Unsupervised Deep Slow Feature Analysis for Change Detection in Multi-Temporal Remote Sensing Images

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

Abstract—Change detection has been a hotspot in remote sensing technology for a long time. With the increasing availability of multi-temporal remote sensing images, numerous change detection algorithms have been proposed. Among these methods, image transformation methods with feature extraction and mapping could effectively highlight the changed information and thus has better change detection performance. However, changes of multi-temporal images are usually complex, existing methods are not effective enough. In recent years, deep network has shown its brilliant performance in many fields including feature extraction and projection. Therefore, in this paper, based on deep network and slow feature analysis (SFA) theory, we proposed a new change detection algorithm for multi-temporal remote sensing images called Deep Slow Feature Analysis (DSFA). In DSFA model, two symmetric deep networks are utilized for projecting the input data of bi-temporal imagery. Then, the SFA module is deployed to suppress the unchanged components and highlight the changed components of the transformed features. The CVA pre-detection is employed to find unchanged pixels with high confidence as training samples. Finally, the change intensity is calculated with chi-square distance and the changes are determined by thresholding algorithms. The experiments are performed on two real-world data sets. The overall detection accuracies of our proposed method on two experiments are 97.64% and 94.32%, respectively. The visual comparison and quantitative evaluation have both shown that DSFA could outperform the other state-of-the-art algorithms, including other SFA-based algorithms.

Index Terms—Change detection, Deep network, Slow feature analysis, Remote sensing images.

I. INTRODUCTION

C

Transportation detection is defined as the process of identifying differences in the state of an object or phenomenon by observing it at different times [1]. With the rapid development of remote sensing technology, more remote sensing images of the earth surface are now available [2]–[4]. The multi-temporal remote sensing images covering the same area could help to detect land-cover and land-use changes, so that change detection could be better applied to diverse real-world applications, such as deforestation monitoring, damage assessment, vegetation phenology variation study, and disaster monitoring [5]–[10].

Generally, change detection algorithms could be divided into the following categories: 1) Image algebra methods mainly include image difference, image ratio, image regression, and change vector analysis [11], [12]. These methods directly calculate the difference between multi-temporal remote sensing images; 2) Image transformation algorithms extract the effective features of multi-temporal remote sensing images by transforming and combining their feature bands, and mainly include Principle Component Analysis (PCA) [13], Multivariate Alteration Detection (MAD) [14], [15], Gramm-Schmidt transformation (GS) [16] and Independent Component Analysis [17]; 3) Classification methods mainly include post-classification and compound classification, which are both based on classification to obtain land-use categories [18]–[21]; 4) Other advanced methods contains the algorithms based on wavelet, Markov random field, and local gradual descent, etc. [22]–[25]. Among all these kinds of change detection algorithms, image transformation methods have been widely studied and applied. The basic idea of image transformation is projecting the original multiband images into a new feature space to better separate changed and unchanged pixels. In this process, the most crucial work is to find an effective projecting algorithm to extract the determinative features.

Changed pixels in multi-temporal remote sensing images always have the feature differences with diverse change directions, while the features of unchanged pixels are supposed to be generally invariant [1]. However, owing to the atmospheric conditions, illumination and sensor calibration and so on, those unchanged pixels always have slight differences [26], [27]. Compared with changed pixels, changes of unchanged pixels usually have the consistent direction. By minimizing the feature variation of unchanged pixels, changed pixels could also be highlighted and separated. Inspired by this idea, slow feature analysis is proposed for detecting real changes and obtained satisfactory performance [28], [29].

SFA is a feature learning algorithm that extracts invariant and slowly varying features from input signals [30], [31]. And it has been successfully applied to solve diverse real-world problems, such as human action recognition, dynamic texture recognition and time series analysis, etc [32]–[35]. In change detection problems, changed and unchanged pixels correspond to quickly and slowly varying features in SFA, respectively.
Based on this theory, Wu, et al. [28] used SFA to suppress the spectral difference between slowly varying unchanged pixels, so that the changed pixels can be highlighted and well detected. By solving SFA problems, the proposed algorithms in [28] could get the projecting matrices to map original data, so that the unchanged components could be suppressed. All these algorithms have shown their good performance in some real-world remote sensing images. However, limited by the feature representative ability, linear SFA algorithms are sometimes not able to separate the changed and unchanged pixels [36]. The potential solutions include projecting original feature into a higher-dimensional complex feature space to improve the model’s complexity and feature representation ability.

Deep networks have been proved to have a powerful ability of representing non-linear functions, and thus can project original features into a more complex feature space [37], [38]. Due to the growing availability of both data and computing resources, deep neural networks have been resurging in these years. Numerous kinds of networks have been developed to complete different tasks, such as classification [39], detection [40], segmentation [41], and feature mapping [38], etc. Besides, in recent years, deep networks have also been applied to learn non-linear transformations of highly correlated datasets, and performed well [42].

Therefore, inspired by the idea of utilizing deep network learning non-linear transformations, we propose a new algorithm called Deep Slow Feature Analysis (DSFA) in this paper. In DSFA, two deep networks are used to extract and represent the features of remote sensing images obtained at different times, respectively. The transformed features by deep networks are then taken as the inputs of SFA to obtain the projecting matrix. The projecting matrix could extract the most invariant component of multi-temporal remote sensing images, so the changed pixels could be accentuated. We formulate the loss function for DSFA model to make sure that the transformed features can represent the original data better. The intention of DSFA is to extract the invariant components of input features, which means that utilizing unchanged pixels as the inputs will help accelerating the training process and improving the final performance. However, in fact, labeled data are usually rare in remote sensing problems. Therefore, in DSFA, we use CVA to make a pre-detection and find unchanged pixel pairs as the inputs for training process. When the deep network is converged, the transformed features will be calculated by passing original features through trained networks. Then the difference of transformed features in SFA space is calculated. Finally, the change intensity map is calculated with chi-square distance, and the binary change map is obtained with thresholding algorithms.

