SCIENCE CHINA
Information Sciences

April 2020, Vol. 63 140302:1–140302:16
https://doi.org/10.1007/s11432-019-2805-y

• RESEARCH PAPER •
Special Focus on Deep Learning in Remote Sensing Image Processing

A new sensor bias-driven spatio-temporal fusion model based on convolutional neural networks

Yunfei LI[1], Jun LI[1*], Lin HE[2*], Jin CHEN[3] & Antonio PLAZA[4]

[1]*Guangdong Provincial Key Laboratory of Urbanization and Geo-simulation, School of Geography and Planning, Sun Yat-sen University, Guangzhou* 510275*, China;*
[2]*School of Automation Science and Engineering, South China University of Technology, Guangzhou* 510640*, China;*
[3]*State Key Laboratory of Earth Surface Processes and Resource Ecology, Institute of Remote Sensing Science and Engineering, Faculty of Geographical Science, Beijing Normal University, Beijing* 100875*, China;*
[4]*Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, Escuela Politécnica, University of Extremadura, Cáceres* E-10071*, Spain*

Received 1 November 2019/Revised 18 January 2020/Accepted 19 February 2020/Published online 9 March 2020

Abstract Owing to the tradeoff between scanning swath and pixel size, currently no satellite Earth observation sensors are able to collect images with high spatial and temporal resolution simultaneously. This limits the application of satellite images in many fields, including the characterization of crop yields or the detailed investigation of human-nature interactions. Spatio-temporal fusion (STF) is a widely used approach to solve the aforementioned problem. Traditional STF methods reconstruct fine-resolution images under the assumption that changes are able to be transferred directly from one sensor to another. However, this assumption may not hold in real scenarios, owing to the different capacity of available sensors to characterize changes. In this paper, we model such differences as a bias, and introduce a new sensor bias-driven STF model (called BiaSTF) to mitigate the differences between the spectral and spatial distortions presented in traditional methods. In addition, we propose a new learning method based on convolutional neural networks (CNNs) to efficiently obtain this bias. An experimental evaluation on two public datasets suggests that our newly developed method achieves excellent performance when compared to other available approaches.

Keywords spatio-temporal fusion (STF), convolutional neural networks (CNNs), sensor bias-driven STF

Citation Li Y F, Li J, He L, et al. A new sensor bias-driven spatio-temporal fusion model based on convolutional neural networks. Sci China Inf Sci, 2020, 63(4): 140302, https://doi.org/10.1007/s11432-019-2805-y

1 Introduction

Remotely sensed images with high temporal and spatial resolution are very important for change detection applications, such as the characterization of crops [1], vegetation monitoring [2], or the detailed investigation of human-nature interactions [3]. In the aforementioned scenarios, changes need to be measured at very fine temporal and spatial scales (especially in heterogeneous regions). However, there are currently no satellite instruments technically capable to provide such images, mainly owing to the tradeoff between the scanning swath width and pixel size [4]. As a matter of fact, images with high temporal resolution usually exhibit low spatial resolution (e.g., MODIS or AVHRR imagery), while images with high spatial resolution are often sparse in frequency (e.g., Landsat and SPOT images). In order to obtain images with high spatial and temporal resolution, spatio-temporal fusion (STF) which fuses the two aforementioned kinds of images has been adopted in the literature as a feasible and effective strategy [5].

* Corresponding author (email: lijun48@mail.sysu.edu.cn, helin@scut.edu.cn)

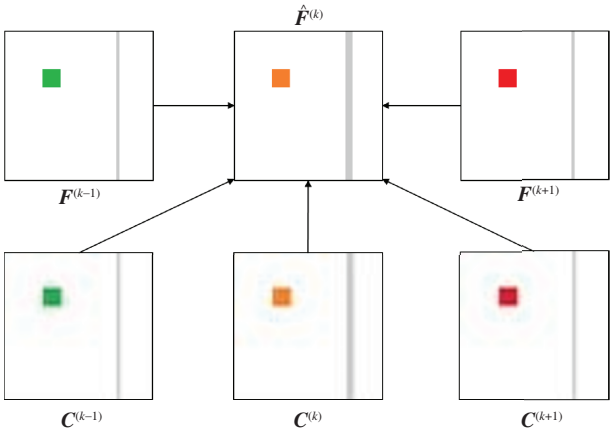

Figure 1 (Color online) Graphical illustration of the main goal of the spatio-temporal fusion (STF) task.

