Spatiotemporal Satellite Image Fusion Through One-Pair Image Learning

Article *in* IEEE Transactions on Geoscience and Remote Sensing · April 2013

DOI: 10.1109/TGRS.2012.2213095

2 authors:

Huihui Song

49 PUBLICATIONS **2,867** CITATIONS

SEE PROFILE

Bo Huang

The Chinese University of Hong Kong

295 PUBLICATIONS **8,379** CITATIONS

SEE PROFILE

Spatiotemporal Satellite Image Fusion Through One-Pair Image Learning

Huihui Song and Bo Huang, *Associate Member, IEEE*

***Abstract*—This paper proposes a novel spatiotemporal fusion model for generating images with high-spatial and high-temporal resolution (HSHT) through learning with only one pair of prior images. For this purpose, this method establishes correspondence between low-spatial-resolution but high-temporal-resolution (LSHT) data and high-spatial-resolution but low-temporal-resolution (HSLT) data through the superresolution of LSHT data and further fusion by using high-pass modulation. Specifically, this method is implemented in two stages. In the first stage, the spatial resolution of LSHT data on prior and prediction dates is improved simultaneously by means of sparse representation; in the second stage, the known HSLT and the superresolved LSHTs are fused via high-pass modulation to generate the HSHT data on the prediction date. Remarkably, this method forms a unified framework for blending remote sensing images with temporal reflectance changes, whether phenology change (e.g., seasonal change of vegetation) or land-cover-type change (e.g., conversion of farmland to built-up area) based on a two-layer spatiotemporal fusion strategy due to the large spatial resolution difference between HSLT and LSHT data. This method was tested on both a simulated data set and two actual data sets of Landsat Enhanced Thematic Mapper Plus–Moderate Resolution Imaging Spectroradiometer acquisitions. It was also compared with other well-known spatiotemporal fusion algorithms on two types of data: images primarily with phenology changes and images primarily with land-cover-type changes. Experimental results demonstrated that our method performed better in capturing surface reflectance changes on both types of images.**

***Index Terms*—High-pass modulation, high spatial resolution, high temporal resolution, sparse representation, spatiotemporal fusion.**

I. Introduction

With the growing awareness of the importance of monitoring the dynamics of land cover and ecosystems in understanding the problems related to climate change and policies for sustainable development [1]–[3], an increasing number of remote sensing sensors have been developed over the last decade; these sensors have a higher capability of capturing multitemporal or high-temporal images of large areas of the Earth's surface. However, because of the limitations of hardware technology and budget constraints, a key technological challenge confronting such multitemporal instruments is the tradeoff between spatial resolution and temporal coverage. For example, the instruments of Moderate Resolution Imaging Spectroradiometer (MODIS) on the Terra/Aqua and Système Pour l'Observation de la Terre Vegetation (SPOT-VGT) can be used to collect daily observations, but these instruments have spatial resolutions ranging from 250 to 1000 m, and this is insufficient for monitoring regional or local land-cover and ecosystem changes. On the other hand, remote sensing images from instruments of SPOT and Landsat with spatial resolution ranging from 10 to 30 m are usually suited for land-use/land-cover mapping and the detection of changes, but their long revisit intervals—from half a month to one month–together with the impact of poor weather and atmospheric conditions, limit their use in detecting rapid surface changes associated with seasonal variations or the disturbance caused by human activities [4]. In order to monitor regional land-cover and ecosystem changes with both high-spatial and high-temporal (HSHT) accuracies, one possible cost-effective solution is to combine remote sensing data from sensors with low spatial resolution but high temporal resolution (LSHT) and those with high spatial resolution but low temporal resolution (HSLT) to generate data with HSHT resolution.

During recent decades, many studies have focused on the fusion of one high-resolution panchromatic band and low-resolution multispectral bands from one or more instruments [5]–[8]. The objective of these studies is to inject proportionally the spatial high-frequency information of a panchromatic band into multispectral bands to obtain a high-spectral-resolution and high-spatial-resolution image. In contrast to this spectral and spatial fusion problem, the purpose of spatiotemporal fusion in this study is to predict the unknown high-spatial-resolution image in quantitative accuracy from its LSHT counterpart and a prior LSHT–HSLT pair. Therefore, we can improve the spatial resolution of LSHT data and increase the coverage frequency of HSLT data by means of the spatiotemporal fusion algorithm.

In recent years, several spatiotemporal fusion algorithms have been proposed [9]–[13] under different assumptions and application contexts. Among them, spatial and temporal adaptive reflectance fusion model (STARFM) [9] and Enhanced STARFM (ESTARFM) [10] have drawn considerable attention. Both of them are considered particularly suited for capturing reflectance changes caused by phenology; however, ESTARM is more suited for heterogeneous regions owing to the use of "conversion coefficient" between the Landsat and MODIS data. Nevertheless, they are built under the assumption that land-cover types and the proportion of each land cover type do not change during the observation period. However, disturbance events (e.g., forest fires) or land-cover-type changes (e.g.,

Manuscript received February 27, 2012; revised May 16, 2012 and July 9, 2012; accepted August 4, 2012. Date of publication October 11, 2012; date of current version March 21, 2013. This work was supported in part by the National High-Tech R&D Program (863 program) through Grant 2009AA122004 and in part by the Hong Kong Research Grant Council through Grant CUHK 444612.

The authors are with The Chinese University of Hong Kong, Shatin, Hong Kong (e-mail: bohuang@cuhk.edu.hk; shhsherry@cuhk.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2012.2213095

conversion of grassland to built-up area) are inevitable because of human activities on the Earth's surface. To deal with the transient events which are not recorded in at least one Landsat image, Hilker *et al.* [11] developed a Spatial and Temporal Adaptive Algorithm for mapping Reflectance CHange (STAARCH), which uses the Landsat images at the two end dates of an observation period to detect the spatial changes and a dense series of spatially coincident MODIS images to detect temporal changes. However, since its change detection is based on the Tasseled Cap Transformation [11], the application of this algorithm may be limited to forest disturbances. Moreover, although the optimal disturbance date can be determined by the MODIS series of images in STAARCH, the prediction of the Landsat-like image is implemented based on the STARFM. Thus, STAARCH is similar to STARFM in delineating the changes in the predicted image. Zurita-Milla *et al.* [12] developed a downscaling fusion algorithm based on a linear mixing model to generate images with the spatial resolution of Landsat and the spectral and temporal resolutions of Medium Resolution Imaging Spectrometer. However, it requires a high-resolution land-use database for pixel unmixing in this algorithm, which may limit its applications in some cases. Consequently, a more effective spatiotemporal fusion model is required to deal with the reflectance changes, including both phenology and land-cover-type changes.

Based on two prior Landsat–MODIS pairs as in the ESTARFM scheme, a sparse-representation-based spatiotemporal reflectance fusion model (SPSTFM) has been recently developed [13]. This model treats reflectance changes, including both phenology and land-cover-type changes, in a unified way by training a dictionary pair between the LSHT difference image and the HSLT difference image. However, in many remote sensing applications, only one pair of prior images may be available; this is because of cloud contamination, time inconsistency of image acquisitions, or other reasons. Another fact is that it is very time consuming to find or prepare another pair of known images in some cases. Thus, fusion based on one pair of prior images is more convenient for users even though it is a more challenging task.

In the natural image processing field, the problem of recovering an original high-resolution image from its low-resolution version is called single-image superresolution [14], [15], in which the low-resolution (denoted as X) and the high-resolution (denoted as Y) images are correlated by a degradation model, as in the following:

$$X = SBY + n \tag{1}$$

where the matrices S and B represent downsampling and blurring, respectively, and n is the noise in the generation of X from Y. Single-image superresolution has received increased interest in recent years [14]–[18]. In this paper, we propose a novel spatiotemporal fusion algorithm under the scenario of one pair of prior images available. Its implementation is based on superresolution of the LSHT data (the first stage) and fusion of the improved LSHT data (defined herein as transition image) with the known HSLT data (the second stage). In the first stage, we will correlate the corresponding bands of LSHT data and HSLT data by a similar degradation process (blurring, downsampling, and adding noise) as in natural images, and then, the spatial resolution of LSHT data is improved by using a similar sparse representation method for single-image superresolution in the natural image field [14], [15]. Specifically, we will establish the relationship between the LSHT and HSLT data by training a dictionary pair [14] on the prior LSHT–HSLT image pair, and thereafter, the spatial resolution of LSHT data in the prediction date is improved by utilizing the sparse coding technique. By doing this, the superresolution process of the LSHT data can, to some extent, unmix the heterogeneous pixels of LSHT data. Furthermore, the superresolution execution on two LSHT images can increase their change detection accuracy. In the second stage, we will take full advantage of the detailed information provided by the known HSLT image by fusing the transition images and the known HSLT image based on a high-pass modulation method [8]. Considering the large spatial resolution difference between MODIS and Landsat, the aforementioned two-stage fusion procedure will be employed to improve the spatial resolution of the MODIS image on the prediction date with a scale factor of two or four in the first layer. Then, in the second layer, the improved version in the first layer will be processed by the same two-stage fusion process to improve its spatial resolution so that it will be the same as the original Landsat data. It is notable that the aforementioned two-layer-based spatiotemporal fusion framework can be applied to both phenology and land-cover-type changes, owing to the superresolution and high-pass modulation processes within the framework.

In Section II hereof, we will briefly introduce the theoretical knowledge of sparse representation and the application of this method in single-image superresolution. Then, in Section III, our proposed method will be illustrated in detail. The experimental results with both simulation data and the actual data set will be shown in Section IV. We will then conclude this paper in Section V with a discussion of our findings.

II. Related Studies

A. Sparse Representation

The application of sparse representation in the image processing field arose from the observation that natural images tend to be sparse in some image bases' space [16]. Different from the traditional Fourier transform (FT), Karhunen-Loève transform, or wavelet transforms, which use fixed dictionaries based on mathematical description, sparse representations possess further sparsity by learning overcomplete dictionaries directly on data. More specifically, the structure of complex natural phenomena can be more accurately and adaptively extracted on specific signal content [17], so that the original signals can be sparsely reconstructed or represented through a sparse coding technique from the degraded signals and the learned dictionaries. During the past decade, with the development of dictionary training methods and sparse coding techniques, sparse representation achieved the state-of-the-art results in image denoising, image superresolution, and many other image processing tasks [16], [18].

