IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received June 11, 2020, accepted June 22, 2020, date of publication September 22, 2020, date of current version December 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3011258

Spatiotemporal Remote-Sensing Image Fusion With Patch-Group Compressed Sensing

LEI LI[1,2], PENG LIU[1], JIE WU[1,2], LIZHE WANG[3], (Member, IEEE), AND GUOJIN HE[1,2]

[1]Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
[2]College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China
[3]School of Computer Science, China University of Geosciences, Wuhan 430074, China

Corresponding author: Peng Liu (liupeng@radi.ac.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61731022, Grant 41971397, Grant 41701468, and Grant U1711266, and in part by the Institute of Remote Sensing and Digital Earth (RADI), Chinese Academy of Sciences, Director Youth Funding under Grant Y6XS5900CX and Grant Y5ZZ08101B.

ABSTRACT Generally, it is difficult to acquire remote sensing data whose resolution is both highly spatial and highly temporal from a single satellite. In this paper, a novel compressed sensing (CS)-based spatiotemporal data fusion (CSBS) method is proposed to synthesize such high-spatiotemporal resolution images. With CSBS, a low-spatial resolution remote senisng image is treated as a sampling of the high-spatial resolution image. The down-sampling in the spatial domain of images is modeled as a CS measurement matrix in CSBS. Moreover, continuity constraints in the temporal domain are also introduced into the CSBS object function for CS reconstruction. To better represent the intrinsic features of the data, images are segmented into many small patches and clustered into several groups via K-means. Dictionary training, measurement matrix identification, and high-resolution prediction are carried out group-by-group. Based on features learned from patch groups, the transformational relationship between spatial-temporal images having different resolutions are easily identified. Compared with previous compressed sensing and dictionary learning methods, CSBS is characterized by: (1) the patch-group stratagem in dictionary learning and measurement matrix learning; (2) the combination of continuity in temporal domain and sparsity in spatial domain. The proposed method is then comprehensively compared with different methods using land-surface reflectance data. Experiment results validate the effectiveness and advancement of CSBS for spatiotemporal data fusion.

INDEX TERMS Compressed sensing, landsat, land-surface reflectance, moderate resolution imaging spectroradiometer (MODIS), remote sensing, spatiotemporal data fusion.

I. INTRODUCTION

With the recent developments of earth observational remote-sensing technologies and the increasing amount of free satellite imagery available to the public, more and more volumes of the Earth's land-surface data can be easily acquired. Satellite-based remote-sensing images provide valuable geospatial information for characterizing land cover and land cover dynamics at both global and regional scales. Many high-spatiotemporal resolution remote-sensing imagery applications have thus been promulgated (e.g., land-cover discrimination [1], seasonal vegetation monitoring [2], [3], carbon sequestration modeling [4], crop yield estimation [5], human nature interactions monitoring [6] and ecosystem climate feedback reporting [7]). However, owing to technical and budget limitations, spatial and temporal resolution capabilities are mutually restricted. On one hand, satellites with high-spatial resolution sensors require long revisit cycles, which implies low temporal resolutions. On the other hand, satellites employing short-time revisit cycles often acquire only low- or middle-spatial resolution images. Therefore, it is difficult to acquire high-spatiotemporal resolution data from a single satellite.

Spatiotemporal data fusion is a feasible solution for the above-mentioned problem and has attracted much attention. This technique provides high-spatiotemporal resolution data by fusing frequent low-spatial resolution images with infrequent high-spatial resolution images. To perform data fusion, fusion methods must accurately identify relationships between images with different resolutions.

Over the past decades, a variety of spatiotemporal data-fusion methods have been developed. The existing

The associate editor coordinating the review of this manuscript and approving it for publication was Gulistan Raja.

spatiotemporal data fusion methods can be categorized into different groups: linear mixture, unmixing, bayesian, deep-learning, and sparse approximation methods etc. The **linear mixture-based** method assumes that the land-surface reflectance between Landsat Enhanced Thematic Mapper Plus (ETM+) and Moderate Resolution Imaging Spectroradiometer (MODIS) are basically consistent and that predicted pixel is the mixture of its neighborhood in temporal, spatial or spectral domain. Among the linear mixture based methods, the spatial and temporal adaptive reflectance fusion model (STARFM) [8] developed by Gao in 2006 is the first algorithm of its kind. STARFM can well handle the situation in which pixels in coarse-resolution images are "pure" pixels. However, the empirical weight function and the "pure" pixels assumption make STARFM short at heterogeneous landscapes. STARFM was later modified and applied to other products, such as land surface temperature (LST) [9] and vegetation indices [10]. STARFM was also be intergrated into an operational framework [11] which automatically perform BRDF correction, co-registration, selection of input data pairs. The enhanced STARFM (ESTARFM) [12] introduces a coefficient to improve STARFM's accuracy in heterogeneous landscapes. But ESTARFM can be worse than STARFM for predicting abrupt changes of land cover type [13]. The spatial-temporal adaptive algorithm for mapping reflectance change (STAARCH) [14] improve STARFM's performance by detecting change points from series of coarse images. But it is more suitable for spatial-temporal fusion of vegetated surface. To increase the performance in heterogeneous landscapes of STARFM, ATPPK-STARFM [15] downscales the 500m MODIS images to 250m before implementation of STARFM. To improve the weight function in STARFM, ISKRFM [16] use image inpainting and steering kernel regression to detect the land cover change and decide the weight function. Among these algorithms, only STAARCH has considered the landscape disturbances, whereas others assume that land-cover types remain unchanged. **Unmixing-based** methods assume that the pixels of low-spatial resolution images are linear mixed at the end members of high-spatial resolution images. Thus, it estimates the value of fine pixels by unmixing the coarse pixels. The multi-sensor multi-resolution technique (MMT) [17] is probably the first unmixing-based fusion method. Other unmixing-based methods [18]–[23] have been considered as improvements. STDFA [18] estimates reflectance difference in a moving window using an unmixing manner. A adaptive moving window size is applied by MSTDFA [19] to further improve STDFA. OB-STVIUM [20] proposed a new way to define the endmember fractions for improving the estimation. Zurita-Mill introduces extra constrains into the unmixing process to ensure that the solved reflectance values are positive and within an appropriate range [21]. For the same purpose, The Landsat-MERIS fusion method [22] predefines the endmember reflectance and modifies the cost function to ensure the correctness of the solved endmember reflectance. To account for the within-class NDVI spatial variability, locally calibrated multivariate regression models is introduced into LAC-GAS NDVI integration method [23]. More importantly, these algorithms face the same difficulty in delineating and characterizing the disturbances precisely when estimating the Landsat surface reflectance. Bayesian-based methods treat spatiotemporal data fusion as a maximum a posterior (MAP) problem. The key to bayesian-based data fusion methods is finding the relationships between the input and target predicted images. Bayesian-based methods predict the target images by maximizing its conditional probability relative to the input fine and coarse images [24]. Xue developed a bayesian data fusion approach predict the target image by interpolating the coarse image [25]. A unified fusion method [26] developed by Huang uses the low-pass filtering to model the relationship between coarse and fine images. Recently, **deep-learning** methods have been developing fast. As a typical convolutional neural network (CNN)-based fusion method, the spatiotemporal fusion using deep convolutional neural networks (STFDCNN) [27], comprises of two five-layer CNNs. Deep-learning methods can better learn the feature relationships of spatiotemporal data and have shown promising results. **Sparse approximation** methods retrieve the underlying relationships from the hidden sparse coding space of an image. Based on sparse approximation theory, an image can be sparsely decomposed into dictionary and coefficients matrix. The sparse representation-based spatiotemporal reflectance fusion model (SPSTFM) [28] jointly trains two dictionaries by enforcing coefficient similarities. Following SPSTFM, many improvements are introduced, such as using only one pair of fine- and coarse-resolution images [29], enhancing fusion by structural sparsity [30], enhancing fusion by adding a perturbation on the over-complete dictionary [31], and others. **Compressed sensing** for spatiotemporal fusion (CSSF) [32] was the first to explicitly exploit the relationships between high- and low-resolution images by building a down-sampling mapping process. Although CSSF provided new understandings for spatio-temporal fusion in spatial down-sampling, it should be noted that the original CSSF method also has limitations in exploring the data characteristics of temporal continuity. Over all, most of these mentioned methods have made some progresses in spatio-temporal fusion. However, it is still an open problem that it is hard to accurately establish the complex relation between high- and low-resolution images due to the insufficient prior knowledge. The application of these methods often suffers from two important limitations [33]. First, most spatial-temporal image fusion algorithms assume that land cover type does not change during the data observation period [8], [12], [14]. Second, most spatial-temporal image fusion methods require the coarse- and fine-resolution remotely sensed data from different satellite sensors to be mutually comparable and correlated. Meanwhile, these methods are also challenged by three important factors [34]: (1) Diversity of regions. Such as urban, rural, forest and mountain areas. (2) Long timespan. Usually, in some cloudy area, it cost a relatively long period of time to acquire clear images. (3) Challenging scenarios. Such as the spatial resolution gap between fine and coarse images, the characterization of changes in heterogeneous areas, and the prediction of land-cover changes. In conclusion, it is necessary that develop a more generic spatial-temporal fusion model to account for the significant type changes that challenge current methods, particularly in rapidly changing areas.

In this paper, to utilize more prior knowledge and introduce a constrain in temporal continuity, we propose a new model of compressed sensing-based spatiotemporal (CSBS) data fusion. In the proposed CSBS, both spatial down-sampling and temporal continuity are considered in a single object function. To improve the robustness and stability of CSBS, spatiotemporal images are segmented into many small patches and clustered into several groups via K-means. Our dictionary-learning, measurement matrix training, and CS reconstruction technique are all applied groupwise to deal with different land cover types. Therefore, the sparsity of the representation is enhanced, and the intrinsic mapping relationship of spatiotemporal data is better established. Compared with previous compressed sensing and dictionary learning methods, CSBS has two characteristics: (1) The patch-group stratagem in dictionary learning and measurement matrix estimation. (2) The combination of continuity in temporal domain and sparsity in spatial domain.

The rest of this paper is organized as follows. In Section II, we explain the compressed sensing theory and integrate it into spatial-temporal data fusion to formulate the proposed CSBS. After the patch group is defined, we implement training and prediction group by group. In Section III, we present experiment results based on three experiments. We also compare CSBS with four algorithms and analyze their performances. Our conclusions are presented in Section IV.

