Spatiotemporal Satellite Image Fusion Using Deep Convolutional Neural Networks

Article *in* IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing · February 2018

DOI: 10.1109/JSTARS.2018.2797894

5 authors, including:

Huihui Song

53 PUBLICATIONS **3,940** CITATIONS

SEE PROFILE

Guojie Wang

Nanjing University of Information Science & Technology

142 PUBLICATIONS **3,967** CITATIONS

SEE PROFILE

Renlong Hang

Nanjing University of Information Science & Technology

58 PUBLICATIONS **3,333** CITATIONS

SEE PROFILE

Bo Huang

The University of Hong Kong

355 PUBLICATIONS **13,679** CITATIONS

SEE PROFILE

Spatiotemporal Satellite Image Fusion Using Deep Convolutional Neural Networks

Huihui Song, Qingshan Liu, Guojie Wang, Renlong Hang, and Bo Huang

***Abstract*—We propose a novel spatiotemporal fusion method based on deep convolutional neural networks (CNNs) under the application background of massive remote sensing data. In the training stage, we build two five-layer CNNs to deal with the problems of complicated correspondence and large spatial resolution gaps between MODIS and Landsat images. Specifically, we first learn a nonlinear mapping CNN between MODIS and low-spatial-resolution (LSR) Landsat images and then learn a super-resolution CNN between LSR Landsat and original Landsat images. In the prediction stage, instead of directly taking the outputs of CNNs as the fusion result, we design a fusion model consisting of high-pass modulation and a weighting strategy to make full use of the information in prior images. Specifically, we first map the input MODIS images to transitional images via the learned nonlinear mapping CNN and further improve the transitional images to LSR Landsat images via the fusion model; then, via the learned SR CNN, the LSR Landsat images are supersolved to transitional images, which are further improved to Landsat images via the fusion model. Compared with the previous learning-based fusion methods, mainly referring to the sparse-representation-based methods, our CNNs-based spatiotemporal method has the following advantages: 1) automatically extracting effective image features; 2) learning an end-to-end mapping between MODIS and LSR Landsat images; and 3) generating more favorable fusion results. To examine the performance of the proposed fusion method, we conduct experiments on two representative Landsat–MODIS datasets by comparing with the sparse-representation-based spatiotemporal fusion model. The quantitative evaluations on all possibleprediction dates and the comparison of fusion results on one key date in both visual effect and quantitative evaluationsdemonstrate that the proposed method can generate more accurate fusionresults.**

***Index Terms*—Convolutional neural network (CNN), nonlinear mapping (NLM), spatial resolution, temporal resolution.**

Manuscript received May 28, 2017; revised October 3, 2017; accepted January 18, 2018. This work was supported in part by the Natural Science Foundation of China under Grant 41501377, Grant 61532009, and Grant 91546117, in part by the Foundation of Jiangsu Province of China under Grant BK20150906 and Grant 15KJA520001, in part by the Key Project of National Social and Scientific Fund Program under Grant 16ZDA047, and in part by the HKRGC General Research Fund under Grant 14606315. *(Corresponding author: Huihui Song.)*

H. Song, Q. Liu, and R. Hang are with the Jiangsu Key Laboratory of Big Data Analysis Technology, Jiangsu Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: hhsongsherry@gmail.com; qsliu@nuist.edu.cn; renlong_hang@163.com).

G. Wang is with the Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: gwang_nuist@163.com).

B. Huang is with the Chinese University of Hong Kong, Hong Kong (e-mail: hhsongsherry@gmail.com).

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/JSTARS.2018.2797894

I. Introduction

With the intensification of global change, the satellite remote sensing technology has received increasing attention over the past decades. For example, remote sensing data play an indispensable role in crop growth monitoring, the change detection of land use/cover, and disaster monitoring. Because continuous observations are basic requirement for dynamic monitoring, high temporal resolution turns into an important property of the demanded remote sensing data in application fields of dynamic monitoring [1]. At the same time, the fragmentation of global terrestrial landscape makes these applications demand remote sensing data with higher spatial resolutions [2]. However, remote sensing data with both high spatial and temporal resolutions are difficult to capture by current satellite platforms due to the constraints in technology and budget. For example, the spatial resolutions of remote sensing images from satellites of the Landsat series, SPOT, and IRS range from 6 to 30 m, which are suitable for applications of land use/cover mapping, change detection [3], and the dynamic monitoring of ecosystems [4]. However, the long revisit circles of these satellites (Landsat/TM: 16 days, SPOT/HRV: 26 days, IRS: 24 days) together with frequent cloud cover and other poor atmospheric conditions limit their applications in detecting rapid changes of land surface (e.g., the intraseasonal disturbances of ecosystems and the phonology change). On the other hand, the sensors of MODIS on the Terra/Aqua, SPOT-VGT and NOAA/AVHRR are capable of providing daily remote sensing observations. However, the spatial resolutions of these sensors range between 250 and 1000 m, which are inadequate to monitor the changes of land covers and ecosystems with high heterogeneity. To exploit the full potential of current remote sensing data in dynamic monitoring of land surface, spatiotemporal fusion of remote sensing data is a feasible and cost-effective way.

Spatiotemporal fusion aims to integrate two types of remote sensing data with similar spectral information, including the number of bands and the bandwidths. One type is featured by high spatial resolution but low temporal resolution (HSLT) and the other type by low spatial resolution but high temporal resolution (LSHT). Given one pair or two pairs of HSLT–LSHT images on prior dates and one or more LSHT images on prediction dates, spatiotemporal fusion models integrate these images to produce high-spatial-resolution images on prediction dates. Besides, spatiotemporal fusion can be implemented under unified fusion frameworks [5], [6]. Gao *et al.* [7] proposed the first spatiotemporal fusion model, named the spatial and temporal adaptive reflectance fusion model (STARFM), to predict

daily Landsat surface reflectance by blending the Landsat and MODIS surface reflectance. Since then, some spatiotemporal fusion methods have been proposed. Generally, we classify these methods into three categories: 1) filter based; 2) unmixing based; and 3) learning based.

In the filter-based methods [7]–[12], each pixel is predicted from a filtering model, which is a weighted sum of spectrally similar neighboring pixels from the input images. Generally, these methods are different mainly in the ways of modeling the relationship between HSLT and LSHT pixels, selecting similar neighboring pixels and calculating the weights. The classic STARFM builds a simple approximate relationship between HSLT and LSHT pixels and searches similar neighboring pixels according to the spectral difference, the temporal difference, and the location distance. Based on STARFM, several improved algorithms have since been developed. Considering the existence of disturbance events or land-cover-type changes not recorded in at least one Landsat image, Hilker *et al.* [8] proposed a spatial and temporal adaptive algorithm for mapping reflectance change, which detects the temporal changes from a dense series of spatially coincident MODIS images. To enhance the predictions in heterogeneous landscapes, Zhu *et al.* [9] developed an enhanced STARFM by assigning different conversion coefficients for homogeneous and heterogeneous pixels. To improve the way of calculating weights, Shen *et al.* [10] considered the sensor observation differences of varied land cover types by referring to a prior high-spatial-resolution classification map.