Suppose that some $\sqrt{n} \times \sqrt{n}$ image patches are represented by $\{x_1, x_2, \ldots, x_N\}$, $x_i \in R^n$ after lexicographically stacking the pixel values. Then, sparse representation assumes that these signals can be sparsely represented by an overcomplete dictionary matrix D and corresponding coefficient vector α, i.e., $x = D\alpha$, $D \in R^{n \times m}$, $n < m$. Here, each column of the dictionary D stands for a basic atom, and the sparse representation coefficient vector α includes very few nonzero components. To achieve this representation, the dictionary D needs to be trained on samples which have similar characteristics with the signal x, and the coefficient vector α needs to be solved efficiently with respect to the dictionary D through sparse coding algorithms. Thus, x can be sparsely represented by the linear combination of a few atoms in D.

In sparse representation theory and algorithms, many researchers committed themselves to developing dictionary training and sparse coding algorithms with both high accuracy and fast speed. For dictionary training methods, the most recent methods focus on l_0 and l_1 sparsity because of their simple formulation and the ease of using highly efficient sparse coding techniques [17], [19]. In this paper, we design the dictionary by employing the K-Singular Value Decomposition (K-SVD) algorithm [19], due to its simplicity and efficiency. To find the best dictionary D to represent the given training samples $X = [x_1, x_2, \ldots, x_N]$, the K-SVD algorithm solves the following objective function:

$$\min_{D,\Lambda} \left\{ \|X - D\Lambda\|_F^2 \right\} \quad \text{s.t.} \quad \forall i, \|\alpha_i\|_0 \leq K_0 \tag{2}$$

where $\Lambda = [\alpha_1, \alpha_2, \ldots, \alpha_N]$, $\|\alpha\|_0$ denotes the number of nonzero elements in α and K_0 is a fixed and predetermined number of nonzero entries of α. The main feature of the K-SVD algorithm is that it alternates an atom-by-atom dictionary updating step based on a simple *singular value decomposition* and the associated sparse coefficient updating step (also called sparse coding in sparse representation theory) based on an *orthogonal matching pursuit* (OMP) [20]. Further details about the K-SVD algorithm can be found in [19].

To represent a specific signal x with respect to the trained dictionary D in the aforementioned equation, our aim is to find representation coefficient α of x with the fewest nonzero elements through a sparse coding algorithm. This sparsity problem can be described by the following sparse approximation formulation:

$$\hat{\alpha} = \arg\min_{\alpha} \|\alpha\|_0 \quad \text{s.t.} \quad \|x - D\alpha\|_2^2 \leq \varepsilon \tag{3}$$

where ε is the error tolerance. This sparse approximation problem, which is known as a nondeterministic polynomial-time hard problem, can be efficiently solved by using several available approximating algorithms, such as basis pursuit [21], FOcal Underdetermined System Solver (FOCUSS) [22], OMP [20], and others. In this paper, we used OMP, because it was found to lead to an overall far more efficient algorithm.

B. Sparse Representation for Single-Image Superresolution

The development of machine learning techniques motivated the advance of learning-based algorithms for single-image superresolution. Sparse representation is one of these learning-based methods in single-image superresolution. Yang *et al.* [14] first proposed a superresolution algorithm based on sparse representation. By jointly training two dictionaries of high-frequency feature patches of low- and high-resolution images, this algorithm can enforce the similarity of sparse representations between each low-resolution and high-resolution patch pair with respect to their own dictionaries. Later, Zeyde *et al.* [15] substantially improved Yang's algorithm in the following way. Instead of training the low- and high-resolution dictionaries simultaneously, it trained the low-resolution dictionary first by using K-SVD algorithm, which is introduced in Section II-A. Then, the high-resolution dictionary can be directly solved by multiplication involving the high-resolution training sample matrix and the same sparse coefficient matrix obtained during training the low-resolution dictionary. Because of the simplification and efficiency of the K-SVD approach, the overall computational complexity of the algorithm architecture in [15] is greatly reduced. In both of these algorithms, the local sparse-land model on image patches is served as a regularization term or prior information.

In recent years, sparse representation has received increased attention in remote sensing image processing, such as in hyperspectral image classification [23], pansharpening [6], [7], SAR image filtering [24], and so on. In this paper, we use sparse representation to improve the spatial resolution of LSHT data by dint of the K-SVD dictionary training method (as in [15]) and the OMP sparse coding technique, which will be presented in detail in Section III. High-pass modulation is also used to complement the superresolution process so that the fusion of satellite images with high-resolution difference can also be accomplished. It should be noted that, in all the aforementioned work, the dictionary is constructed on numerous image pairs, whereas the dictionary learning for the satellite image fusion in this paper is based on one pair of images, due to the high similarity existed in the sequence of remote sensing images.

III. Proposed Methodology

In our method, the Landsat Enhanced Thematic Mapper Plus (ETM+) surface reflectance data are selected as the image examples with HSLT, and the MODIS surface reflectance data are selected as the image examples with LSHT. Our task is to utilize a given Landsat–MODIS image pair on t_1 (a prior date) and one MODIS image on t_2 (the prediction date) to predict the Landsat image on t_2. Considering that different land-cover types show different spectral features and different reflectance ranges in different bands, we carry out the spatiotemporal fusion process in band-by-band form. In the following, we will first show how we apply sparse representations to improve the spatial resolution of LSHT data to obtain a *transition image*. Then, we will present the fusion process of the *transition image* and HSLT image via a high-pass modulation based on the linear temporal change model between two transition images. Subsequently, we will show how the aforementioned two stages are integrated into the overall spatiotemporal fusion algorithm in a two-layer framework. It is noted that the superresolution algorithm used in this section was also adopted in [15] and it

is the first time that we proposed the application of high-pass modulation and the two-layer framework for the spatiotemporal fusion problems.

A. Superresolution of LSHT Data

Because of the large spatial resolution difference between MODIS (i.e., LSHT data) and Landsat (i.e., HSLT data), directly fusing them will cause large prediction errors. A straightforward solution for this problem is first to improve the spatial resolution of MODIS data and then to fuse the MODIS with an improved spatial resolution and the original Landsat data. If we correlate the corresponding bands of MODIS data and Landsat data by similar degradation processes (blurring, downsampling, and adding noise), as in (1), then the spatial resolution of the MODIS data can be improved by using a similar sparse representation method for single-image superresolution in the natural image field [14], [15]. In fact, it is rational to adopt the degradation model between the corresponding bands of MODIS and Landsat sensors. On the one hand, the parameter setting of sensors and the imaging environment of MODIS and Landsat-7 are very similar in the following respects: equal orbital parameters, close viewing (near nadir) and solar geometries, close time of platform crossing the equator (there is about 30-min time difference), and similar corresponding bandwidths [9]. On the other hand, suppose that the MODIS data and the Landsat data are ideally preprocessed by radiometric calibration, geometric rectification, and atmospheric correction and thereafter put the system biases into the noise term; then, the corresponding bands of MODIS and Landsat can be correlated by a similar degradation model in (1). However, there are two main different aspects between single natural image superresolution and the superresolution of LSHT data in our problem. On the one hand, due to the effects of weather, atmosphere, terrain, and many other complex factors during remote sensing image capture, it is harder to build the accurate correspondence between Landsat and MODIS sensors than it is with natural images. It is therefore more difficult to reconstruct the original high-resolution images as in a single natural image superresolution. On the other hand, in natural image superresolution, the training samples can be extracted from many other image sources, whereas in our problem, we suppose the toughest case when there is only one Landsat–MODIS pair for training. Based on the aforementioned features in our problem, the spatial resolution of the reconstructed image should be in the middle resolution between that of the MODIS and that of the Landsat images. We define this improved MODIS version with middle spatial resolution as the *transition image*, which can provide much more accurate change ratio between the prediction and prior dates for the next step of high-pass modulation.

Denote the MODIS image, the Landsat image, and the predicted transition image on t_i as M_i, L_i, and T_i, respectively. The superresolution of LSHT data contains two steps: the dictionary-pair training on known M_1 and L_1 and the transition image prediction. For training a dictionary pair, the high-resolution image features and low-resolution image features are extracted from difference image space of $L_1 - M_1$ and gradient feature space of M_1 in patch form (e.g., 5×5), respectively. Stacking these feature patches into columns forms the training sample matrices Y and X, where Y and X stand for high-resolution samples and low-resolution samples, respectively, and their columns are in correspondence. First, the low-resolution dictionary D_l is derived by applying the K-SVD [19] training procedure on X via optimizing the following objective function:

$$\{D_l^*, \Lambda^*\} = \arg\min_{D_l, \Lambda} \left\{ \|X - D_l \Lambda\|_F^2 \right\} \quad \text{s.t.} \quad \forall i, \|\alpha_i\|_0 \le K_0 \tag{4}$$

where Λ is a column combination of representation coefficients corresponding to every column in X. To establish correspondence between high-resolution and low-resolution training samples, the high-resolution dictionary is constructed by minimizing the approximation error on Y with the same sparse representation coefficients Λ^* in (4), i.e.,

$$D_h^* = \arg\min_{D_h} \|Y - D_h \Lambda^*\|_F^2. \tag{5}$$

The solution of this problem can be directly derived from the following pseudoinverse expression (given that Λ^* has full row rank):

$$D_h = Y(\Lambda^*)^+ = Y\Lambda^{*T}(\Lambda^*\Lambda^{*T})^{-1}. \tag{6}$$

To predict the transition image T_2 from M_2, the same gradient features X_2 are extracted from M_2 as in the training process. Denote the ith column of X_2 as x_{2i}; then, its sparse coefficient α_i with respect to dictionary D_l can be obtained by employing the sparse coding technique OMP. Because the corresponding high-resolution sample and low-resolution sample are enforced represented by the same sparse coefficients with respect to D_h and D_l, respectively, the corresponding ith middle-resolution patch column y_{2i} can be predicted by $y_{2i} = D_h * \alpha_i$. The other middle-resolution patch columns can be predicted by this same process. After transforming all columns y_{2i} into a patch form, the difference image Y_2 between T_2 and M_2 is predicted. Thus, T_2 is reconstructed by $T_2 = Y_2 + M_2$. For the fusion procedure in the next stage, the transition image T_1 is also predicted in the same procedure. Here, the transition images T_1 and T_2 have the same size and extent as that of L_1 and L_2.