II. APPROACHES

In this section, we first briefly describe compressed sensing theory, then we apply it to spatiotemporal data fusion by considering a low resolution image as an observation sample of a high-resolution image. To sparsely represent the data and adaptively establish a sampling relationship in the feature space, a dictionary and a measurement matrix are trained using a patch-group model. To introduce a temporal constrain, a linear weighted strategy is applied. Finally, we provide the training and prediction algorithms.

A. COMPRESSED SENSING

Compressed sensing [35] efficiently acquires and reconstructs a signal from a series of sample measurements. For an original remote senisng signal $x \in R^M$, the observation $y \in R^N$ can be represented as

$$y = \Phi x + \epsilon \tag{1}$$

where Φ is the measurement matrix mapping from R^M to R^N, N is typically much smaller than M, and ϵ is the noise. Matrix Φ represents a dimensionality reduction. Restoring x from y is very challenging. For signal x, it is sparse in some domains and can be represented by the basis D:

$$x = D\alpha \tag{2}$$

where α are the coefficients of x in the basis D. D refers to the dictionary. Thus, observation y is expressed as

$$y = \Phi D\alpha + \epsilon \tag{3}$$

Finally, the objective function of a compressed sensing problem is

$$\min_{\alpha} \|\alpha\|_0, \quad \text{subject to } y = \Phi D\alpha. \tag{4}$$

To solve (4) means solve a ℓ_0-norm minimization problem which is unsolvable in polynomial time. But the solutions of ℓ_0-norm minimization can be converted to ℓ_1-norm problem when Φ and D satisfies restricted isometry property (RIP) [36]–[38]. The ℓ_1-norm problem is solvable in polynomial time. By relaxing the equality constraint, imposing ℓ_2-norm on the data-fitting term, and applying a Lagrangian form, (4) becomes

$$\min_{\alpha} \|\alpha\|_1 + \lambda \|y - \Phi D\alpha\|_2^2 \tag{5}$$

For the spatiotemporal data fusion problem, the low-spatial resolution image y is regarded as a measurement of the corresponding high-spatial resolution image x at the same time. The high-spatial resolution image can be reconstructed from the low-spatial resolution image when (5) is solvable. In the pure CS problem, RIP makes sure that (4) is equal to (5). But here in spatial-temporal data fusion, due to the situation that images are acquired by different sensors in different conditions, it is difficult to ensure RIP. Therefore, extra conditions should be introduced to make sure the validity of the solution to (5). In the following section, we discuss how to achieve spatial-temporal data fusion using compressed sensing theory.

B. SPATIOTEMPORAL FUSION WITH COMPRESSED SENSING

Based on compressed sensing theory, spatiotemporal data fusion can be regarded as reconstruct high-resolution image from the down-sampled low-resolution image. As in Fig. 1(a), our goal is to predict an image of the spatial resolution of Landsat ETM+ at t_2 with two pairs of Landsat ETM+ and MODIS images (acquired at t_1 and t_3) and one MODIS image (acquired at t_2). As shown in Fig. 1(a), the proposed CSBS method uses five images to predict one high-spatiotemporal image, which we call it a two-pairs model. Another method uses three images to predict one high-spatiotemporal image as in Fig. 1(b), which we call it a one-pair model. It is easy to prepare data for a one-pair fusion model. However, it is sometimes difficult for a one-pair model to improve its accuracy, because it utilizes less information than a two-pairs model. The object function, training, and prediction steps of the two-pairs model fusion are all more complex than that of

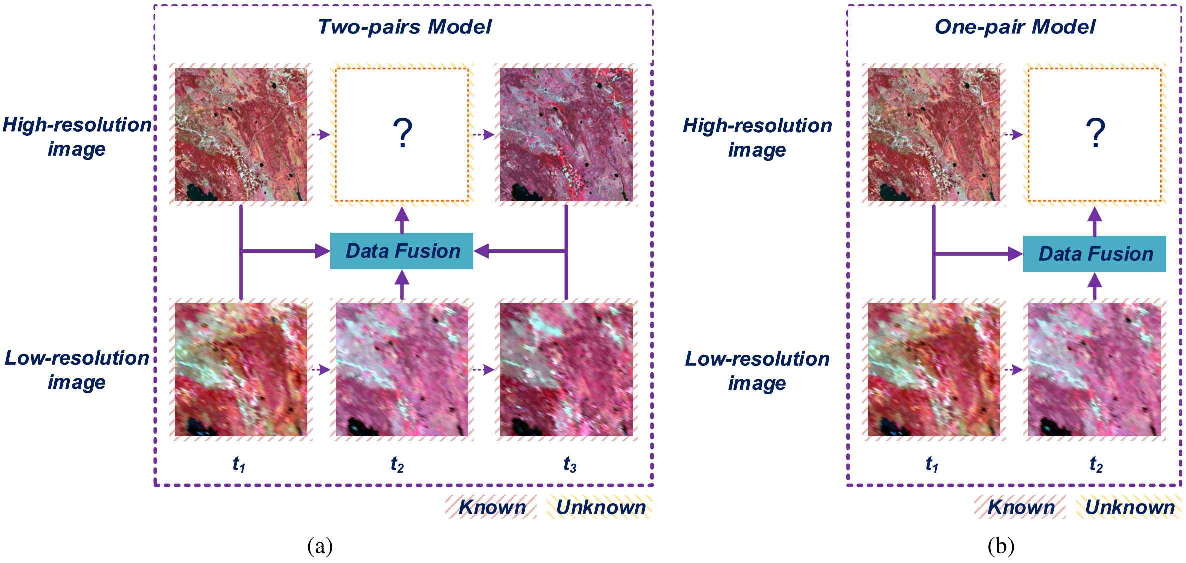


FIGURE 1. Two kinds of spatial-temporal data fusion model. (a) Two-pairs fusion model. (b) One-pair fusion model.

the one-pair fusion model. As a result, there are more possibilities for improvements for the two-pairs model, because more information is engaged in the fusion process. Thus, we use the two-pairs model in the proposed CSBS framework.

For the same scene, the high-spatial resolution image at time t is defined as y^{h_t}, and the low-resolution images at time t is defined as y^{l_t}. In our two-pairs model framework, the unknown high-spatial resolution image at time t_2 is defined as y^{h_2}. The unknown image y^{h_2} is predicted using $y^{h_1}, y^{l_1}, y^{h_3}, y^{l_3}$, and y^{l_2}. For spatiotemporal data fusion, as in Fig. 1(a), for an arbitrary t, images y^{h_t} and y^{l_t} show similar features with different resolutions, because they are captured at similar times and for same scene. Therefore, the observation model in spatial domain is defined as

$$y^{l_t} = \Phi^t y^{h_t} \tag{6}$$

where the low-resolution image y^{l_t} is defined as the observation down-sampled from y^{h_t}, and Φ^t represents the down-sampling operation. By considering y^{l_t} as an observation of y^{h_t}, we introduce compressed sensing into spatiotemporal data fusion.

According to compressed sensing theory, y^{h_t} should be sparse in some domains according to a dictionary D^t. Remote sensing image y^{h_t}, as a natural image, can usually be sparsely represented according to some orthonormal basis (e.g., wavelet, Fourier) or tight frames (e.g., curvelet, Gabor) [35], [39], [40]. After that, the image y^{h_t} is sparsely represented. To integrate compressed sensing into spatial-temporal data fusion, a suitable measurement matrix Φ^t should be derived. In this paper, to tackle the land cover type diversity, both the dictionary D^t and measurement matrixes Φ^t in (6) is trained from the different group of image patches, which is discussed further in Section II-C.

With dictionary D^{h_2} and measurement matrix Φ^{t_2} for y^{h_2}, we introduce compressed sensing theory into spatiotemporal data fusion. Like (5), the object function for the unknown image y^{h_2} at time t_2 is defined as

$$\begin{aligned} &\min_{\alpha^{h_2}} \left\|\alpha^{h_2}\right\|_1 + \left\|y^{l_2} - \Phi^{t_2} y^{h_2}\right\|_2^2, \\ &\text{subject to} \quad y^{h_2} = D^{h_2}\alpha^{h_2}. \end{aligned} \tag{7}$$

The spatial down-sampling of the measurement matrix Φ^{t_2} in (7) is insufficient for spatiotemporal data fusion. In many cases, the resolution of y^{h_2} is far higher than y^{l_2}. For example, when using Landsat ETM+ and MODIS data, one MODIS pixel represents 17 × 17 Landsat ETM+ pixels. The large resolution gap makes CS reconstruction unreliable, even when α^{h_2} is sparse. To ensure the validation of the solution to (7), both the spatial and temporal relationships between the two image sequences should be considered. In the temporal domain, for a period of time, the relationship between y^{h_2} and y^{h_1} or y^{h_3} can limit the reconstruction of y^{h_2} in a reasonable range.

Therefore, to describe the continuous sequence of images in the temporal domain, the difference between y^{h_2} and y^{h_1} or y^{h_3}, as a constraint, is introduced to the spatiotemporal data fusion. For different time t_1, t_2, t_3, y^h is acquires by the same sensor and the relation between y^h and y^l remains the same. We assume that $D^{h_1} \approx D^{h_2} \approx D^{h_3}$ and $\Phi^{h_1} \approx \Phi^{h_2} \approx \Phi^{h_3}$. The object function then becomes

$$\begin{aligned} \min_{\alpha^{h_2}} &\left\|\alpha^{h_2}\right\|_1 + \lambda_1 \left\|y^{l_2} - \Phi^{t_2} D^{h_2}\alpha^{h_2}\right\|_2^2 \\ &+ \lambda_2 \left\|y^{h_1} - D^{h_2}\alpha^{h_2}\right\|_2^2 \\ &+ \lambda_3 \left\|y^{h_3} - D^{h_2}\alpha^{h_2}\right\|_2^2 \end{aligned} \tag{8}$$

where $y^{h_2} = D^{h_2}\alpha^{h_2}$. There are still no theories to ensure the perfect reconstruction for (8). However, (8) will reconstruct more information than (7) since the more constraint makes the equation steadier. D^{h_2} is trained by the object function

$$\begin{aligned}\min_{\alpha_1^h,\alpha_3^h,D^{h_2}} \quad & \|\alpha^{h_1}\|_1 + \|\alpha^{h_3}\|_1 \\ & + \lambda_2\|y^{h_1} - D^{h_2}\alpha^{h_1}\|_2^2 \\ & + \lambda_3\|y^{h_3} - D^{h_2}\alpha^{h_3}\|_2^2 \end{aligned} \tag{9}$$

The ℓ_1-norm minimization and dictionary learning in CSBS are optimized using the SPArse Modeling Software (SPAMS) [41], which is an open-source toolbox for solving various sparse estimation problems. We generate D^{h_2} by training data from both y_1^h and y_3^h. Similar algorithms, such as EBSCDL [31] and CSSF [32], utilize the relationship of D^{h_t} and D^{l_t} to maintain the similarity of coefficients. Then, the dictionaries are coupled. Coupled dictionaries can avoid the searching of sampling matrix Φ. Others have introduced the transformation matrix between coefficients for D^{h_2} and D^{h_1} or D^{h_3}. These additional constraints have difficulty representing the complex relationship of sparse coefficients of multi-source data, because all coefficients are generated via a non-linear optimization (e.g., K-SVD or OMP [42]). For the same resolution, the difference of representation ability between D^{h_2} and D^{h_1} or D^{h_3} is not obvious. Thus, a transformation matrix for sparse coefficients derived from images having the same resolution is unnecessary. The most important is the fundamental feature relationship in spatial domain and temporal domain but not transformation domain.