Let $\boldsymbol{C}$ and $\boldsymbol{F}$ respectively denote a coarse and a fine spatial resolution image, and let $k-1$, k and $k+1$ denote three different acquisition times. The goal of the fusion task is to predict $\widehat{\boldsymbol{F}}^{(k)}$, with $\boldsymbol{F}^{(k-1)}, \boldsymbol{F}^{(k+1)}$ and $\boldsymbol{C}^{(k-1)}, \boldsymbol{C}^{(k)}, \boldsymbol{C}^{(k+1)}$ available (or just with $\boldsymbol{F}^{(k-1)}, \boldsymbol{C}^{(k-1)}, \boldsymbol{C}^{(k)}$). In this study, we only consider the first case, as shown in Figure 1. In fact, there are many fusion methodologies aimed at predicting $\widehat{\boldsymbol{F}}^{(k)}$. In all these methodologies, a key concern is how to model the land surface reflectance changes including two aspects: phenological changes (e.g., seasonal changes of crops) and land-cover changes [6]. Most available STF methods can be categorized into three groups: (1) weighted function-based, (2) unmixing-based, and (3) learning-based.

(1) Among weighted function-based methods, the spatial and temporal adaptive reflectance fusion model (STARFM) [7] has been one of the most successful benchmarks. It assumes that all pixels in the coarse images are pure, and uses a weighted strategy to add the reflectance changes between two coarse images, i.e., $\boldsymbol{C}^{(k-1)}$ and $\boldsymbol{C}^{(k)}$, to the (prior) fine resolution image $\boldsymbol{F}^{(k-1)}$, so as to obtain an approximate prediction $\widehat{\boldsymbol{F}}^{(k)}$ directly. There are other methods in this category, such as STAARCH [8], ESTARFM [9], SADFAT [10], and bilateral filter method [11]. All these methods adopt a similar strategy as compared to the one used by STARFM. However, there is a significant limitation shared by all weighted function-based methods: because the weights are defined by the the prior fine resolution image, in case that there are significant changes between the two times, these weights may not hold anymore for the predicted time.

(2) Among unmixing-based methods, the flexible spatiotemporal data fusion (FSDAF) [4] assumes that the coarse pixels in both $\boldsymbol{C}^{(k-1)}$ and $\boldsymbol{C}^{(k)}$ are mixed by the classes from the prior fine resolution image $\boldsymbol{F}^{(k-1)}$, and the unmixing differences between the two coarse images are used to represent the reflectance changes between the two times. ESTDFM [12], STDFA [13], MSTDFA [14] and STRUM [15] also belong to this category. Owing to the fact that pixels in the coarse resolution images are likely to be mixed, unmixing-based methods perform better than weighted function-based methods to some degree [15]. However, similar to weighted function-based methods, unmixing-based methods also assume that $\boldsymbol{F}^{(k-1)}$ and $\boldsymbol{F}^{(k)}$ contain the same classes, i.e., there are no significant changes between the two times (which may not be realistic in most cases).

(3) Among learning-based methods, the spatiotemporal satellite image fusion through one-pair image learning method [16] differs from STARFM and FSDAF in the fact that it aims at reducing the gap between the coarse and fine spatial resolutions by improving the spatial resolution of the coarse resolution images. This is done by establishing the correlation of the coarse and fine resolution images via sparse representation. With similar spatial resolutions, the fusion task is much easier, while in [16], this task was performed by a simple high-pass modulation.