The rest of this paper is organized as follows. Section II introduces the SFA theory and the details of SFA in change detection. Section III presents the algorithm details of proposed DSFA. In Section IV, we implement our proposed method and perform experiments on two real-world datasets. And Section V draws the conclusion of this paper.

II. SLOW FEATUE ANALYSIS

In this section, we’ll introduce the mathematical theory of SFA, and how SFA is extended to solve change detection problems. Mathematically, SFA is formulated as follows:

Given a multi-dimensional temporal signal \( s(t) = [s_1(t), s_2(t), \cdots, s_n(t)] \), where \( n \) represents the dimension and \( t \in [t_0, t_1] \), the target of SFA is finding a set of transforming functions \( \{g_1(x), g_2(x), \cdots, g_M(x)\} \) to generate the output signal \( z(t) = [g_1(s), g_2(s), \cdots, g_M(s)] \) and ensuring that transformed signal is time invariant as possible. Mathematically, the objective function of SFA is

\[
\min_{g_j} : \langle (g_j(s))^2 \rangle_t, j \in [1, 2, \cdots, M],
\]

under the following constraints:

\[
\langle g_j(s) \rangle_t = 0, \quad (2)
\]

\[
\langle g_j(s)^2 \rangle_t = 1, \quad (3)
\]

\[
\forall i < j : \langle g_i(s)g_j(s) \rangle_t = 0, \quad (4)
\]

where \( \langle g_j(s) \rangle_t \) denotes the mean signal of \( g_j(s) \) over time \( t \) and \( \dot{g}_j(ss) \) is the first-order derivate of \( g_j(s) \). Therefore, the objective of SFA is minimizing the mean value of the first-order derivate of transformed signal. Among these constraints, Constraint (2) is to simplify the process of solving the optimization problem. Constraint (3) ensures that each output signal could contain certain information. And Constraint (4) is presented to eliminate the correlation between output signals and force each signal carries different type of information.

In the linear case, the transforming function could be expressed as a mapping matrix:

\[
g_j(s) = w_j^T s,
\]

where \( w_j^T \) denotes the transposition of \( w_j \). And the objective function and constraints could be reformulated as follows:

\[
\langle (w_j^T \dot{s})^2 \rangle_t = \langle w_j^T \dot{s}s^T \rangle_t w_j = w_j^T A w_j, \quad (6)
\]

\[
\langle (w_j^T s) \rangle_t = 0, \quad (7)
\]

\[
\langle (w_j^T s)(w_j^T \dot{s}) \rangle_t = \langle w_j^T s \dot{s}s^T \rangle_t w_j = w_j^T B w_j = 1, \quad (8)
\]

\[
\langle (w_i^T s)(w_j^T \dot{s}) \rangle_t = \langle w_i^T s \dot{s}s^T \rangle_t w_j = w_i^T B w_j = 0. \quad (9)
\]

In (6), \( A = \langle \dot{s}s^T \rangle_t \) is the expectation of the covariance matrix of the first-order derivate of input signals. (7) represents Constraint (2), and it can be implemented by pre-processing the input data. (8) and (9) denote Constrain (3) and (4), respectively. And \( B = \langle ss^T \rangle_t \) is the expectation of covariance matrix of original input signals.

In SFA theory, (9) can be integrated to (6) as follows:

\[
\langle (w_j^T s)^2 \rangle_t = w_j^T A w_j = \frac{w_j^T A w_j}{w_j^T B w_j} = \frac{\langle (w_j^T s)^2 \rangle_t}{\langle (w_j^T s) \rangle_t}. \quad (10)
\]
And this optimization problem can be solved by the generalized eigenvalue problem:

$$AW = BW\Lambda,$$

(11)

where $W$ and $\Lambda$ is the generalized eigenvector matrix and a diagonal matrix of eigenvalues, respectively. According to (10) and (11), the most invariant component of the output signal has the smallest eigenvalue.

In pixel-based change detection problems, the input signals are raw pixels of remote sensing images, which are discrete. In consequence, SFA need to be reconstructed to cope with discrete cases. As shown in Figure 1, the objective of SFA in change detection problems is suppressing unchanged pixels to highlight changed ones, so that they could be separated much easier. Mathematically, let $x_i, y_i \in \mathbb{R}^m$ denote corresponding pixels in bi-temporal remote sensing images, where $m$ is the number of bands, and $n$ is the pixel number. After normalizing the input data, the objective of SFA is reformulated as

$$\min_{w_j} : \frac{1}{n} \sum_{i=1}^{n} (w_j^T x_i - w_j^T y_i)^2,$$

(12)

where $n$ is the total number of pixels. And constraints are rewritten as

$$\frac{1}{2n} \left[ \sum_{i=1}^{n} w_j^T x_i + \sum_{i=1}^{n} w_j^T y_i \right] = 0,$$

(13)

$$\frac{1}{2n} \left[ \sum_{i=1}^{n} (w_j^T x_i)^2 + \sum_{i=1}^{n} (w_j^T y_i)^2 \right] = 1,$$

(14)

$$\frac{1}{2n} \left[ \sum_{i=1}^{n} (w_j^T x_i)(w_j^T x_i) + \sum_{i=1}^{n} (w_j^T y_i)(w_j^T y_i) \right] = 0.$$  

(15)

In the generalized eigenvalue problem of SFA, $A$ and $B$ in (11) are reformulated as follows:

$$A = \frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)(x_i - y_i)^T,$$

(16)

$$B = \frac{1}{2n} \left[ \sum_{i=1}^{n} x_i x_i^T + \sum_{i=1}^{n} y_i y_i^T \right].$$

(17)

When $A$ and $B$ are obtained, the eigenvector matrix $W$ will be solved. By normalizing $W$, the final mapping matrix is obtained.

$$\tilde{w}_j = \frac{w_j}{\sqrt{w_j^T B w_j}}.$$ 

(18)

Then the change detection result, the difference between transformed bi-temporal images, is calculated as $D_j = \tilde{w}_j^T x_j - \tilde{w}_j^T y_j$.