By now, a CSBS data-fusion model has been built using (8) and (9). To fully apply CSBS, weighting parameters λ in (8) and (9) should be determined. Based on our prior experiment, we found that the intensity of the change is associated with the time difference. We assume that a short period of time leads to few landscape changes. Thus, the near observation earns more weight in the objective function in CSBS. A smaller $t_2 - t_1$ could mean fewer landscape changes between times t_1 and t_2. We propose that the weights of time parameters should be reflected in parameters λ_2 and λ_3 in (8) and (9). This strategy was also applied by STARFM [8]. Finally, the regularization parameters in (8) and (9) are determined by

$$\lambda_1 = \frac{1}{2}, \lambda_2 = \frac{t_3 - t_2}{t_3 - t_1}, \lambda_3 = \frac{t_2 - t_1}{t_3 - t_1} \tag{10}$$

Note that even λ_2 and λ_3 is decided in a linear manner, Equation. (8) and Equation. (9) are still partially linear as $D_i^{h_2}$ is trained in patch-group way. When a large image is segmented into many small patches, only the patches belong to the same cluster center will be taken as a locally linear changes. It doesn't mean a totally linearity for all features. Furthermore, $\left\|\alpha^{h_2}\right\|_1 + \lambda_1\left\|y^{l_2} - \Phi^{t_2}D^{h_2}\alpha^{h_2}\right\|_2^2$ will also provide non-linear information for fusion.

In this section, we provide the framework of the proposed CSBS model. By regarding both down-sampling in the spatial domain and the continuity of the temporal domain, we construct an object function of spatial-temporal data fusion by compressed sensing with temporal domain constraints. However, with real-world data, the implementation of spatial-temporal data fusion is still very difficult. To improve the robustness and stability of CSBS, we must optimize the process of training and predicting in spatial-temporal data fusion. This is discussed in next section.

C. TRAINING AND PREDICTING IN THE PATCH GROUP

In this paper, both the training of the dictionary D^{h_2} and the prediction of the high-spatial resolution image y^{h_2} are based upon the patch-group method. The high-spatial resolution image at time t_1 is defined as y^{h_1}. For an arbitrary kth pixel in y^{h_1}, its $\sqrt{n} \times \sqrt{n}$ neighborhood is defined as a patch, denoted as $\mathfrak{y}_k^{h_1}$. Based on the definition of the patch, image y^{h_1} is represented as patch set $\mathcal{Y}^{h_1} = \{\mathfrak{y}_k^{h_1}\}_{k=1}^{s}$, where $\mathcal{Y}^{h_1} \in R^{n\times s}$. Images y^{h_1}, y^{h_2}, y^{h_3}, y^{l_1}, y^{l_2}, and y^{l_3} have similar patch definitions as $\mathcal{Y}^{h_1}$. To balance feature diversity and sparsity, each patch dataset $\mathcal{Y}^{h_t}(t = 1, 2, 3)$, is clustered into I classes according to the Euclidean distance, as shown in Fig. 2. Here in CSBS, for simplicity, K-Means is used to classify data. Then $\mathcal{Y}^{h_t} = \{Y_i^{h_t}\}_{i=1}^{I}$ $(I \leq s)$, where the subclass or cluster $Y_i^{h_t} = \{\mathfrak{y}_j^{h_t}\}_{j=k_i}^{s_i}$. Thus, $\mathfrak{y}_{s_i}^{h_1}, \cdots, \mathfrak{y}_{k_i}^{h_1}$ are patches belonging to the ith cluster $Y_i^{h_t}$ in the high-spatial resolution dataset $\mathcal{Y}^{h_1}$ of time t_1. For the low-resolution image y^{l_t}, $Y_i^{l_t}$ is defined in a similar way as $Y_i^{h_t}$, where $t = 1, 2, 3$. For the low-spatial resolution image y^{l_t}, we have a similar patch set $\mathcal{Y}^{l_t} = \{Y_i^{l_t}\}_{i=1}^{I}$ $(I \leq s)$. However, as shown in Fig. 2, each cluster $Y_i^{l_t}$ is created by allocating each patch $\mathfrak{y}_k^{l_t}$ to its nearest cluster center in $\mathcal{Y}^{h_t}$. For each subclass $Y_i^{h_t}$ in $\mathcal{Y}^{h_t}$, we search for a corresponding dictionary $D_i^{h_t}$. The sampling matrix Φ_i is trained using the cluster pair $\langle Y_i^{h_t}, Y_i^{l_t}\rangle$. The localization dictionary more sparsely represents the data. The localization sampling matrix is adaptively sampled so that the condition of reconstruction for CS is more easily satisfied.

After K-means clustering, each class $Y_i^{h_t}$ is normalized by subtracting its mean value $M_i^{h_t}$, and dividing its standard deviation $S_i^{h_t}$. The normalized $\hat{Y}_i^{h_t}$ is represented as

$$\hat{Y}_i^{h_t} = \frac{Y_i^{h_t} - M_i^{h_t}}{S_i^{h_t}} \tag{11}$$

The low-spatial resolution image patch set $\hat{Y}_i^{l_t}$, has a similar definition as $\hat{Y}_i^{h_t}$. The arbitrary ith subclass from the K-means object function is

$$\begin{aligned}\min_{\alpha_i^{h_2}} \|\alpha_i^{h_2}\|_1 + \lambda_1\|\hat{Y}_i^{l_2} - \Phi_i D_i^{h_2}\alpha_i^{h_2}\|_2^2 & \\ + \lambda_2\|\hat{Y}_i^{h_1} - D_i^{h_2}\alpha_i^{h_2}\|_2^2 & \\ + \lambda_3\|\hat{Y}_i^{h_3} - D_i^{h_2}\alpha_i^{h_2}\|_2^2 & \end{aligned} \tag{12}$$

In (12), $D_i^{h_2}$ is trained by

$$\begin{aligned}\min_{\alpha_i^{h_1},\alpha_i^{h_3},D_i^{h_2}} \quad & \|\alpha_i^{h_1}\|_1 + \|\alpha_i^{h_3}\|_1 \\ & + \lambda_2\|\hat{Y}_i^{h_1} - D_i^{h_2}\alpha_i^{h_1}\|_2 \\ & + \lambda_3\|\hat{Y}_i^{h_3} - D_i^{h_2}\alpha_i^{h_3}\|_2 \end{aligned} \tag{13}$$