Owing to the fact that learning-based methods are theoretically able to work in any scenario, including cases with significant changes, which are difficult to handle for the weighted function-based and unmixing-based methods, in recent years there has been a significant interest in this kind of approaches. Other than sparse representation-based approaches [16–21], techniques such as regression trees [22], random

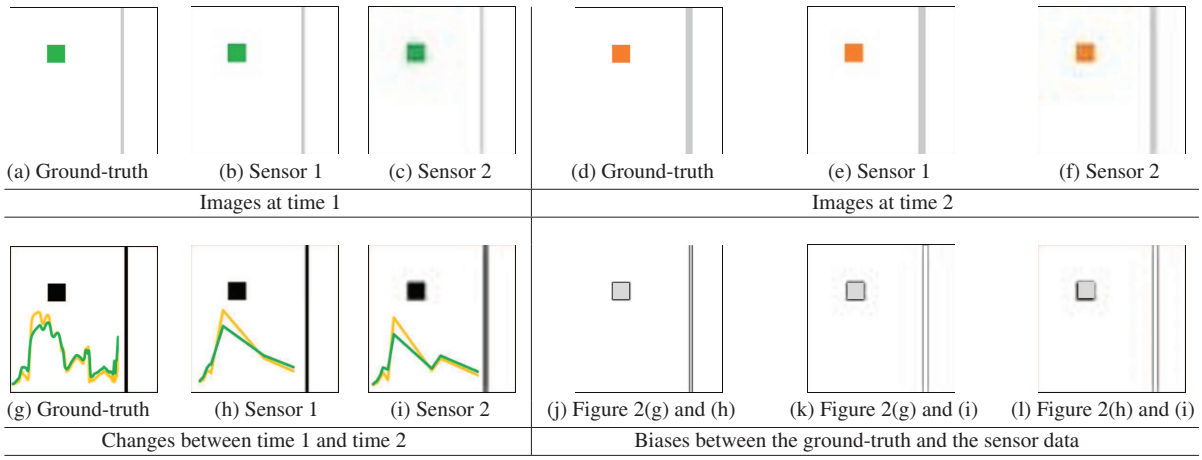

Figure 2 (Color online) Toy example illustrating the impact of the bias. (a)–(c) The images collected at time 1; (d)–(f) the images collected at time 2; (g)–(i) show the changes; and (j)–(l) illustrate the bias maps, with (j) showing the bias between the ground-truth and sensor 1, (k) showing the bias between the ground-truth and sensor 2, and (l) showing the bias between sensors 1 and 2, respectively. It can be seen that the bias between sensors 1 and 2 (i.e., (l)) is significant, which is expected to play an essential role in the STF process.

forests [23] and extreme learning machines [24] have been widely used.

Recently, convolutional neural networks (CNNs) have been widely used in image processing tasks, such as super-resolution [25,26] or pansharpening [27,28], obtaining remarkable results. CNNs have also been used for STF, an area in which the spatiotemporal satellite image fusion using deep convolutional neural networks (STFDCNN) [29] was a pioneering approach exhibiting remarkable performance.

In summary, all of the aforementioned methods can be formulated as follows:

$$\begin{aligned}\widehat{\boldsymbol{F}}^{(k)} &= \mathrm{STF}(\boldsymbol{F}^{(k-1)}, \boldsymbol{F}^{(k+1)}, \Delta\boldsymbol{F}) \\ &\approx \mathrm{STF}(\boldsymbol{F}^{(k-1)}, \boldsymbol{F}^{(k+1)}, \Delta\boldsymbol{C}|(\boldsymbol{C}^{(k-1)}, \boldsymbol{C}^{(k)}, \boldsymbol{C}^{(k+1)})),\end{aligned} \tag{1}$$

where $\mathrm{STF}(\cdot)$ refers to a spatio-temporal fusion model, $\Delta\boldsymbol{F} = \{\Delta\boldsymbol{F}^{(k-1,k)}, \Delta\boldsymbol{F}^{(k+1,k)}\}$ denotes the changes between times $(k-1, k+1)$ and k for the fine resolution images, and $\Delta\boldsymbol{C} = \{\Delta\boldsymbol{C}^{(k-1,k)}, \Delta\boldsymbol{C}^{(k+1,k)}\}$ represents those for the coarse resolution images. The core idea of existing methods is to use $\Delta\boldsymbol{C}$ to approximate $\Delta\boldsymbol{F}$. This is reasonable under the assumption that, for a specific area and period, the change information can be directly transferred from one sensor to another, owing to the fact that spatial and temporal land-surface changes are instrument-independent. However, the assumption may not hold in real situations because different sensors usually exhibit different capacities to capture land-surface information owing to their specific design, such as bandwidth coverage, spectral response function, and spatial resolution. These characteristics, along with atmospheric conditions, can result in very different abilities for change characterization. Such differences, referred hereinafter as bias, are expected to play an important role in STF.