In the unmixing-based methods, there are usually the following three steps:

1) unsupervised classification of HSLT images;
2) spectral unmixing of LSHT images via a linear model by using the proportion information obtained in the first step; and
3) generation of the fused image through replacing the spectral information of HTLT image by the unmixed spectral information of LSHT image.

Zurita-Milla *et al.* [13] proposed an unmixing-based spatial and spectral fusion method to produce the synthetic images with the spatial information of Landsat/TM data and the spectral information of medium-resolution imaging spectrometer (MERIS) data. Based on this paper, they further developed a spatiotemporal fusion method in [14] to generate images with the spatial resolution of Landsat/TM and the spectral and temporal resolutions of MERIS.

Based on sparse representation theory, learning-based methods [15]–[18] were proposed in recent years. Given two Landsat–MODIS image pairs, Huang and Song [15] proposed to build a correspondence between the different images of Landsat and MODIS via learning a dictionary pair and then predict the fused image through weighting the predictions from the two end dates of the observation period. To deal with the one Landsat–MODIS image-pair case, they further developed a spatiotemporal fusion framework [16] by designing a two-layer fusion framework including super-resolution (SR) and high-pass-modulation steps. Although with higher complexity, learning-based methods are expected to perform better than the other two categories of methods by exploring more information from the prior data. However, there are several limitations in the previous sparse-representation-based methods. First, the image features need to be designed manually, which brings the complexity into algorithms and the instability into performances. Second, during algorithm implementations, the steps of feature extraction, dictionary learning, sparse coding, and image reconstruction are carried out separately. Third, the algorithms were developed and verified for small-scale study areas and did not consider the massive remote sensing data in reality.

To solve the above three problems, we propose a novel spatiotemporal fusion model using deep convolutional neural networks (CNNs). As one kind of typical deep learning technique, the CNNs have achieved superior performance in object recognition, object detection, image classification, image denoising, image SR, etc., [19]–[22]. For example, Dong *et al.* [20] proposed a deep-convolutional-network-based SR method, which directly takes the low-resolution images and the corresponding high-resolution images as the input and output, respectively. By consisting of three convolutional layers, i.e., patch extraction and representation, nonlinear mapping (NLM), and reconstruction, this method was formulated with a simple network structure, and yet provided superior performance. These successes of CNNs can be attributed to three aspects:

1) the deep architecture of CNNs is more effective in capturing the features of large scale images;
2) the proposal of fast and efficient training methods, such as the rectified linear unit (ReLU) [23], batch normalization [24], residual learning [25], etc.; and
3) the training speed is significantly accelerated via the parallel computation on powerful GPUs.

Specifically, the proposed model consists of three steps: NLM, SR, and image fusion, where the first two steps are achieved via a five-layer CNN. Compared with the shallow sparse representation models in [15] and [16], this deep convolutional network model can extract and express the rich information in massive remote sensing data more effectively via constructing the representation model with multiple hidden layers. Besides, the networks learn the image features automatically from the data themselves and optimize the steps of feature extraction, NLM, and image reconstruction jointly via backpropagation.

The remaining of this paper is organized as follows. In Section II, we introduce the proposed method in detail. The experimental results are shown in Section III, and Section IV concludes this paper with some discussions.

II. Methodology

As usually did in the previous spatiotemporal fusion models [7]–[10], [15], [16], we select Landsat Enhanced Thematic Mapper Plus (ETM+) and MODIS images as the examples of HSLT and LSHT data, respectively, to demonstrate the proposed fusion method. Since there are no constraints on the changes of land cover types or the proportion of each land cover type, the proposed method is capable of dealing with both phenology and land-cover-type changes. Although the proposed method is able to handle the case with one pair of prior images, we assume there

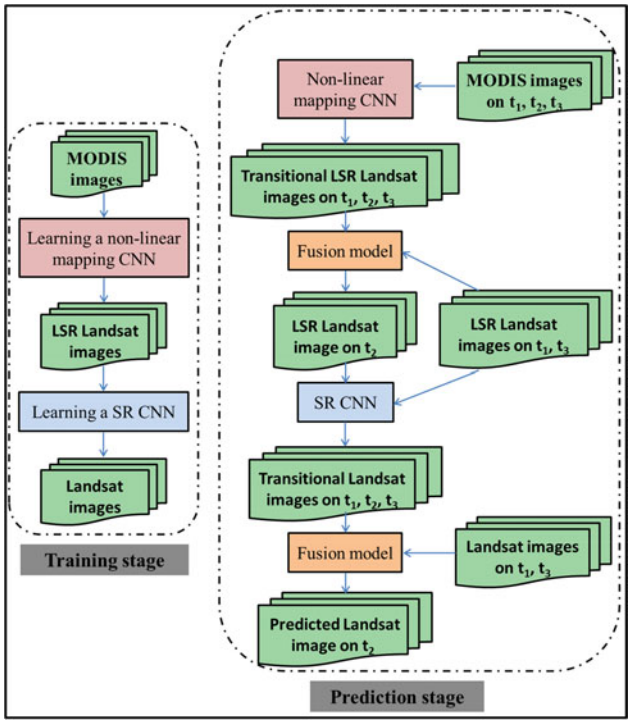


Fig. 1. Flowchart of the proposed method.

are two pairs of prior Landsat–MODIS images considering the application background of massive remote sensing data.

The flowchart of the proposed fusion method is demonstrated in Fig. 1. Generally, the method consists of training stage and prediction stage. Since the training and prediction are implemented separately on all bands in the same way, we take Landsat and MODIS images on one band (e.g., red band) to illustrate the proposed method. In the training stage, we first learn a NLM model between MODIS (500 m) and low-spatial-resolution (LSR) Landsat (250 m) images and then learn an SR model between LSR and original Landsat images. The former step is inspired by the facts that Landsat and MODIS data come from different sensors and undergo different atmospheric and geometric corrections and that there are large spatial resolution gaps between them (25 m versus 500 m). Motivated by the successful applications of CNNs in image classification and image SR, we learn an NLM CNN and an SR CNN, respectively. In the prediction stage, a fusion model consisting of high-pass modulation and weighting is adopted to make full use of the available information. Specifically, the MODIS images in prior and prediction dates are fed into the learned NLM CNN to obtain the transitional LSR Landsat images, which are then fed into the fusion model to get the LSR Landsat image on prediction date; together with the simulated LSR Landsat images on prior dates, the fusion result in the last step is fed into the learned SR CNN to obtain the transitional Landsat images, which are then fed into the fusion model to get the final fusion result on the prediction date.