B. Fusion of HSLT and Transition Images via High-Pass Modulation

Because the training image pair L_1–M_1 has many similarities with the image pair L_2–M_2 in regard to phenology or land-cover-type changes in most remote sensing cases, the predicted transition images are more like the Landsat images. Hence, we assume that the pixel purity between L_1 (or L_2) and T_1 (or T_2) is the same. Therefore, if we establish a linear temporal change model between T_1 and T_2 in the following:

$$T_2(i,j) = aT_1(i,j) + b \tag{7}$$

where (i, j) is a given pixel location and a and b are linear regression coefficients for relative temporal change from time

t_1 to t_2, it is very reasonable to assume that this temporal change model also applies to L_1 and L_2, i.e.,

$$L_2(i,j) = aL_1(i,j) + b. \tag{8}$$

From (7) and (8), we can obtain

$$L_2(i,j) = T_2(i,j) + a\left[L_1(i,j) - T_1(i,j)\right]. \tag{9}$$

Equation (9) shows that the HSLT pixel on t_2 equals the sum of the transition pixel on t_2 and the scaled pixel change between HSLT and the transition image on t_1. To compute in matrix form, (9) can be written in an approximate manner, as follows:

$$L_2 = T_2 + \left(\frac{T_2}{T_1}\right)[L_1 - T_1]. \tag{10}$$

The operations in the aforementioned equation are at the pixel level. There are two main advantages of taking (10) as the prediction formulation: The matrix formulation can greatly reduce the amount of computation for scenes with large extent, and because T_1 and T_2 are reconstructed from the same high-resolution dictionary, they intend to have the same level error in one specific image patch, so that the divide operation of T_2/T_1 intends to reduce the prediction error in (10). Equation (10) shows that it transfers the high-frequency information of high-resolution image L_1 to transition image T_2 with modulation coefficients of the ratio between transition images T_2 and T_1. In other words, the high-frequency information on t_1 is proportionally transferred to t_2 by using the high-pass modulation equation (10), which is also widely used in the fusion of panchromatic images and multispectral images [8]. Because T_1 and T_2 have already been predicted and L_1 is known, the image L_2 can be predicted in an easy and efficient way.

C. Overall Two-Layer Spatiotemporal Fusion Framework

Considering that there exists a large spatial resolution difference between Landsat and MODIS data (it is 8–16 times in our experiments), the spatiotemporal fusion procedure is executed in a two-layer framework. In every layer, the aforementioned two stages—1) the superresolution of LSHT data and 2) the fusion of two transition images and an HSLT image via high-pass modulation—are utilized. In order to set the magnification factors in each layer, we consider the following factors: The resolution difference is larger between low- and high-resolution images, and the prediction error in the reconstruction phase is greater; the input low-resolution data for training and reconstruction in the first layer are MODIS images, while the input low-resolution data for training in the second layer is the Landsat downsampled version. Apparently, with the same magnification factor, the superresolution procedure in the first layer is harder than that in the second layer. We therefore magnify the MODIS images in the first layer with a scale factor of two (for a case with about eight times resolution difference between Landsat and MODIS) or four (for a case with about 16 times resolution difference between Landsat and MODIS). Then, the output of the first layer is perceived as the low-resolution input

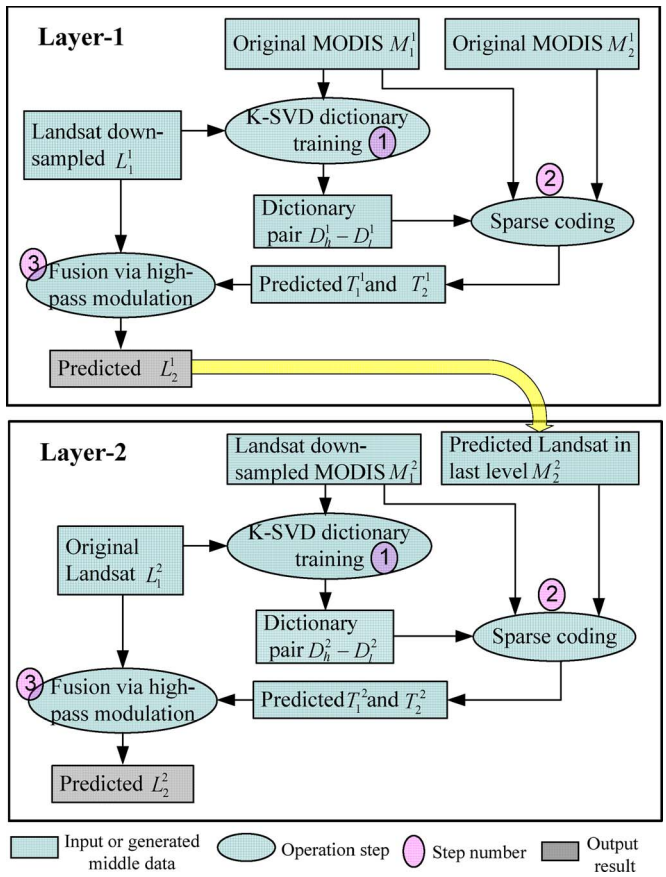

Fig. 1. Flowchart of the two-layer-based spatiotemporal fusion framework.

to the second layer, in which the spatial resolution of the input data is then increased to the same as the original Landsat image with a scale factor of four. The flowchart of the overall two-layer-based spatiotemporal fusion framework is shown in Fig. 1.

Several points need to be clarified for the overall two-layer spatiotemporal fusion framework in Fig. 1: 1) the superscripts of variables in Fig. 1 denote the layer number; 2) the high-spatial-resolution image L_1^1 for training and fusion in the first layer is the downsampled version of the original Landsat by utilizing the bicubic downsampling method in Matlab software, so that the scale factor between L_1^1 and M_1^1 is two or four; for example, if the spatial resolution of M_1^1 is 60 m, then the spatial resolution of L_1^1 is 120 (for a case with the scale factor of two) or 240 m (for a case with the scale factor of four); 3) the predicted high-resolution image L_2^1 in the first layer is as the second input low-resolution image M_2^2 in the second layer; 4) the input low-resolution image M_1^2 in the second layer, which has the same spatial resolution as M_2^2, is the downsampled version of the original Landsat image L_1, so that the scale factor between M_1^2 and L_1 is four; and 5) the combination of steps 1 and 2 is the superresolution process of the LSHT data in Section III-A, and step 3 is the fusion of the HSLT image and transition images via high-pass modulation in Section III-B.

It should be noted that both the aforementioned algorithm and the two-pair-based fusion algorithm, i.e., SPSTFM [13], are developed based on dictionary-pair learning. However,

the dictionaries for SPSTFM are learned for building the relationship between the differences of two high-resolution and two low-resolution images, while the dictionaries in this paper are learned between the high- and low-resolution images directly. These two algorithms also differ in that they are assisted by different strategies in their fusion process. SPSTFM predicts a high-resolution difference image based on its low-resolution counterpart through the dictionary pair constructed and a weighting strategy, whereas the algorithm in this paper predicts a high-resolution image, rather than a difference image, through high-pass modulation and a two-layer framework. The advantage and disadvantage of these two algorithms will be further illustrated in the experimental section (i.e., Section IV).

IV. Experimental Results and Comparisons

In this section, our proposed method is compared to the well-known STARFM algorithm by using both a simulated data set and two actual data sets of Landsat-7 ETM+ and MODIS images. To further illustrate the applicability of this algorithm, it is compared with ESTARFM and SPSTFM under the scenario of two-pair prior images that are required by ESTARFM and SPSTFM. In the preprocessing step, the Landsat-7 ETM+ images were radiometrically and atmospherically corrected by using the MODIS 6S approach, whereas the MODIS images were obtained directly from the Land Processes Distributed Active Archive Center (https://lpdaac.usgs.gov/lpdaac). To ensure that the comparisons are fair, we adjusted the parameters of the involved algorithms to achieve the best overall prediction effect. The parameter values are reported in each specific experiment described hereinafter.

A. Quality Assessment Indices for Image Fusion Results

To evaluate the quality of fusion results quantitatively and visually, we employed several representative metrics. To directly reflect the deviation of the predicted reflectance, we adopted the average absolute difference (AAD) between the predicted reflectance and the actual reflectance. Root-mean-square error (rmse), which is widely applied in the quantitative assessment of image qualities, was chosen as the second index for our assessments. To evaluate the overall fusion result, we utilized the *erreur relative global adimensionnelle de synthèse* (ERGAS) [25]. If ERGAS is smaller and closer to zero, a better fusion result is achieved. To examine the spectral distortion of the fusion result, the spectral angle mapper (SAM) [28] was employed. For an ideally fused image, the SAM should be zero.