The toy example in Figure 2 shows a scenario containing a road and a tree, and includes two types of changes: a spatial change in the shape of the road, and a spectral change in the phenology of the tree. Figures 2(a)–(c) show the ground image and the ones collected by sensors 1 and 2, at time 1, respectively. Figures 2(d)–(f) show the ground image and the ones collected by sensors 1 and 2, respectively, at time 2. Figures 2(g)–(i) show the real changes and the ones captured by sensors 1 and 2, respectively. As shown by Figure 2, the changes captured by the two sensors are different from the real changes, leading to a bias between the ground-truth and the remotely sensed images. As expected, the information captured by the two sensors is different (see Figure 2(h) and (i)) from both the spectral and spatial viewpoints, which leads to a bias between the two sensors as well, as illustrated in Figure 2(l).

As mentioned before, existing methods approximate the changes in the fine resolution images (e.g., Figure 2(i)), by using the change information from the coarse resolution images (e.g., Figure 2(h)), ignoring the bias between these two sensors (e.g., Figure 2(l)). This is the main reason for the spatial

distortion (originated by land-cover changes, and also for the spectral distortion originated by phenological changes). In order to mitigate these issues and achieve a better spatio-temporal prediction, a natural solution is to include the bias for STF modeling.

Inspired by the aforementioned ideas, in this paper we develop a new sensor-bias driven STF model (called BiaSTF) specifically aimed at exploiting the bias information in the fusion of different sensors, which is a main innovative contribution of this study. In other words, our STF method is the first one that considers both the bias and the changes. Similar to existing methods, our proposed model still assumes that, for a specific area and period, the change information can be transferred from one sensor to another (conditionally, not directly). In this regard, an essential consideration exploited in this study is that the sensor bias also should be transferred from one sensor to another, apart from the direct changes. Another important contribution of this study is the design of a new CNN-based method to efficiently learn the bias. Therefore, the proposed method actually belongs to the learning-based category. From a theoretical viewpoint, the proposed CNN-based BiaSTF[1)] exhibits lower losses when compared to traditional methods such as STFDCNN [29] (which is also a CNN-based STF method that adopts the traditional STF model). Another advantage of the proposed approach is that, with the inclusion of the bias in the model, the spectral and spatial distortions are significantly reduced, resulting in remarkable performance in terms of STF.

The remainder of this paper is organized as follows. Section 2 introduces the BiaSTF fusion model and the BiaSTF-CNN bias learning, with some theoretical model analyses also presented and discussed. Section 3 evaluates the proposed method via experiments with remotely sensed data sets collected by Landsat and MODIS, respectively. Section 4 concludes the paper with some remarks and hints at plausible future research lines.

2 Methodology

Assuming that $\boldsymbol{F}^{(k-1)}$, $\boldsymbol{F}^{(k+1)}$, $\boldsymbol{C}^{(k-1)}$, $\boldsymbol{C}^{(k)}$, $\boldsymbol{C}^{(k+1)}$ are available, the STF task consists of predicting $\widehat{\boldsymbol{F}}^{(k)}$, which can be obtained as follows:

$$\widehat{\boldsymbol{F}}^{(k)} = f(\widehat{\boldsymbol{F}}^{(k-1,k)}, \widehat{\boldsymbol{F}}^{(k+1,k)}), \tag{2}$$

where $f(\cdot)$ is a fusion model (generally a weighted function), $\widehat{\boldsymbol{F}}^{(k-1,k)}$ and $\widehat{\boldsymbol{F}}^{(k+1,k)}$ are two transitional predicted images from $\boldsymbol{F}^{(k-1)}$ and $\boldsymbol{F}^{(k+1)}$. This function can be defined using different strategies. For instance, the SPSTFM [17] uses the absolute average change of the sum between the normalized difference vegetation index and the normalized difference built-up index to determine the weights, while the STFDCNN obtains the weights by measuring the similarity of pixels of the improved coarse resolution images. $\widehat{\boldsymbol{F}}^{(k-1,k)}$ and $\widehat{\boldsymbol{F}}^{(k+1,k)}$ are respectively given by