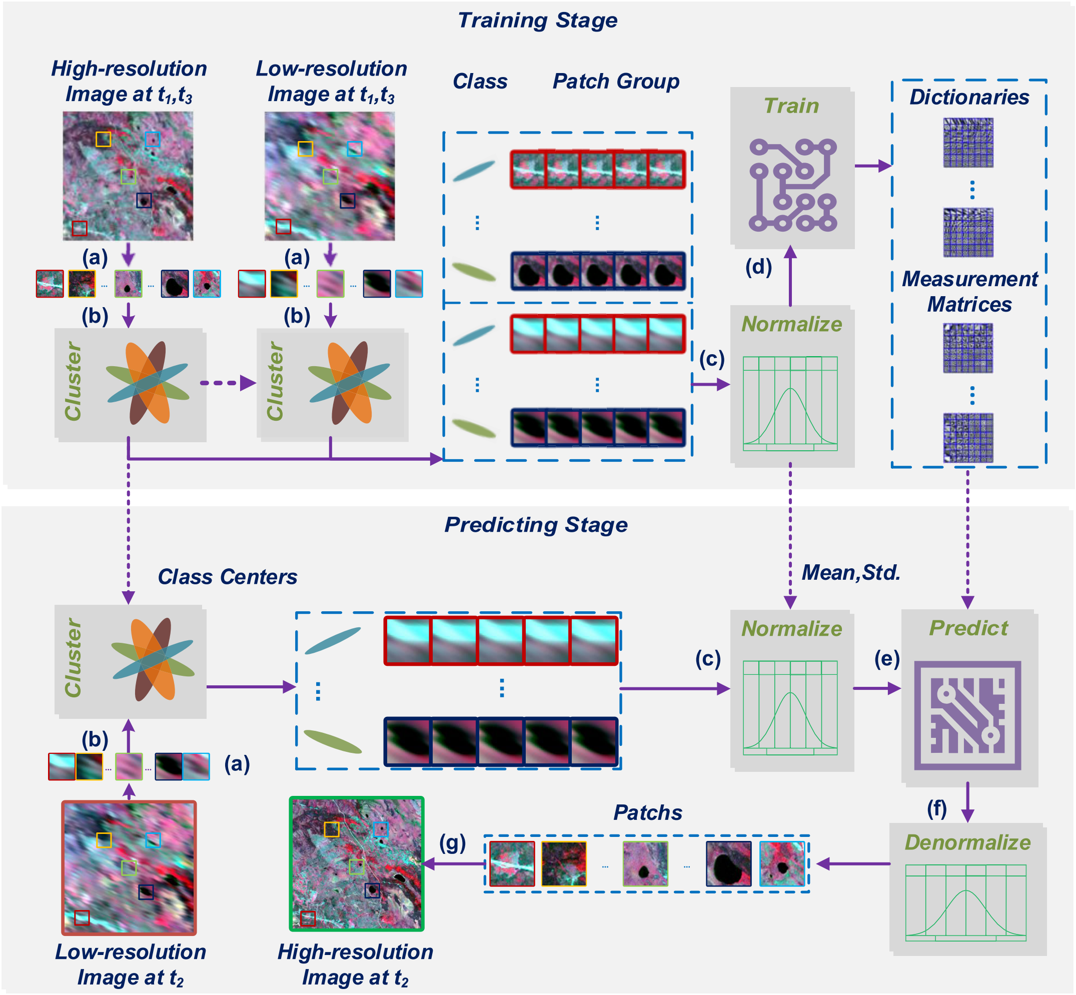


FIGURE 2. Proposed CSBS fusion consisting of two stages: training stage and predicting stage. In the training stage, the input high-resolution image at t_1 and t_3 along with low-resolution image at t_1 and t_3 are first sliced into patches (a), then the patches are grouped using K-means (b). The low-resolution image patche groups are created by allocating each patch to its nearest cluster center in the corresponding high-resolution image patch groups. After K-means clustering, each class is normalized (c) and is used to train dictionaries and measurement matrices (d). During the predicting stage, the input low-resolution image at t_2 is also sliced into patches (a). Then the patches are grouped using cluster centers generated from the training stage (b). After the group process, each class is normalized (c) using mean values and standard deviations derived from the training stage and is used to predict the desired image with dictionaries and measurement matrices from the training stage (e). After the denormalization (f), the result patches are finally used to build the desired high-resolution image (g). The dotted line in this figure demarks the data transfer. The solid line shows the fusion process.

As spatiotemporal fusion is a blind-inverse problem, we cannot obtain the down-sampling matrix Φ by directly solving $y_1^l = \Phi y_1^h$. Differences of acquisition time, bandwidth, data processing, and geolocations lead to small biases. To obtain the intrinsic relationship between remotely sensed data from different satellite sensors, we use principal component analysis (PCA) to calculate the measurement matrix. For the ith cluster, the object function for Φ_i is

$$\min_{\Phi_i} \lambda_2\|\Phi_i\hat{Y}_i^{h_1} - \hat{Y}_i^{l_1}\|_2 + \lambda_3\|\Phi_i\hat{Y}_i^{h_3} - \hat{Y}_i^{l_3}\|_2 \tag{14}$$

The measurement matrix need to be obtained for each class. Decomposing the covariance matrix $H_iH_i^T$ by using SVD, where $H_i = \lambda_2\hat{Y}_i^{h_1} + \lambda_3\hat{Y}_i^{h_3}$. The symmetric matrix can be expressed as

$$H_iH_i^T = W_i\Sigma W_i^T \tag{15}$$

where W_i is taken as a column vector matrix, and Σ is a diagonal matrix. Normally, to reduce the dimension of data from m to n, the first n principal components are selected as the transition matrix. However, with CSBS, we select n components of W_i to form $W_{(n_i)}$ to minimize

$$f(W_{(n_i)}^T) = \|W_{(n_i)}^T H_i - L_i\|_2 \tag{16}$$

where $L_i = \lambda_2\hat{Y}_i^{l_1} + \lambda_3\hat{Y}_i^{l_3}$, and $W_{(n_i)}^T \in R^{n\times m}$ is the desired measurement matrix Φ_i

$$\Phi_i = W_{(n_i)}^T \tag{17}$$

The whole Φ for all clusters is composed as $\Phi = \{\Phi_1, \cdots, \Phi_i, \cdots, \Phi_I\}$. The training steps are summarized in Algorithm 1.

Algorithm 1 Train of CSBS

```
Require: $y^{h_1}, y^{l_1}, y^{h_3}, y^{l_3}$.
    $\mathcal{Y}^{h_1}, \mathcal{Y}^{l_1}, \mathcal{Y}^{h_3}, \mathcal{Y}^{l_3} \leftarrow$ Slice $(y^{h_1}, y^{l_1}, y^{h_3}, y^{l_3})$
    $C \leftarrow$ Cluster $\mathcal{Y}^{h_1}$ use K-means
    Cluster $\mathcal{Y}^{l_1}, \mathcal{Y}^{h_3}, \mathcal{Y}^{l_3}$ use class centers $C$
    for all $Y_i^{h_1} \subset \mathcal{Y}^{h_1}, Y_i^{h_3} \subset \mathcal{Y}^{h_3}, Y_i^{l_1} \subset \mathcal{Y}^{l_1}, Y_i^{l_3} \subset \mathcal{Y}^{l_3}$ do
        $M_i, S_i, \hat{Y}_i^{h_1} \leftarrow$ Normalize $(Y_i^{h_1})$ by solving (11)
        $\hat{Y}_i^{l_1}, \hat{Y}_i^{h_3}, \hat{Y}_i^{l_3} \leftarrow$ Normalize $Y_i^{l_1}, Y_i^{h_3}, Y_i^{l_3}$ by solving (11) use $M_i, S_i$
        $D_i \leftarrow$ Solve (13) use $\hat{Y}_i^{h_1}, \hat{Y}_i^{h_3}$
        $W_i \leftarrow$ Solve (15) use $\hat{Y}_i^{h_1}, \hat{Y}_i^{h_3}$
        $\Phi_i \leftarrow$ Solve (16) use $\hat{Y}_i^{h_1}, \hat{Y}_i^{h_3}, \hat{Y}_i^{l_1}, \hat{Y}_i^{l_3}$
    end for
Ensure: $D, \Phi, M, S, C$
```

As mentioned above, (13) can be solved using ℓ_1-norm minimizaiton. Φ_i can be found using (16). After obtaining dictionary $D_i^{h_2}$ and sampling matrix Φ_i, for a given low-resolution image y^{l_2}, by solving (12), we can obtain all clusters $\hat{Y}_i^{h_2}$. For convenience, (12) can be solved as

$$\min_{\alpha_i^{h_2}} \lambda \left\|\alpha_i^{h_2}\right\|_1 + \| \begin{bmatrix} \sqrt{\lambda_1}\hat{Y}_i^{l_2} \\ \sqrt{\lambda_2}\hat{Y}_i^{h_1} \\ \sqrt{\lambda_3}\hat{Y}_i^{h_3} \end{bmatrix} - \begin{bmatrix} \sqrt{\lambda_1}\Phi_i D_i^{h_2} \\ \sqrt{\lambda_2} D_i^{h_2} \\ \sqrt{\lambda_3} D_i^{h_2} \end{bmatrix} \alpha_i^{h_2} \|_2^2 \tag{18}$$

After that, the final clusters are obtained, CSBS denormalize $\hat{Y}_i^{h_2}$ to $Y_i^{h_2}$ by

$$Y_i^{h_2} = \hat{Y}_i^{h_2} \times S_i^{h_2} + M_i^{h_2} \tag{19}$$

The final fusion image y^{h_2} is obtained by resetting all the patches in each $Y_i^{h_2}$ back to their original position in the image. The prediction steps are summarized in Algorithm 2

Algorithm 2 Predict of CSBS

```
Require: $y^{l_2}, y^{h_1}, y^{h_3}, \Phi, D, M, S, C$.
    $\mathcal{Y}^{l_2}, \mathcal{Y}^{h_1}, \mathcal{Y}^{h_3} \leftarrow$ Slice $(y^{l_2}, y^{h_1}, y^{h_3})$
    Cluster $\mathcal{Y}^{l_2}, \mathcal{Y}^{h_1}, \mathcal{Y}^{h_3}$ use class centers $C$
    for all $Y_i^{l_2} \subset \mathcal{Y}^{l_2}, Y_i^{h_1} \subset \mathcal{Y}^{h_1}, Y_i^{h_3} \subset \mathcal{Y}^{h_3}, \Phi_i \subset \Phi, D_i \subset D$ do
        $\hat{Y}_i^{l_2}, \hat{Y}_i^{h_1}, \hat{Y}_i^{h_3} \leftarrow$ Normalize $Y_i^{l_2}, Y_i^{h_1}, Y_i^{h_3}$ by solving (11) use $M_i, S_i$
        $\hat{Y}_i^{h_2} \leftarrow$ Solve (12) use $\hat{Y}_i^{l_2}, \hat{Y}_i^{h_1}, \hat{Y}_i^{h_3}, D_i, \Phi_i$
        $Y_i^{h_2} \leftarrow$ Denormalize $\hat{Y}_i^{h_2}$ by solving (19) use $M_i, S_i$
    end for
    $y^{h_2} \leftarrow$ Rearrange $\mathcal{Y}^{h_2}$
Ensure: $y^{h_2}$
```

In conclusion, the theoretical improvement of our proposed CSBS can be summarized as: (1) patch-group stratagem in dictionary learning and measurement matrix estimation, (2) The combination of continuity of temporal feature and sparsity in spatial features. Other dictionary learning method such as EBSCDL [31], its dictionary Learning is not based on patch-group. As in Fig. 2, the patches representing similar land-cover type can be clustered into one patch-group, so that the dictionaries learned from patch-group are more powerful in modeling complex land surfaces. Other compressed sensing method such as CSSF [32], its measurement matrix is global for the whole image so that it is hard to manipulate the relationship between dictionary and measurement matrix. However, our measurement matrix is a localization one that only need to search within current patch-group. Furthermore, the temporal continuity in CSSF [32] is reflected in the coefficients of sparse representation of temporal image. Their temporal continuity cannot be well explored since the coefficient relationship is far from the real temporal feature correspondence. Overall, in theory, the proposed CSBS is more concise and reasonable than EBSCDL [31] and CSSF [32].

TABLE 1. Image acquisition time of three datasets. In the simulation experiment, we predict the target Landsat image at t_2 using Landsat images and MODIS images at t_1, t_3 and MODIS image acquired at t_2.

Dataset	t_1	t_2	t_3
DATA-I	2001-08-03	**2001-08-12**	2001-10-06
DATA-II	2018-01-04	**2018-02-05**	2018-03-25
DATA-III	2018-10-01	**2018-12-04**	2019-01-21

III. EXPERIMENTS AND RESULTS

In this section, we apply the CSBS algorithm to the Landsat ETM+/OLI and MODIS imagery. We first introduce details about experimental datasets. Then, we briefly explain the experimental schemes and some parameter settings. Finally, the performance and computational cost of proposed CSBS are evaluated on three datasets: DATA-I, DATA-II, and DATA-III.