A. Training Stage

1) Nonlinear Mapping CNN: To build an NLM model between MODIS and Landsat images, we first downsample the spatial resolution of Landsat images to be similar with that of MODIS images. To narrow the resolution gap in the next SR step, the spatial resolution of Landsat images is reduced by ten times. Denote the training samples of MODIS and LSR Landsat images as $\mathbf{X}$ and $\mathbf{Y}^l$, respectively, we expect to learn an NLM function $F^M(\cdot)$ such that $F^M(\mathbf{X})$ approximates $\mathbf{Y}^l$. Considering that $\mathbf{X}$ and $\mathbf{Y}^l$ are largely similar, we learn the residual image between $\mathbf{X}$ and $\mathbf{Y}^l$ from $\mathbf{X}$, so that we can focus on learning the high-frequency details of $\mathbf{Y}^l$. The residual image hereof is defined as $\mathbf{R} = \mathbf{Y}^l - \mathbf{X}$. Thus, we expect to learn a mapping function $F^M(\cdot)$ such that $F^M(\mathbf{X})$ approximates $\mathbf{R}$. After predicting the residual image, the ground truth LSR Landsat image is obtained by the sum of the input MODIS image and the residual image. The recent works in [26] and [27] have demonstrated the effectiveness of residual learning in image denoising and image SR. As demonstrated in Fig. 2, the structure of the NLM CNN consists of five layers, the input layer, three convolutional hidden layers, and the output layer, where three hidden layers correspond to three operations: feature extraction; NLM; and reconstruction. Next, we describe these three operations in detail.

To extract the features of the input MODIS images, we apply n_1 filters with kernel size of $k_1 \times k_1$ to each of them. To speed up the convergence of the network while ensuring the accuracy, the rectified linear units [ReLU, $\max(0, \cdot)$] are adopted for nonlinearity in the filter responses. Then, n_1 MODIS feature maps for each input image are obtained from the first hidden layer. This step can be expressed by the following equation:

$$F_1^M(\mathbf{X}) = \max\left(0, \mathbf{W}_1 * \mathbf{X} + \mathbf{b}_1\right) \tag{1}$$

where $\mathbf{W}_1 \in \mathbb{R}^{k_1 \times k_1 \times n_1}$ denotes the weights of filters, $\mathbf{b}_1 \in \mathbb{R}^{n_1}$ is a biases vector corresponding to all filters, and $*$ denotes the convolution operation. Compared with the sparse representation methods in [15] and [16], which explicitly extract the image features to obtain the dictionary atoms, the CNN automatically extracts the most effective features via optimizing $\mathbf{W}_1$ and $\mathbf{b}_1$.

To establish correspondences between MODIS and LSR Landsat images, we map the extracted MODIS features onto the residual image features. To this end, we apply n_2 filters with kernel size of $k_2 \times k_2$ to the MODIS feature maps. After the nonlinearity conversion of ReLU, n_2 feature maps of residual image are obtained from the second hidden layer. This step is expressed mathematically as follows:

$$F_2^M(\mathbf{X}) = \max\left(0, \mathbf{W}_2 * F_1^M(\mathbf{X}) + \mathbf{b}_2\right) \tag{2}$$

where $\mathbf{W}_2 \in \mathbb{R}^{n_1 \times k_2 \times k_2 \times n_2}$ denotes the weights of filters and $\mathbf{b}_2 \in \mathbb{R}^{n_2}$ is a biases vector corresponding to all filters.

In above feature extractions, the filter kernels convolving with the images are equivalent to extracting features in form of patches. To ensure the continuity among patches, the filters extract image patches with overlapping. Thus, the mapped feature patches of residual images are also overlapped. In sparse-representation-based spatiotemporal fusion methods [15], [16], the overlapped prediction patches are processed by taking the averages of them, which is just an approximate estimate and cannot guarantee the optimal results. In order to improve the

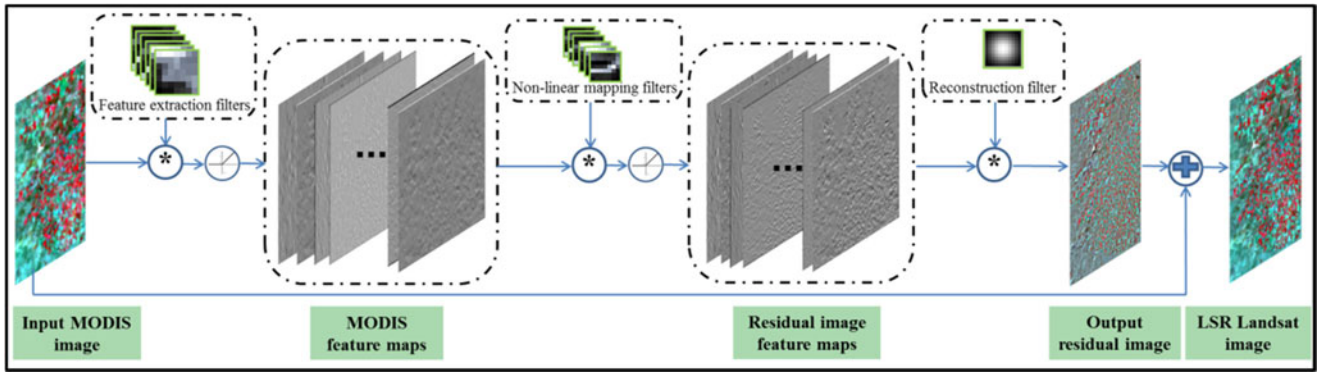


Fig. 2. Structure of NLM CNN.

prediction accuracy of the overlapped areas, we apply a filter to the overlapping areas to get the prediction values. This means we can add a convolutional layer to reconstruct the residual images. Suppose the size of this reconstruction filter is $k_3 \times k_3$, the formula of this step is

$$F^M(\mathbf{X}) = \mathbf{W}_3 * F_3^M(\mathbf{X}) + b_3 \tag{3}$$

where $\mathbf{W}_3 \in \mathbb{R}^{n_2 \times k_3 \times k_3 \times 1}$ denotes the weights of filters and b_3 is the filter bias. Via optimizing $\mathbf{W}_3$ and b_3, we can obtain the optimal predictions for the overlapped areas.

Combining (1)–(3), it can be observed that F^M is a function of network parameters $\Theta = \{\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3, \mathbf{b}_1, \mathbf{b}_2, b_3\}$. To solve Θ, we minimize the loss between the reconstructed residual images $F^M(\mathbf{X}; \Theta)$ and the corresponding ground truth residual images $\mathbf{R}$. Given N MODIS and LSR Landsat training samples $\{(\mathbf{X}_i, \mathbf{Y}_i^l)\}_{i=1}^N$, the mean squared error (MSE) is adopted as the loss function as

$$L(\Theta) = \frac{1}{N}\sum_{i=1}^{N} \|F^M(\mathbf{X}_i; \Theta) - (\mathbf{Y}_i^l - \mathbf{X}_i)\|^2. \tag{4}$$

This loss is minimized by adopting stochastic gradient descent with the standard backpropagation [28]. We empirically set the learning rate to 10^{-4} for the first two hidden layers and 10^{-5} for the last hidden layer.

