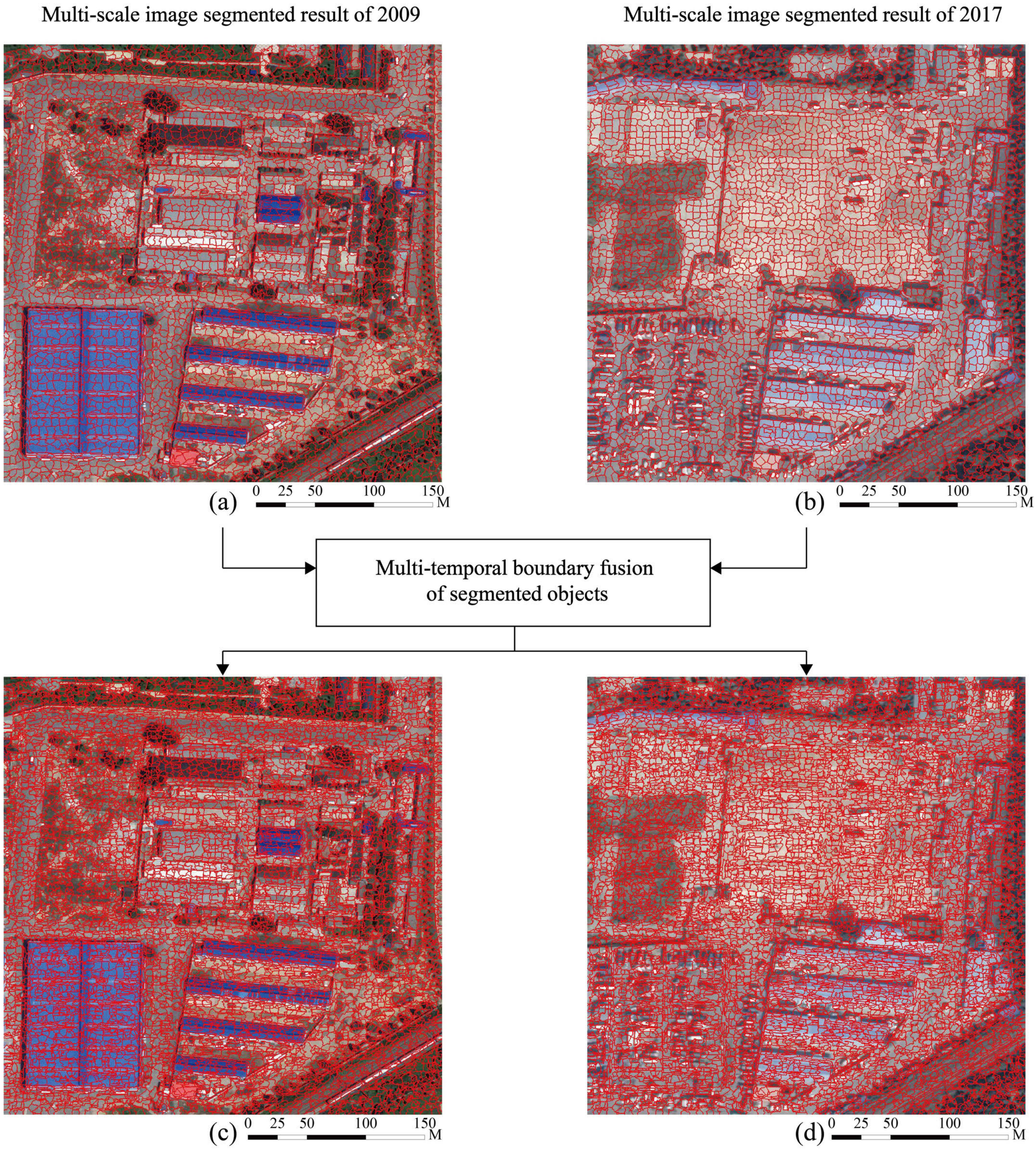
Multi-scale image segmented result of 2009
Multi-scale image segmented result of 2017
(a)
0 25 50 100 150 M
(b)
0 25 50 100 150 M
Multi-temporal boundary fusion of segmented objects
(c)
0 25 50 100 150 M
(d)
0 25 50 100 150 M

FIGURE 4. The boundary fusion of segmented objects for: (a) 2009 before boundary fusion, (b) 2017 before boundary fusion, (c) 2009 after boundary fusion, and (d) 2017 after boundary fusion.

converted by map projection. For the ratio of overlap area, larger region of the overlapped objects is selected as the denominator to ensure a moderate object fusion. The overlap area is used as the numerator only if the ratio is greater than 90% and the labels remain unchanged. Otherwise, the objects are reassigned to different labels, as shown in the equation [1]:

$$P = \begin{cases} R, & area\,(P_1 \cap P_2) \geq 90\% \\ R', & otherwise \end{cases} \tag{1}$$

where P represents segmented objects with the same label, R and R’ represent the old and new labels, respectively.

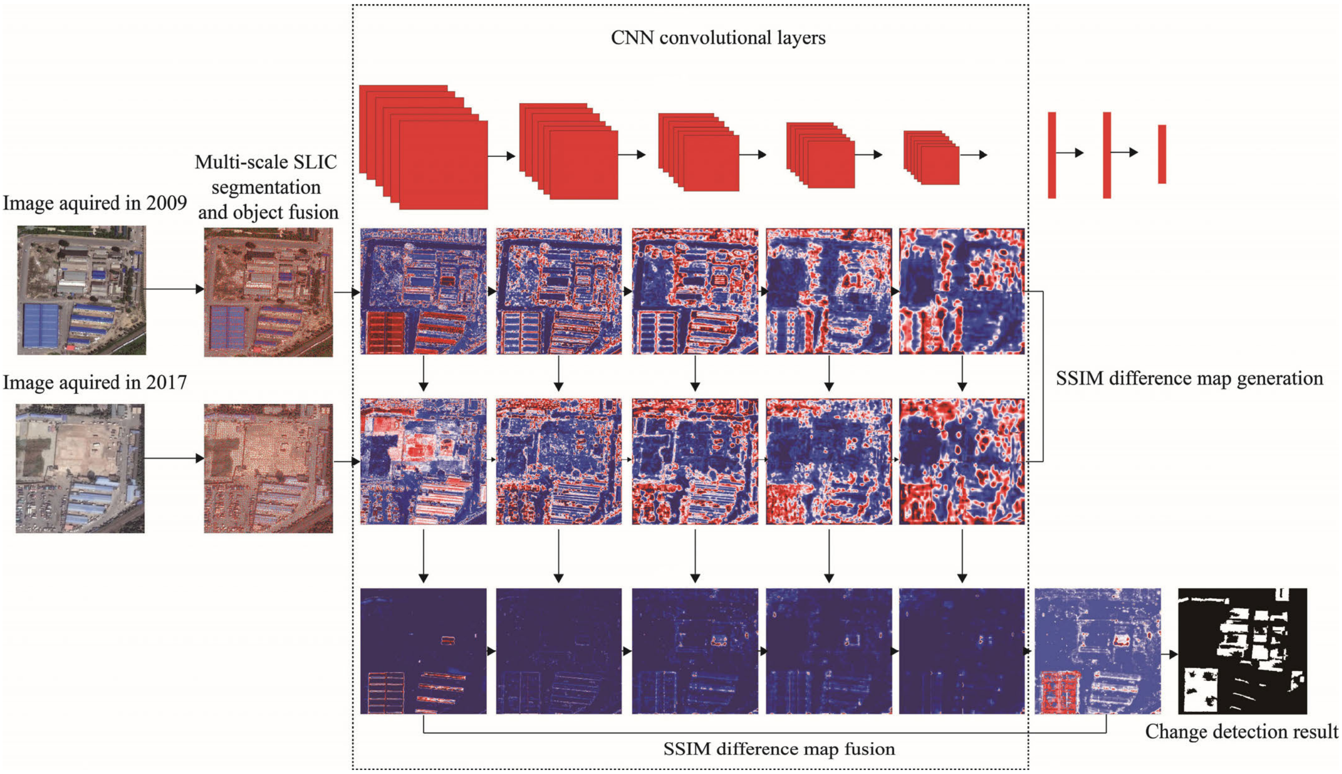
CNN convolutional layers
Multi-scale SLIC segmentation and object fusion
Image aquired in 2009
Image aquired in 2017
SSIM difference map generation
Change detection result
SSIM difference map fusion

FIGURE 5. Change map generation via the CNN architecture. In the SSIM difference maps, the red and blue areas correspond to high and low values in the feature maps, respectively.

2) CHANGE MAP GENERATION BASED ON THE CNN

Handcrafted features are typically used in change detection methods [26]. However, the extraction procedure relies on professional knowledge and parameter setting, such as size, scale, and orientation. And it is cumbersome to manipulate, which is not conducive to the implementation of change detection [27]. With segmentation results obtained from the multi-scale SLIC segmentation, a fine-tuned CNN architecture is used to automatically extract ground object features. The CNN architecture consists of a series of repeated convolutional layers, maximum pooling layers, and several fully connected layers. The final result is obtained through feed-forward process [28]. The convolutional layer outputs a series of feature maps, which are obtained by point-multiplication between trained filters and corresponding local receptive fields. The pooling layer calculates the maximum, minimum, or mean values of local regions to acquire down-sampled feature maps. Fully-connected layer synthesizes features from upper layers with fully connected nodes. With the above functional layers, hierarchical feature maps can be automatically extracted (Figure 5).