III. METHODOLOGY

As mentioned above, those existing SFA-based change detection algorithms are all linear. In order to improve the representing ability of features and final change detection performance, in this section, we propose Deep Slow Feature Analysis (DSFA). The main structure of DSFA is shown in Figure 2.

As we can see in Figure 2, the input of DSFA is pairwise pixels of multi-temporal imagery. Then DSFA could be roughly divided to two parts: Deep Network module and SFA constraint. In the Deep Network module, two symmetric networks, whose layers are all Full Connected Layer, are used to project original input data into a new complex high-dimensional feature space. In Figure 2, the red nodes denote the nodes of input layer, the blue nodes represent the nodes of hidden layers and the yellow nodes are used to represent output layers. Each hidden layer of the Deep Network module has the same number of nodes. After the original data is transformed, we use the SFA constraint to suppress the invariant components and highlight the changed components of transformed features. We formulate the loss function of DSFA so that the parameters of deep networks could be solved based on gradient-based optimization algorithms.
A. Formulation

Mathematically, DSFA is defined as follows: Assuming the original bi-temporal remote sensing images are $X, Y \in \mathbb{R}^{m \times n}$, where $m$ and $n$ respectively denote the number of feature bands and pixels. For clarity, let $h_i$ denotes the number of nodes of the $i$-th hidden layer of the networks, and $o$ is the number of nodes of the output layer. Given an instance $X$, the output of the first hidden layer could be formulated as

$$f_1^i(X) = s(w_1^i X + b_1^i),$$

where $w_1^i \in \mathbb{R}^{h_i \times m}$ and $b_1^i \in \mathbb{R}^{h_i}$ denote the weight matrix and the bias vector, respectively. And $s(\cdot)$ represents the activation function. The output of the subsequent layers is calculated in the same way. For a network with $l$ hidden layers, the output of the last hidden layer is $f_l^i(X) = s(w_l^i f_{l-1}^i(X) + b_l^i)$, where $w_l^i \in \mathbb{R}^{h_l \times h_{l-1}}$ and $b_l^i \in \mathbb{R}^{h_l}$. After that, $f_l^i(X)$ will be mapped by the output layer.

Finally, the final transformed feature of this network is

$$X_\phi = f(\theta_1, X) = s(w_0^1 X + b_0^1),$$

where $w_0^1 \in \mathbb{R}^{o \times h_1}$ and $b_0^1 \in \mathbb{R}^o$ are the weight matrix and bias vector, respectively. And $\theta_1$ is the set of all the parameters in the network, including $w_1^1, \ldots, w_l^1, w_0^1$ and $b_1^1, \ldots, b_l^1, b_0^1$. And for another instance $Y$, $Y_\phi$ has a symmetric expression and meaning.

$$Y_\phi = f(\theta_2, Y) = s(w_0^2 f_1^2(Y) + b_0^2).$$

When the original given data is mapped into a new high dimensional feature space by deep networks, let $\hat{X}_\phi = X_\phi - \frac{1}{n} 1_n X_\phi$ and $\hat{Y}_\phi = Y_\phi - \frac{1}{n} 1_n Y_\phi$ denote the centralized $X_\phi$ and $Y_\phi$, respectively, where $1 \in \mathbb{R}^{o \times o}$ is a matrix whose elements are all 1. Then the covariance matrix of transformed data will be calculated.

$$\Sigma_{XX} = \hat{X}_\phi \hat{X}_\phi^T + r \ast I,$$  \hspace{1cm} (22)

$$\Sigma_{YY} = \hat{Y}_\phi \hat{Y}_\phi^T + r \ast I,$$  \hspace{1cm} (23)

$$\Sigma_{XY} = (\hat{X}_\phi - \hat{Y}_\phi)(\hat{X}_\phi - \hat{Y}_\phi)^T + r \ast I,$$  \hspace{1cm} (24)

where $I$ denotes the identity matrix and $r$ is a regularization constant. Assume that $r > 0$, so that $\Sigma_{XX}, \Sigma_{YY}$ and $\Sigma_{XY}$ are positive definite. Therefore, in DSFA problem, the generalized eigenvalue problem to be solved is formulated as:

$$A_\phi W = B_\phi W \Lambda \Leftrightarrow B_\phi^{-1} A_\phi W = W \Lambda,$$  \hspace{1cm} (25)

where $A_\phi = \Sigma_{XX}$ and $B_\phi = \frac{1}{2} (\Sigma_{XX} + \Sigma_{YY})$. According to (22–24), the final form of this problem is

$$\frac{1}{2} (\Sigma_{XX} + \Sigma_{YY})^{-1} \Sigma_{XY} W = W \Lambda.$$  \hspace{1cm} (26)

Based on SFA theory, the most invariant component has the smallest eigenvalue. Thus, the objective of DSFA could be designed as minimizing the total square of all eigenvalues, so that the variance of unchanged pixels can be suppressed and changed pixels are much easier to be detected. The loss function of DSFA then could be formulated as follows:

$$L(\theta_1, \theta_2) = tr[(B_\phi^{-1} A_\phi)^2],$$  \hspace{1cm} (27)

where $tr(\cdot)$ denotes the trace of a matrix. Utilizing (27), the loss value of DSFA could be calculated and the parameters of networks $\theta_1$ and $\theta_2$ can be obtained with gradient-based optimization algorithm.