In addition to quantitative assessment, a visual assessment index—structural similarity (SSIM) [26]—was also employed to measure the similarity of the overall structure, between the predicted and actual images. SSIM is obtained by computing the mean and variance of the predicted and actual comparison images, i.e.,

$$\mathrm{SSIM}(L,\hat{L}) = \frac{(2\mu_L\mu_{\hat{L}} + C_1)(2\sigma_{L\hat{L}} + C_2)}{\left(\mu_L^2 + \mu_{\hat{L}}^2 + C_1\right)\left(\sigma_L^2 + \sigma_{\hat{L}}^2 + C_2\right)} \quad (11)$$

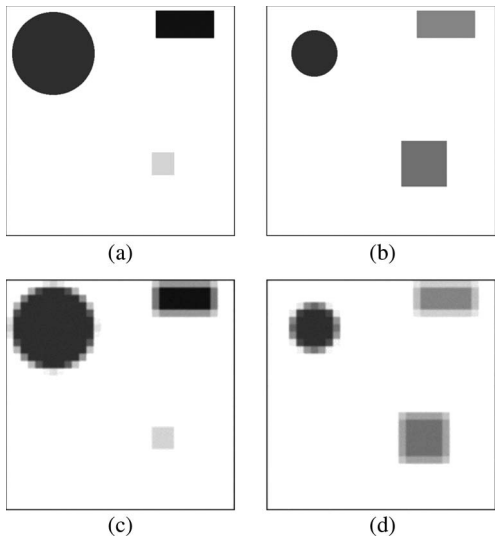

Fig. 2. Two pairs of simulated objects with three kinds of changes. (a) and (b) are the Landsat-like images, and (c) and (d) are their corresponding MODIS-like images. The (a)–(c) pair is acquired on t_1, and the (b)–(d) pair is acquired on t_2.

where μ_L and $\mu_{\hat{L}}$ stand for the means of the comparison images, σ_L and $\sigma_{\hat{L}}$ are the variances of them, $\sigma_{L\hat{L}}$ is the covariance of $\hat{L}$ and L, and C_1 and C_2 are small constants used to avoid instability when $\mu_L^2 + \mu_{\hat{L}}^2$ or $\sigma_L^2 + \sigma_{\hat{L}}^2$ is very close to zero. The comparison rule is based on that the much closer SSIM to one, the more similar $\hat{L}$ to L.

Finally, the scatter plots of the predicted against those of the actual images for each band were shown to provide an intuitive comparison between the estimated and actual reflectances on the approximation extent of distribution.

B. Experiments With Simulated Data

The simulated reflectance data allow us to examine the behavior and performance of the proposed algorithm both on phenology and on type changes of land cover. In this simple case, we set only three objects (i.e., a lake represented by a circle, a grassland represented by the upper right rectangle, and a built-up area represented by the lower right square) as shown in Fig. 2. Fig. 2(a) and (b) shows two Landsat-like images with a 30-m spatial resolution, and Fig. 2(c) and (d) shows their corresponding MODIS-like images with a 480-m spatial resolution, which were aggregated from the Landsat-like images and then were added by white Gaussian noise with a signal-to-noise ratio of 35 dB. It was assumed that the three objects changed from t_1 to t_2 in the following ways: The circular object kept a constant reflectance and reduced its radius from 1800 to 1000 m, the rectangular object remained in the same size and changed its reflectance from 20 to 150, and the square object changed its reflectance from 220 to 130 and changed its side length from 1000 to 2000 m simultaneously. We used the Landsat-like and MODIS-like image pair in Fig. 2(a) and (c) and the MODIS-like image in Fig. 2(b) to predict the Landsat-like image in Fig. 2(d). Then, the prediction results were compared to the original image. In the STARFM algorithm, the window size was fine-tuned to be 23. In our method, the MODIS-like image with

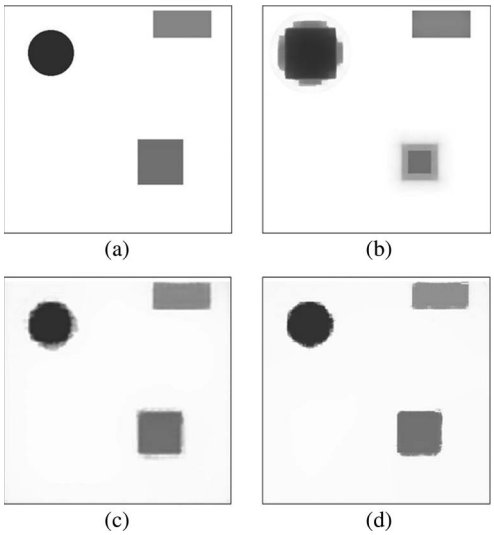

Fig. 3. Comparisons between the original image and the predicted results for the simulated data. (a) Original Landsat-like image on t_2. (b) Predicted image on t_2 by using STARFM. (c) Output image of the first layer by our method on t_2. (d) Output image of the second layer by our method on t_2, i.e., the final predicted image by using our method.

TABLE I
QUANTITATIVE COMPARISON OF OUR METHOD WITH STARFM ON SIMULATED IMAGES

Method	AAD	RMSE	SSIM
STARFM	0.0185	0.136	0.9546
Our method	**0.0002**	**0.015**	**0.9770**

a 480-m spatial resolution at t_2 was magnified to have a 120-m spatial resolution in the first layer and then to have a 30-m spatial resolution in the second layer. Given the relatively more noise in the MODIS-like images, the output images of the two stages (superresolution and fusion based on high-pass modulation) in both layers were then processed by utilizing the bilateral filtering technique [27], which could smooth out the noise while preserving the edges. Fig. 3 shows the prediction results via our method and the STARFM algorithm. In order to make a visual comparison with the original image and other prediction results, the output result of the first layer by our method was upscaled to be the same size as the original image by using the bicubic interpolation method. The outputs of the first layer and the second layer in our method are shown in Figs. 3(c) and (d), from which we can see that the quality of the output image from the first layer has greatly been improved compared with its original MODIS version in Fig. 2(d), but it is still not as good as that of the final output from the second layer. Apparently, the final predicted image of our algorithm [Fig. 3(d)] is very similar to the original image both in reflectance and in shape, whereas on the image predicted by STARFM [Fig. 3(b)], there appears to be a serious error in the regions with shape changes [the circular and square objects in Fig. 3(b)]. The quantitative comparisons of our method with STARFM in terms of AAD, rmse, and SSIM are shown in Table I, from which we can see that there are smaller prediction errors in our fusion result than in the STARFM fusion result.

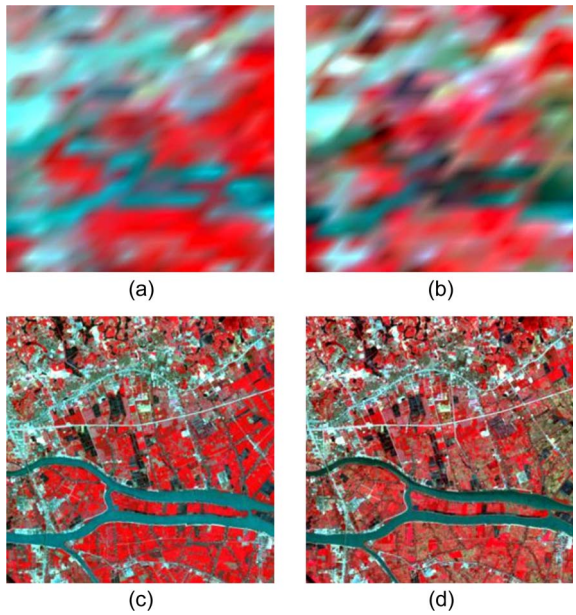

Fig. 4. (Upper row) MODIS composited surface reflectance and (lower row, 300 × 300 pixels) Landsat composited surface reflectance. The (a)–(c) pair was acquired on September 14, 2000, and the (b)–(d) pair was acquired on November 1, 2000.

C. Experiments With Satellite Data—Phenology Changes Over Farmland Region

In this experiment, we used two Landsat-7 ETM+ images and a series of MODIS images covering an area of 24 km × 24 km in Guangzhou, China, to predict a series of Landsat-like surface reflectance data and further applied the fused images to derive the normalized difference vegetation index (NDVI). This data set contains two Landsat–MODIS pairs acquired in September and November 2000 and other ten MODIS images acquired in each month from June to December in 2000 and from January to May in 2001. For the MODIS data set, we utilized the eight-day composite surface reflectance with 500-m spatial resolution, which includes all seven MODIS bands primarily with high-quality nadir view observations. The corresponding Landsat data have spatial resolution of 30 m with an image size of 800 × 800. To show the fusion results and retrieve the NDVI values, we used green, red, and Near-InfraRed (NIR) bands, which correspond with bands 2, 3, and 4 of Landsat and bands 4, 1, and 2 of MODIS. This data set is characterized by phenological changes with the major change in the farmland region, which was chosen as our study area with the size of 300 × 300 pixels. Fig. 4 shows the scenes with an NIR–red–green as red–green–blue composite for both MODIS (upper row) and Landsat (lower row) surface reflectances in September and November 2000.

For our method, we trained the dictionary pair from the image pair in September [Fig. 4(a) and (c)] and utilized the image pair in November [Fig. 4(b) and (d)] as the validation data. Furthermore, we predicted the Landsat-like images from corresponding MODIS images in other months. The number of dictionary atoms was set empirically as follows: 900 in the first layer and 1500 in the second layer. For STARFM, we utilized the given image pair in September as a reference to predict the Landsat-like image on other dates from their corresponding