As shown in Figure 5, images fed into CNN architecture must be resampled to conform to the size requirement of input layer. However, direct resampling to whole image will result in a loss of feature details. Besides, in order to better persist feature edges, segmentation objects are transformed into image patches, which are inputted into the network to obtain feature maps of convolutional layers. More features (e.g., edges and corners) with detailed information are extracted from shallow layers (e.g., feature maps of the 1st convolutional layer). With the increase of layer depth, similar spectral and texture features are gradually merged, resulting in a gradual dominance of shape features.

Variations of contrast, brightness, and hue exist in VHR satellite images reduce robust when generating change map. Also, unchanged objects are dominant in multi-temporal images. Therefore, based on feature maps extracted from convolutional layers, the structural similarity index measure (SSIM) algorithm [29] is introduced to generate change maps. The SSIM algorithm uses intensity (s), brightness (l), and contrast (c), as depicted in equation [2]:

$$\begin{cases} l(X_1^n, X_2^n) = \dfrac{2\mu_{X_1^n}\mu_{X_2^n} + C_1}{\mu_{X_1^n}^2\mu_{X_2^n}^2 + C_1} \\ c(X_1^n, X_2^n) = \dfrac{2\sigma_{X_1^n}\sigma_{X_2^n} + C_2}{\sigma_{X_1^n}^2\sigma_{X_2^n}^2 + C_2} \\ s(X_1^n, X_2^n) = \dfrac{\sigma_{X_1^n X_2^n} + C_3}{\sigma_{X_1^n}\sigma_{X_2^n} + C_3} \end{cases} \tag{2}$$

In equation [2], μx_1^n, μx_2^n, σx_1^n, σx_2^n, and $\sigma x_1^n x_2^n$ represent mean, standard deviation, and covariance between various time phase images, respectively. Additionally, to avoid zero denominator, C_1, C_2, and C_3 are set to constants, $C_1 = (K_1 * L)^2$, $C_2 = (K_2 * L)^2$, and $C_3 = C_2/\ 2$ and, in general,

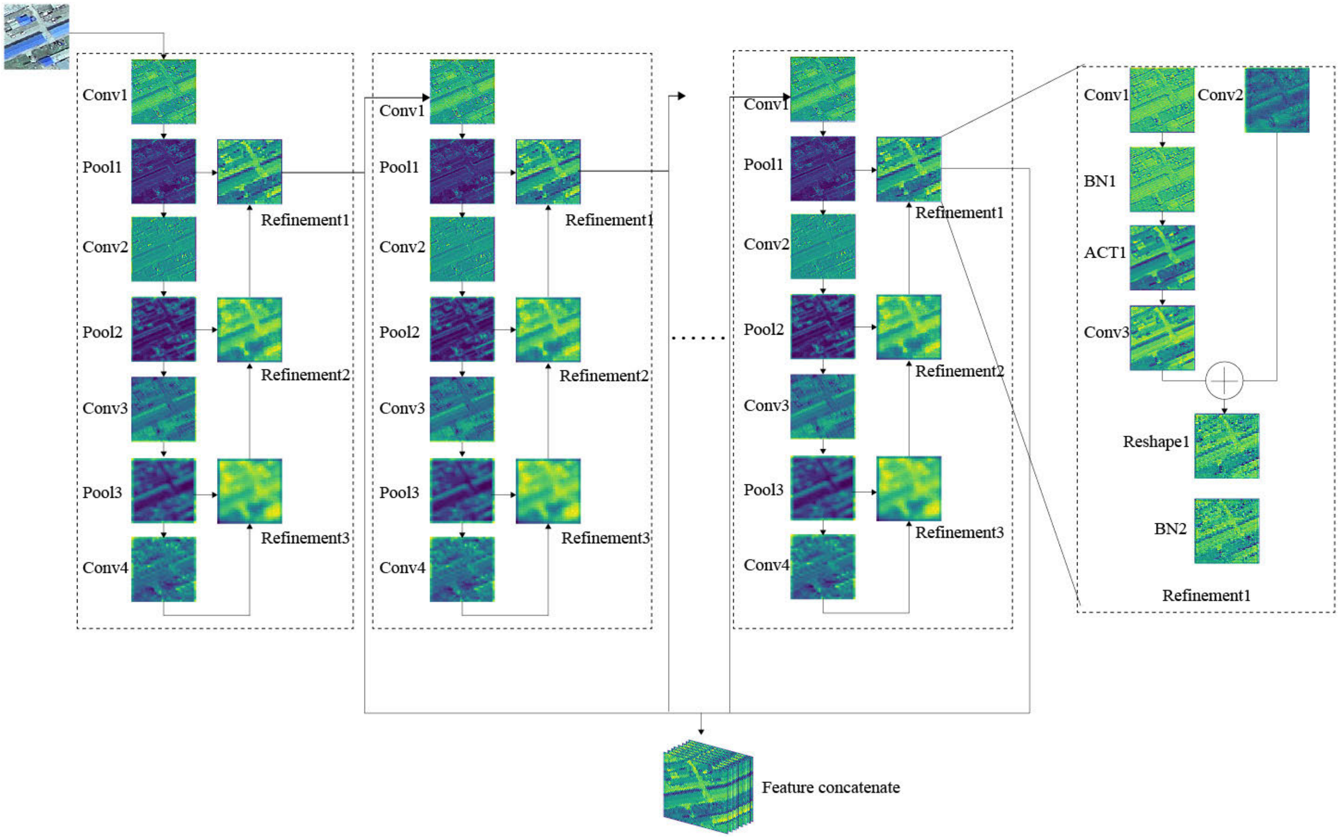

FIGURE 6. The architecture of the SCAE layer.

$K_1 = 0.01$, $K_2 = 0.03$, and L is determined according to grayscale, for 8bit google earth images, $L = 255$ [30]. We adopt Gaussian functions to calculate mean, standard deviation, and covariance to increase the execution speed. With l, c, and s, change maps could be generated with equation [3]:

$$X_d^n = [l(X_1^n, X_2^n)]^\alpha [c(X_1^n, X_2^n)]^\beta [s(X_1^n, X_2^n)]^\gamma \tag{3}$$

where n is the band number of image, and α, β, γ are used to evaluate the importance of each measure (i.e., s, l, and c, respectively) on the final calculation result (X_d^n). To integrate the influence of s, l, and c, we set $\alpha = \beta = \gamma = 1$.

Feature map pairs extracted from convolutional layers are used as input data and difference images are generated with the SSIM algorithm. Subsequently, a fusion process is conducted to fully use the difference images. While influenced by the convolutional and pooling operations, difference images are in different sizes. Therefore, the bilinear interpolation algorithm is used for resampling. Additionally, the difference images are in different dimensions, so, a summation operation is performed to reduce dimensionality. After resampling and dimensionality reduction, the difference images with same size are connected. Because of low computational complexity and automatic segmentation ability of foreground and background, the Otsu algorithm [31] is used to divide the difference images into change and non-change regions. Finally, change map is generated.

B. SCAE FEATURE EXTRACTION

1) SCAE ARCHITECTURE

In general, neural network architecture requires manual labeled dataset to train parameters before classification. However, data annotation requires abundant (albeit subjective) expert experience, which is time-intensive. Therefore, a self-learning SCAE architecture [32] is adopted for feature learning (Figure 1). The SCAE is a stacked structure, with a substructure that consists of coding and decoding components. The coding component learns compressed feature information of input images, after which the decoding component reconstructs features to generate predicted results of the input images (Figure 6).

As shown in Figure 6, the coding component mainly consists of convolutional layers, pooling layers, batch normalized layers, and activation layers. The refinement layer exists in decoding component of each substructure, which performs convolution, batch normalization, and activation operations on the feed-forward input images. Operation results are added with the output from pooling layers at the same size and level. Afterwards, the added results are up-sampled to the original size layer-by-layer. Since the SCAE is a self-learning architecture, a mean standard error (MSE) calculated from the images before and after reconstruction is used as the loss function to train the network. The feed-forward output of each substructure is used as the input for subsequent substructures, which trains the entire SCAE architecture, as expressed

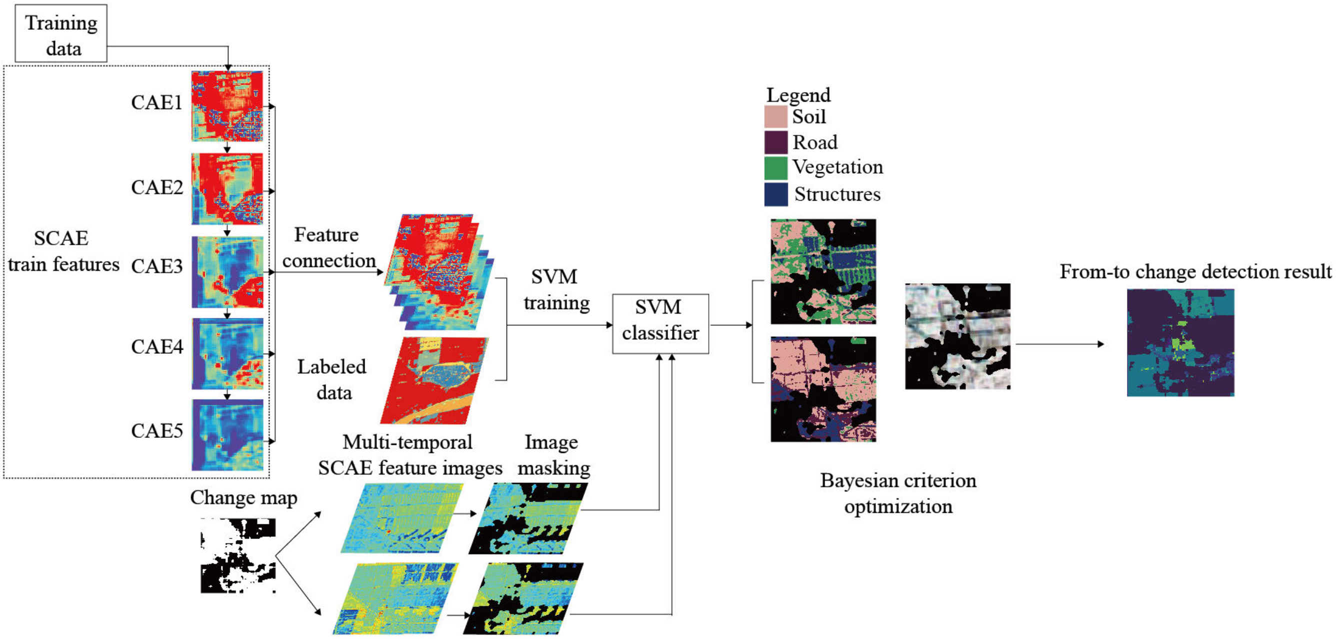

FIGURE 7. The "from-to" change information extraction based on an SCAE-SVM classifier. The various colors in the feature maps indicate values of extracted features. The white areas in change map indicate regions of change while the black indicates unchanged regions. The colors in the labeled data, Bayesian criterion optimization, and from-to change detection result indicate the types of ground objects.