2) Super-Resolution CNN: To correlate the LSR Landsat images and original Landsat images, we learn a SR CNN model between them. Although with spatial resolution gap of ten times (250 m versus 25 m), we build a five-layer network similar to NLM CNN due to the simpler degradation procedure between them than that between MODIS and LSR Landsat images. Similar to NLM CNN, we learn a residual network considering the large similarity between LSR Landsat and original Landsat images. Denote the training samples of LSR Landsat and original Landsat images as $\mathbf{Y}^l$ and $\mathbf{Y}$, respectively, we expect to learn a mapping function $F^S(\cdot)$ such that $F^S(\mathbf{Y}^l)$ approximates their residual images $\mathbf{R}^s$, where $\mathbf{R}^s$ is defined as $\mathbf{R}^s = \mathbf{Y} - \mathbf{Y}^l$. The structure of this SR CNN consists of five layers, the input LSR Landsat images, three convolutional hidden layers, and the output residual images, where three hidden layers correspond to three operations, feature extraction, NLM, and reconstruction.

Denote the network parameters of three convolutional layers as $\Theta' = \{\mathbf{W}_1', \mathbf{W}_2', \mathbf{W}_3', \mathbf{b}_1', \mathbf{b}_2', b_3'\}$, we aim to solve Θ' by minimizing the loss between the reconstructed residual images $F^S(\mathbf{Y}^l; \Theta')$ and the corresponding ground truth residual images $\mathbf{R}^s$. Given M LSR Landsat and original Landsat training samples $\{(\mathbf{Y}_i^l; \mathbf{Y}_i)\}_{i=1}^M$, the loss function based on MSE is expressed as follows:

$$L(\Theta') = \frac{1}{M}\sum_{i=1}^{M} \|F^S(\mathbf{Y}_i^l; \Theta') - (\mathbf{Y}_i - \mathbf{Y}_i^l)\|^2. \tag{5}$$

We minimize this loss by using stochastic gradient descent with the standard backpropagation [28]. Similar to the training of NLM CNN, the learning rate of the first two hidden layers is set to 10^{-4} and the last hidden layer to 10^{-5}.

B. Prediction Stage

Given two pairs of prior Landsat–MODIS images and the MODIS image on prediction date, we aim to fuse them to predict the Landsat-like image on prediction date. Denote the prior dates as t_1 and t_3, the prediction date as t_2, and the corresponding MODIS and Landsat images as $\mathbf{M}_i$ and $\mathbf{L}_i$ (i = 1, 2, 3), respectively, we predict $\mathbf{L}_2$ based on the learned NLM CNN and SR CNN. For the input MODIS images, we first map them to the LSR Landsat images via the NLM CNN and further super-resolve the LSR Landsat images to Landsat images via the SR CNN. However, it is difficult to build accurate correspondences between MODIS and LSR Landsat images due to the effects of atmosphere, weather, terrain, and many other complex factors during capturing remote sensing images. At the same time, accurate reconstruction from LSR Landsat images to Landsat images is hard because of the existence of large spatial resolution gap. Therefore, we define the predicted images from NLM CNN and SR CNN as *transitional images* and further improve them by utilizing available information via a fusion model.

The implementation procedures of fusion model from transitional images to LSR Landsat images are demonstrated in Fig. 3. With the learned NLM CNN, we first map the input MODIS images $\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3$ to transitional images, which are denoted as $\mathbf{T}_i^l$ (i = 1, 2, 3). Since the spatial resolutions of transitional images and LSR Landsat images are very similar, the temporal change information of transitional images from t_1

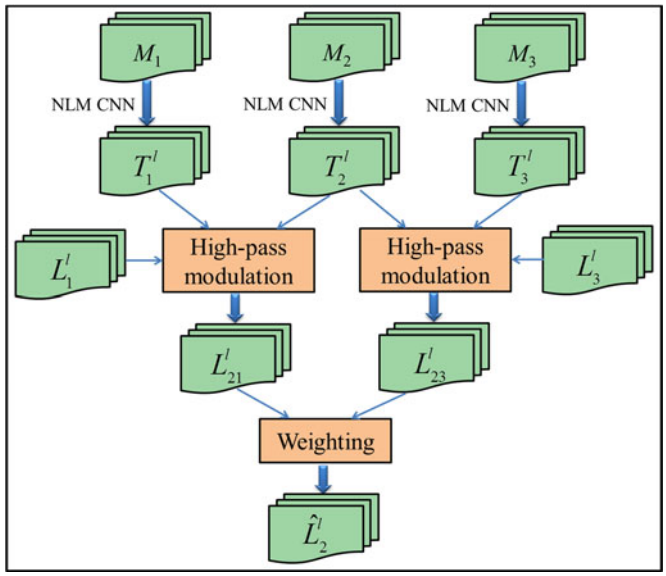


Fig. 3. Illustration of the fusion model.

to t_2 is utilized to predict the LSR Landsat image at t_2 by applying high-pass modulation [29]. Denote the LSR Landsat images downsampled from Landsat image via bicubic interpolation as $\mathbf{L}_i^l$ $(i = 1, 3)$, the fusion of transitional images and the LSR Landsat image from t_1 end is achieved through the following high-pass modulation equation:

$$\mathbf{L}_{21}^l = \mathbf{L}_1^l \left(\frac{\mathbf{T}_2^l}{\mathbf{T}_1^l} \right). \tag{6}$$

The operations in (6) are in pixel level. Similarly, this high-pass modulation can also be applied to t_3 end and get another prediction as follows:

$$\mathbf{L}_{23}^l = \mathbf{L}_3^l \left(\frac{\mathbf{T}_2^l}{\mathbf{T}_3^l} \right). \tag{7}$$

To combine the predictions from two ends, we apply a weighting strategy to them. To calculate the weighting coefficients, we employ the change information between transitional images as follows:

$$w_i = \frac{\frac{1}{|\mathbf{T}_2^l - \mathbf{T}_i^l|}}{\frac{1}{|\mathbf{T}_2^l - \mathbf{T}_1^l|} + \frac{1}{|\mathbf{T}_2^l - \mathbf{T}_3^l|}}, \quad i = 1, 3. \tag{8}$$

Then, the LSR Landsat image at t_2 is predicted as follows:

$$\hat{\mathbf{L}}_2^l = w_1 \mathbf{L}_{21}^l + w_3 \mathbf{L}_{23}^l. \tag{9}$$

To reduce the prediction errors, (8) and (9) are implemented in patch form (e.g., 9×9) with overlapping. The fusion procedures from LSR Landsat images to Landsat images based on SR CNN are the same to the above ones as demonstrated in the prediction stage of Fig. 1.