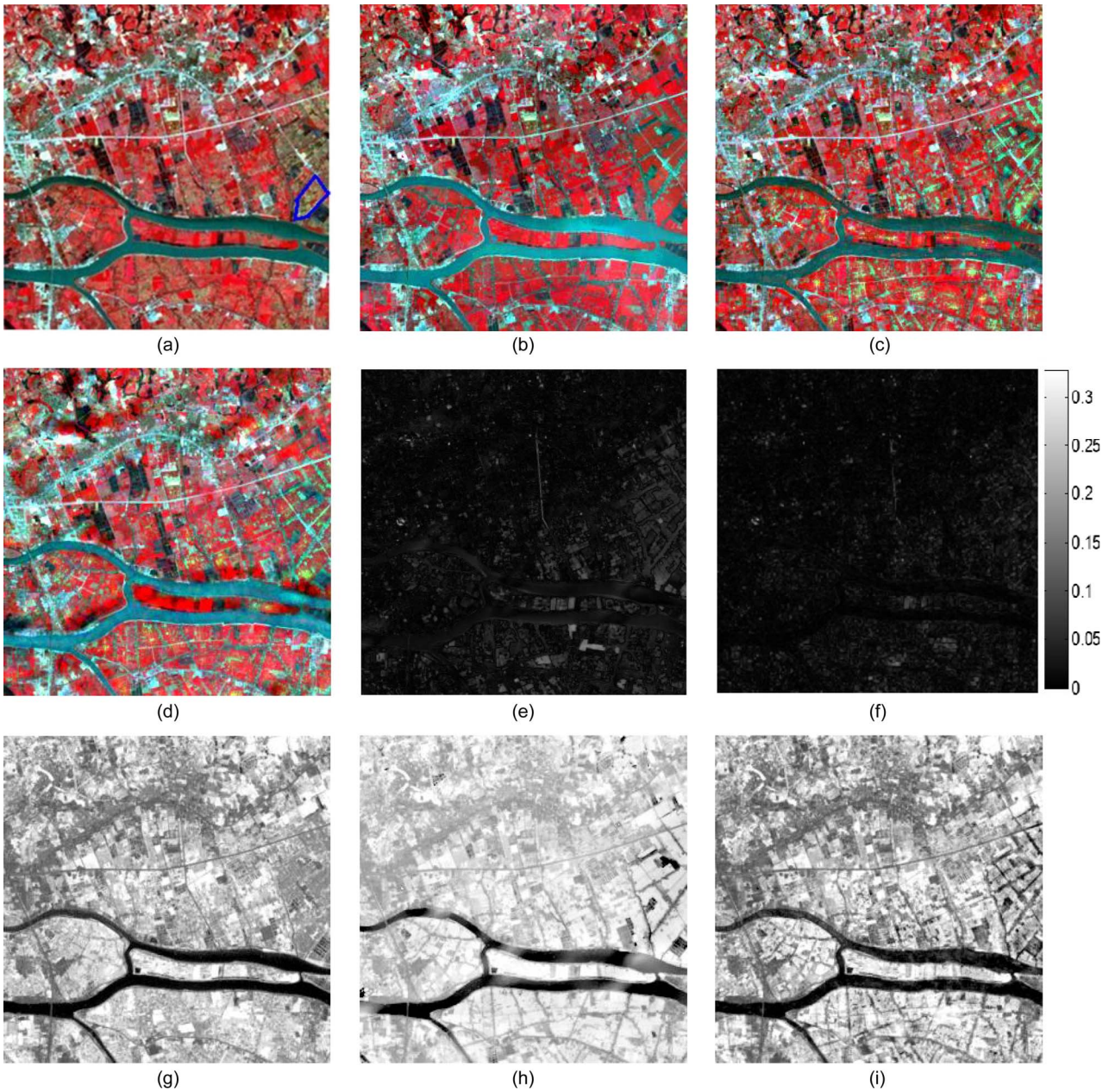

Fig. 5. Comparisons between the actual and the predicted surface reflectances with focus on seasonal changes: (a) Actual surface reflectance, (b) predicted surface reflectance obtained by STARFM, (c) predicted surface reflectance obtained by our method, (d) predicted surface reflectance obtained by our method under the one-layer framework, (e) rmse map of the fusion result obtained by STARFM, (f) rmse map of the fusion result obtained by our method, (g) actual NDVI map, (h) NDVI map obtained from the predicted fusion result of STARFM, and (i) NDVI map obtained from the predicted fusion result of our method.

MODIS image, and the window size in STARFM was fine-tuned to be 50. In our method, the spatial resolution of MODIS on the prediction date was magnified from 500 to 120 m in the first layer and then to 30 m in the second layer. Fig. 5 shows the STARFM prediction [Fig. 5(b)] and the prediction by our method [Fig. 5(c)] in comparison with the actual Landsat ETM+ surface reflectance [Fig. 5(a)]. The fusion result of our method under one-layer framework is shown in Fig. 5(d), from which we can see that there is bigger loss of information in both spatial details and spectral colors than the fusion result under the two-layer framework. Fig. 5(e) and (f) shows the average rmse maps of three bands in the fusion results of STARFM and our method, respectively, which show that the prediction errors mainly exist in the farmland area with major phenology changes (the lower right part). However, we can see that our method has smaller prediction error than STARFM in this area, which reveals that the proposed method is more robust than STARFM in dealing with phenology changes. As mentioned in Section I, this can be attributed to one of the limitations of STARFM, i.e., it does not consider the conversion coefficient between the sensors of Landsat and MODIS as in the ESTARFM algorithm, thereby causing spectral distortion.

The comparison of the three bands (NIR–red–green) between STARFM and our method in terms of AAD, rmse, SSIM, ERGAS, and SAM is listed in Table II. The average AAD values of the three bands for STARFM and our method are 0.021

TABLE II
QUANTITATIVE COMPARISON OF OUR METHOD WITH STARFM ON THE DATA PRIMARILY WITH SEASONAL CHANGES

Method	AAD			RMSE			SSIM			ERGAS	SAM
	NIR	Red	Green	NIR	Red	Green	NIR	Red	Green		
STARFM	0.0032	0.0172	0.0114	0.0495	0.0244	0.0156	0.7767	0.8356	0.9104	0.4390	7.1266
Our method	**0.0205**	**0.0132**	**0.0081**	**0.0300**	**0.0204**	**0.0118**	**0.8269**	**0.8396**	**0.9232**	**0.3121**	**4.4477**

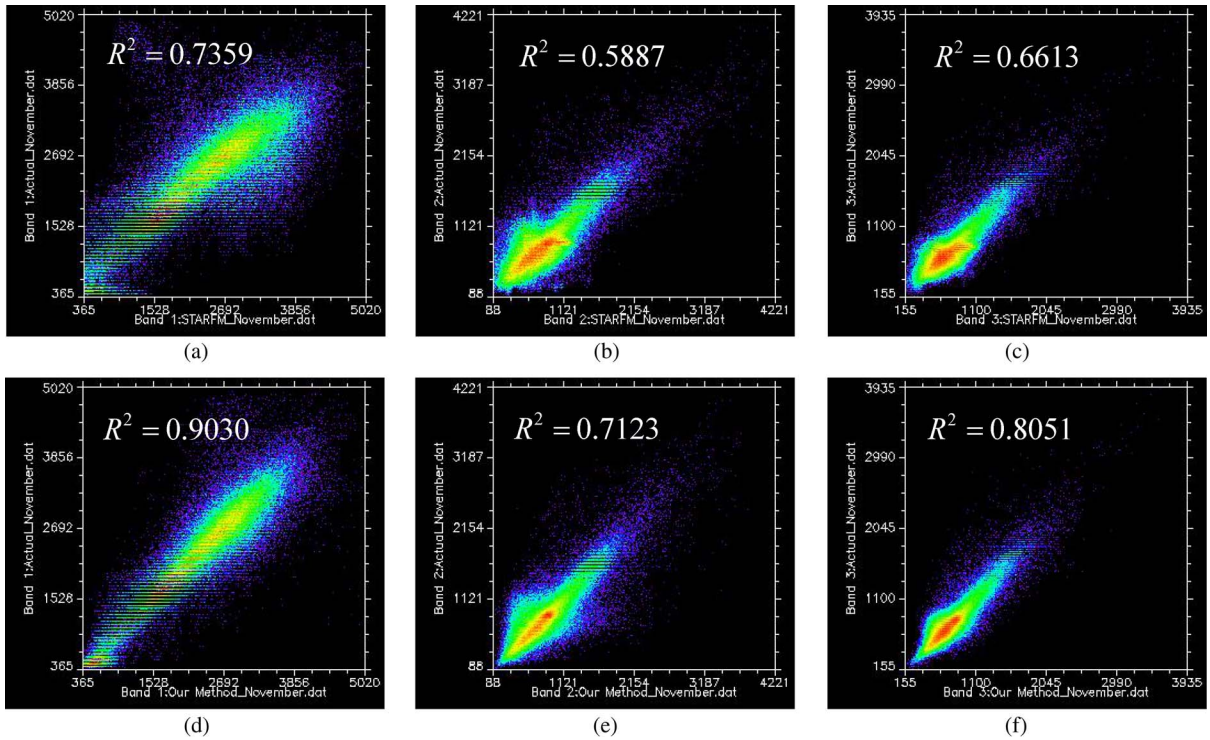

Fig. 6. Scatter plots of predicted against actual reflectances for NIR–red–green bands from left to right. (a)–(c) are the scatter plots of the predicted reflectance by using STARFM against the actual reflectance. (d)–(f) are the scatter plots of the predicted reflectance by using our method against the actual reflectance.

and 0.013, respectively, and the average rmse values of three bands for the two models are 0.029 and 0.020, respectively. These indicate that our method can reconstruct the Landsat surface reflectance more precisely than STARFM. The average SSIM values of the three bands for the two methods are 0.8409 and 0.8632, respectively, and this indicates that our method can retrieve more precise structural details on the surface reflectance than STARFM with smaller reflectance deviations. Also, the ERGAS and SAM values reveal that the spatial details and spectral colors of our fusion result are better than those of STARFM (0.4390 versus 0.3121 and 7.1266 versus 4.4477). The scatter plots with R^2 in Fig. 6 (where the scale factor is 10 000) show the reflectance distribution between the predicted and the actual surface reflectances of the three bands: from left to right, in NIR, red, and green, respectively. Fig. 6(a)–(c) shows the scatter plots of predicted reflectance obtained by using the STARFM method against the observed reflectance, and Fig. 6(d)–(f) shows the scatter plots of predicted reflectance obtained by using our method against the observed reflectance. In Fig. 6, we can see that our method achieves better effects than STARFM with higher R^2 in all three bands.