in equation [4]:

$$H^k = \sigma(W^k * H^{k-1}) \tag{4}$$

where k represents the k-th substructure in the SCAE network; $H^0 = X$, where X represents the input image; σ is a nonlinear activation operation (ReLU); and W^k is the hidden layer of the k-th substructure. This type of architectural organization allows the SCAE to learn deeper image features [33].

2) FEATURE EXTRACTION BASED ON SCAE

Target images are inputted into the trained SCAE architecture to obtain the results of reconstruction. The feature maps shown in Figure 6 are resampled to obtain a uniform size for display. Taking the feature maps of front substructure as an example, in the convolution layer, the convolutional operation toward input image can automatically obtain convolutional results in different main directions and colors. Subsequently, high-dimensional features can be obtained after an activation operation. The pooling layers compress features and reduce redundant parameters. From Pool1 to Pool3, feature compression happens with a gradual coarsening of ground object feature information. Refinement3–Refinement1 layers combining with pooling layers reconstruct image features layer-by-layer. The reconstructed features of lower-level substructures are used as the input for deeper substructures. And the SCAE features of ground objects are extracted as training data for image classifier.

C. CHANGE DETECTION BASED ON CHANGE MAPS AND SCAE FEATURES

Features extracted from the SCAE architecture are combined with corresponding labels to train a support vector machine (SVM) classifier, which is a classic machine learning algorithm first used to solve classification problems. The core idea of the SVM is to find an optimal hyperplane that can distinguish different classes and maximize the classification distance between different training sample classes [34]. A small amount of labelled data with high-dimensional image features extracted by the SCAE architecture are used to train an SVM classifier. With the classifier, "from-to" change information could be obtained based on previously generated change maps (Figure 7).

As shown in Figure 7, the labelled training dataset is inputted into the SCAE architecture to obtain feature maps of the first substructure, which is subsequently used as the input data for the second substructure, layer-by-layer, until feature maps of the last substructure are finally obtained. These feature maps are concatenated, and the dimension is expanded to 320. With the concatenated feature maps and labelled data, an SVM classifier is trained. Additionally, the images to be detected are inputted into the SCAE architecture to acquire feature maps. To improve the efficiency of change detection, the binary change map generated from multi-scale SLIC-CNN is used to cover the entire study area, and the generated SCAE features are masked by the change map to exclude unchanged regions. Afterwards, the trained SVM classifier is used to classify the masked SCAE features. An optimization method based on the Bayesian information criterion is proposed to optimize the ''from-to'' change information results.

In most cases, the ''from-to'' change information is acquired based on direct comparison. However, the correlations between classification results are often neglected, resulting in error accumulation. To address the problem, time correlation of land cover types is considered in this paper. And thus, the Bayesian information criterion is introduced to optimize ''from-to'' change information results. The optimization procedure finds a combination of classification labels (l_1, l_2) that allows multi-temporal features f_1 and f_2 to achieve maximum posterior probability [35], as defined in equation [5]:

$$\max_{l_{1i},l_{2j}}\{P(l_{1i}, l_{2j}|f_1,f_2)\} \tag{5}$$

According to Bayesian information criterion, equation [5] can be rewritten as:

$$\max_{l_{1i},l_{2j}}\{\frac{P(f_1,f_2|l_{1i}, l_{2j})P(l_{1i}, l_{2j})}{P(f_1,f_2)}\} \to \max_{l_{1i},l_{2j}}\{\frac{P(f_1,f_2|l_{1i}, l_{2j})P(l_{1i}, l_{2j})P(l_{2j})}{P(f_1,f_2)}\} \tag{6}$$

where $P(f_1,f_2)$ is determined by the image itself, independent of classification labels l_1 and l_2, and therefore, does not contribute to the optimization of label pairs. Additionally, image features of ground objects are only related to the ground object types at a single time phase. Therefore, $P(f_1, f_2 |l_{1i}, l_{2j})$ can be rewritten as $P(f_1 |l_{1i})P(f_2|l_{2j})$. Equation [6] is then defined in equation [7]:

$$\max_{l_{1i},l_{2j}}\{\frac{P(l_{1i}|f_1)P(f_1)}{P(l_{1i})} \cdot \frac{P(l_{2i}|f_2)P(f_2)}{P(l_{2i})} \cdot P(l_{1i}|l_{2j})P(l_{2j})\} \tag{7}$$

where $P(f_1)$ and $P(f_2)$ are determined by spectral characteristics; $P(l_{1i})$ is invariant within a given image. The final label optimization rule based on the Bayesian information criterion can be written as equation [8]:

$$\max_{l_{1i},l_{2j}}\{P(l_{1i}|f_1)P(l_{2j}|f_2)P(l_{1i}|l_{2j})\} \tag{8}$$

where $P(l_{1i}|f_1)$ and $P(l_{2j}|f_2)$ are the conditional probabilities of classification results obtained from trained SCAE architecture, and $P(l_{1i}|l_{2j})$ is the transition probability of change among the ground objects, which is acquired by the change map generated from SLIC-CNN architecture.

III. RESULTS

The multi-temporal, Google Earth (v7.1.8) 17-level images of Beijing, Wuhan in Hubei province, and Binzhou in Shandong province were selected as image data sources (Table 1). The amount of change ground objects in the three study areas is different. In Beijing study area, there are more complex change types. In Wuhan study area, changes of industrial area exist. And in Binzhou study area, changes happen in large amount of low-rise buildings.

TABLE 1. Times and spatial resolutions of the image data per study area.

Study area	Time phase 1	Time phase 2	Spatial resolution (m)
Beijing	2015.9	2016.10	0.26
Wuhan	2015.1	2016.2	0.26
Binzhou	2015.5	2016.5	0.26

Overall accuracy (OA), user's accuracy, producer's accuracy and classification Kappa coefficient (clKC) are used to evaluate the results of ''from-to'' classification. The False Positive (FP) and False Negative (FN) rate (FP (%), FN (%)), overall error (OE) and overall error rate (OE (%)), change detection Kappa coefficient (cdKC), precision rate (Pr (%)), recall rate (Re (%)) and F1-score are used to quantitatively analyze the change detection results, which are defined in equation [9]–[14].

$$\mathrm{FP}(\%) = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} \times 100 \tag{9}$$

$$\mathrm{FN}(\%) = \frac{\mathrm{FN}}{\mathrm{FN}+\mathrm{TP}} \times 100 \tag{10}$$

$$\mathrm{OE} = \mathrm{FP}+\mathrm{FN}$$

$$\mathrm{OE}(\%) = \frac{\mathrm{FP}+\mathrm{FN}}{\mathrm{TN}+\mathrm{FP}+\mathrm{TP}+\mathrm{FN}} \times 100 \tag{11}$$

$$\mathrm{Pr}(\%) = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \times 100 \tag{12}$$

$$Re(\%) = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \times 100 \tag{13}$$

$$\mathrm{F1} = \frac{2\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})+(\mathrm{TP}+\mathrm{FP})} \tag{14}$$

where FP represents the false positive; FN is regarded as the false negative; TP and TN represent the correct changed and unchanged pixels compared with the true value, respectively.

A. PARAMETER SELECTION

The parameter sequence $p = [\sigma, \mathrm{k}]$ in multi-scale SLIC segmentation should be estimated, where σ denotes the scale parameter and k denotes the number of segmented objects. Ground objects at different scales show various characteristics, which will affect segmentation results, therefore, a scale selection procedure is necessary. To obtain image pairs under different scale values, multi-temporal images of original resolution are resampled by the bi-cubic interpolation algorithm. Subsequently, combined with field survey data, a statistical separability between different ground objects was calculated as well as scale-dispersion curves (Figure 8).