$$\widehat{\boldsymbol{F}}^{(k-1,k)} = \boldsymbol{F}^{(k-1)} + \Delta\boldsymbol{F}^{(k-1,k)}, \tag{3}$$

$$\widehat{\boldsymbol{F}}^{(k+1,k)} = \boldsymbol{F}^{(k+1)} + \Delta\boldsymbol{F}^{(k+1,k)}. \tag{4}$$

Recall that $\Delta\boldsymbol{F}^{(k-1,k)}$ and $\Delta\boldsymbol{F}^{(k+1,k)}$ are the reflectance changes from times $k-1$ and $k+1$ to k of the fine resolution images, respectively, and $\Delta\boldsymbol{F} = \{\Delta\boldsymbol{F}^{(k-1,k)}, \Delta\boldsymbol{F}^{(k+1,k)}\}$. Then, the fusion model in (2) turns to

$$\widehat{\boldsymbol{F}}^{(k)} = f(\boldsymbol{F}^{(k-1)}, \boldsymbol{F}^{(k+1)}, \Delta\boldsymbol{F}). \tag{5}$$

Notice that, if we have enough training images, a model such as CNN-based learning is likely able to retrieve the change information $\Delta\boldsymbol{F}^{(k-1,k)}$ and $\Delta\boldsymbol{F}^{(k+1,k)}$ from the fine resolution images only. However, limited training information is generally available (especially for the fine images, i.e., the ones that need to be reconstructed). Therefore, it is very difficult (or even impossible) to obtain $\Delta\boldsymbol{F}^{(k-1,k)}$ and

1) For simplicity, hereinafter we use the term BiaSTF to refer both to the model and to the proposed CNN-based learning method.

$\Delta \boldsymbol{F}^{(k+1,k)}$ in most cases. In the literature, an alternative solution is to replace $\Delta \boldsymbol{F}$ with an altered $\Delta \boldsymbol{C} = \{\Delta \boldsymbol{C}^{(k-1,k)}, \Delta \boldsymbol{C}^{(k+1,k)}\}$. In this case, we have

$$\widehat{\boldsymbol{F}}^{(k-1,k)} = \boldsymbol{F}^{(k-1)} + \boldsymbol{G}(\Delta \boldsymbol{C}^{(k-1,k)}), \tag{6}$$

$$\widehat{\boldsymbol{F}}^{(k+1,k)} = \boldsymbol{F}^{(k+1)} + \boldsymbol{G}(\Delta \boldsymbol{C}^{(k+1,k)}), \tag{7}$$

where $\boldsymbol{G}(\cdot)$ refers to the operations used to alter the original $\Delta \boldsymbol{C}$. Remind that $\Delta \boldsymbol{C}^{(k-1,k)}$ and $\Delta \boldsymbol{C}^{(k+1,k)}$ are the reflectance changes in the coarse images from times $k-1$ and $k+1$ to k. Therefore, the fusion model in (5) becomes

$$\widehat{\boldsymbol{F}}^{(k)} = f(\boldsymbol{F}^{(k-1)}, \boldsymbol{F}^{(k+1)}, \boldsymbol{G}(\Delta \boldsymbol{C})). \tag{8}$$

As mentioned before, the main task of STF is to obtain $\boldsymbol{G}(\cdot)$. For instance, STARFM introduces the similarity information of the (prior) fine resolution image into the corresponding $\Delta \boldsymbol{C}$, while FSDAF exploits umixing-based information. In turn, learning-based methods such as SPSTFM directly build a relation between the $\Delta \boldsymbol{F}$ and the corresponding $\Delta \boldsymbol{C}$ via parameter learning.