A. DATA COLLECTION

For this study, we use three datasets to test CSBS: DATA-I, DATA-II, and DATA-III. DATA-I is a subset of the dataset used in [8], and this dataset is widely used in [12], [31], [32]. To fully test the algorithms' performance in different situations, dataset from [43] are used to compose DATA-II and DATA-III. Due to their different sources, there is a little difference between DATA-I and DATA-II, DATA-III. Therefore, the interpolation methods of MODIS images are different in Fig.3, 4, 5. All the dataset are composed of Landsat ETM+/OLI and MODIS images, which with similar orbital parameters, solar geometries and corresponding bandwidths. The bands usage are nir, red, and green. DATA-I is composed of Landsat 7 ETM+ data and MODIS data, which is acquired for the Boreal Ecosystem-Atmosphere Study (BOREAS) southern study area (54 ° N, 104 ° W). The growing season is short and phenology changes are extreme. Trees such as spruce, pine, aspen, and birch

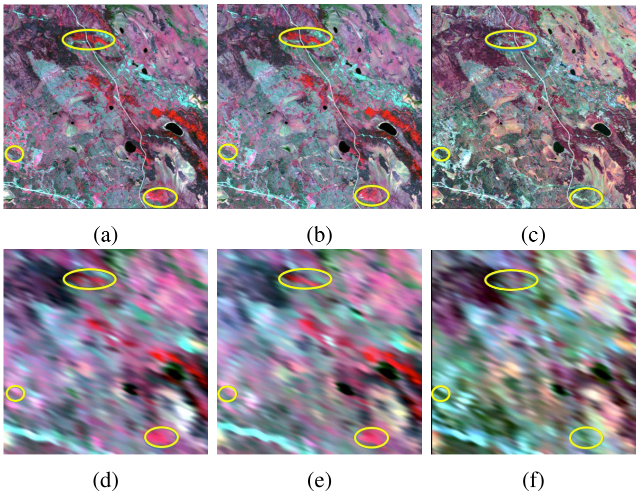


FIGURE 3. Landsat surface reflectance (upper row (a)-(c)) and MODIS composited surface reflectance (lower row (d)-(f)) images in DATA-I on 2001-08-03, 2001-08-12, 2001-10-06, accordingly. The areas with obvious changes are marked by the yellow circle. Note that there are notable land-cover changes between 2001-08-12 and 2001-10-06 images. The main purpose of DATA-I is to test the accuracy of spatial-temporal date fusion methods in the task of detecting land-cover type changes in forest area.

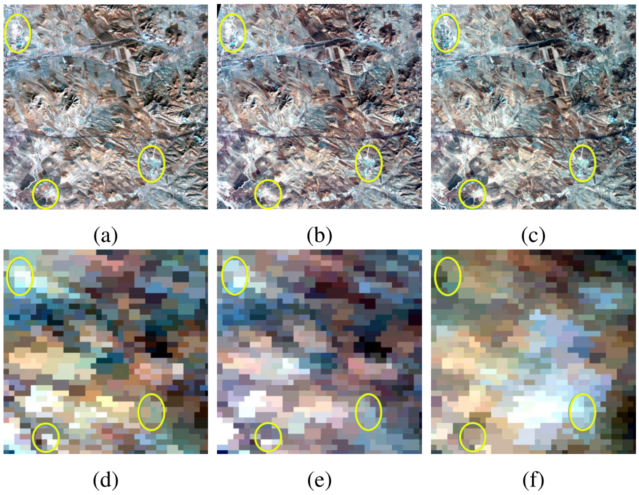


FIGURE 4. Landsat surface reflectance (upper row (a)-(c)) and MODIS composited surface reflectance (lower row (d)-(f)) images in DATA-II on 2018-01-04, 2018-02-05, 2018-03-25, separately. The areas with obvious phenological changes are marked by the yellow circle. The main purpose of DATA-II is to test the performance of data fusion methods in the application of detecting phenology changes in rural area.

dominate the landscape. DATA-II and DATA-III are composed by Landsat 8 OLI and MODIS data [43]. DATA-II is collected over Ar Horqin Banner of Inner Mongolia province, and DATA-III is collected over Tianjin city. The Ar Horqin Banner (43.3619°N, 119.0375°E) is located in the northeast of China. The major industries of Ar Horqin Banner are agriculture and animal boundary. Owing to the growth of crops and other kinds of vegetation, this area experienced significant phenological changes. DATA-II is mainly used to test the performance of data fusion methods in the application of detecting phenology changes in rural area. Tianjin city (39.8625° N, 117.8591° E) is a municipality in the north of China with clear seasonal changes during the year. Therefore, DATA-III is mainly used to evaluate the data fusion methods in the application of detecting phenology changes in urban area. In DATA-I, the 30-m land-surface reflectance data of Landsat-7 ETM+ C1 Level-2 are taken as high-spatial but low-temporal resolution images (y^{h_t}), and the standard 500-m daily surface reflectance data (MOD09GA) of MODIS are taken as low-spatial but high-temporal resolution images (y^{l_t}). During preprocessing for fusion, the MODIS daily surface reflectance data are projected and resampled to the Landsat ETM+ resolution by using MODIS Reprojection Tools (MRT) [44]. DATA-II and DATA-III are provided by [43]. The near-infrared (NIR), red and green bands are used to constitute fake-color images for visual comparisons. The image acquisition time of the three datasets are listed in Table. 1. After identical linear stretches, DATA-I, DATA-II and DATA-III are shown in Figs.3, 4, and 5, respectively using a NIR-red-green as red-green-blue composite. Note that the Landsat ETM+/OLI and MODIS land-surface reflectance images are very similar for near-day observations. The major land-cover type changes are from season variations. DATA-I, DATA-II, and DATA-III are collected to test algorithms' performance in different applications. Images in DATA-I were acquired during summer and autumn, and a sharp cooling faded the leaves and bare soil revealed between Figs. 3(b) and 3(c). Thus, DATA-I is mainly used to test the algorithms in the task of detecting land-cover changes. DATA-II is collected in Inner Mongolia province. Owing to the growth of crops and other kinds of vegetation, this area experienced significant phenological changes. Therefore, DATA-II aims to test the algorithms in the task of detecting phenology

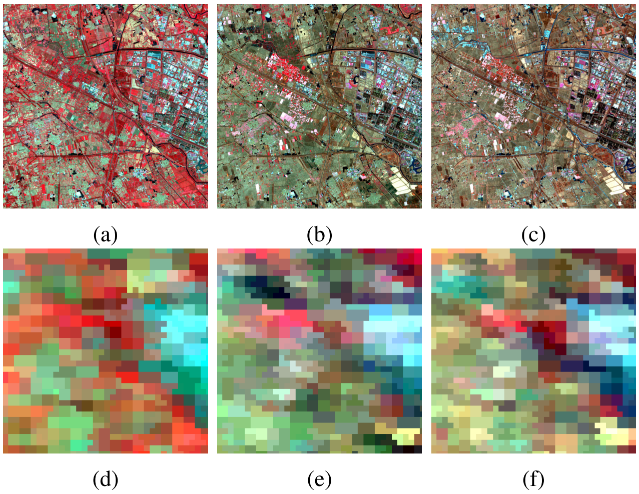


FIGURE 5. Landsat surface reflectance (upper row (a)-(c)) and MODIS composited surface reflectance (lower row (d)-(f)) images in DATA-III on 2018-10-01, 2018-12-04, 2019-01-21, separately. Phenology changes happened during 2018-10-01 and 2019-01-21. This dataset is mainly used to evaluate the data fusion methods in the application of detecting phenology changes in urban area.

changes in rural area. Meanwhile, DATA-III is collected in Tianjin city, which aims to evaluate the data fusion methods in the application of detecting phenology changes in urban area. In the next section, we will illustrate the experiment details.

B. EXPERIMENTAL SCHEMES

To validate CSBS's performance, as well as considering the availability of the source code, we compared it with four competitive algorithms: STARFM [8], ESTARFM [12], EBSCDL [31], and CSSF [32]. They each take two pairs of Landsat - MODIS images and one MODIS image as input. All algorithms are two-pairs models as shown in Fig. 1(a). Other algorithms that use one-pair model that shown in Fig. 1(b) (e.g., flexible spatial-temporal data fusion [45]) are not compared, because they use different numbers of input images or belong to different prediction schemes.

We designed three experiments using DATA-I, DATA-II, and DATA-III to test algorithmic performances in different applications, such as land-cover type changes, phenology changes in rural area and urban area. The experiment using DATA-I is to test the accuracy of spatial-temporal date fusion methods in the task of detecting land-cover type changes in forest area. DATA-II is mainly used to test the accuracy of spatial-temporal data fusion methods in the task of detecting phenology changes in urban area. DATA-III is mainly used to evaluate the data fusion methods in the application of detecting phenology changes in urban area. Regarding the features of different images from different datasets, according our prior statistics, we set the cluster number empirically as 10. The cluster number is an important parameter, a large cluster number may cause low fusion precision while a small cluster number may cause the lost of land cover type diversity. The patch size is all set to 7 × 7.

Three indicators are used to compare the performance in quantity: root mean-square error (RMSE), correlation coefficient (CC) and ERGAS. The RMSE of the fused images are used to measure performances by comparing the predicted images with the real images. RMSE is always nonnegative. An RMSE value of 0 indicates a perfect fit to the data. Generally, a lower RMSE is better than a higher one. RMSE is also related to the pixel scale. The land-surface reflectance should be within a range of 0 to 1. Thus, we set the scale factor equal to 10,000 as in STARFM [8]. The CC is a measure of the linear correlation between two variables. It has a value between $+1$ and -1, where $+1$ is total positive linear correlation, 0 is no linear correlation, and -1 is a total negative linear correlation. The ERGAS measure [46] is an error index that offers a global picture of the quality of a fused image. According to the authors, an ERGAS value lower than 3 denotes a satisfactory quality, while an ERGAS value greater than 3 means a poor quality. To visually validate the performance, we also present the zoomed regions. Experiments are explained in Section III-C, III-D, and III-E separately. Besides, to evaluate the computational cost of CSBS, we also present the running time cost comparison in section III-F.