The values in scale-dispersion curves are related to the separability of ground objects, high values indicate a candidate scale. In Figure 8, the curves of most objects show a notable trough at the scale of 0.46 m, and then rise. A reason for this is that, as the scale parameter increases, the number of mixed pixels between ground objects increases rapidly, resulting in a decrease in the separability of the ground objects, representing a decreasing dispersion value. As the scale parameter further increases, the influence of mixed pixels on separability is reduced. Along with the spectral variation

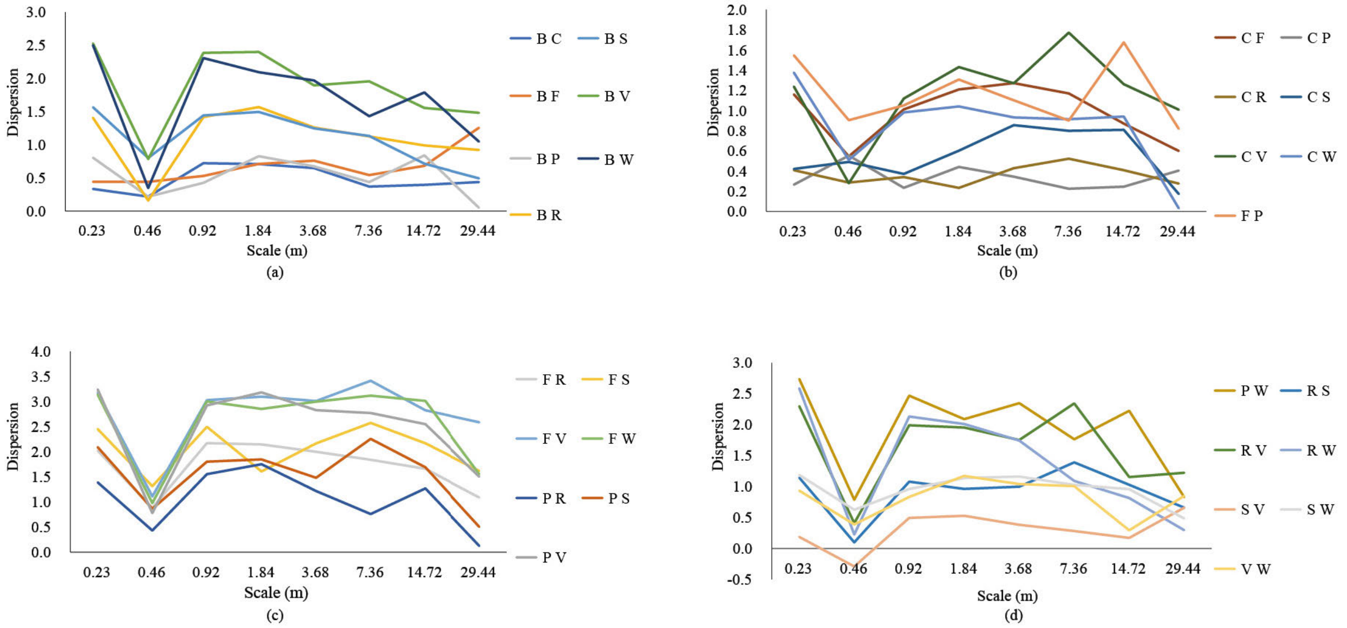

FIGURE 8. Scale-dispersion curves. In the figure, the B, C, F, P, R, S, V, and W represent buildings, structures, farmlands, parking lots, roads, bare soil, vegetation, and bodies of water, respectively. (a) Buildings with C, F, P, R, S, V, and W; (b) Structures with F, P, R, S, V, W, and farmlands with P; (c) Farmlands with R, S, V, W, and parking lots with R, S, V; (d) Parking lots with W, roads with S, V, W, bare soil with V, W, vegetation with W.

enhancement among ground objects, there is an increase in dispersion values. This happens to the adjacent ground objects in urban areas. While building-farm, structure-parking lot, and structure-bare soil show different tendencies. As the scale parameter continues increasing, there is no substantial change in dispersion. This is possibly due to the weak spatial adjacency within the ground objects, making scale parameter a limited contribution to separability. Considering the change characteristics of scale-dispersion curves and the actual distribution of ground objects, the scale parameters are selected as 0.23, 1.84, and 7.36 m. And the transformed images are segmented by the SLIC algorithm.

In the SLIC algorithm, the number of segmented objects k needs to be specified. If the value of k is too small, under-segmented results will be generated. And corresponding spectral and spatial features cannot be fully utilized, which result in weakened boundaries of segmented objects. On the versa, excessive parameter values affect the execution efficiency of algorithm. Therefore, the rate of change (ROC) [36] was introduced to select the k values. The initial k value was set to 10, with an increasing step of 10 to segment the images of selected scales (0.23, 1.84, and 7.36 m). And local variance-ROC (LV-ROC) curves were generated (Figure 9).

Local variance (LV) reflects homogeneity between segmented objects. As the number of segmented objects increases, the LV value decreases. However, the LV curve doesn't vary obviously with changing numbers of segmented objects. Therefore, it is difficult to use LV as basis for selecting k-values. Whereas, the Rate of Change (ROC) could amplify the changes in LV. The maximum ROC represents an abrupt change of LV on both sides of the k value. If the k-value is selected as candidate, notable difference between segmented objects can be maintained [37]. As shown in Figure 9, the convergence speed of LV-ROC curves differs, it is affected by different study areas and image scales. The ROC curve of Beijing converges at $k = 380$, 330 and 430 (corresponding to spatial resolutions of 0.23, 1.84, and 7.36 m, respectively). For Binzhou, the ROC curve converges at $k = 530$, 530 and 380 at 0.23, 1.84, and 7.36 m, respectively. The ROC curve of Wuhan converges at $k = 570$, 570 and 390 at 0.23, 1.84, and 7.36 m, respectively. The multi-scale segmentation with obtained parameter sequences was performed on multi-temporal images of the study areas. And the results were inputted into the CNN architecture, with SSIM algorithm, change map could be generated.

Because of good performance in classification tasks, AlexNet [14] is selected as the CNN architecture, whose weights are pre-trained based on the ImageNet dataset [38], which contains 1.2 million images. To adjust the architecture more suitable for remote sensing applications, the Land Use Dataset from the University of California (UC) Merced, USA [39] was used to fine-tune the whole architecture. The UC Merced Land Use Dataset includes land cover types that can be recognized by high-resolution satellite images. It includes 21 land cover types (i.e., agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis court), each of which includes 100 images with spatial resolution of 0.3 m and size of 256 × 256. For each image in the dataset, its unique name is combined with a corresponding label

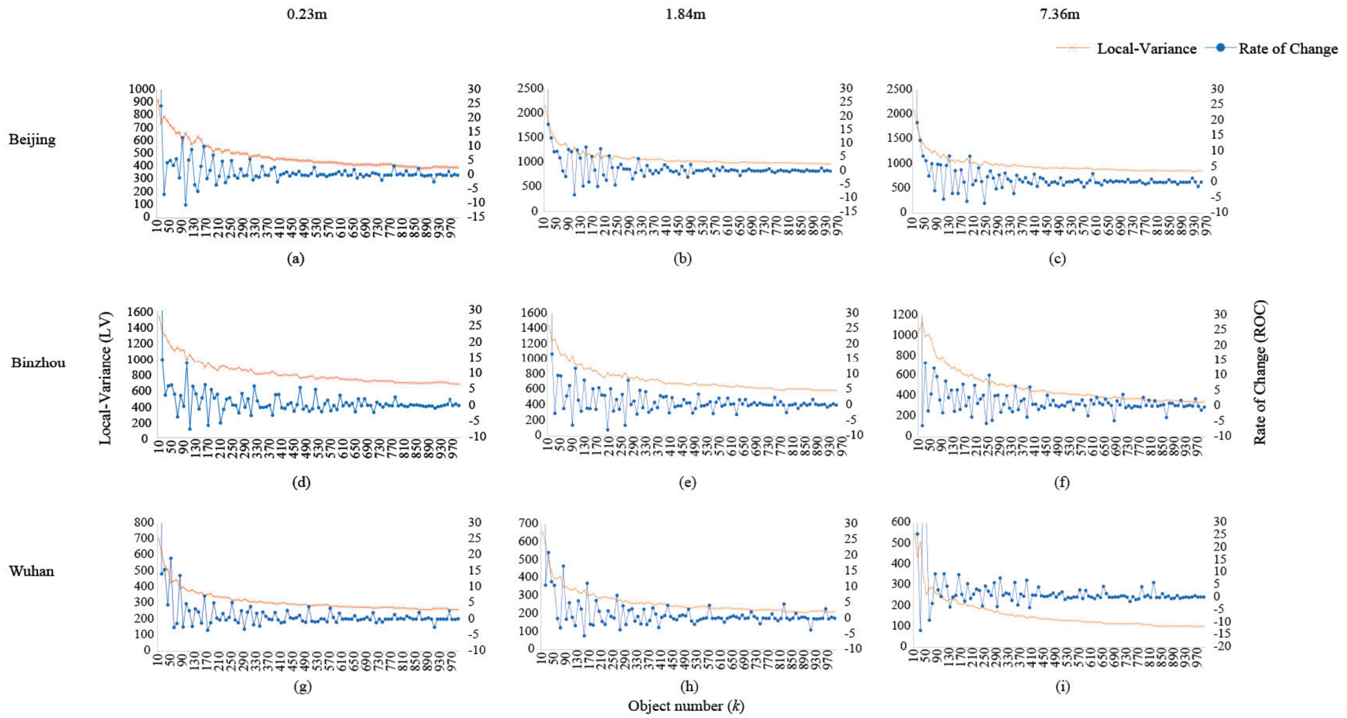

FIGURE 9. **Local variance-rate of change (LV-ROC) curves. (a)-(c) are curves of Beijing, (d)-(f) are curves of Binzhou, and (g)-(i) are curves of Wuhan at spatial resolutions of 0.23, 1.84, and 7.36 m respectively.**

from 0-20. The images in the dataset were enhanced by rotating, mirroring, and noise addition. And the dataset was finally expanded to contain 50,000 images.

The 50,000 images were divided into training and testing datasets with a 9:1 ratio to ensure an adequate number of images to update parameters and avoid over-fitting [40]. We choose softmax loss as loss function [14] and Adam optimizer for parameter optimizing [41]. The batch size is set to 10, which means that ten images are taken from the training dataset for each iteration. The initial learning rate is set to 10^{-5}, and divided by 10 every 2,000 iterations. To prevent over-fitting, the parameter value of weight decay is set to 0.02. The momentum parameter is 0.9 to accelerate gradient descent procedure.