III. Experiments

In this section, we compare the proposed method to the representative learning-based sparse representation method in [16]. We select two Landsat–MODIS datasets due to the following two reasons. First, these two datasets consist of times series image pairs (14 pairs and 17 pairs, respectively), which fits in the application background of massive remote sensing data. Second, these two datasets are featured by spatial heterogeneity and temporal dynamics, respectively, which are capable of fully testing the practicality of the methods. For simplicity, we name the proposed method as spatiotemporal fusion using deep convolutional neural networks (STFDCNN). For the comparison method in [16], we name it spatiotemporal fusion based on sparse representation.

A. Study Sites and Datasets

The study sites and datasets used in this paper are the same as those used in [30]. The first study site is the Coleambally irrigation area (CIA) located in southern New South Wales, Australia, and covering an area of 2193 km^2. In CIA, there are 17 cloud-free Landsat–MODIS pairs available in 2001–2002 austral summer growing season; all Landsat images were acquired by Landsat-7 ETM+ and were atmospherically corrected by using MODTRAN4 [31]. The other study site is the Lower Gwydir Catchment (LGC) located in northern New South Wales, Australia, and covering an area of 5440 km^2. In LGC, there are 14 cloud-free Landsat–MODIS pairs available from April 2004 to April 2005; all Landsat images were acquired by Landsat-5 TM and were atmospherically corrected by using the algorithm in [32]. For both study sites, we also use the same MODIS Terra MOD09GA Collection 5 data as in [30]. These data were upsampled to the same spatial resolution (25 m) as the Landsat data using a nearest neighbor algorithm. To achieve subpixel accuracy, each Landsat–MODIS pair was coregistered to within one 25-m pixel by defining the optimal offset required to maximize the correlation function between the two images. For CIA dataset, there are 6 bands with an image size of 1720×2040; for LGC dataset, there are 6 bands with an image size of 3200×2720. The bands 1, 2, 3, 4, 5, and 7 are selected for Landsat images and the bands 1, 2, 3, 4, 6, and 7 are selected for MODIS images. For fusion purpose, the band orders of MODIS images are adjusted accordingly to match those of Landsat images.

For CIA dataset, the temporal dynamics mainly associate with crop phenology over a single growing season, but the surrounding agricultural and woodland areas vary less over time. On the other hand, the field sizes in CIA are relatively small. Thus, we can consider CIA as a more spatially heterogeneous site. For LGC dataset, the temporal extent is about one year. In mid-December 2004, there occurred a large flood, which caused inundation over large areas (about 44%). Because the flooding event leads to different spatial and temporal variations, the LGC can be considered a temporally dynamic site.

B. Quantitative Evaluation Indices

To evaluate the fusion results on prediction dates, we compare them to the real observed Landsat images. Several statistical indices are utilized to give a quantitative evaluation. The first index is the root-mean-square error (RMSE), which gives a global depiction of the radiometric differences between the fusion result

and the real observed image. It is defined as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{R}\sum_{j=1}^{C}(L(i,j)-\hat{L}(i,j))^2}{R\times C}} \tag{10}$$

where L and $\hat{L}$ denote the real observed image and the fusion result, respectively, and R and C denote the image height and width, respectively. A smaller RMSE indicates a better prediction.

The second index is the *erreur relative global adimensionnelle de synthese* (ERGAS) [33], which measures the overall fusion result. ERGAS is defined as follows:

$$\text{ERGAS} = 100\frac{h}{l}\sqrt{\frac{1}{M}\sum_{i=1}^{M}[\text{RMSE}(L_i)^2/(\mu_i)^2]} \tag{11}$$

where h and l denote the spatial resolutions of Landsat and MODIS images, respectively; M is the total number of bands; L_i denotes the ith band image; and μ_i represents the mean value of the ith band image. When ERGAS is smaller and closer to zero, a better fusion result is achieved.

The third index is the spectral angle mapper (SAM) [34], which measures the spectral distortion of the fusion result. It is defined as follows:

$$\text{SAM} = \frac{1}{N}\sum_{i=1}^{N}\arccos\frac{\sum_{j=1}^{M}(L_i^j\hat{L}_i^j)}{\sqrt{\sum_{j=1}^{M}(L_i^j)^2\sum_{j=1}^{M}(\hat{L}_i^j)^2}} \tag{12}$$

where N denotes the total number of pixels in the predicted image. A smaller SAM means a better result.

The fourth metric is the structural similarity (SSIM) [35], which assesses the spatial quality of the fusion results by measuring the similarity of the overall structures between the predicted and real images. It is defined as follows:

$$\text{SSIM} = \frac{(2\mu_L\mu_{\hat{L}}+C_1)(2\sigma_{L\hat{L}}+C_2)}{(\mu_L^2+\mu_{\hat{L}}^2+C_1)(\sigma_L^2+\sigma_{\hat{L}}^2+C_2)} \tag{13}$$

where σ_L and $\sigma_{\hat{L}}$ denote the variances of the observed and predicted images, respectively; $\sigma_{L\hat{L}}$ represents the covariance between the observed and predicted images; μ_L and $\mu_{\hat{L}}$ are the means of the observed and predicted images, respectively; 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 valid range of SSIM is $[-1, 1]$ and a larger SSIM means a better result.

C. Experimental Results

For both CIA and LGC datasets, we arrange all Landsat–MODIS pairs in chronological order. For the proposed STFDCNN, we set the filter number and the filter size by referring to [20]. In general, the performance would improve if we increase these two parameters but as the cost of running time. We, thus, set the parameters to achieve the best tradeoff between performance and speed. In both NLM CNN and SR CNN, we set the parameters as follows: $n_1 = 64$, $n_2 = 32$, $k_1 = 9$, $k_2 = 5$, $k_3 = 5$. In training stage, we select the 1st, the 6th, and the 14th image pairs as the training data and set the size of training subimages as 33 for both CIA and LGC datasets. For CIA dataset, we extract 12 288 subimages in learning NLM CNN and 33 664 subimages in learning SR CNN. For LGC dataset, we extract 38 016 subimages in learning NLM CNN and 87 168 subimages in learning SR CNN. Considering the different wavelength centers and different bandwidths between MODIS and Landsat bands, we learn an NLM CNN for each band. Considering that the LSR Landsat images are obtained via downsampling the original Landsat images, we learn an SR CNN for all bands. For STFSR, we adopt the same weighting strategy as the proposed method for the two predictions from two neighboring dates and set the optimal parameters as suggested in [16].