C. APPLICATION ON LAND-COVER TYPE CHANGES

As is listed in Table. 1. In this experiment, Landsat ETM+ and MODIS image pairs acquired from 2001-08-03 and 2001-10-06 along with MODIS image acquired from 2001-08-12 are used as inputs to predict a high-spatiotemporal resolution image from 2001-08-12. As is shown in Fig. 3, there are notable land-cover type changes between 2001-08-12 and 2001-10-06 images. In this experiment, all algorithms are tested on DATA-I to validate their performance in the task of detecting land-cover type changes in forest area.

The RMSE results are listed in Table. 2 and the CC results are listed in Table. 3. According to the results, we observe that the proposed CSBS has the smallest RMSE and the highest CC of all bands. In Fig. 6, we know that the CSBS line features are closer to the ground truth in Fig. 6(f) than other methods. The EBSCDL method in Fig. 6(c) and the CSSF method in 6(d) also show good line features in the zoom area, but they show obvious spectral distortions. Spectral distorts in CSBS are very slight and are consistent with good CC performance, as shown in Table. 3.The STARFM method of Fig. 6(a) lost many details. This can be caused by STARFM's strategy that use the weighted sum of the surrounding pixels' reflectance to calculate the central pixel's reflectance. The ESTARFM method in Fig. 6(b) shows some fake texture features in the zoomed region, although its RMSE is low. ERGAS values in Table. 4 again shows that CSBS fusion can better reconstruct high-spatial resolution images. The experimental results reveal that CSBS is more robust than STARFM, ESTARFM, EBSCDL, and CSSF in dealing with land-cover type changes.

D. APPLICATION ON PHENOLOGY CHANGES IN RURAL AREA

This experiment is tested on DATA-II. As is listed in Table 1, in this experiment, the Landsat OLI and MODIS image pairs acquired from 2018-01-04 and 2018-03-25 along with MODIS image acquired from 2018-02-05 are used as inputs to predict a high-spatiotemporal resolution from 2018-02-05.

TABLE 2. RMSE of different algorithms using DATA-I.

Band	Algorithms				
	STARFM	ESTARFM	EBSCDL	CSSF	CSBS
NIR	0.01444	0.00967	0.01361	0.01299	**0.00909**
red	0.00959	0.00543	0.00584	0.00566	**0.00370**
green	0.00424	0.00330	0.00461	0.00443	**0.00328**

TABLE 3. CC of different algorithms using DATA-I.

Band	Algorithms				
	STARFM	ESTARFM	EBSCDL	CSSF	CSBS
NIR	0.93913	0.97378	0.95295	0.96414	**0.98590**
red	0.68149	0.88447	0.87061	0.87927	**0.93631**
green	0.86293	0.91500	0.88508	0.89718	**0.92769**

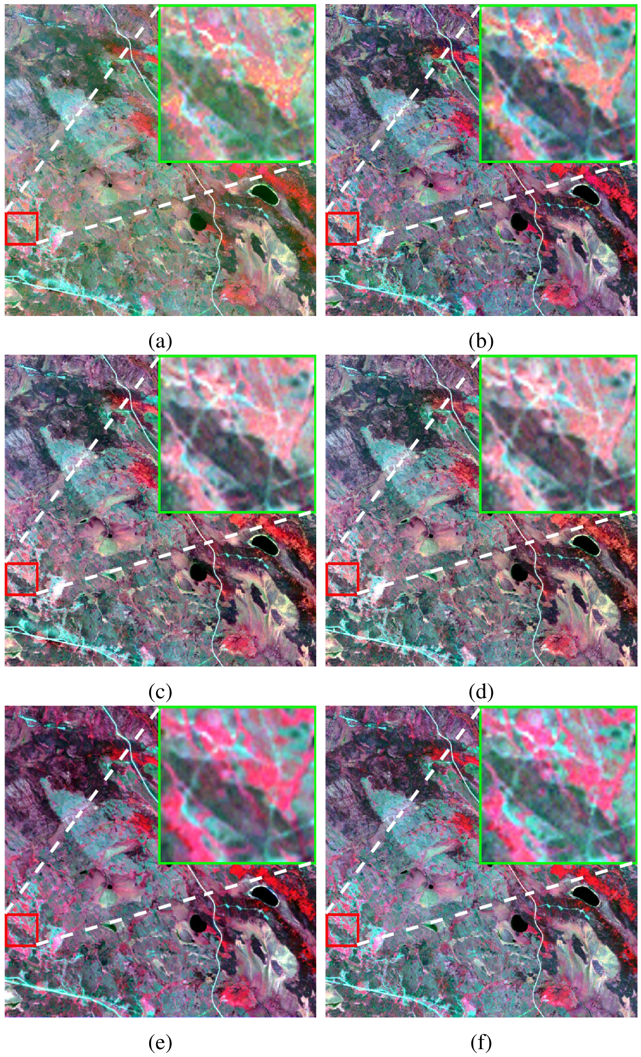

(a) (b) (c) (d) (e) (f)

FIGURE 6. Result images by different algorithms using DATA-I. Sub-figures (a)-(e) are the reconstructed image of STARFM, ESTARFM, EBSCDL, CSSF, and CSBS accordingly. Sub-figure (f) is the actual Landsat ETM+ land surface reflectance image at 2001-08-12. In the zoomed area, leaves faded and bare soil revealed from 2001-08-03 to 2001-10-06. CSBS more accurately predicted the changes compared with other algorithms.

TABLE 4. ERGAS of different algorithms using DATA-I.

	Algorithms				
	STARFM	ESTARFM	EBSCDL	CSSF	CSBS
ERGAS	1.03486	0.61167	0.69797	0.67448	**0.51039**

DATA-II is collected over Ar Horqin Banner of Inner Mongolia province. Owing to the growth of crops and other kinds of vegetation, this area experienced significant phenological changes. Therefore, this experiment is mainly used to test the performance of data fusion methods in the application of detecting phenology changes in rural area.

The RMSE results are listed in Table. 5 and the CC results are listed in Table. 6. CSBS has the smallest errors and largest CC values for all three bands. In Table. 7, CSBS has the best ERGAS value. The result images for all methods are shown in Fig.7. The result of STARFM in 7(a) looks like a mosaic. This may caused by that STARFM reconstruct images block by block. The results of ESTARFM, EBSCDL and CSSF show some spectral distortions compared to the ground truth to some extent. Results validate that the combination of sparsity in spatial domain and constrains in temporal domain successfully reconstruct the target image.

TABLE 5. RMSE of different algorithms using DATA-II.

Band	Algorithms				
	STARFM	ESTARFM	EBSCDL	CSSF	CSBS
NIR	0.42969	0.03309	0.01023	0.00891	**0.00698**
red	0.43191	0.02987	0.01994	0.01952	**0.01384**
green	0.43391	0.04235	0.02916	0.02895	**0.02178**

TABLE 6. CC of different algorithms using DATA-II.

Band	Algorithms				
	STARFM	ESTARFM	EBSCDL	CSSF	CSBS
NIR	0.10867	0.63954	0.89264	0.91344	**0.95699**
red	0.09317	0.60147	0.85282	0.87199	**0.95251**
green	0.08257	0.55003	0.77223	0.79682	**0.93554**

TABLE 7. ERGAS of different algorithms using DATA-II.

	Algorithms				
	STARFM	ESTARFM	EBSCDL	CSSF	CSBS
ERGAS	2.62495	3.31373	2.82121	2.77480	**1.78198**

E. APPLICATION ON PHENOLOGY CHANGES IN URBAN AREA

This experiment is tested on DATA-III. As is listed in Table 1, in this experiment, the Landsat OLI and MODIS image pairs acquired from 2018-10-01 and 2019-01-21 along with MODIS image acquired from 2018-12-04 are used as inputs to predict a high-spatiotemporal resolution from 2018-12-04. DATA-III is collected over Tianjin city, which is a municipality in the north of China with clear seasonal changes during the year. Therefore, this experiment is mainly used to test the performance of data fusion methods in the application of detecting phenology changes in urban area.

The RMSE results are listed in Table. 8, the CC results are listed in Table. 9, and the ERGAS values are listed in Table. 10. From the result, CSBS has the best indicator values for all three bands. The result images for all methods are shown in Fig.8. The result of STARFM in 8(a) shows spectral distortion to some extent. In the zoomed region, 8(b), 8(c), and 8(d) all show some fake textures in the high reflectance region. In this experiment, the results again validate the performance of CSBS.

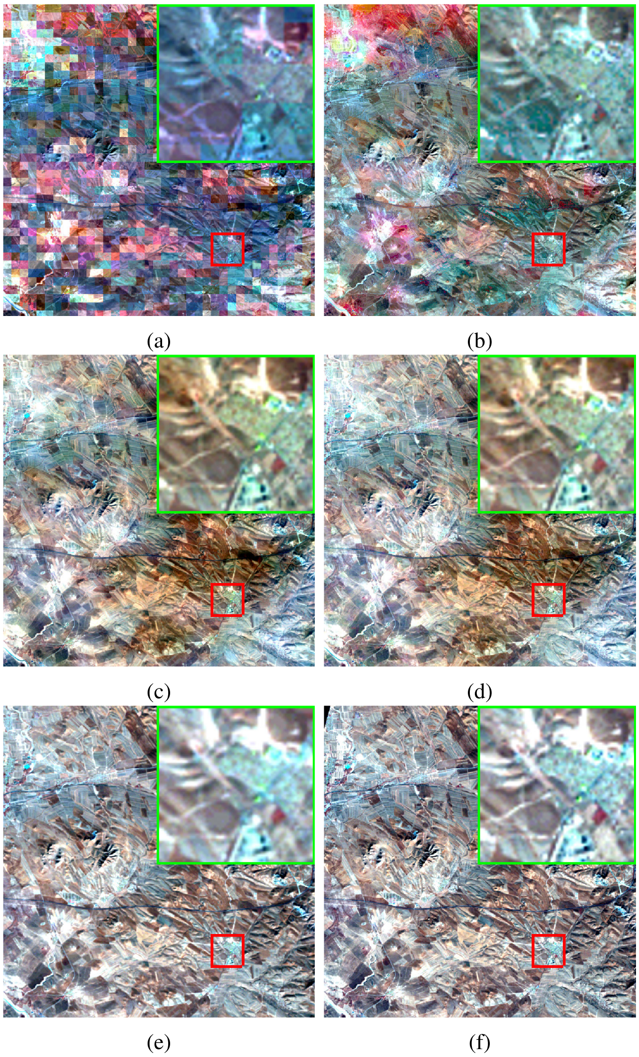


FIGURE 7. Result images by different algorithms using DATA-II. Sub-figures (a)-(e) are the reconstructed image of STARFM, ESTARFM, EBSCDL, CSSF, and CSBS accordingly. Sub-figure (f) is the actual Landsat ETM+ land surface reflectance image at 2018-02-05. According to the results, CSBS shows best result among the algorithms.