2.1 Problem formulation

Let $\boldsymbol{B} = \{\boldsymbol{B}^{(k-1,k)}, \boldsymbol{B}^{(k+1,k)}\}$ be the bias, and let $\boldsymbol{B}^{(k-1,k)}$ and $\boldsymbol{B}^{(k+1,k)}$ be the corresponding biases at times $k-1$ and $k+1$ to k, which are defined as follows:

$$\boldsymbol{B}^{(k-1,k)} = \Delta \boldsymbol{F}^{(k-1,k)} - \boldsymbol{G}(\Delta \boldsymbol{C}^{(k-1,k)}) \tag{9}$$

and

$$\boldsymbol{B}^{(k+1,k)} = \Delta \boldsymbol{F}^{(k+1,k)} - \boldsymbol{G}(\Delta \boldsymbol{C}^{(k+1,k)}). \tag{10}$$

Therefore, Eqs. (6) and (7) turn to

$$\widehat{\boldsymbol{F}}^{(k-1,k)} = \boldsymbol{F}^{(k-1)} + \boldsymbol{G}(\Delta \boldsymbol{C}^{(k-1,k)}) + \boldsymbol{B}^{(k-1,k)} \tag{11}$$

and

$$\widehat{\boldsymbol{F}}^{(k+1,k)} = \boldsymbol{F}^{(k+1)} + \boldsymbol{G}(\Delta \boldsymbol{C}^{(k+1,k)}) + \boldsymbol{B}^{(k+1,k)}. \tag{12}$$

The fusion model in (2) turns to

$$\widehat{\boldsymbol{F}}^{(k)} = f(\boldsymbol{F}^{(k-1)}, \boldsymbol{F}^{(k+1)}, \boldsymbol{G}(\Delta \boldsymbol{C}), \boldsymbol{B}). \tag{13}$$

As shown in (13), in our model we need to learn both $\boldsymbol{G}(\Delta \boldsymbol{C})$ and $\boldsymbol{B}$ for the reconstruction of $\widehat{\boldsymbol{F}}^{(k)}$.

2.2 CNN-based learning

Let $\boldsymbol{X}_l$ and $\boldsymbol{Z}_l$ be the input and output of the lth CNN layer, respectively. The response of a convolutional layer in a CNN is given by

$$\boldsymbol{Z}_l = \varphi(\boldsymbol{W}_l * \boldsymbol{X}_l + \boldsymbol{\delta}_l), \tag{14}$$

where $*$ denotes the convolution operation, $\boldsymbol{W}_l$ and $\boldsymbol{\delta}_l$ are the weight and model bias metrics, respectively, and $\varphi(\cdot)$ represents the activation function. Owing to its computational simplicity and ability to mitigate gradient vanishing, the rectified linear unit (ReLU) [30] is commonly used in CNNs. Its input-output relation is $\boldsymbol{Z}_l = \max(0, \boldsymbol{X}_l)$ [31–33].

Let N_p be the number of available training image pairs, denoted by $\{(\boldsymbol{C}_i, \boldsymbol{F}_i)\}_{i=1}^{N_p}$, and let $\boldsymbol{\theta}$ represent the free parameters to be optimized in the CNN context. The proposed BiaSTF method minimizes two terms: (i) the loss between the reconstructed change information $\widehat{\boldsymbol{G}}(\boldsymbol{X};\boldsymbol{\theta})$ and the real changes $\Delta \boldsymbol{F}$, and (ii) the loss between the reconstructed bias $\widehat{\boldsymbol{B}}(\boldsymbol{X};\boldsymbol{\theta})$ and the corresponding ground truth $\boldsymbol{B} = \Delta \boldsymbol{F} - \widehat{\boldsymbol{G}}(\Delta \boldsymbol{C})$. With these definitions in mind, our method adopts the following loss function, which minimizes the residual between the reconstructed images $\widehat{\boldsymbol{F}}(\boldsymbol{X};\boldsymbol{\theta})$ and the corresponding ground-truth image $\boldsymbol{F}$:

$$\ell(\boldsymbol{\theta}) = \|\widehat{\boldsymbol{F}}(\boldsymbol{X};\boldsymbol{\theta}) - \boldsymbol{F}\|_F^2, \tag{15}$$