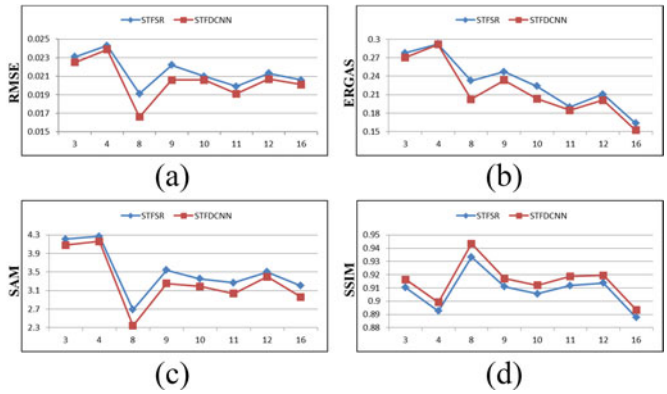


Fig. 4. Quantitative evaluation results for eight prediction dates at the CIA site: (a) RMSE; (b) ERGAS; (c) SAM; (d) SSIM.

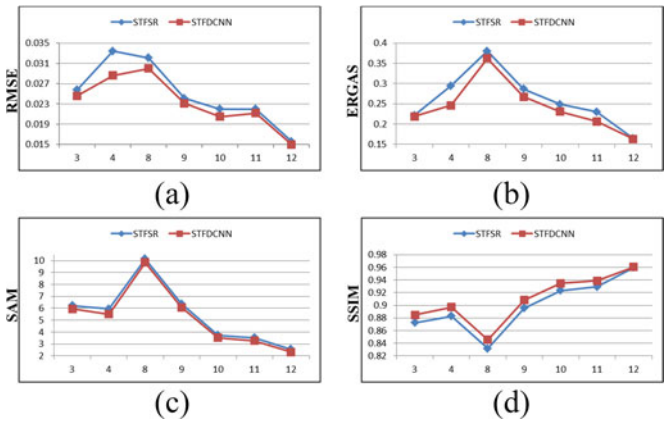


Fig. 5. Quantitative evaluation results for seven prediction dates at the LGC site: (a) RMSE; (b) ERGAS; (c) SAM; (d) SSIM.

In prediction stage, we predict the Landsat-like image on a certain date using the corresponding MODIS image and two Landsat–MODIS pairs that were the nearest temporal neighbors to the prediction date, one before and one after. Excluding the image pairs for training, we predict the 3rd, the 4th, the 8th, the 9th, the 10th, the 11th, the 12th, and the 16th Landsat images for CIA dataset and predict the 3rd, the 4th, the 8th, the 9th, the 10th, the 11th, and the 12th Landsat images for LGC dataset. For CIA and LGC datasets, the quantitative evaluations of the average of all bands in terms of RMSE, ERGAS, SAM, and SSIM for the fusion results from STFSR and STFDCNN are demonstrated in Figs. 4 and 5, respectively. From these two figures, we can observe that the proposed STFDCNN achieves

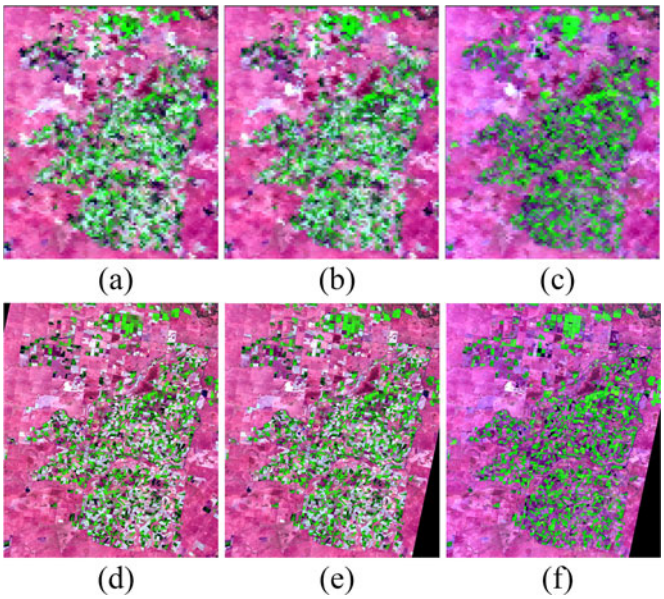


Fig. 6. Illustration of CIA image pairs on the 7th, 8th, and 9th dates. (a)–(c) MODIS images on the 7th, 8th, and 9th dates, respectively, and (d)–(f) Landsat images on the 7th, 8th, and 9th dates, respectively.

lower RMSE, ERGAS, and SAM values and higher SSIM values than STFSR for all eight prediction dates at the CIA site and all seven prediction dates at the LGC site. This indicates that the proposed method is capable of producing fusion results with higher fidelity from the radiometric, spatial structure and spectral aspects. By comparing the results in Figs. 4 and 5, we can conclude that the improvements of STFDCNN over STFSR are generally more at the CIA site than at the LGC site in terms of ERGAS, SAM, and SSIM. This may attribute to the loss of complex change information of land cover types at the LGC site, which is difficult to predict from the LSR MODIS images.

To demonstrate more details of the fusion results, we display the fusion result on one key date for both study sites. For CIA site, we select the 8th prediction due to the color turning in the sporadic irrigation fields through the 8th date to the 9th date, as shown in Fig. 6. For LGC site, we also select the 8th prediction because of the occurrence of a large flood on the 8th date causing temporal dynamics and abnormal change of land surface as shown in Fig. 7. In Figs. 6 and 7, the MODIS and Landsat images on the 7th, 8th, and 9th dates are shown, where the Landsat images are displayed with bands 5, 4, and 3 as R-G-B and the MODIS images with bands 6, 2, and 1 as R-G-B. Assuming that the Landsat image on the 8th date was not known, we predict it from other input images for both CIA and LGC sites.

The fusion results from STFSR and STFDCNN for CIA and LGC sites are demonstrated in Figs. 8 and 9, respectively. The actual observed Landsat images and the fusion results are shown in the first rows and their zoomed details for the parts in the black rectangle are demonstrated in the second rows. From Fig. 8, we can observe that both methods are capable of generally grasping the phenology changes between the prediction and prior dates. However, for some special heterogeneous regions, such as the

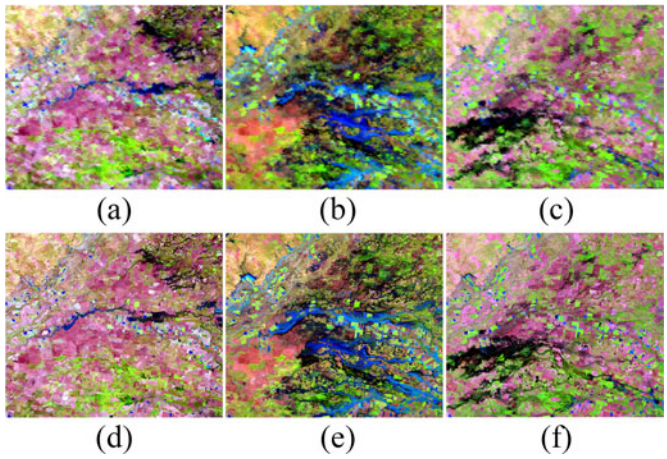


Fig. 7. Illustration of LGC image pairs on the 7th, 8th, and 9th dates. (a)–(c) MODIS images on the 7th, 8th, and 9th dates, respectively, and (d)–(f) Landsat images on the 7th, 8th, and 9th dates, respectively.