B. Optimization

To calculate the gradient of $L(\theta_1, \theta_2)$ with respect to all the $w_i^1$ and $b_i^1$, we could use the back-propagation algorithm, which requires computing the gradient of $L(\theta_1, \theta_2)$ with respect to $\hat{X}_\phi$ and $\hat{Y}_\phi$. 

![Fig. 2. A schematic of DSFA, consisting two deep networks.](image-url)
According to the reference [43], and using the fact that $A_\phi$ and $B_\phi$ are both symmetric, we could then have:

$$\nabla_A = \frac{\partial L(\theta_1, \theta_2)}{\partial A_\phi} = 2B^{-1}_\phi A_\phi B^{-1}_\phi,$$  \hspace{1cm} (28)

$$\nabla_B = \frac{\partial L(\theta_1, \theta_2)}{\partial B_\phi} = -2B^{-1}_\phi A_\phi B^{-1}_\phi A_\phi B^{-1}_\phi.$$ \hspace{1cm} (29)

Utilizing the derivation in [42], we could have the gradient of $A_\phi$ with respect to each element of $X_\phi$:

$$\frac{\partial A_{ab}^{ij}}{\partial X_{ij}^\phi} = \frac{1}{n}(\xi_{a(i)}X_{b(j)}^\phi + \xi_{b(i)}X_{a(j)}^\phi) - \frac{1}{n}(\xi_{a(i)}Y_{b(j)}^\phi + \xi_{b(i)}Y_{a(j)}^\phi),$$ \hspace{1cm} (30)

where $\xi_{(c)}$ represents the indicator function. If $e$ is true, then $\xi_{(c)} = 1$, otherwise $\xi_{(c)} = 0$. Similarly, the gradient of $B_\phi$ with respect to each element of $X_\phi$ is computed as follows:

$$\frac{\partial B_{ab}^{ij}}{\partial X_{ij}^\phi} = \frac{1}{2n}(\xi_{a(i)}X_{b(j)}^\phi + \xi_{b(i)}X_{a(j)}^\phi),$$ \hspace{1cm} (31)

Integrating (28)-(31), the gradient of $L(\theta_1, \theta_2)$ with respect to $X_{ij}^\phi$ is:

$$\frac{L(\theta_1, \theta_2)}{X_{ij}^\phi} = \sum_{ab}^{\nabla A_{ij}^{ab}}(\theta_1, \theta_2) + \sum_{ab}^{\nabla B_{ij}^{ab}}(\theta_1, \theta_2),$$ \hspace{1cm} (32)

where $\nabla A_{ij}^{ab}$ and $\nabla B_{ij}^{ab}$ are symmetric. Then the change intensity of bi-temporal images could be calculated. In order to eliminate the differences in the scale of each feature bands, in this paper, we use chi-square distance to measure the intensity of changes, which is calculated as

$$chi^2 = \sum_{i=1}^{m} \frac{(D_i)^2}{\sigma^2_i}. \hspace{1cm} (35)$$

In (35), $m$ is the number of feature bands, and $\sigma^2$ is variance of each bands. Thresholding algorithms, such as Ostu method and Kmeans method, are then employed to get the final binary change map. The whole detailed process of training and generating binary change map for DSFA is summarized in Algorithm 1.

### Algorithm 1 Process of training and generating binary change map for DSFA.

**Input:**
- Multi-temporal input images $I^1$ and $I^2$;

**Output:**
- The binary change map $D$;

1: Standardize $I^1$ and $I^2$ using $z$ – score method;
2: Employ CVA pre-detection to generate training samples $X$ and $Y$;
3: Initialize the network’s parameters $\{\theta_1, \theta_2\}$;
4: while $i < max\_epochs$ do
5: Calculate the projected features $X_\phi = f_1(X, \theta_1)$ and $Y_\phi = f_2(Y, \theta_1)$;
6: Calculate the loss value:
$$L(\theta_1, \theta_2) = tr[(B^{-1}_\phi A_\phi)^2];$$
7: Calculate the gradient: $\partial L(\theta_1, \theta_2)/\partial \theta_1$ and $\partial L(\theta_1, \theta_2)/\partial \theta_2$;
8: Updating the parameters using Gradient Descent algorithm;
9: $i++$;
10: end while
11: Calculate the mapped features $I^1_\phi$ and $I^2_\phi$ of $I^1$ and $I^2$;
12: Solve SFA problem to obtain projecting matrix $w_\phi$;
13: Calculate the difference map:
$$\Delta I = w^T_\phi I^1_\phi - w^T_\phi I^2_\phi$$;
14: Thresholding to get the binary change map $D$;
15: return $D$;

**IV. EXPERIMENT**

To evaluate the performance of DSFA, in this section, we implement DSFA on TensorFlow and perform experiments on two multi-temporal remote sensing image data sets. Data sets
used in our experiment include two Enhanced Thematic Mapper (ETM) data sets. The first one is Taizhou dataset, covering the city of Taizhou, China, acquired in 2000 and 2003. And the second is Nanjing dataset, which are respectively acquired in 2000 and 2002. Both data sets were obtained by the Landsat 7 Enhanced Thematic Mapper Plus (ETM+) sensor with a spatial resolution of 30 m. And 6 spectral bands (1-5 and 7) are selected for our experiments.

### A. Experiment settings

In the DSFA model, the weight and bias matrices of each layer are initialized randomly, and need to be optimized. The other values, including the number of layers and nodes in each view and the DSFA regularization parameter in (22-24) are hyperparameters. As for the DSFA regularization parameter, we tuned it over the range $[10^{-6}, 10^{-1}]$, and eventually found that $10^{-4}$ is quiet a proper value for our proposed model.