For the SCAE architecture, input layer size is $32 \times 32 \times 3$ for finer details. To generate training and testing dataset, the images in large size were randomly cropped to yield a cropped dataset, and two-thirds images of the study areas were used. The dataset had 50,000 images with the same size of input layer. 45,000 images were randomly separated as the training dataset and the remaining 5,000 images were used as the testing dataset. Mean-squared-error is utilized as loss function and an Adam optimizer is used to speed up the convergence process [41]. The initial learning rate 10^{-5} is set to be divided by 10 to continue training if the test loss doesn't decrease for 5 consecutive iterations. The CNN and SCAE architectures were trained for approximately 4,000 epochs. The training process terminated when the test loss didn't decrease for 10 consecutive iterations. The implementation of the architectures is based on Keras [42] framework with Tensorflow [43] backend running on an NVIDIA GeForce GTX TITAN video card equipped with 12 GB of memory.

B. COMPARISON WITH OTHER METHODS ON CHANGE DETECTION

Google Earth 20-level images (spatial resolution: 0.11 m, the acquisition date was similar, no more than one month before and after those of 17-level images) were used as the reference images. Binary and ''from-to'' reference maps were generated by manually vectorizing to obtain change information and change type information. Therefore, change detection accuracy of different methods can be quantitatively evaluated.

To validate the potential prospect of proposed method in change detection, a traditional non-deep learning method (IR-MAD) is introduced. Moreover, two deep learning methods U-Net and SegNet are employed in comparison.

IR-MAD: IR-MAD has the ability of capturing image change status efficiently with multivariate characteristics. And it is widely used in change detection [44], [45].

U-Net: U-Net was originally proposed for biomedical image segmentation [46]. It consists of an encoder part and a decoder part. The encoder part extracts features of the image with convolution and pooling layers. The decoder part whereas recovers the image details with upsampling and short connections. Because of the high accuracy on large image segmentation, U-Net has been used in many remote sensing tasks [47], [48].

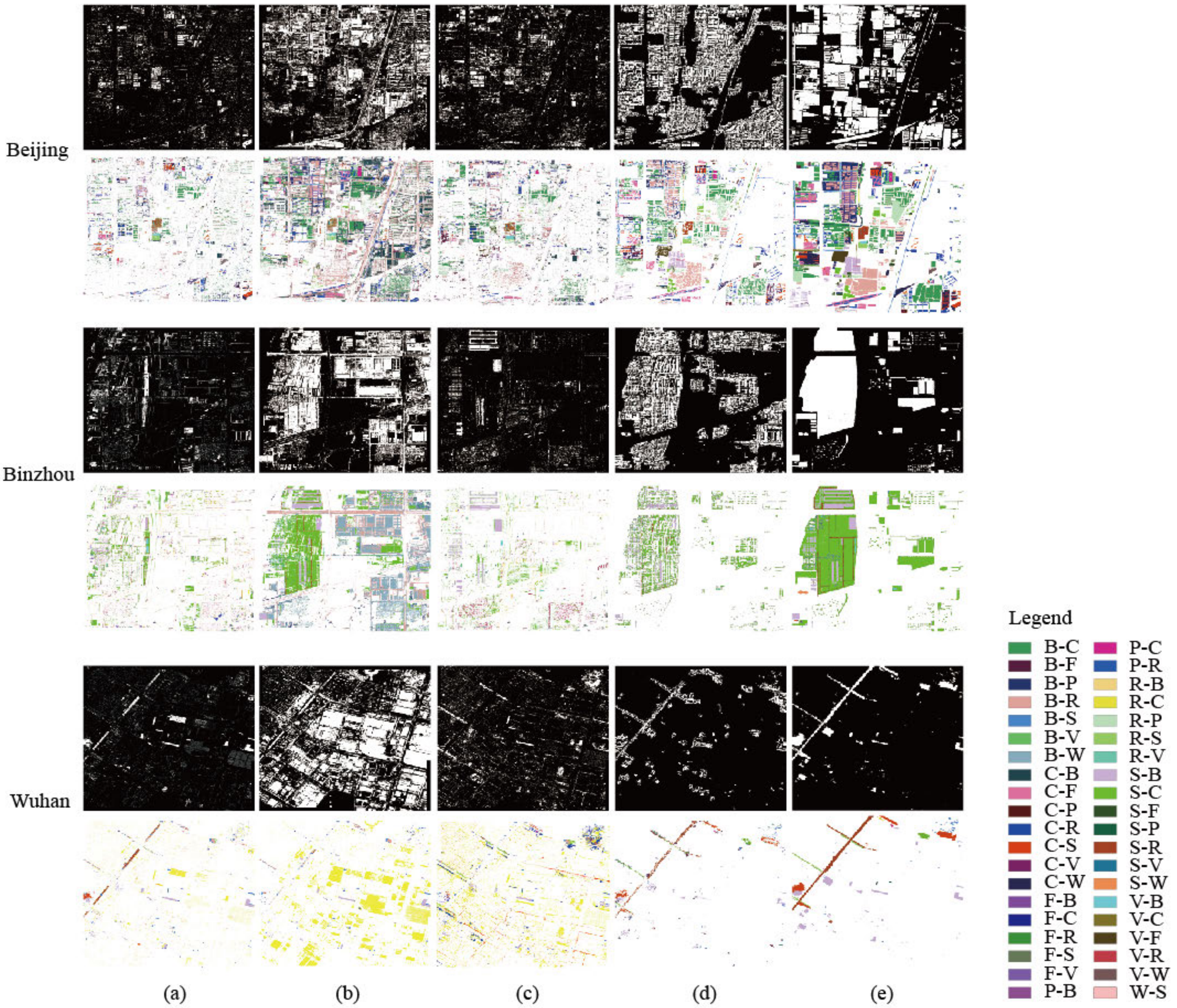

FIGURE 10. Change detection results of (a) IR-MAD, (b) U-Net, (c) SegNet, (d) The proposed method. And (e) Reference map.

SegNet: Badrinarayanan *et al.* [49] proposed SegNet for road scene segmentation, it is a symmetric encoder-decoder architecture with shortcut connections, and partly based on VGG architecture. Compared to other architectures, SegNet expends less memory and affords high accuracy.

The "from-to" change information of IR-MAD was obtained with an SVM classifier. And change detection results of four methods are shown in Figure 10.

As shown in Figure 10, there are FPs and FNs in the results of each change detection methods, where the U-Net extracts most false positive patches in each study area. Compared with U-Net, the amount of false positive patches is reduced in the results of IR-MAD and SegNet. Similarly, false negative patches exist in the results of both methods. Change regions are not effectively extracted. The visual effect of proposed method is more consistent with the reference maps. For the results of "from-to" change information extraction, IR-MAD directly used the SVM classifier trained from spectral and textural features to classify multi-temporal images separately. Broken transformation of land cover types is in the results. In reverse, the deep learning based methods obtain uniform and consistent change detection results. A confusion matrix was also produced to quantitatively evaluate the performance of the SVM and SCAE classifiers (Figure 11 and Table 2).

As shown in Figure 11 and Table 2, the accuracy of same classifier in each study area changes slightly in the multi-temporal images, indicating a relatively stable classification performance. Overall, the accuracy of SCAE classifier is higher than that of the SVM classifier in each study area. The accuracies of the SVM and SCAE classifiers in Beijing