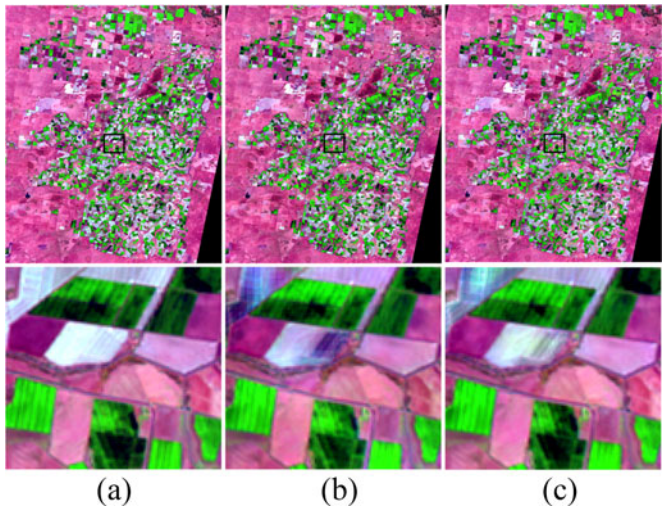


Fig. 8. Illustration of fusion results on the 8th date for CIA site. The first row shows the observed Landsat image and the fusion results and the second row shows the zoomed details in the black rectangle of images in the first row. (a) Observed Landsat image; (b) fusion result from STFSR; (c) fusion result from STFDCNN.

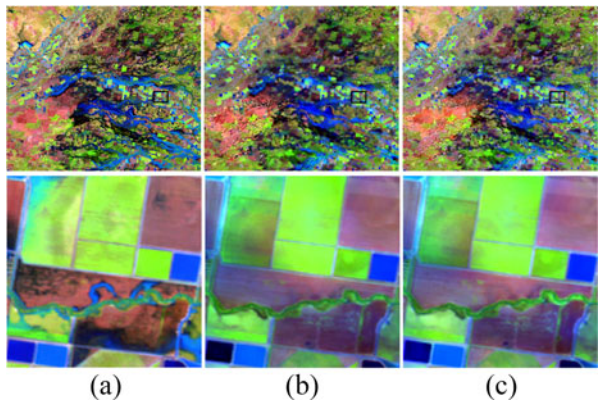


Fig. 9. Illustration of fusion results on the 8th date for LGC site. The first row shows the observed Landsat image and the fusion results and the second row shows the zoomed details in the black rectangle of images in the first row. (a) Observed Landsat image; (b) fusion result from STFSR; (c) fusion result from STFDCNN.

TABLE I
QUANTITATIVE EVALUATIONS FOR THE FUSION RESULTS IN FIG. 8, WHERE BOLD INDICATES BETTER RESULT

Index	Bands	STFSR	STFDCNN
RMSE	B1	0.0093	**0.0073**
	B2	0.0115	**0.0112**
	B3	0.0184	**0.0143**
	B4	0.0251	**0.0222**
	B5	0.0267	**0.0231**
	B6	0.0239	**0.0213**
ERGAS		0.2325	**0.2026**
SSIM	B1	0.9679	**0.9775**
	B2	0.9647	**0.9700**
	B3	0.9423	**0.9556**
	B4	0.9205	**0.9309**
	B5	0.9059	**0.9175**
	B6	0.8986	**0.9096**
SAM		2.6879	**2.3393**

TABLE II
QUANTITATIVE EVALUATIONS FOR THE FUSION RESULTS IN FIG. 9, WHERE BOLD INDICATES BETTER RESULT

Index	Bands	STFSR	STFDCNN
RMSE	B1	0.0152	**0.0141**
	B2	0.0203	**0.0195**
	B3	0.0256	**0.0246**
	B4	0.0344	**0.0317**
	B5	0.0542	**0.0514**
	B6	0.0426	**0.0389**
ERGAS		0.3801	**0.3622**
SSIM	B1	0.9372	**0.9482**
	B2	0.9178	**0.9284**
	B3	0.8948	**0.9042**
	B4	0.8517	**0.8589**
	B5	0.6741	**0.6916**
	B6	0.7135	**0.7401**
SAM		10.1531	**9.8586**

zoomed regions in the second row of Fig. 8, STFDCNN is better than STFSR in predicting the spectral information of the actual Landsat images. From Fig. 9, we can observe that both methods are not able to accurately predict the seriously flooded areas due to the loss of change information in LSR MODIS images, but predicted most of the areas well in both spectral information and spatial structures considering the dramatic changes of land cover types as shown in Fig. 7. From the zoomed details in the second row of Fig. 9, we can observe that both fusion results lose some spatial details and have some degree of spectral distortion compared with the observed Landsat images, but STFDCNN is better than STFSR in predicting the spectral information of the actual Landsat images.

The quantitative evaluations in terms of RMSE, ERGAS, SSIM, and SAM for the fusion results of Figs. 8 and 9 are demonstrated in Tables I and II, respectively. These two tables indicate that STFDCNN performs better than STFSR on all bands of the fusion results.

IV. CONCLUSION

Based on deep CNNs, this paper proposed a spatiotemporal fusion method to combine the spatial information of Landsat images and the temporal information of MODIS images. Considering the complex correspondence relationship and the large spatial resolution gap between MODIS and Landsat images, we first learn an NLM CNN between MODIS and LSR Landsat images. Then, we learn an SR CNN between LSR Landsat and original Landsat images. Via residual learning and back-propagation, a five-layer NLM CNN and a five-layer SR CNN are learned in training stage. In prediction stage, we predict the Landsat image on prediction date from two given prior Landsat–MODIS image pairs and the corresponding MODIS image. To fully utilize the prior information, we define the predicted images from the NLM CNN and SR CNN as transitional images and then adopt a high-pass modulation to integrate the information of prior LSR Landsat or original Landsat images. By conducting experiments on two datasets featured by spatial heterogeneity and temporal dynamics, respectively, the superiority of the proposed method is validated compared with the sparse-representation-based spatiotemporal fusion method. For the CIA dataset mainly with phenology changes, the proposed method achieved satisfactory performance; for the LGC dataset mainly with land cover type changes, the proposed method generally performs well but cannot predict the lost spatial details in LSR MODIS images. It is worth mentioning that although the proposed method is presented for two pairs of prior images, it is also applied to the case of one pair of prior images by only taking the prediction from one end in the prediction stage.