To further validate the effectiveness of the fusion results, we applied all predicted Landsat-like images in 12 months (i.e., from June 2000 to May 2001) to derive the temporal variation of NDVI. To assure pure pixels on MODIS images, we chose a small piece of farmland [marked by the blue polygon in Fig. 5(a)] with high spatial homogeneity and major temporal changes. This chosen area is located in a square area with 50 $\times$ 50 pixels (1.5 km by 1.5 km) on the Landsat images, and the central 2 $\times$ 2 region of its corresponding area on MODIS images was chosen to compute the NDVIs. For comparisons, the three temporal NDVI profiles obtained from MODIS images, the prediction results of STARFM, and the prediction results of our method are shown respectively in Fig. 7. For NDVI values in September, they were derived from the actual Landsat image that was available. The observed NDVI map in November and the derived NDVI maps from predicted Landsat-like images are shown in Fig. 5(g)–(i), respectively. By comparing the predicted NDVI maps in Fig. 5(h) and (i) with the actual map and the predicted NDVI values in November and August with the observed values in Fig. 7, we can see that our fusion result can retrieve more accurate NDVI than STARFM. By comparing the change rate of temporal NDVI profiles in Fig. 7, it can be easily observed that our fusion result can reflect the variability of NDVI to a larger extent, owing to the improved spatial resolution. Due to the existence

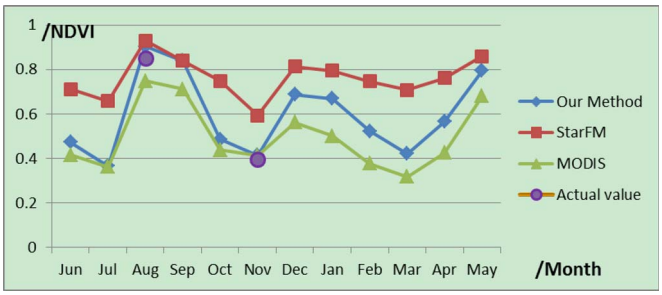

Fig. 7. Temporal NDVI profiles of 12 months computed from (the light green line) the interpolated MODIS images, (the red line) the Landsat-like images predicted from STARFM, and (the blue line) the Landsat-like images predicted from our method. (Purple dot) NDVI values in August and November were obtained from the actual Landsat images and used as the validation values.

of water area around the region of interest [see the black part around the polygon in Fig. 5(a)], the overall NDVI of MODIS is lower than that of the predicted Landsat-like images. On the other hand, because STARFM calculates the central pixel's reflectance on the prediction date based on the weighted sum of the surrounding pixels' reflectance in the reference image, the neighboring pixel values in the reference image have stronger effect on the predicted image values. In this case, due to the high NDVI value in September (i.e., the reference image), the overall NDVI of STARFM fusion results is higher than that of our fusion results.

D. Experiments With Satellite Data—Land-Cover-Type Changes Over Urbanized Area

Land-cover-type changes are another important category in the Earth's surface dynamics which result from disturbances. In order to examine the performance of our algorithm in incorporating land-cover-type changes, we used the Landsat-7 ETM+ and MODIS images of an area of 15 km $\times$ 15 km in Shenzhen, China, to predict a Landsat-like surface reflectance. This data set contains three Landsat–MODIS pairs acquired in the same month but in different years, i.e., on November 1, 2000, November 7, 2002, and November 8, 2004, respectively. We chose bands 2, 3, and 4 for Landsat surface reflectance, all with the spatial resolution of 30 m and the image size of 500 $\times$ 500 pixels. Their corresponding MODIS bands are bands 4, 1, and 2, and the spatial resolutions and image sizes are 500, 250, and 250 m and 30 $\times$ 30, 60 $\times$ 60, and 60 $\times$ 60 pixels, respectively. For convenience of processing, the spatial resolution of the MODIS band with 500 m was upscaled to 250 m by using the bicubic interpolation method. Fig. 8 shows the scenes with an NIR–red–green as red–green–blue composite both for MODIS (upper row) and for Landsat (lower row) surface reflectances. It can be observed that most of the vegetation regions remained similar from 2000 to 2004; however, some vegetation regions were converted into built-up areas or vice versa from 2000 to 2004.

Similar to Section IV-C, we used the image pair on November 1, 2000, and the MODIS image on November 7, 2002, to predict the Landsat surface reflectance on November 7, 2002. The number of dictionary atoms was set empirically in our method as follows: 1000 in the first layer and 2000 in the second layer. In STARFM, the window size was fine-tuned to be 23. In our method, the spatial resolution of MODIS on the prediction date is magnified from 250 to 120 m in the first layer and then to 30 m in the second layer. The predictions using STARFM [Fig. 9(b)] and our method [Fig. 9(c)] are shown in Fig. 9 in comparison with the actual Landsat ETM+ surface reflectance [Fig. 9(a)]. The fusion result of our method under one-layer framework is shown in Fig. 9(d). Compared with the two-layer fusion result, more spectral distortions appear in the area with land-cover-type changes. Fig. 9(e) and (f) shows the average rmse maps of three bands for the fusion results of STARFM and our method, respectively, which show that the prediction errors mainly exist in the area with land-cover-type changes [marked by the blue circle in Fig. 9(a)]. Clearly, we can see that our method has smaller prediction error than STARFM in these areas, which reveals that our proposed method is more robust than STARFM in dealing with land-cover-type changes. However, there are slight deviations at the edges of the change areas in our predicted surface reflectance, and these are mainly caused by the slight geometric mismatch and the large resolution differences between the Landsat ETM+ and MODIS images as well as by the complex change structures in this data set.

The comparisons in terms of AAD, rmse, SSIM, ERGAS, and SAM are listed in the upper two rows of Table III. The average AAD values of the three bands for STARFM and our method are 0.0132 and 0.0107, respectively, and the average rmse values of the three bands for the two models are 0.0208 and 0.0163, respectively. These indicate that our method can reconstruct the Landsat surface reflectance more precisely than STARFM. The average SSIM values of three bands for the two methods are 0.7636 and 0.7799, respectively, which show that our method captured more structural details on the surface reflectance than STARFM. Also, the ERGAS and SAM values indicate that the spatial details and spectral colors of our fusion result are better than those of the STARFM fusion result (1.4687 versus 1.9229 and 2.6645 versus 3.0838). The scatter plots (the scale factor is 10 000) with R^2 of the predicted surface reflectances by using STARFM and our method in the three bands (NIR–red–green) are shown in Fig. 10, respectively. Fig. 10(a)–(c) (upper row) shows scatter plots between the predicted reflectance by using STARFM and the actual reflectance in the NIR, red, and green bands, respectively, whereas Fig. 10(d)–(f) (lower row) shows those between the predicted reflectance by using our method and the actual reflectance in the corresponding three bands. The comparison between the upper and lower rows of the scatter plots in Fig. 10 shows that our method achieves a better fit to the 1 : 1 line with higher R^2 in all three bands, thereby indicating that our method retrieved more precise Landsat surface reflectance with smaller derivations than STARFM. This should be attributed to the rationalities of superresolution of MODIS and the fusion based on high-pass modulation under a two-layer framework, which result in a better characterization and delineation of the type changes.

To further validate the effectiveness of the proposed algorithm, we performed experiments with two-pair prior images and compared the results with SPSTFM [13] and ESTARFM [10]. We used the image pairs on November 1, 2000, and November 8, 2004, and the MODIS image on November 7,

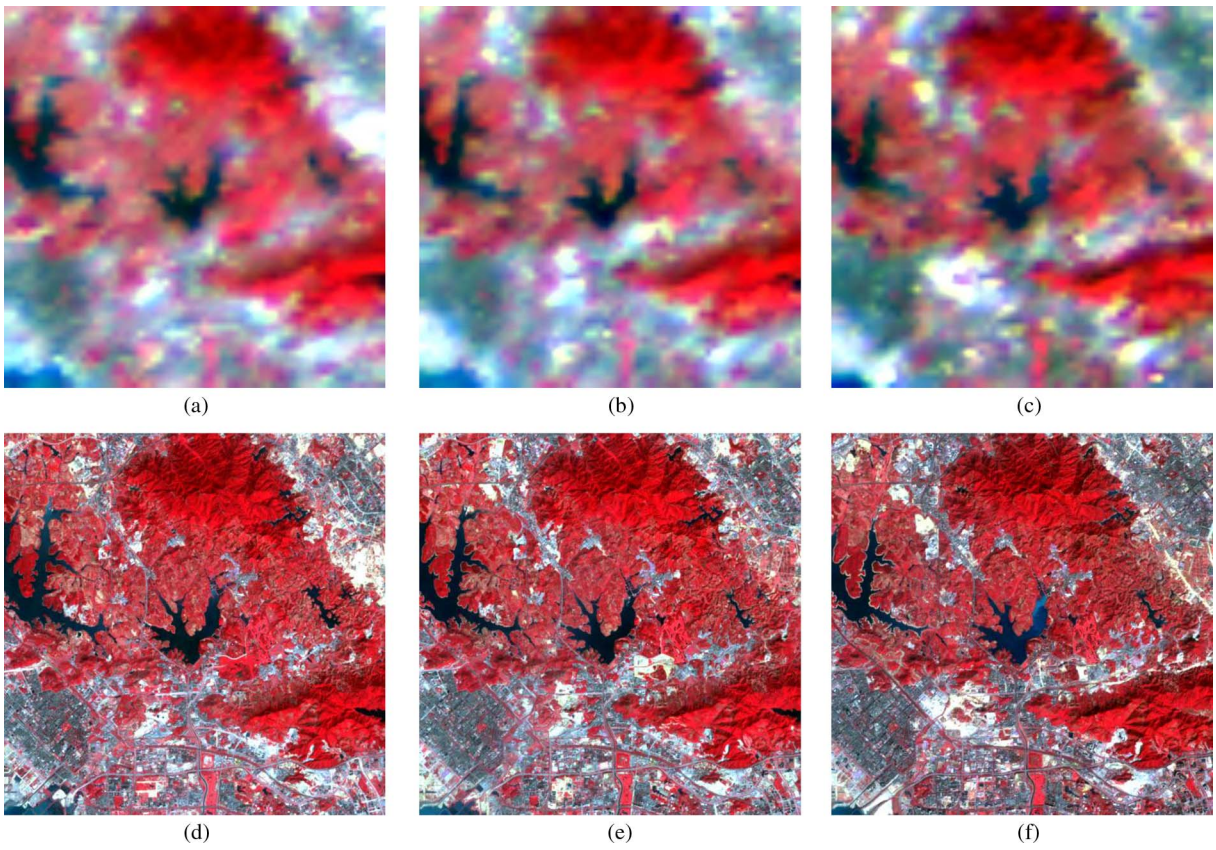

Fig. 8. (Upper row) MODIS composited surface reflectance and (lower row, 500 × 500 pixels) Landsat composited surface reflectance of an area in Shenzhen. From left to right, the MODIS–Landsat image pair was acquired in November 2000, 2002, and 2004, respectively.