Some other conventional and SFA-based change detection algorithms are also implemented for comparison, including CVA [13], MAD [14], IRMAD [44], USFA [28], and ISFA [28]. All of them are unsupervised algorithms. Before calculating the difference map, PCA uses Principal Component Analysis method to project original data into a new lower dimensional feature space. MAD is a change detection method based on the established theory Canonical Correlation Analysis (CCA), which is firstly proposed in [45]. It utilizes CCA to maximize the correlation between the features of multi-temporal images. IRMAD is an iteratively weighted extension of MAD. It firstly calculates the original MAD variates. In the following iterations, it applies different weights to each pixels or regions to emphasize the changed parts of images. USFA and ISFA are proposed in [28]. Based on the SFA theory, USFA computes a projecting matrix to suppress the unaltered components of input data to highlight changed components. And ISFA is an iteratively weighted extension of USFA, and has the same way to calculate weights as IRMAD.

For all these algorithms, we choose all of the output feature bands to calculate the change intensity.

### B. Experiments on Taizhou ETM data set

The study area of the first data set is Taizhou city, Jiangsu Province, China. The image size is $400 \times 400$. Figure 3 shows the pseudo color and ground truth images of this data set. (a) and (b) are the pseudo color images acquired in 2002 and 2003, respectively. And (c) is the sampled ground truth image of changed and unchanged regions of Taizhou city, where the green pixels represent the unchanged regions, red pixels represent changed regions, and grey denotes undefined. The changed area contains 4227 pixels, and unchanged area contains 17163 pixels.

In the experiment of DSFA on Taizhou data set, 4000 pixels, which are about 2.5% of the total number of pixels, are randomly selected from the unchanged region of CVA pre-detected image for training to get the parameters of the networks and the projecting matrix of SFA. Due to the use of random initialization, for DSFA, we take the sum of change intensity of 10 independent runs as the final change intensity map, and the presented values of evaluation criteria are the results of the summed intensity map.

Figure 4 shows the change intensity maps of Taizhou dataset by (a) CVA, (b) PCA, (c) MAD, (d) IRMAD, (e) USFA, (f) ISFA, (g) DSFA-64-2, (h) DSFA-128-2, and (i) DSFA-256-2. The presented values of evaluation criteria are the results of the summed intensity map.

| TABLE I | CHANGE DETECTION RESULTS OF TAIZHOU DATASET USING OTSU. |
|---------|----------------------------------------------------------|
|         | OTSU | OA\_CHG | OA\_UN | OA | Kappa | AUC |
| CVA     | 0.8439 | 0.9970 | 0.9667 | 0.8890 | 0.9323 |
| PCA     | 0.7755 | 0.9961 | 0.9525 | 0.8374 | 0.8979 |
| MAD     | 0.8855 | 0.9474 | 0.9352 | 0.8030 | 0.9313 |
| IRMAD   | 0.9056 | 0.9818 | 0.9667 | 0.8942 | 0.9536 |
| USFA    | 0.7093 | 0.9922 | 0.9363 | 0.7773 | 0.8764 |
| ISFA    | 0.8077 | 0.9991 | 0.9612 | 0.8684 | 0.9266 |
| DSFA-64-2 | 0.8294 | 0.9982 | 0.9648 | 0.8819 | 0.9195 |
| DSFA-128-2 | 0.8985 | 0.9954 | **0.9763** | **0.9227** | **0.9519** |
| DSFA-256-2 | 0.8450 | 0.9966 | 0.9667 | 0.8888 | 0.9218 |

As the Table I shows, IRMAD and ISFA have the best performance in OA\_CHG and OA\_UN, respectively. However, on detecting unchanged pixels, IRMAD has the second worst performance. And ISFA also performs bad on detection changed regions. And it is worth noting that DSFA-128-2 outperforms the other algorithms in OA, which indicates that it has a higher accuracy in both changed and unchanged part of remote sensing images. And other DSFA-based methods also have very good performance in OA, especially compared with USFA and ISFA. Besides, in Kappa coefficient, all DSFA-based methods have better performances than USFA and ISFA. The Kappa coefficient of DSFA-128-2 is 0.9227, which is much better than the other change detection methods. IRMAD have an AUC value of 0.9536, which is slightly better
than DSFA-128-2. In general, among all SFA-based methods, DSFA could outperform USFA and ISFA in most aspects. Considering the total detection accuracy of all changed and unchanged pixels, Kappa coefficient, and AUC value, DSFA-128-2 is the best method, and IRMAD is slight better than ISFA and other DSFA-based methods.

The change detection results obtained by Kmeans method are presented in Table II. As we can see from this table, all of these methods don’t show obvious differences in performance when using different thresholding algorithm. And this suggests that these methods, including our proposed DSFA-based algorithms, are very stable to different thresholding methods. The results in Table II are very similar to those in Table I. IRMAD has the best performance in OA
detected unchanged regions, but has low accuracy in detecting changed pixels. For both changed and unchanged regions, DSFA-128-2 has a detection accuracy of 97.64%, which is the highest among all methods. DSFA-128-2 also has the highest Kappa coefficient and the second highest AUC value. Generally, all of DSFA-based algorithms have pretty good performance. And among all the methods, DSFA-128-2 is still the best one.

In Table III, we present the best change detection results of Taizhou dataset by the traversal of all thresholds. In this table, we could see that all DSFA-based methods could outperform the other algorithms exclude CVA and ISFA. And among all DSFA-based methods, DSFA-128-2 has best performances in all evaluation criteria. ISFA has almost the same performances with DSFA-128-2. Besides, it’s worth noting that the best change detection results of USFA and ISFA are much better than those obtained with OTSU and Kmeans method, while DSFA-based methods’ best results are very close to those using OTSU and Kmeans method. We can conclude that though the best results of ISFA are very close to DSFA, the latter has much better discriminability than the former.