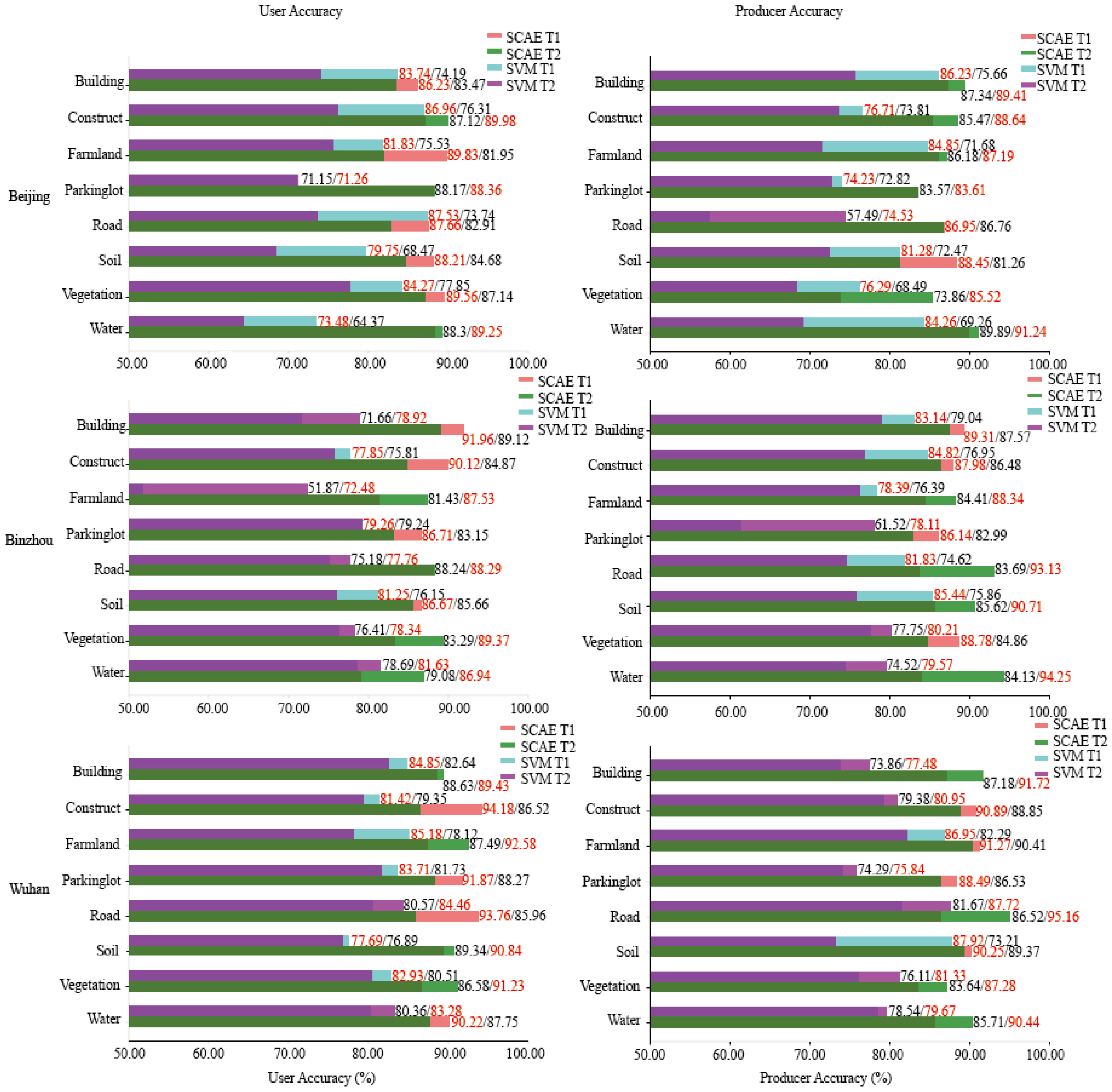

FIGURE 11. Confusion matrices of the SVM and SCAE classifiers for the user's and producer's accuracy in each study area at time phases T1 and T2.

TABLE 2. Overall accuracy and classification kappa coefficient of svm and scae classifiers in each study area at time phases T1 and T2.

		Beijing		Binzhou		Wuhan	
		T1	T2	T1	T2	T1	T2
SVM	OA	79.43	72.86	78.67	76.61	81.59	79.98
	clKC	0.64	0.61	0.56	0.54	0.73	0.68
SCAE	OA	87.75	89.86	90.25	92.68	91.56	92.24
	clKC	0.79	0.76	0.81	0.85	0.85	0.83

are lower than those in other study areas, because the distribution and types of land cover in Beijing are more complex. Table 3 lists the accuracy of the change detection results.

Table 3 shows that, due to the large number of changed ground objects in Beijing, as well as their complex types, each change detection algorithm has a poor performance.

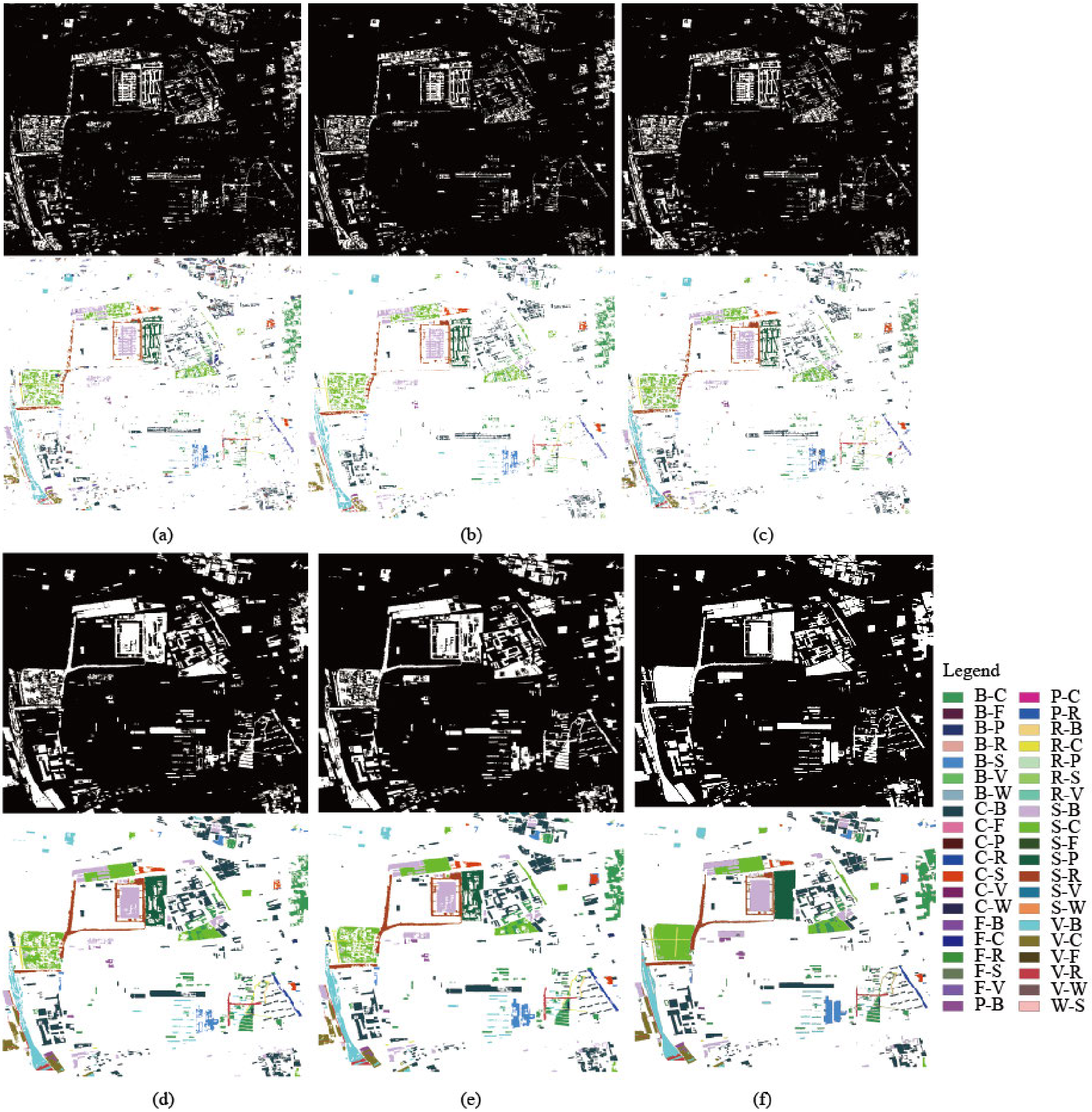

FIGURE 12. Comparison results of controlled experiments. (a) Method 1, (b) Method 2, (c) Method 3, (d) Method 4 and (e) The proposed method. (f) Reference map. Binary images are the detected change maps and color images are the results of the "from-to" change detection.

The method based on IR-MAD algorithm achieves the lowest FP rate but has the highest FN rate. The method based on U-Net gains the highest FP rate and a high FN rate. Meanwhile, the method based on SegNet algorithm achieves a relative low FP rate and FN rate. Finally, the proposed method has the lowest OE value and the highest cdKC, Pr (%), Re (%), F1 values.

There are numerous areas changing from bare soil to buildings or other man-made structures in Binzhou. The method based on IR-MAD algorithm has the lowest FP rate in study area. However, large numbers of undetected objects result in a high FN rate. The method based on U-Net achieves the lowest FN rate, but a high FP rate. While the SegNet obtains low FP rate and the generated FN rate is moderate, resulting in a better performance than U-Net. The proposed method achieves the highest change detection accuracy, i.e., the highest F1 value.

In Wuhan, all four change detection methods performed well, but the proposed method achieves the highest accuracy. However, since the SVM classifier used in IR-MAD

TABLE 3. Quantitative evaluation of the change detection results based on the method employed, i.e., the IR-MAD, U-Net, SegNet, and the proposed method, for each locality. Values in bold represent the most accurate classifications, where parenthetical values are the percentages of corresponding pixels.

	Beijing						
	FP (%)	FN (%)	OE (%)	cdKC	Pr(%)	Re(%)	F1
IR-MAD	**4,685,309** **(3.99)**	35,175,981 (29.92)	39,861,290 (33.91)	0.4190	56.5	42.3	0.4951
U-Net	23,967,791 (20.39)	25,685,594 (21.85)	49,653,385 (42.24)	0.5018	50.9	65.7	0.6494
SegNet	15,876,453 (12.95)	**18,021,199** **(15.17)**	33,897,652 (28.12)	0.6217	83.4	71.2	0.7546
Proposed method	14,449,756 (12.29)	18,179,278 (15.46)	**32,629,034** **(27.75)**	**0.7164**	**86.3**	**79.8**	**0.8040**
	Binzhou						
	FP (%)	FN (%)	OE (%)	cdKC	Pr(%)	Re(%)	F1
IR-MAD	**4,231,875** **(3.51)**	25,687,696 (21.30)	29,919,571 (24.81)	0.5302	58.3	44.9	0.5684
U-Net	27,704,020 (22.97)	**12,097,142** **(10.03)**	39,801,162 (33.00)	0.7113	75.1	82.6	0.7860
SegNet	10,413,048 (8.39)	16,456,387 (14.87)	26,869,435 (23.26)	0.7643	**83.9**	75.7	0.8156
Proposed method	14,033,617 (11.64)	13,664,791 (11.33)	**27,698,408** **(22.97)**	**0.7841**	77.2	**85.4**	**0.8285**
	Wuhan						
	FP (%)	FN (%)	OE (%)	cdKC	Pr(%)	Re(%)	F1
IR-MAD	4,797,340 (4.07)	13,388,713 (11.35)	18,186,053 (15.42)	0.6323	70.8	58.2	0.6365
U-Net	10,467,335 (8.87)	3,614,341 (3.06)	14,081,676 (11.93)	0.8504	82.7	**92.1**	0.8537
SegNet	**3,430,254** **(2.91)**	3,844,083 (3.26)	7,274,337 (6.17)	0.9071	94.3	86.4	0.9087
Proposed method	4,380,912 (3.71)	**2,658,829** **(2.25)**	**7,039,741** **(5.96)**	**0.9443**	**96.0**	90.8	**0.9458**

only took advantage of the spectral features without other features, the classifier presents a poor performance. This ultimately leads to an accumulation of error in ''from-to'' change information extraction, which further causes low detection accuracies.