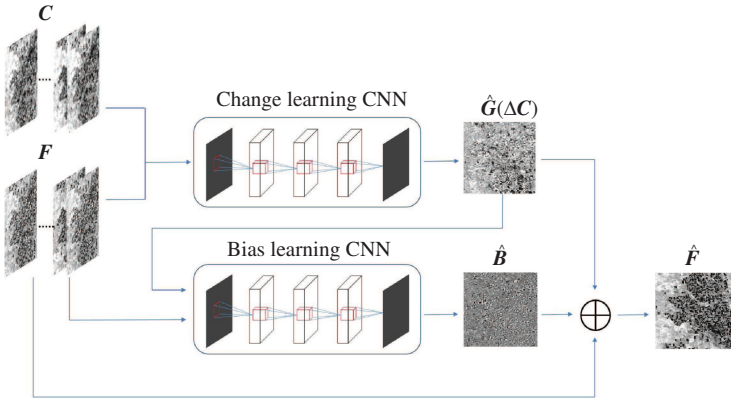

Figure 3 (Color online) Flowchart of the proposed BiaSTF method.

where $\|\cdot\|_F$ stands for the Frobenius norm. As shown in the fusion model in (13), in order to reconstruct $\widehat{\boldsymbol{F}}$, we need to learn $\widehat{\boldsymbol{G}}(\boldsymbol{X};\boldsymbol{\theta})$ and $\widehat{\boldsymbol{B}}(\boldsymbol{X};\boldsymbol{\theta})$. Owing to the fact that it is difficult to learn $\widehat{\boldsymbol{G}}(\boldsymbol{X};\boldsymbol{\theta})$ and $\widehat{\boldsymbol{B}}(\boldsymbol{X};\boldsymbol{\theta})$ simultaneously with a set of joint parameters $\boldsymbol{\theta}$, we approximate $\ell(\boldsymbol{\theta})$ via two subproblems related to $\widehat{\boldsymbol{G}}(\boldsymbol{X};\boldsymbol{\theta})$ and $\widehat{\boldsymbol{B}}(\boldsymbol{X};\boldsymbol{\theta})$ as follows:

- For the first subproblem, we use the following loss function for $\widehat{\boldsymbol{G}}(\cdot)$:

$$\begin{aligned}\ell_1(\boldsymbol{\theta}) &= \|\widehat{\boldsymbol{G}}(\boldsymbol{X};\boldsymbol{\theta}) - \Delta\boldsymbol{F}\|_F^2 \\ &= c\sum_{i=1}^{N_p-1}\sum_{j=2}^{N_p}\|\widehat{\boldsymbol{G}}^{(i,j)}(\boldsymbol{X}^{(i,j)};\boldsymbol{\theta}) - \Delta\boldsymbol{F}^{(i,j)}\|_F^2,\end{aligned} \tag{16}$$

where $c = \frac{1}{2}\frac{1}{N_p}\frac{1}{N_p-1}$, $\Delta\boldsymbol{F}^{(i,j)} = \boldsymbol{F}^{(i)} - \boldsymbol{F}^{(j)}$.

- Then, concerning the second subproblem, we have

$$\begin{aligned}\ell_2(\boldsymbol{\theta}) &= \|\widehat{\boldsymbol{B}}(\boldsymbol{X};\boldsymbol{\theta}) - \boldsymbol{B}\|_F^2 \\ &= c\sum_{i=1}^{N_p-1}\sum_{j=2}^{N_p}\|\widehat{\boldsymbol{B}}^{(i,j)}(\boldsymbol{X}^{(i,j)};\boldsymbol{\theta}) - \boldsymbol{B}^{(i,j)}\|_F^2,\end{aligned} \tag{17}$$

with

$$\boldsymbol{B}^{(i,j)} = \Delta\boldsymbol{F}^{(i,j)} - \widehat{\boldsymbol{G}}(\Delta\boldsymbol{C}^{(i,j)}). \tag{18}$$

For the learning, we adopt the well-known Adam algorithm [34] as the optimizer to minimize the loss functions. This algorithm has been proven to be very efficient for training purposes. For illustrative purposes, Figure 3 shows a flowchart depicting the proposed BiaSTF method. From Figure 3, we can observe that two CNNs are used: one learns the changes given in the first subproblem (17) and the other learns the bias given in the second subproblem (18). The final prediction is obtained by considering the observations, the changes and the bias. At this point, we would like to emphasize that there are tools which should be able to learn the changes and bias more efficiently and effectively. However, because the main purpose of this study is presenting our new bias-driven STF model, we adopt the same naive learning strategy. Nevertheless, as highlighted in Section 3, the proposed BiaSTF achieves very good performance.