In Figure 5, we present the binary change maps obtained by OTSU method of (a) CVA, (b) PCA, (c) MAD, (d) IRMAD, (e) USFA, (f) ISFA, (g) DSFA-64-2, (h) DSFA-128-2, and (i) DSFA-256-2. In this figure, green, red, white, and purple regions represent unchanged pixels that are detected as unchanged, changed pixels that are detected as changed, changed pixels that are detected as unchanged, and unchanged pixels that are detected as changed, respectively. And we could refer them as true negative, true positive, false negative, and false positive samples. As Figure 5 presents, intuitively,
DSFA-128-2 have the best performance. And compared with DSFA-128-2, the results of MAD-based methods have more false positive pixels than other algorithms. CVA, PCA and two SFA-based methods tend to classify changed pixels as unchanged. The other DSFA-based methods, DSFA-64-2 and DSFA-256-2, are prone to judge some specific changed regions as unchanged.

C. Experiments on Nanjing ETM data set

The second experiment is carried on the Nanjing ETM data set. Nanjing data set includes two 6 spectral bands remote sensing images with a size of 800×800, which are acquired...
in 2000 and 2002, respectively. Figure 6 presents the pseudo color images of Nanjing city obtained in (a) 2000, (b) 2002, and (c) is the ground truth of sampled changed and unchanged areas. The red part of (c) represents the sampled changed area of Nanjing city, which includes 2363 pixels. And the green part is the sampled unchanged area and includes 12393 pixels.

In the experiment on Nanjing data set, we randomly select 8000 pixels from unchanged area pre-detected by CV A to train our DSFA model. Like the experiment on Taizhou data set, the presented results of each evaluation criteria of DSFA are based on the total change intensity map of 10 runs.

Figure 7 shows the change intensity maps of Nanjing dataset by (a) CV A, (b) PCA, (c) MAD, (d) IRMAD, (e) USFA, (f) ISFA, (g) DSFA-64-2, (h) DSFA-128-2, and (i) DSFA-256-2. In this figure, brighter regions have bigger change probabilities. As we can see from this figure, USFA and ISFA have less bright area, which means that they tend to detect much less changed pixels that other change detection algorithms. And CV A, MAD and IRMAD have more bright area. This represents that they are prone to think these unchanged pixels also have a certain probability to change. DSFA-128-2 and DSFA-256-2 have very close results to each other. Both them have very good discriminability of changed and unchanged pixels. In addition, the result of PCA is also very close to
Fig. 6. The pseudo-color images of Nanjing city obtained in (a) 2000, (b) 2002, and (c) ground truth.

| TABLE IV | CHANGE DETECTION RESULTS OF NANJING DATASET USING OTSU. |
|----------|----------------------------------------------------------|
| OTSU     | OA_CHANGE | OA_UN | OA | Kappa | AUC  |
| CVA      | 0.8595    | 0.9168 | 0.9076 | 0.6933 | 0.8901 |
| PCA      | 0.8625    | 0.9363 | 0.9244 | 0.7398 | 0.8932 |
| MAD      | 0.9534    | 0.8530 | 0.8691 | 0.6236 | 0.8854 |
| IRMAD    | 0.9530    | 0.8922 | 0.9019 | 0.6987 | 0.9148 |
| USFA     | 0.5959    | 0.9680 | 0.9084 | 0.6234 | 0.7751 |
| ISFA     | 0.6416    | 0.9760 | 0.9224 | 0.6816 | 0.8133 |
| DSFA-64-2| 0.7288    | 0.9817 | 0.9412 | 0.7647 | 0.8574 |
| DSFA-128-2| 0.7465   | 0.9806 | 0.9431 | 0.7747 | 0.8692 |
| DSFA-256-2| 0.7360   | 0.9793 | 0.9403 | 0.7633 | 0.8674 |

DSFA-64-2. But the distinction between their changed and unchanged regions is not very obvious. On the whole, visually, the result of DSFA-128-2 is the best in calculating the change intensity.

In Table IV, we present the change detection results of Nanjing dataset utilizing OTSU method, the best values of each evaluation criteria are highlighted with bold in this table. As we can see, in general, DSFA-based methods, especially DSFA-128-2, have the best performance among all these methods. DSFA-128-2 could outperform other algorithms in OA_UN, OA, and Kappa coefficient. And in these criteria, all DSFA-based methods are much better than others. MAD and IRMAD have the best performance in OA_CHANGE and AUC value, respectively, which is consistent with their change intensity results. However, they have the worst performance in detecting unchanged area. Similar to MAD and IRMAD, CVA and PCA have very high values in OA_CHANGE and AUC, but are far worse than DSFA-based methods in OA_UN, OA and Kappa coefficient. On the contrary, USFA and ISFA do well in detecting unchanged pixels, but have the lowest accuracy in OA_CHANGE and AUC value.

Table V shows the evaluation results of the experiment on Nanjing data set using Kmeans method. Similar to the results of OTSU, compared with MAD-based and SFA-based methods, DSFA is still better in detecting unchanged and changed areas, respectively. On the whole, DSFA-based algorithms have higher overall accuracy and Kappa value than others. Though DSFA has lower AUC value than MAD-based method, CVA and PCA, owing to its good performance in OA and Kappa, DSFA-based methods still could be regarded as the best one among all the methods.