Additionally, the computational efficiency performance of four methods on Beijing study area are shown in Table 4,

Table 4 lists the training, testing and total time. For a fair comparison, the batch size parameters for deep learning based methods are set to 4. The table shows that SegNet consumes most time compared to other three methods. Although the proposed method takes more time than U-Net, it outperforms other methods. Overall, the proposed method balances accuracy and time consumption.

IV. DISCUSSION

The proposed method consists of image segmentation, change map generation, and change information extraction based on SCAE features. Any part has potential influence on the final accuracy of change detection results. Additionally, the parameters of CNN and SCAE used in proposed method were trained separately, while many change detection methods based on deep learning are trained through end-to-end manner. To further validate the proposed method in this study, controlled experiments were conducted with adjusted architectures and end-to-end training manner.

1) Method 1: Without image segmentation, features extracted from the CNN architecture and SSIM algorithm are directly used to generate the change map. The other components in this method keep the same as those used in the proposed method. Thereby, the influence that image segmentation on change detection results is analyzed.
2) Method 2: Based on the segmentation results of multi-scale SLIC, a change map is obtained using the SSIM algorithm without CNN features. The other components in this method are the same as those in the proposed method to analyze the influence of CNN features on the change detection results.
3) Method 3: Based on the change map generated from the multi-scale SLIC-CNN algorithm, an SVM classifier is trained by adopting spectral and textural features of sample points from the study areas. The trained SVM classifier is then used to classify the multi-temporal images masked by change maps to analyze the influence of SCAE features on change detection.
4) Method 4: With separately pre-trained weights of CNN and SCAE by the proposed method, the parameters are further trained through an end-to-end manner, and the training parameters are refer to [50], a SGD optimizer and MSE loss are used. The initial learning rate is 0.01 and a reducing factor of 0.1 after 2000 iterations.

TABLE 4. Time performances of different methods.

Method	Training (1 epoch)	Testing	Total
IR-MAD	-	340.8	340.8
U-Net	1332.9	11.1	1344
SegNet	2843.4	12.8	2856.2
Proposed method	2644.1	13.7	2657.8

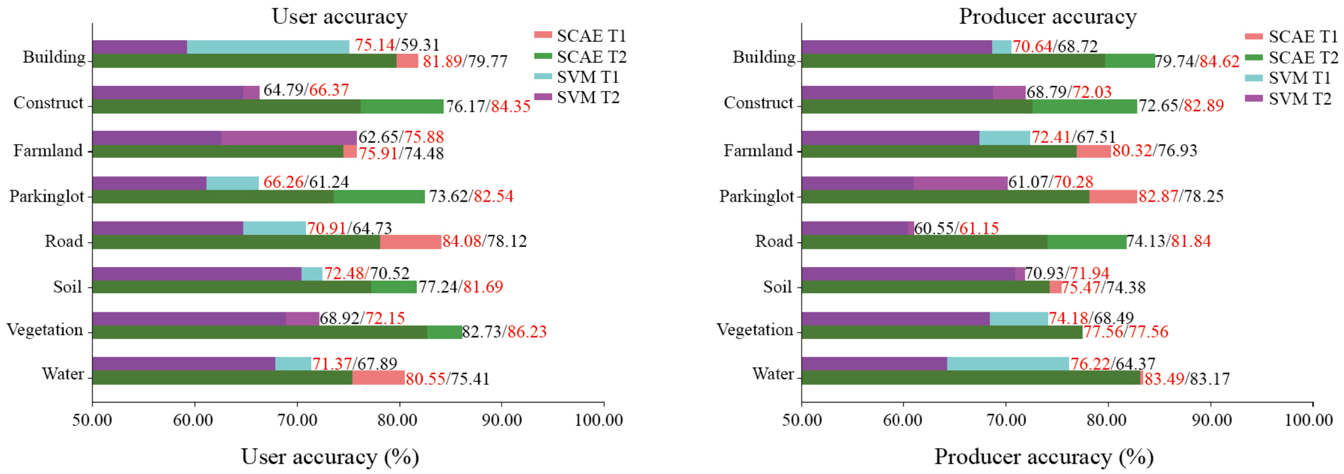

FIGURE 13. Confusion matrices of the SVM and SCAE for the user's and producer's accuracy of controlled experiments in the T1 and T2 time phases.

TABLE 5. Overall accuracy and classification kappa coefficient of svm and scae classifiers in each study area of controlled experiments at time phases T1 and T2.

		T1	T2
SVM	OA	60.06	61.71
	clKC	0.55	0.56
SCAE	OA	82.15	81.86
	clKC	0.75	0.73

The momentum parameter is set to 0.9 and the weight decay parameter 0.005. The training process continued for about 8500 epochs until val-loss converged. Other parts keep same with the proposed method.

Based on the above methods, Figure 12 was generated to show the change detection results.

As shown in Figure 12, the visual results of proposed method are superior to those of other four methods, with fewer FPs and FNs, as well as more consistent results with the reference maps. Method 1 uses pixel-based features extracted from CNN architecture to generate change maps, where more defects exist along edges of ground objects. In Method 2, the multi-scale SLIC and SSIM algorithms are used directly to calculate the change maps. Figure 12 shows that the details of generated change maps are relatively broken, with a generation of numerous falsely extracted objects. The change detection procedure is conducted directly on the masked multi-temporal images with a trained SVM classifier in Method 3. And an overlay analysis is performed to obtain changes in land cover types. There are more FPs and FNs generated in the results of this method. The results obtained by Method 4 is similar to the proposed method, because they are more consistent in architecture but the training manner. While there are more undetected change objects exist. To compare the accuracy of classifiers used in above methods, confusion matrices were introduced to quantitatively evaluate the performance of the SVM and SCAE classifiers (Figure13 and Table 5).

As shown in Figure 13 and Table 5, since the trained SVM classifier only uses spectral and textural information of sample points selected from the study areas, its classification accuracy is relatively low at the T1 and T2 time phases. The accuracy of the SCAE classifier is higher than that of the SVM classifier, where the value of kappa coefficient also increases from 0.5 to 0.7. The accuracy of change detection is shown in Table 6.

Table 6 indicates that the accuracy of the proposed method is higher than those of the other four methods. Although Method 4 applies an end-to-end training manner with pre-trained weights, its accuracy is still lower than that of the proposed method. Except that, the additional training procedure spends more time. Method 3 obtains the lowest

TABLE 6. Change detection evaluation of controlled experiments in the study areas via false positive (FP), false negative (FN), overall error (OE), change detectison kappa coefficient (cdKC), precision, recall rate (pr(%), re(%)) and the F1 score (F1) tests. Values in bold represent the most accurate classifications, where parenthetical values are the percentages of corresponding pixels.

	FP (%)	FN (%)	OE (%)	cdKC	Pr(%)	Re(%)	F1
Method 1	1,910,744 (1.63)	11,782,017 (10.02)	13,692,761 (11.65)	0.8727	93.8	86.0	0.8957
Method 2	**752,686** **(0.64)**	11,936,944 (10.15)	12,689,630 (10.79)	0.8798	88.3	85.1	0.8716
Method 3	1,571,716 (1.34)	11,726,679 (9.97)	13,298,395 (11.31)	0.8766	91.1	78.7	0.8692
Method 4	3,286,347 (2.71)	4,726,940 (4.52)	8,013,287 (7.23)	0.9498	91.3	**95.7**	0.9626
Proposed method	2,855,352 (2.43)	**4,211,564** **(3.58)**	**7,066,916** **(6.01)**	**0.9571**	**97.6**	90.2	**0.9640**

accuracy, because the SVM classifier only uses spectral and textural features without SCAE features during the training process, which affected the change detection accuracy. Method 1 generates change maps without image segmentation, resulting in high FP, FN rates and lower F1 value compared to the proposed method. It proves that image segmentation is contribute to performance improvement of change detection, which is in accordance with the conclusion presented by Lei *et al.* [6]. Method 2 does not use the CNN image features but only an SSIM algorithm to generate the change maps, resulting in a high percentage of FNs.

V. CONCLUSION

A novel method of change detection with VHR satellite images is proposed in this study. The method combines multi-scale SLIC-CNN and SCAE features that effectively address problems associated with change detection using VHR satellite images. Compared with existing change detection methods, the proposed method uses the self-learning SCAE architecture as the feature extractor to integrate multi-scale, spectral, geometric, textural, and deep structural features to enhance the characteristics of ground objects in images. Additionally, the SLIC image segmentation algorithm and CNN architecture avoid poor performance in the ‘‘from-to’’ change information extraction due to direct classification. The results of controlled experiments demonstrate that without image segmentation, the change detection results are more fragmented and the FP and FN rates increase, indicating that the multi-scale SLIC image segmentation affects the integrity of change detection results. Merely using the SSIM algorithms to generate change maps without features of CNN architecture results in broken objects in change maps, indicating that the features of CNN architecture affect the consistency of the change maps. When the SVM classifier, trained with spectral and texture features without the SCAE features, is directly applied to the multi-temporal images, the change detection accuracy is low, indicating that the SCAE features also affect the performance of the SVM model. The comparison results with other methods indicate that the proposed method can obtain more accurate change detection results in areas with clear features and less interference. Simultaneously, the deep learning feature extractor can enhance the extracting ability of ground objects. In future studies, we will more focus on the development of change detection methods based on the deep learning features of multi-temporal images from various sensors.