2002, to predict the Landsat surface reflectance on November 7, 2002. Our method adopted the same parameter settings as those in the one-pair case and the same weighting strategy as in [13], i.e., during the reconstruction step, the Landsat-like images in 2002 were first predicted from 2000 to 2004, respectively, and then these two predictions were weighted to obtain the final prediction. For the ESTARFM algorithm, the window size was fine-tuned to be 23. The fusion results of ESTARFM, SPSTFM, and our method are shown in Fig. 9(g)–(i), respectively. Clearly, the fusion results of SPSTFM and our method are very similar to the actual Landsat image and are better than ESTARFM both visually and in the detection of change areas (see blue circles). The comparisons in terms of AAD, rmse, SSIM, ERGAS, and SAM are listed in the lower three rows of Table III (due to a better preprocessing of MODIS data, the evaluation results of SPSTFM in this paper are better than those in [13]). Generally, we can observe that the fusion results based on two-pair prior images are better than those based on one-pair prior images due to the use of more prior information. The average AAD values of the three bands for ESTARFM, SPSTFM, and our method are 0.0127, 0.0106, and 0.0099, respectively, and the average rmse values of the three bands for these three methods are 0.0196, 0.0154, and 0.0147, respectively. These indicate that our method can reconstruct the Landsat surface reflectance more precisely than the other two methods; this can be further illustrated through the comparisons on ERGAS and SAM. The average SSIM values of three bands for the three methods are 0.7926, 0.8247, and 0.8098, respectively, which show that SPSTFM captured more structural details on the surface reflectance than ESTARFM and our method. This might be due to the fact that SPSTFM is based on the prediction of difference images; thus, more structural details are accurately predicted. On the other hand, our method is based on the direct prediction of the high-resolution image from its low-resolution counterpart with the aid of high-pass modulation under a two-layer framework; thus, the overall reflectance accuracy is better.

V. Conclusion

Based on the superresolution of high-temporal-resolution data and a high-pass modulation, this paper has presented a new spatiotemporal fusion model. Given that only one pair of LSHT–HSLT data is available in many cases, this fusion model predicts the Landsat version of the corresponding MODIS data on the prediction date. With the assumption that there exists a degradation process between the Landsat and the MODIS data, a correspondence is built between them based on the known Landsat–MODIS pair. Then, the MODIS images are magnified to be transition images, whose spatial resolution is much closer to the Landsat images than the MODIS ones. To further utilize the prior Landsat data, the high-pass modulation is applied to the fusion of the two transition images and the prior Landsat data. Therefore, the final Landsat data are predicted based on the aforementioned superresolution stage and the high-pass modulation stage. Because the Landsat–MODIS correspondence in the first stage and the linear temporal change model in

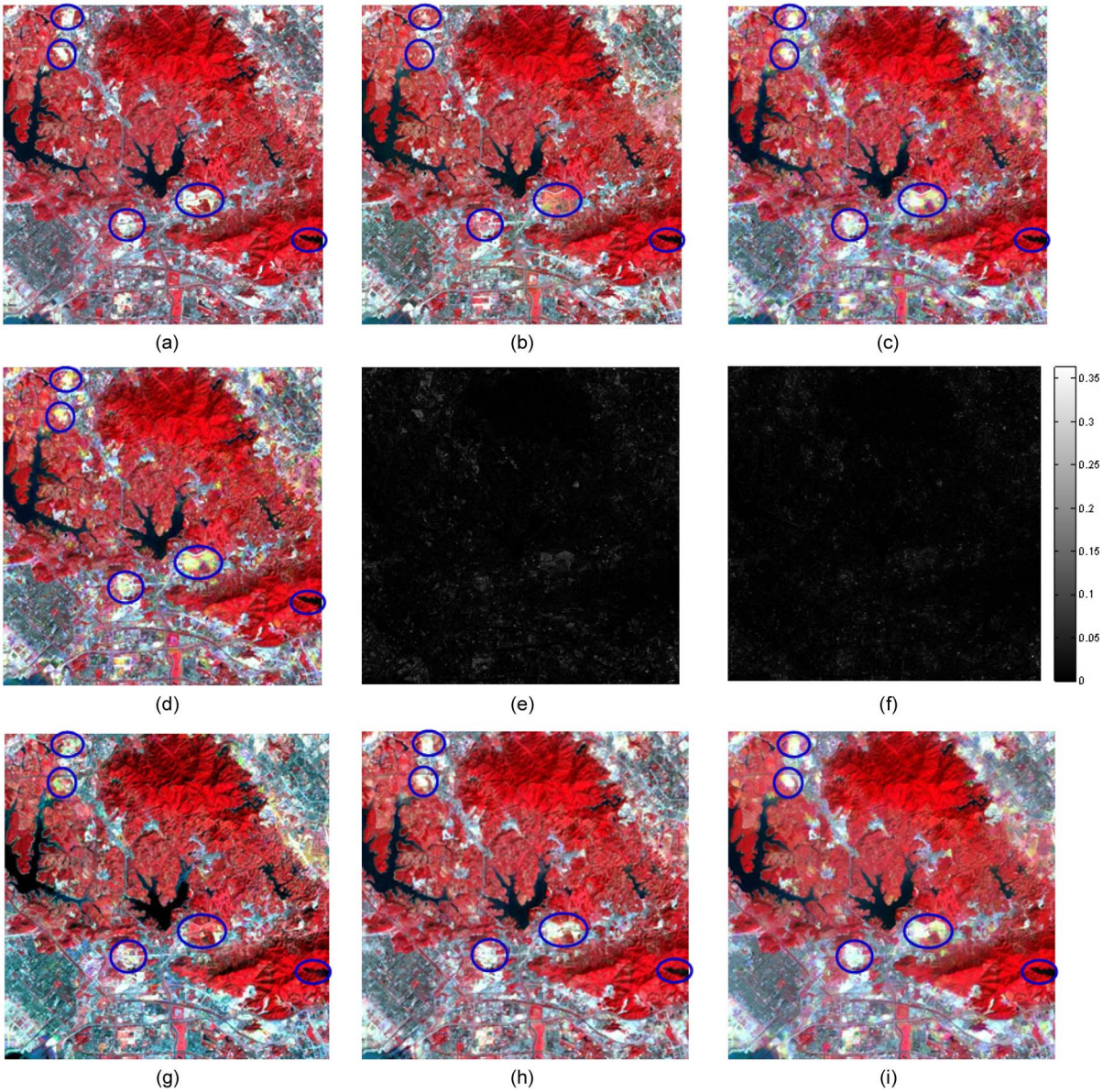

Fig. 9. Comparisons between the actual and the predicted surface reflectances with focus on type changes: (a) Actual surface reflectance in 2002, (b) predicted surface reflectance in 2002 obtained by STARFM, (c) predicted surface reflectance in 2002 obtained by our method, (d) predicted surface reflectance obtained by our method under the one-layer framework, (e) RMSE map of the fusion result obtained by STARFM, (f) RMSE map of the fusion result obtained by our method, and (g)–(i) fusion results with two-pair prior images [(g) predicted surface reflectance in 2002 obtained by ESTARFM, (h) predicted surface reflectance in 2002 obtained by SPSTFM, and (i) predicted surface reflectance in 2002 obtained by our method].

the second stage apply to both data with phenology changes and data with land-cover-type changes, this spatiotemporal fusion algorithm applies to both categories of remote sensing data. Compared to the classic STARFM algorithm and its extended works, the major advantage of our method is that it improves the change delineation accuracy in the prediction image by increasing the spatial resolution of MODIS data. Certainly, our method takes more computational time due to dictionary training.

It is worth mentioning that, although our derived algorithm is based on MODIS and Landsat sensors, it is also applicable to other sensors where the parameter setting and the imaging environment of two sensors are very similar. To ensure that the trained dictionary pair in our algorithm could be applied to the prediction date, it should be confirmed that similarities exist between high-frequency information space of remote sensing data at the prior date and that at the prediction date. For example, in seasonal data which consist mainly of phenology changes, the overall structures show almost no change, whereas in other data which consist mainly of land-cover-type changes, the land-cover types on the prediction date can be found on the prior date and the study area is sufficiently large.