In Table V, we present the best thresholding result of each changed detection methods by traversing all values. We could see from this table that DSFA-based methods are still the best in OA and Kappa, and have the second-best value in AUC. Actually, the AUC values of DSFA-based method are merely lower than IRMAD to a small degree. IRMAD and ISFA have high values in AUC, but are much worse in OA and Kappa than DSFA. Besides, it’s also worth noting that the best results of DSFA-based methods are very close to the results obtained by OTSU and Kmeans, which could be an evidence of the good discriminability of DSFA’s results. On
the contrary, thresholding results and the best results of USFA and ISFA have a sensible difference. And the best results of CVA, PCA and MAD-based methods are also much better than their thresholding results in both OA and Kappa coefficient.

Figure 8 shows the binary change maps of (a) CVA, (b) PCA, (c) MAD, (d) IRMAD, (e) USFA, (f) ISFA, (g) DSFA-64-2, (h) DSFA-128-2, and (i) DSFA-256-2, which are segmented by OTSU method. According to this figure, we could see that the binary change result of DSFA with different net structure are almost the same. Obviously, compared with DSFA’s results, results of MAD and IRMAD have much more purple pixels, which represent the false positive samples. On the contrary, results of USFA and ISFA contain more false negative pixels, which are colored with white. The results of CVA and PCA are close to DSFA’s results, but still has less true negative and more false positive samples than the latter.

V. CONCLUSION
In this paper, we proposed a novel change detection algorithm called DSFA for multi-temporal remote sensing images.
In the DSFA model, two deep networks are used to project the bi-temporal original input data into a new feature space. Then, SFA is used to extract the most invariant components of unchanged pixels and suppress them in changed regions to highlight changed components. We formulated the SFA process and loss function of DSFA model, and present the derivation of computing gradient of loss. Our proposed algorithm is unsupervised, which means it doesn’t need prior labeled pixels for the training process.

We implemented our algorithm and performed experiments on two real world data sets. The visual and quantitative results have both shown that our method could outperform the other state-of-the-art methods, including other SFA-based algorithms. The overall accuracy and kappa coefficient using Kmeans thresholding method of the proposed algorithm on two data sets are 97.64% and 0.9232, 94.32% and 0.7748, respectively. The overall accuracy and kappa coefficient by traversing all thresholds are 97.83% and 0.9304, 94.50% and 0.7915, respectively. In addition, it’s worth noting that the best results and thresholding results of our method are almost the same, which suggests that the changed and unchanged pixels in the results of DSFA have a better discriminability than other methods.

Our proposed method focuses on differentiating the changed
and unchanged regions in bi-temporal remote sensing imagery. The future work is required to explore DSFA’s potential in detecting multi-class changes.

**APPENDIX A**

**DERIVATION OF GRADIENT OF LOSS**

Here we will present the detailed deduction process of computing the gradient of $\mathcal{L}(\theta_1, \theta_2)$ with respect to $X_{\phi}$. From the reference [43], we have the following equations:

$$\begin{align}
\frac{\partial r(ABA^T C)}{\partial A} &= CAB + C^T A B^T, \\
\frac{\partial (X^{-1})_{kl}}{\partial X_{ij}} &= -(X^{-1})_{kl}(X^{-1})_{ji}.
\end{align}$$

(36)

(37)

Based on (36) and the fact that $A_{\phi}$ and $B_{\phi}$ are both symmetric, we could obtain:

$$\nabla A = \frac{\partial \mathcal{L}(\theta_1, \theta_2)}{\partial A_{\phi}} = 2B_{\phi}^{-1} A_{\phi} B_{\phi}^{-1},$$

(38)

$$\frac{\partial \mathcal{L}(\theta_1, \theta_2)}{\partial B_{\phi}^{-1}} = 2A_{\phi} B_{\phi}^{-1} A_{\phi}$$

$$\Leftrightarrow \left( \frac{\partial \mathcal{L}(\theta_1, \theta_2)}{\partial B_{\phi}^{-1}} \right)_{kl} = 2(A_{\phi} B_{\phi}^{-1} A_{\phi})_{kl},$$

(39)

Then, combining (37), $\nabla_B = \partial \mathcal{L}(\theta_1, \theta_2)/\partial B_{\phi}$ is calculated as the following equation:

$$\nabla_B = -2 \sum_{kl} (A_{\phi} B_{\phi}^{-1} A_{\phi})_{kl}(B_{\phi}^{-1})_{kl}(B_{\phi}^{-1})_{ji}$$

$$= -2 \sum_{kl} (B_{\phi}^{-1})_{ik}(A_{\phi} B_{\phi}^{-1} A_{\phi})_{kl}(B_{\phi}^{-1})_{ji}$$

$$= -2(B_{\phi}^{-1} A_{\phi} B_{\phi}^{-1} A_{\phi} B_{\phi}^{-1}).$$

(40)

We could expand the expression of $A_{\phi}$ out:

$$A_{\phi} = \sum_{XY} Y = \frac{1}{n}(\hat{X}_\phi - \hat{Y}_\phi)(\hat{X}_\phi - \hat{Y}_\phi)^T$$

$$= \frac{1}{n}(\hat{X}_\phi \hat{X}_\phi^T + \hat{Y}_\phi \hat{Y}_\phi^T - \hat{X}_\phi \hat{Y}_\phi^T - \hat{X}_\phi \hat{Y}_\phi^T).$$

(41)

First, based on the derivation in the appendix of [42], we have:

$$\frac{\partial (\hat{X}_\phi \hat{X}_\phi^T)_{ab}}{\partial X_{ij}^\phi} = \begin{cases} \frac{2}{n}(\hat{X}^\phi_{ij} - \frac{1}{n} \sum_k \hat{X}^\phi_{ik}), a = i, b = i \\ \frac{1}{n}(\hat{X}^\phi_{bj} - \frac{1}{n} \sum_k \hat{X}^\phi_{bk}), a = i, b \neq i \\ \frac{1}{n}(\hat{X}^\phi_{ij} - \frac{1}{n} \sum_k \hat{X}^\phi_{ik}), a \neq i, b = i \\ 0, a \neq i, b \neq i \end{cases}$$

$$= \frac{1}{n}(\xi(a = i) \hat{X}^\phi_{bj} + \xi(b = i) \hat{X}^\phi_{aj}).$$

(42)