REFERENCES

[1] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, ‘‘Deep learning in remote sensing applications: A meta-analysis and review,’’ *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 166–177, Jun. 2019.

[2] 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.

[3] M. Hussain, D. Chen, A. Cheng, H. Wei, and D. Stanley, ‘‘Change detection from remotely sensed images: From pixel-based to object-based approaches,’’ *ISPRS J. Photogramm. Remote Sens.*, vol. 80, pp. 91–106, Jun. 2013.

[4] M. D. Hossain and D. Chen, ‘‘Segmentation for object-based image analysis (OBIA): A review of algorithms and challenges from remote sensing perspective,’’ *ISPRS J. Photogramm. Remote Sens.*, vol. 150, pp. 115–134, Apr. 2019.

[5] 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, May 2017.

[6] Y. Lei, X. Liu, J. Shi, C. Lei, and J. Wang, ‘‘Multiscale superpixel segmentation with deep features for change detection,’’ *IEEE Access*, vol. 7, pp. 36600–36616, 2019.

[7] M. Hao, M. Zhou, J. Jin, and W. Shi, ‘‘An advanced superpixel-based Markov random field model for unsupervised change detection,’’ *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1401–1405, Aug. 2020.

[8] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, ‘‘Change detection in synthetic aperture radar images based on deep neural networks,’’ *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 1, pp. 125–138, Jan. 2016.

[9] W. Wiratama, J. Lee, S.-E. Park, and D. Sim, ‘‘Dual-dense convolution network for change detection of high-resolution panchromatic imagery,’’ *Appl. Sci.*, vol. 8, no. 10, p. 1785, Oct. 2018.

[10] C. Zhang, S. Wei, S. Ji, and M. Lu, ‘‘Detecting large-scale urban land cover changes from very high resolution remote sensing images using CNN-based classification,’’ *ISPRS Int. J. Geo-Inf.*, vol. 8, no. 4, p. 189, Apr. 2019.

[11] W. Wiratama, J. Lee, and D. Sim, ‘‘Change detection on multispectral images based on feature-level U-Net,’’ *IEEE Access*, vol. 8, pp. 12279–12289, 2020.

[12] G. E. Hinton, S. Osindero, and Y.-W. Teh, ‘‘A fast learning algorithm for deep belief nets,’’ *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

[13] R. Collobert and S. Bengio, ‘‘Links between perceptrons, MLPs and SVMs,’’ presented at the 21st Int. Conf. Mach. Learn. (ICML), Banff, AB, Canada, 2004, doi: 10.1145/1015330.1015415.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ‘‘ImageNet classification with deep convolutional neural networks,’’ in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, vol. 1, 2012, pp. 84–90.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.

[16] B. Wang, X. Lu, X. Zheng, and X. Li, "Semantic descriptions of high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1274–1278, Aug. 2019.

[17] M. Yang, L. Jiao, F. Liu, B. Hou, and S. Yang, "Transferred deep learning-based change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6960–6973, Sep. 2019.

[18] B. Hou, Y. Wang, and Q. Liu, "Change detection based on deep features and low rank," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2418–2422, Dec. 2017.

[19] P. Zhang, M. Gong, L. Su, J. Liu, and Z. Li, "Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 116, pp. 24–41, Jun. 2016.

[20] M. Wang, K. Tan, X. Jia, X. Wang, and Y. Chen, "A deep siamese network with hybrid convolutional feature extraction module for change detection based on multi-sensor remote sensing images," *Remote Sens.*, vol. 12, no. 2, pp. 18–205, Jan. 2020.

[21] S. H. Khan, X. He, F. Porikli, and M. Bennamoun, "Forest change detection in incomplete satellite images with deep neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5407–5423, Sep. 2017.

[22] Y. Sun, X. Zhang, J. Huang, H. Wang, and Q. Xin, "Fine-grained building change detection from very high-spatial-resolution remote sensing images based on deep multitask learning," *IEEE Geosci. Remote Sens. Lett.*, early access, Sep. 7, 2020, doi: 10.1109/LGRS.2020.3018858.

[23] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[24] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[25] X. Lv, D. Ming, Y. Chen, and M. Wang, "Very high resolution remote sensing image classification with SEEDS-CNN and scale effect analysis for superpixel CNN classification," *Int. J. Remote Sens.*, vol. 40, no. 2, pp. 506–531, Jan. 2019.

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

[27] C. E. Woodcock, T. R. Loveland, M. Herold, and M. E. Bauer, "Transitioning from change detection to monitoring with remote sensing: A paradigm shift," *Remote Sens. Environ.*, vol. 238, Mar. 2020, Art. no. 111558.

[28] Y. Chen, Y. Yang, W. Wang, and C.-C.-J. Kuo, "Ensembles of feedforward-designed convolutional neural networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Taipei, Taiwan, Sep. 2019, pp. 3796–3800.

[29] A. Łoza, L. Mihaylova, D. Bull, and N. Canagarajah, "Structural similarity-based object tracking in multimodality surveillance videos," *Mach. Vis. Appl.*, vol. 20, no. 2, pp. 71–83, Feb. 2009.

[30] H. Zhuang, K. Deng, H. Fan, and S. Ma, "A novel approach based on structural information for change detection in SAR images," *Int. J. Remote Sens.*, vol. 39, no. 8, pp. 2341–2365, Apr. 2018.

[31] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.

[32] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *Proc. Int. Conf. Artif. Neural Netw.*, Espoo, Finland, 2011, pp. 52–59.

[33] R. Kemker and C. Kanan, "Self-taught feature learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2693–2705, May 2017.

[34] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[35] R. Touati, M. Mignotte, and M. Dahmane, "Multimodal change detection in remote sensing images using an unsupervised pixel pairwise-based Markov random field model," *IEEE Trans. Image Process.*, vol. 29, pp. 757–767, 2020.

[36] L. Drăgut, D. Tiede, and S. R. Levick, "ESP: A tool to estimate scale parameter for multiresolution image segmentation of remotely sensed data," *Int. J. Geographical Inf. Sci.*, vol. 24, no. 6, pp. 859–871, Apr. 2010.

[37] L. Drăgut, O. Csillik, C. Eisank, and D. Tiede, "Automated parameterisation for multi-scale image segmentation on multiple layers," *ISPRS J. Photogramm. Remote Sens.*, vol. 88, pp. 119–127, Feb. 2014.

[38] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[39] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst. (GIS)*, San Jose, CA, USA, 2010, pp. 270–279.

[40] A. Romero, C. Gatta, and G. Camps-Valls, "Unsupervised deep feature extraction for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1349–1362, Mar. 2016.

[41] S. Tang, C. Shen, D. Wang, S. Li, W. Huang, and Z. Zhu, "Adaptive deep feature learning network with nesterov momentum and its application to rotating machinery fault diagnosis," *Neurocomputing*, vol. 305, pp. 1–14, Aug. 2018.

[42] D. Graziotin and P. Abrahamsson, "A Web-based modeling tool for the SEMAT essence theory of software engineering," *J. Open Res. Softw.*, vol. 1, no. 1, p. e4, 2013.

[43] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2016, *arXiv:1603.04467*. [Online]. Available: http://arxiv.org/abs/1603.04467

[44] B. Wang, J. Choi, S. Choi, S. Lee, P. Wu, and Y. Gao, "Image fusion-based land cover change detection using multi-temporal high-resolution satellite images," *Remote Sens.*, vol. 9, no. 8, p. 804, Aug. 2017.

[45] A. A. Nielsen, "The regularized iteratively reweighted MAD method for change detection in Multi- and hyperspectral data," *IEEE Trans. Image Process.*, vol. 16, no. 2, pp. 463–478, Feb. 2007.

[46] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assisted Intervent.*, Munich, Germany, 2015, pp. 234–241.

[47] X. Zhao, Y. Yuan, M. Song, Y. Ding, F. Lin, D. Liang, and D. Zhang, "Use of unmanned aerial vehicle imagery and deep learning UNet to extract rice lodging," *Sensors*, vol. 19, no. 18, p. 3859, Sep. 2019.

[48] N. He, L. Fang, and A. Plaza, "Hybrid first and second order attention UNET for building segmentation in remote sensing images," *Sci. China Inf. Sci.*, vol. 63, no. 4, pp. 1–140305, Apr. 2020.

[49] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[50] Z. Huang, Z. Pan, and B. Lei, "Transfer learning with deep convolutional neural network for SAR target classification with limited labeled data," *Remote Sens.*, vol. 9, no. 9, p. 907, Aug. 2017.

RAN JING received the Ph.D. degree in cartography and geographical information system from the College of Resource Environment and Tourism, Capital Normal University, Beijing, China, in 2020.

He is currently a Lecturer with the School of Geosciences, Yangtze University, Wuhan, China. His research interests include image classification and segmentation, pattern recognition, and computer vision methods with applications in remote sensing.

ZHAONING GONG received the Ph.D. degree in ecohydrology from the Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Beijing, China, in 2006.

She is currently a Professor with the College of Resource Environment and Tourism, Capital Normal University, Beijing. Her current research interests include GIS and remote sensing and biophysical and biochemical parameters extraction.

HONGLIANG GUAN received the M.Sc. degree in applied animal science from Tokyo University, Tokyo, Japan, in 2001, and the Ph.D. degree in remote sensing image processing from the College of Resource Environment and Tourism, Capital Normal University, Beijing, China, in 2010.

His current research interests include monitoring technology and evolution mechanism of urban-ground subsidence.