REFERENCES

[1] T. Hilker *et al.*, "Generation of dense time series synthetic landsat data through data blending with modis using a spatial and temporal adaptive reflectance fusion model," *Remote Sens. Environ.*, vol. 113, no. 9, pp. 1988–1999, 2009.

[2] K. J. Ranson, K. Kovacs, G. Sun, and V. I. Kharuk, "Disturbance recognition in the boreal forest using radar and Landsat-7," *Can. J. Remote Sens.*, vol. 29, no. 2, pp. 271–285, 2003.

[3] C. E. Woodcock and M. Ozdogan, "Trends in land cover mapping and monitoring," in *Land Change Science: Observing, Monitoring and Understanding Trajectories of Change on the Earth's Surface*. New York, NY, USA: Springer-Verlag, 2012.

[4] J. G. Masek *et al.*, "North American forest disturbance mapped from a decadal Landsat record," *Remote Sens. Environ.*, vol. 112, no. 6, pp. 2914–2926, 2008.

[5] P. Wu, H. Shen, L. Zhang, and F. M. Gottsche, "Integrated fusion of multi-scale polar-orbiting and geostationary satellite observations for the mapping of high spatial and temporal resolution land surface temperature," *Remote Sens. Environ.*, vol. 156, pp. 169–181, 2015.

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

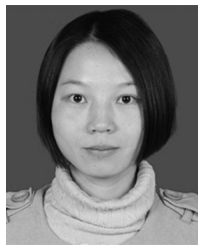
[7] F. Gao, J. Masek, M. Schwaller, and F. Hall, "On the blending of the landsat and modis surface reflectance: Predicting daily landsat surface reflectance," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 8, pp. 2207–2218, Aug. 2006.

[8] T. Hilker *et al.*, "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, 2009.

[9] X. L. Zhu, C. Jin, G. Feng, 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, 2010.

[10] H. Shen, P. Wu, Y. Liu, T. Ai, Y. Wang, and X. Liu, "Spatial and temporal reflectance fusion model considering sensor observation differences," *Int. J. Remote Sens.*, vol. 34, no. 12, pp. 4367–4383, 2013.

[11] 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.
[12] Q. Cheng, H. Liu, H. Shen, P. Wu, and L. Zhang, "A spatial and temporal nonlocal filter-based data fusion method," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4476–4488, Aug. 2017.
[13] R. Zurita-Milla, J. G. P. W. 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.
[14] R. Zuritamilla, 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.," *Remote Sens. Environ.*, vol. 113, no. 9, pp. 1874–1885, 2009.
[15] 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.
[16] 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.
[17] M. Guo, H. Zhang, J. Li, L. Zhang, and H. Shen, "An online coupled dictionary learning approach for remote sensing image fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 4, pp. 1284–1294, Apr. 2014.
[18] C. Jiang, H. Zhang, H. Shen, and L. Zhang, "Two-step sparse coding for the pan-sharpening of remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 5, pp. 1792–1805, May 2014.
[19] V. Jain and H. S. Seung, "Natural image denoising with convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 769–776.
[20] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.
[21] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. 2014 IEEE Comput. Vis. Pattern Recognit.*, 2013, pp. 580–587.
[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[23] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 807–814.
[24] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[26] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising.," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.
[27] W. Yang *et al.*, "Deep edge guided recurrent residual learning for image super-resolution," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5895–5907, Dec. 2017.
[28] Y. Lcun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[29] 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.
[30] I. V. Emelyanova, T. R. Mcvicar, T. G. V. Niel, L. T. Li, and A. I. J. M. V. Dijk, "Assessing the accuracy of blending landsatcmodis surface reflectances in two landscapes with contrasting spatial and temporal dynamics: A framework for algorithm selection," *Remote Sens. Environ.*, vol. 133, no. 12, pp. 193–209, 2013.
[31] A. Berk *et al.*, "Modtran4 radiative transfer modeling for atmospheric correction," *Proc. SPIE*, vol. 3756, pp. 348–353, 1999.
[32] F. Li, D. L. B. Jupp, S. Reddy, and L. Lymburner, "An evaluation of the use of atmospheric and BRDF correction to standardize landsat data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 3, no. 3, pp. 257–270, Sep. 2010.
[33] 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.
[34] R. Yuhas, A. F. H. Goetz, and J. W. Boardman, "Descrimination 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.
[35] 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.



Huihui Song received the B.S. degree in technology and science of electronic information from the Ocean University of China, Qingdao, China, in 2008, the Master's degree in communication and information system from the University of Science and Technology of China, Hefei, China, in 2011, and the Ph.D. degree in geography and resource management from the Chinese University of Hong Kong, Hong Kong, in 2014.

She is a Professor with the Jiangsu Key Laboratory of Big Data Analysis Technology, Nanjing University of Information Science and Technology, Nanjing, China. Her research interests include remote sensing image processing and image fusion.

Qingshan Liu received the M.S. degree from the Department of Auto Control, South-East University, Nanjing, China, in 2000, and the Ph.D. degree from the National Laboratory of Pattern Recognition, Chinese Academic of Science, Beijing, China, in 2003.

He is a Professor with the School of Information and Control Engineering, Nanjing University of Information Science and Technology, Nanjing, China. He was an Assistant Research Professor with the Department of Computer Science, Computational Biomedicine Imaging and Modeling Center, State University of New Jersey, Rutgers, NJ, USA, from 2010 to 2011. Before he joined Rutgers University, he was an Associate Professor with the National Laboratory of Pattern Recognition, Chinese Academic of Science, and an Associate Researcher with the Multimedia Laboratory, Chinese University of Hong Kong, during June 2004 and April 2005. His research interests include image and vision analysis including face image analysis, graph and hypergraph-based image and video understanding, medical image analysis, event-based video analysis, etc.

Guojie Wang received the Ph.D. degree in meteorology direction from the Department of Earth Sciences, Vrije University Amsterdam, Holland, the Netherlands, in 2010.

He is a Professor with the School of Geography, Nanjing University of Information Science and Technology. His research interests include global and regional scale water circulation of satellite remote sensing and land surface atmosphere interaction.

Renlong Hang received the Ph.D. degree in meteorological information technology from the Nanjing University of Information Science and Technology, Nanjing, China, in 2017.

He is currently a Lecturer with the School of Information and Control, Nanjing University of Information Science and Technology, Nanjing, China. His research interests include machine learning and pattern recognition.

Bo Huang is a Professor with the Department of Geography and Resource Management and an Associate Director with the Institute of Space and Earth Information Science, The Chinese University of Hong Kong, Hong Kong. Prior to this, he held faculty positions with the Schulich School of Engineering, University of Calgary, Canada, during 2004–2006 and with the Department of Civil Engineering, National University of Singapore, during 2001–2004. His research interests include most aspects of geoinformation science, specifically spatiotemporal image fusion for environmental monitoring, spatial/spatiotemporal statistics for land cover/land use change modeling, and spatial optimization for sustainable urban and land use planning.