Also,

$$\frac{\partial (\hat{Y}_\phi \hat{Y}_\phi^T)_{ab}}{\partial X_{ij}^\phi} = \frac{1}{n}(\hat{Y}^\phi_{bj} - \frac{1}{n} \sum_k \hat{Y}^\phi_{bk}) = \frac{1}{n} \xi(a = i) \hat{Y}^\phi_{bj}.$$  

(43)

Integrating (42) and (43) into (41):

$$\frac{\partial A_{\phi}^b}{\partial X_{ij}^\phi} = \frac{\partial (\hat{X}_\phi \hat{X}_\phi^T)_{ab}}{\partial X_{ij}^\phi} \frac{\partial (\hat{Y}_\phi \hat{Y}_\phi^T)_{ab}}{\partial X_{ij}^\phi} = \frac{\partial (\hat{X}_\phi \hat{Y}_\phi^T)_{ab}}{\partial X_{ij}^\phi}$$

$$= \frac{1}{n}(\xi(a = i) \hat{X}^\phi_{bj} + \xi(b = i) \hat{X}^\phi_{aj})$$

$$- \frac{1}{n}(\xi(b = i) \hat{Y}^\phi_{aj} + \xi(a = i) \hat{Y}^\phi_{bj}).$$

(44)

Similarly, with respect to $B_{\phi}$, we have:

$$\frac{\partial B_{\phi}^b}{\partial X_{ij}^\phi} = \frac{\partial \sum_{XY} A_{\phi}^b}{\partial X_{ij}^\phi} + \frac{\partial \sum_{XY} B_{\phi}^b}{\partial X_{ij}^\phi}$$

$$= \frac{1}{n} (\sum_b \nabla_{A_{\phi}} \hat{X}^\phi_{bj} + \sum_a \nabla_{A_{\phi}} \hat{Y}^\phi_{aj})$$

$$+ \frac{1}{n} (\sum_b \nabla_{B_{\phi}} \hat{X}^\phi_{bj} + \sum_a \nabla_{B_{\phi}} \hat{Y}^\phi_{aj})$$

$$= \frac{1}{n} (\nabla_A \hat{X}_\phi + \nabla_A \hat{Y}_\phi - \nabla_A \hat{Y}_\phi)_{ij}$$

$$+ \frac{1}{n} (\nabla_B \hat{X}_\phi + \nabla_B \hat{Y}_\phi)_{ij}.$$  

(45)

Putting (44) and (45) together, the gradient of $\mathcal{L}(\theta_1, \theta_2)$ with respect to $X_{ij}$ is then computed as:

$$\frac{\partial \mathcal{L}(\theta_1, \theta_2)}{\partial X_{ij}^\phi} = \sum_{ab} \nabla_{A_{\phi}} \frac{\partial A_{\phi}^b}{\partial X_{ij}^\phi} + \sum_{ab} \nabla_{B_{\phi}} \frac{\partial B_{\phi}^b}{\partial X_{ij}^\phi}$$

$$= \frac{1}{n} (\sum_b \nabla_{A_{\phi}} \hat{X}^\phi_{bj} + \sum_a \nabla_{A_{\phi}} \hat{Y}^\phi_{aj})$$

$$+ \frac{1}{n} (\sum_b \nabla_{B_{\phi}} \hat{X}^\phi_{bj} + \sum_a \nabla_{B_{\phi}} \hat{Y}^\phi_{aj})$$

$$= \frac{1}{n} (\nabla_A \hat{X}_\phi + \nabla_A \hat{Y}_\phi - \nabla_A \hat{Y}_\phi)_{ij}$$

$$+ \frac{1}{n} (\nabla_B \hat{X}_\phi + \nabla_B \hat{Y}_\phi)_{ij}.$$  

(46)

Obviously, $\nabla_A$ and $\nabla_B$ are both symmetric matrices. Therefore,

$$\frac{\mathcal{L}(\theta_1, \theta_2)}{X_{ij}^\phi} = \frac{2}{n} (\nabla_A \hat{X}_\phi - \nabla_A \hat{Y}_\phi)_{ij} + \frac{1}{n} (\nabla_B \hat{X}_\phi)_{ij}.$$  

(47)

Finally, we could obtain the gradient of $\mathcal{L}(\theta_1, \theta_2)$ with respect to $X_{\phi}$:

$$\frac{\mathcal{L}(\theta_1, \theta_2)}{X_{\phi}} = \frac{2}{n} (\nabla_A \hat{X}_\phi - \nabla_A \hat{Y}_\phi) + \frac{1}{n} \nabla_B \hat{X}_\phi.$$  

(48)

**TABLE VI**

Best Change Detection Results of Nanjing Dataset.

|       | BEST | OA    | Kappa | AUC   |
|-------|------|-------|-------|-------|
| CVA   | 0.9248 | 0.7178 | 0.8952 |
| PCA   | 0.9341 | 0.7518 | 0.8955 |
| MAD   | 0.9227 | 0.7244 | 0.8921 |
| IRMAD | 0.9229 | 0.7340 | 0.9162 |
| USFA  | 0.9164 | 0.6997 | 0.8813 |
| DSFA-64-2 | 0.9450 | 0.7915 | 0.9086 |
| DSFA-128-2 | 0.9439 | 0.7850 | 0.9100 |
| DSFA-256-2 | 0.9409 | 0.7664 | 0.8951 |
[45] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, “Canonical correlation analysis: An overview with application to learning methods,” *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.