TABLE 8. RMSE of different algorithms using DATA-III.

Band	Algorithms				
	STARFM	ESTARFM	EBSCDL	CSSF	CSBS
NIR	4.92688	0.08243	0.11280	0.11267	**0.04494**
red	0.03042	0.02315	0.02333	0.02325	**0.02057**
green	0.02547	0.08264	0.02215	0.02208	**0.02159**

TABLE 9. CC of different algorithms using DATA-III.

Band	Algorithms				
	STARFM	ESTARFM	EBSCDL	CSSF	CSBS
NIR	0.32495	0.66482	0.82161	0.82537	**0.85178**
red	0.64252	0.81607	0.81852	0.81911	**0.83088**
green	0.76453	0.27601	0.84743	0.84743	**0.85136**

TABLE 10. ERGAS of different algorithms using DATA-III.

	Algorithms				
	STARFM	ESTARFM	EBSCDL	CSSF	CSBS
ERGAS	2.09021	1.86956	4.23646	4.23320	**0.96636**

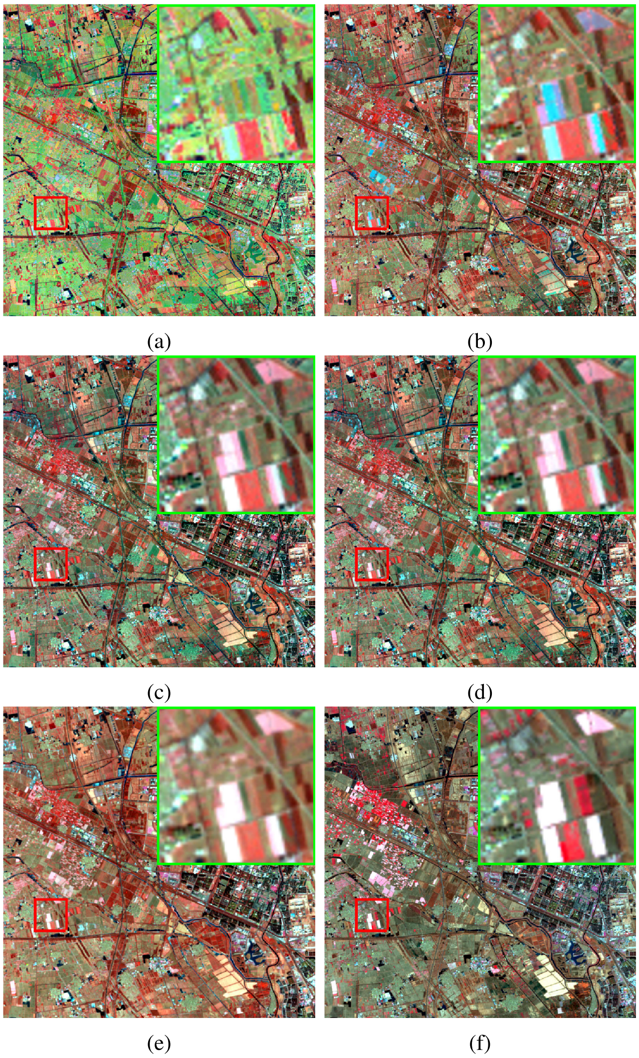


FIGURE 8. Result images by different algorithms using DATA-III. Sub-figures (a)-(e) are the reconstructed image of STARFM, ESTARFM, EBSCDL, CSSF, and CSBS accordingly. Sub-figure (f) is the actual Landsat ETM+ land surface reflectance image at 2018-12-04. ESTARFM, EBSCDL, and CSSF all show fake textures in the high reflectance region.

F. COMPUTATIONAL COST

When the computational cost is concerned, CSBS still has an advantage. In the test, all the algorithms were tested on a Windows laptop, which has one Intel Core i7-4700MQ CPU @ 2.40GHz and 16.0GB RAM. In the test, STARFM is coded in C with multithreads supported (OpenMP). ESTARFM is coded in IDL. EBSCDL, CSSF and CSBS are coded in MATLAB. The algorithms were tested using DATA-I, DATA-II and DATA-III. The average running time is listed in Table 11. The efficient programming language and the simplicity of computational model make STARFM the fastest algorithm. Except for STARFM, CSBS outperforms other algorithms. The concise objective function makes CSBS more efficient than EBSCDL and CSSF.

TABLE 11. Programming language and running time of different algorithms. The time in this table is the mean time of band nir, red and green on DATA-I, DATA-II and DATA-III accordingly.

	Algorithms				
	STARFM	ESTARFM	EBSCDL	CSSF	CSBS
Language	C	IDL	Matlab	Matlab	Matlab
DATA-I	3.1	484.4	983.9	4736.2	473.8
DATA-II	4.3	585.0	1437.8	4882.9	472.6
DATA-III	2.1	427.0	947.2	4464.8	379.3

Unit:second.

IV. CONCLUSION

We proposed a new spatiotemporal data-fusion model based on compressed sensing. In our CSBS model, the low-spatial resolution images were taken as observations of high-spatial resolution images based on CS theory. Then, the temporal continuity was reasonably introduced into the CS object function. Training and predicting were thus implemented using the patch-group model. Therefore, the spatial-temporal features were well-explored by the patch-group of the CS reconstruction. In conclusion, our proposed CSBS model is characterized by the patch-group stratagem in dictionary learning and measurement matrix estimation and the combination of continuity of temporal feature and sparsity in spatial features. We compared the proposed CSBS method with four other algorithms in terms of both quantity and quality. In terms of quantity, the proposed CSBS showed larger CC and smaller RMSE and ERGAS for different datasets in most cases. From the perspective of quality, CSBS showed more fine features of reconstructed high-spatiotemporal images in visual methods than others. Experiments showed CSBS can handle land-cover type changes and phenology changes well. From the perspective of computational cost, CSBS also showed advantages. Experimental results confirmed the effectiveness of the spatial-temporal CS reconstruction for fusion based on patch-group model.

ACKNOWLEDGMENT

The authors would like to thank Dr. Xiaodao Chen for the insightful discussion. The authors also want to thank the reviewers for their suggestions.

REFERENCES

[1] A. Schneider, "Monitoring land cover change in urban and peri-urban areas using dense time stacks of Landsat satellite data and a data mining approach," *Remote Sens. Environ.*, vol. 124, pp. 689–704, Sep. 2012.

[2] M. Shen, Y. Tang, J. Chen, X. Zhu, and Y. Zheng, "Influences of temperature and precipitation before the growing season on spring phenology in grasslands of the Central and Eastern Qinghai-Tibetan Plateau," *Agricult. Forest Meteorol.*, vol. 151, no. 12, pp. 1711–1722, Dec. 2011.

[3] F. Gao, T. Hilker, X. Zhu, M. Anderson, J. Masek, P. Wang, and Y. Yang, "Fusing Landsat and MODIS data for vegetation monitoring," *IEEE Geosci. Remote Sens. Mag.*, vol. 3, no. 3, pp. 47–60, Sep. 2015.

[4] K. J. Lees, T. Quaife, R. R. E. Artz, M. Khomik, and J. M. Clark, "Potential for using remote sensing to estimate carbon fluxes across Northern Peatlands–a review," *Sci. Total Environ.*, vol. 615, pp. 857–874, Feb. 2018.

[5] M. D. Johnson, W. W. Hsieh, A. J. Cannon, A. Davidson, and F. Bédard, "Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods," *Agricult. Forest Meteorol.*, vols. 218–219, pp. 74–84, Mar. 2016.

[6] X. Li, Y. Zhou, G. R. Asrar, J. Mao, X. Li, and W. Li, "Response of vegetation phenology to urbanization in the conterminous United States," *Global Change Biol.*, vol. 23, no. 7, pp. 2818–2830, Jul. 2017.

[7] Q. Liu, Y. H. Fu, Z. Zhu, Y. Liu, Z. Liu, M. Huang, I. A. Janssens, and S. Piao, "Delayed autumn phenology in the Northern Hemisphere is related to change in both climate and spring phenology," *Global Change Biol.*, vol. 22, no. 11, pp. 3702–3711, Nov. 2016.

[8] F. Gao, J. Masek, M. Schwaller, and F. Hall, "On the blending of the Landsat and MODIS surface reflectance: Predicting daily landsat surface reflectance," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 8, pp. 2207–2218, Aug. 2006.

[9] Q. Weng, P. Fu, and F. Gao, "Generating daily land surface temperature at Landsat resolution by fusing Landsat and MODIS data," *Remote Sens. Environ.*, vol. 145, pp. 55–67, Apr. 2014.

[10] C. Liao, J. Wang, I. Pritchard, J. Liu, and J. Shang, "A spatio-temporal data fusion model for generating NDVI time series in heterogeneous regions," *Remote Sens.*, vol. 9, no. 11, p. 1125, Nov. 2017.

[11] P. Wang, F. Gao, and J. G. Masek, "Operational data fusion framework for building frequent landsat-like imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 7353–7365, Nov. 2014.

[12] X. Zhu, J. Chen, F. Gao, X. Chen, and J. G. Masek, "An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions," *Remote Sens. Environ.*, vol. 114, no. 11, pp. 2610–2623, Nov. 2010.

[13] I. V. Emelyanova, T. R. McVicar, T. G. Van Niel, L. T. Li, and A. I. J. M. van Dijk, "Assessing the accuracy of blending landsat–modis surface reflectances in two landscapes with contrasting spatial and temporal dynamics: A framework for algorithm selection," *Remote Sens. Environ.*, vol. 133, pp. 193–209, Jun. 2013.

[14] T. Hilker, M. A. Wulder, N. C. Coops, J. Linke, G. McDermid, J. G. Masek, F. Gao, and J. C. White, "A new data fusion model for high spatial- and temporal-resolution mapping of forest disturbance based on Landsat and MODIS," *Remote Sens. Environ.*, vol. 113, no. 8, pp. 1613–1627, Aug. 2009.