Although our algorithm is developed under the rigorous assumption that only one pair of LSHT–HSLT data is available, this spatiotemporal fusion model can be applied to other application contexts. For instance, if two pairs of Landsat–MODIS data are available, as also illustrated in our experiments, then the dictionary pair could be trained on these two pairs of data. Similarly, as in ESTARFM, a Landsat-like image can be predicted from both known time points, and so, the final prediction could be obtained by the temporal weighting of these

TABLE III
QUANTITATIVE COMPARISONS OF OUR METHOD WITH STARFM UNDER ONE-PAIR PRIOR IMAGES AND WITH ESTARFM AND SPSTFM UNDER TWO-PAIR PRIOR IMAGES, RESPECTIVELY, ON THE DATA PRIMARILY WITH TYPE CHANGE

Method	AAD			RMSE			SSIM			ERGAS	SAM
	Green	Red	NIR	Green	Red	NIR	Green	Red	NIR		
STARFM	0.0097	0.0145	0.0153	0.0157	0.0237	0.0231	0.7783	0.7204	0.7920	1.9229	3.0838
Our method[a]	**0.0083**	**0.0108**	**0.0129**	**0.0132**	**0.0171**	**0.0187**	**0.7897**	**0.7447**	**0.8053**	**1.4687**	**2.6645**
ESTARFM	0.0093	0.0138	0.0151	0.0153	0.0213	0.0223	0.8002	0.7653	0.8122	1.3460	3.0187
SPSTFM	0.0078	0.0110	0.0129	0.0121	0.0163	0.0179	**0.8435**	**0.7849**	**0.8457**	1.0997	2.4439
Our method[b]	**0.0076**	**0.0099**	**0.0122**	**0.0117**	**0.0152**	**0.0171**	0.8250	0.7746	0.8297	**0.5084**	**2.4318**

a. The prediction result of our method based on one-pair prior images
b. The prediction result of our method based on two-pair prior images

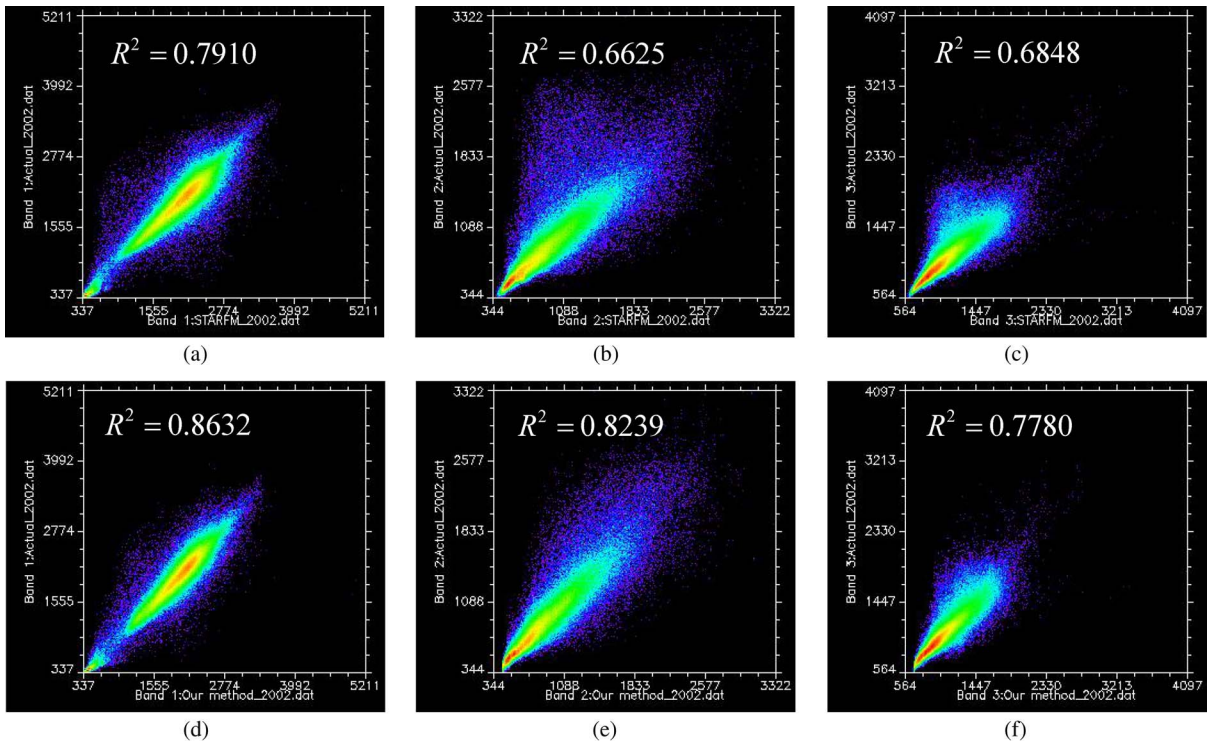

Fig. 10. Scatter plots of the predicted reflectance against the actual reflectance in 2002. (a)–(c) are scatter plots of the predicted reflectance by using STARFM against the actual reflectance in the NIR, red, and green bands, respectively. (d)–(f) are scatter plots of the predicted reflectance by using our method against the actual reflectance in the same three bands.

two predictions. In an expanded perspective, when one pair of LSHT–HSLT data and a series of LSHT data are available, our spatiotemporal fusion algorithm could be employed to generate a series of HSHT data.

ACKNOWLEDGMENT

The authors would like to thank Dr. Y. Xu for providing the Guangzhou data set.

REFERENCES

[1] C. E. Woodcock and M. Ozdogan, "Trends in land cover mapping and monitoring," in *Land Change Science*, Gutman, Ed. New York: Springer-Verlag, 2004, pp. 367–377.
[2] J. E. Vogelmann, S. M. Howard, L. Yang, C. R. Larson, B. K. Wylie, and J. N. Van Driel, "Completion of the 1990's National Land Cover Data Set for the conterminous United States from Landsat Thematic Mapper data and ancillary data sources," *Photogramm. Eng. Remote Sens.*, vol. 67, no. 6, pp. 650–662, Jun. 2001.
[3] J. G. Masek, C. Q. Huang, R. Wolfe, W. Cohen, F. Hall, J. Kutler, and P. Nelson, "North American forest disturbance mapped from a decadal Landsat record," *Remote Sens. Environ.*, vol. 112, pp. 2914–2926, 2008.
[4] J. C. Price, "How unique are spectral signatures?," *Remote Sens. Environ.*, vol. 49, no. 3, pp. 181–186, Sep. 1994.
[5] V. P. Shah, N. H. Younan, and R. L. King, "An efficient pan-sharpening method via a combined adaptive PCA approach and contourlets," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1323–1335, May 2008.
[6] S. Li and B. Yang, "A new pan-sharpening method using a compressed sensing technique," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 2, pp. 738–746, Feb. 2011.
[7] Y. Zhang, "Understanding image fusion," *Photogramm. Eng. Remote Sens.*, vol. 70, no. 6, pp. 657–661, 2004.
[8] C. Thomas, T. Ranchin, L. Wald, and J. Chanussot, "Synthesis of multispectral images to high spatial resolution: A critical review of fusion methods based on remote sensing physics," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1301–1312, May 2008.
[9] 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.
[10] X. Zhu, J. Chen, F. Gao, X. H. 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.
[11] T. Hilker, M. A. Wulder, N. C. Coops, J. Linke, J. McDermid, J. G. Masek, F. Gao, and J. C. White, "A new data fusion model for high spatial- and temporal-resolution mapping of forest based on Landsat and MODIS," *Remote Sens. Environ.*, vol. 113, no. 8, pp. 1613–1627, Aug. 2009.
[12] R. Zurita-Milla, G. Kaiser, J. G. P. W. Clevers, W. Schneider, and M. E. Schaepman, "Downscaling time series of MERIS full resolution data to monitor vegetation seasonal dynamics," *Environ. Monitoring Assessment*, vol. 113, no. 9, pp. 1874–1885, Sep. 2009.
[13] 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.
[14] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.
[15] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proc. Curves Surf.*, Avignon, France, Jun. 24–30, 2010, pp. 711–730.
[16] M. Elad, M. A. T. Figueiredo, and Y. Ma, "On the role of sparse and redundant representations in image processing," *Proc. IEEE*, vol. 98, no. 6, pp. 972–982, Jun. 2010.
[17] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proc. IEEE*, vol. 98, no. 6, pp. 1045–1057, Jun. 2010.
[18] W. Dong, L. Zhang, G. Shi, and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 1838–1857, Jul. 2011.
[19] M. Aharon, M. Elad, and A. M. Bruckstein, "The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.
[20] G. Davis, S. Mallat, and M. Avellaneda, "Adaptive greedy approximations," *Constructive Approx.*, vol. 13, no. 1, pp. 57–98, Mar. 1997.
[21] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Rev.*, vol. 43, no. 1, pp. 129–159, 2001.
[22] I. Gorodnitsky and B. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, Mar. 1997.
[23] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification using dictionary-based sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3973–3985, Oct. 2011.
[24] S. Foucher, "SAR image filtering via learned dictionaries and sparse representations," in *Proc. IEEE IGARSS*, 2008, vol. 1, pp. 229–232.
[25] M. M. Khan, L. Alparone, and J. Chanussot, "Pansharpening quality assessment using the modulation transfer functions of instruments," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 11, pp. 3880–3891, Nov. 2009.
[26] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[27] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. 6th Int. Conf. Comput. Vis.*, Jan. 1998, pp. 839–846.
[28] R. H. Yuhas, A. F. H. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm," in *Proc. Summaries 3rd Annu. JPL Airborne Geosci. Workshop*, 1992, pp. 147–149.

Huihui Song received the B.S. degree in technology and science of electronic information from the Ocean University of China, Qingdao, China, in 2008 and the M.S. degree in communication and information system from the University of Science and Technology of China, Hefei, China, in 2011. She is currently working toward the Ph.D. degree in the Department of Geography and Resource Management, The Chinese University of Hong Kong, Shatin, Hong Kong.

Her research interests include remote sensing image processing and image fusion.

Bo Huang (A'12) received the Ph.D. degree in remote sensing and mapping from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China, in 1997.

He is currently a Professor with the Department of Geography and Resource Management, The Chinese University of Hong Kong, Shatin, Hong Kong, where he is also the Associate Director of the Institute of Space and Earth Information Science. Prior to this, he held faculty positions with the Schulich School of Engineering, University of Calgary, Calgary, AB, Canada, from 2004 to 2006, and the Department of Civil and Environmental Engineering, National University of Singapore, Singapore, from 2001 to 2004. He serves as the Executive Editor of *Annals of GIS* and on the editorial boards of several international journals, including the *International Journal of Geographical Information Science* and the *Journal of Remote Sensing*. His research interests are broad, covering most aspects of geoinformation science, specifically spatiotemporal image fusion for environmental monitoring, spatial/spatiotemporal statistics for land-cover/land-use change modeling, and multiobjective spatial optimization for sustainable land-use planning. He is currently exploring along the line of precision remote sensing, convinced that this new paradigm will revolutionize the way how multisensor remotely sensed data are fused and exploited to improve the performance and quality of various applications in the future.