[15] Q. Wang, Y. Zhang, A. O. Onojeghuo, X. Zhu, and P. M. Atkinson, "Enhancing spatio-temporal fusion of MODIS and Landsat data by incorporating 250 m MODIS data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 9, pp. 4116–4123, Sep. 2017.

[16] B. Wu, B. Huang, K. Cao, and G. Zhuo, "Improving spatiotemporal reflectance fusion using image inpainting and steering kernel regression techniques," *Int. J. Remote Sens.*, vol. 38, no. 3, pp. 706–727, Feb. 2017.

[17] B. Zhukov, D. Oertel, F. Lanzl, and G. Reinhackel, "Unmixing-based multisensor multiresolution image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 3, pp. 1212–1226, May 1999.

[18] Z. Niu, "Use of MODIS and landsat time series data to generate high-resolution temporal synthetic landsat data using a spatial and temporal reflectance fusion model," *J. Appl. Remote Sens.*, vol. 6, no. 1, Mar. 2012, Art. no. 063507.

[19] M. Wu, W. Huang, Z. Niu, and C. Wang, "Generating daily synthetic landsat imagery by combining Landsat and MODIS data," *Sensors*, vol. 15, no. 9, pp. 24002–24025, Sep. 2015.

[20] M. Lu, J. Chen, H. Tang, Y. Rao, P. Yang, and W. Wu, "Land cover change detection by integrating object-based data blending model of Landsat and MODIS," *Remote Sens. Environ.*, vol. 184, pp. 374–386, Oct. 2016.

[21] R. Zurita-Milla, J. Clevers, and M. E. Schaepman, "Unmixing-based landsat TM and MERIS FR data fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 3, pp. 453–457, Jul. 2008.

[22] J. Amorós-López, L. Gómez-Chova, L. Alonso, L. Guanter, R. Zurita-Milla, J. Moreno, and G. Camps-Valls, "Multitemporal fusion of Landsat/TM and Envisat/MERIS for crop monitoring," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 23, pp. 132–141, Aug. 2013.

[23] F. Maselli and F. Rembold, "Integration of LAC and GAC NDVI data to improve vegetation monitoring in semi-arid environments," *Int. J. Remote Sens.*, vol. 23, no. 12, pp. 2475–2488, Jan. 2002.

[24] H. Shen, X. Meng, and L. Zhang, "An integrated framework for the spatio–temporal–spectral fusion of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7135–7148, Dec. 2016.

[25] J. Xue, Y. Leung, and T. Fung, "A Bayesian data fusion approach to spatio-temporal fusion of remotely sensed images," *Remote Sens.*, vol. 9, no. 12, p. 1310, Dec. 2017.

[26] B. Huang, H. Zhang, H. Song, J. Wang, and C. Song, "Unified fusion of remote-sensing imagery: Generating simultaneously high-resolution synthetic spatial–temporal–spectral Earth observations," *Remote Sens. Lett.*, vol. 4, no. 6, pp. 561–569, Jun. 2013.

[27] H. Song, Q. Liu, G. Wang, R. Hang, and B. Huang, "Spatiotemporal satellite image fusion using deep convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 3, pp. 821–829, Mar. 2018.

[28] B. Huang and H. Song, "Spatiotemporal reflectance fusion via sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 10, pp. 3707–3716, Oct. 2012.

[29] H. Song and B. Huang, "Spatiotemporal satellite image fusion through one-pair image learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 4, pp. 1883–1896, Apr. 2013.

[30] J. Wei, L. Wang, P. Liu, and W. Song, "Spatiotemporal fusion of remote sensing images with structural sparsity and semi-coupled dictionary learning," *Remote Sens.*, vol. 9, no. 1, p. 21, Dec. 2016.

[31] B. Wu, B. Huang, and L. Zhang, "An error-bound-regularized sparse coding for spatiotemporal reflectance fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6791–6803, Dec. 2015.

[32] J. Wei, L. Wang, P. Liu, X. Chen, W. Li, and A. Y. Zomaya, "Spatiotemporal fusion of MODIS and Landsat-7 reflectance images via compressed sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 7126–7139, Dec. 2017.

[33] X. Li, F. Ling, G. M. Foody, Y. Ge, Y. Zhang, and Y. Du, "Generating a series of fine spatial and temporal resolution land cover maps by fusing coarse spatial resolution remotely sensed images and fine spatial resolution land cover maps," *Remote Sens. Environ.*, vol. 196, pp. 293–311, Jul. 2017.

[34] J. Li, Y. Li, L. He, J. Chen, and A. Plaza, "Spatio-temporal fusion for remote sensing data: An overview and new benchmark," *Sci. China Inf. Sci.*, vol. 63, no. 4, Apr. 2020, Art. no. 140301.

[35] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[36] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[37] D. L. Donoho, "For most large underdetermined systems of equations, the minimal ℓ_1-norm near-solution approximates the sparsest near-solution," *Commun. Pure Appl. Math.*, vol. 59, no. 7, pp. 907–934, Jul. 2006.

[38] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *Comp. Rendus Mathematique*, vol. 346, nos. 9–10, pp. 589–592, May 2008.

[39] E. J. Candès, Y. C. Eldar, D. Needell, and P. Randall, "Compressed sensing with coherent and redundant dictionaries," *Appl. Comput. Harmon. Anal.*, vol. 31, no. 1, pp. 59–73, Jul. 2011.

[40] G. Kutyniok, "Theory and applications of compressed sensing," *GAMM-Mitteilungen*, vol. 36, no. 1, pp. 79–101, Aug. 2013.

[41] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, no. 1, pp. 19–60, 2010.

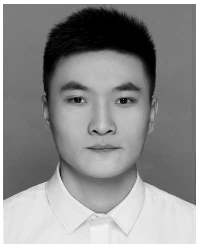
[42] M. A. Davenport, M. F. Duarte, Y. C. Eldar, and G. Kutyniok, "Introduction to compressed sensing," in *Compressed Sensing: Theory and Applications*, Y. C. Eldar and G. Kutyniok, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2012, pp. 1–68.

[43] J. Li, Y. Li, L. He, J. Chen, and A. Plaza, "Spatio-temporal fusion for remote sensing data: An overview and new benchmark," *J. Sci. China Inf. Sci.*, vol. 63, no. 4, 2020, Art. no. 140301, doi: 10.1007/s11432-019-2785-y.

[44] J. J. Qu, W. Gao, M. Kafatos, R. E. Murphy, and V. V. Salomonson, *Earth Science Satellite Remote Sensing: Vol.2: Data, Computational Processing, and Tools*. Berlin, Germany: Springer-Verlag, 2006.

[45] X. Zhu, E. H. Helmer, F. Gao, D. Liu, J. Chen, and M. A. Lefsky, "A flexible spatiotemporal method for fusing satellite images with different resolutions," *Remote Sens. Environ.*, vol. 172, pp. 165–177, Jan. 2016.

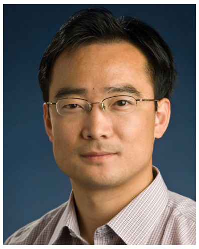
[46] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogramm. Eng. Remote Sens.*, vol. 63, pp. 691–699, Nov. 1997.



LEI LI was born in Nanyang, China, in 1996. He received the B.S. degree in computer science from the China University of Geosciences, Wuhan, China, in 2017. He is currently pursuing the M.S. degree in cartography and geography information system with the University of Chinese Academy of Sciences, Beijing, China.

PENG LIU received the M.S. and Ph.D. degrees in signal processing from the Chinese Academy of Sciences, in 2004 and 2009, respectively. From May 2012 to May 2013, he was with the Department of Electrical and Computer Engineering, George Washington University, as a Visiting Scholar. He is currently an Associate Professor with the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences. He has published more than 30 scientific peer-reviewed articles. His research interests include big data, sparse representation, compressed sensing, and deep learning and their applications to remote sensing data processing. He serves as an Associate Editor for *Frontiers in Environmental Science* and IEEE ACCESS. He also serves as a Reviewer for *Journal of Applied Remote Sensing*, the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, *Neurocomputing*, *Signal Processing*, and so on.

JIE WU was born in Jinzhong, China, in 1993. She received the B.S. degree in geography information system from Tianjin Normal University, Tianjin, China, in 2016. She is currently pursuing the Ph.D. degree in cartography and geography information system with the University of Chinese Academy of Sciences, Beijing, China.



LIZHE WANG (Member, IEEE) received the B.E. and M.E. degrees from Tsinghua University and the D.E. degree *(Magna Cum Laude)* from the Karlsruhe Institute of Technology, Germany. He is currently the Dean of the School of Computer Science, China University of Geosciences. His research interests include remote sensing data processing, digital earth, and big data computing. He is a Fellow of IET and BCS. He was a recipient of the Distinguished Young Scholars of NSFC, the National Leading Talents of Science and Technology Innovation, and the 100-Talents Program of the Chinese Academy of Sciences. He serves as an Associate Editor for *Remote Sensing*, *IJDE*, *ACM Computing Surveys*, the IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, the IEEE TRANSACTIONS ON SUSTAINABLE COMPUTING, and so on.

GUOJIN HE was born in Fujian, China, in 1968. He received the B.Sc. degree in geology from Fuzhou University, Fuzhou, China, in 1989, the M.Sc. degree in remote sensing of geology from the China University of Geosciences, Wuhan, China, in 1992, and the Ph.D. degree in geology from the Institute of Geology, Chinese Academy of Sciences (CAS), Beijing, China, in 1998. From 1992 to 2007, he was with the Information Processing Department, China Remote Sensing Satellite Ground Station (RSGS), CAS, where he was a Deputy Director with the Information Processing Department, in 2001. Since 2004, he has been a Professor and the Director of the Information Processing Department, RSGS. He has also been the Head of the Research Group of the Remote Sensing Information Mining and Intelligent Processing. From 2008 to 2012, he was a Professor and the Director of the Value-Added Product Department and a Deputy Director with the Spatial Data Center, Center for Earth Observation and Digital Earth, CAS. Since 2013, he has been a Professor, the Director of the Satellite Data Based Value-Added Product Department, and a Deputy Director with RSGS, Institute of Remote Sensing and Digital Earth, CAS. A large part of his earlier research dealt with information processing and applications of satellite remote sensing data. His current research interests include optical high-resolution remote sensing image understanding and using information retrieved from satellite remote sensing images in combination with other sources of data to support better understanding of the Earth.

• • •