2.3 Model analysis

In this subsection, we perform an insightful analysis on the model from two viewpoints. On the one hand, we show that the traditional STFDCNN approximates the changes in the fine resolution images by using the changes from the coarse resolution images. On the other hand, we clarify that our newly developed BiaSTF approach exhibits lower loss, i.e., residual in our implementation, than the STFDCNN.

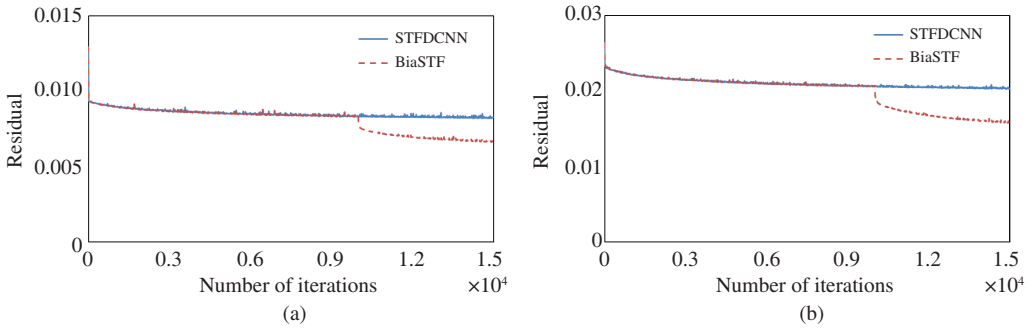

Figure 4 (Color online) Reconstruction residual of the proposed BiaSTF and STFDCNN on two datasets, i.e., (a) CIA and (b) LGC datasets, that will be used for detailed evaluation in Section 3.

First of all, we revisit the STFDCNN method. It minimizes the loss between the reconstructed residual images $\widehat{\boldsymbol{R}}(\boldsymbol{X};\boldsymbol{\theta})$ and the corresponding ground-truth residual image $\boldsymbol{R}=\boldsymbol{F}-\boldsymbol{G}(\boldsymbol{C})$, similar to the deep residual network (ResNet) [35] as follows:

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \|\widehat{\boldsymbol{R}}(\boldsymbol{X};\boldsymbol{\theta})-\boldsymbol{R}\|_F^2\\ &= \frac{1}{N_p}\sum_{i=1}^{N_p}\|\widehat{\boldsymbol{R}}^{(i)}(\boldsymbol{X}^{(i)};\boldsymbol{\theta})+\boldsymbol{G}(\boldsymbol{C}^{(i)})-\boldsymbol{F}^{(i)}\|_F^2,\end{aligned}\tag{19}$$

with $\boldsymbol{\theta}$ being the CNN parameters.

By carefully inspecting (19), we assume $\widehat{\boldsymbol{R}}=\widehat{\boldsymbol{F}}-\boldsymbol{G}(\widehat{\boldsymbol{C}})$ and $Q=\|\widehat{\boldsymbol{R}}-\boldsymbol{R}\|_F^2$. We can thus obtain

$$\begin{aligned}Q &= \|\widehat{\boldsymbol{R}}-\boldsymbol{R}\|_F^2\\ &= \|\widehat{\boldsymbol{F}}-\boldsymbol{G}(\widehat{\boldsymbol{C}})-(\boldsymbol{F}-\boldsymbol{G}(\mathbf{C}))\|_F^2\\ &= \|(\widehat{\boldsymbol{F}}-\boldsymbol{F})-\boldsymbol{G}(\widehat{\boldsymbol{C}}-\boldsymbol{C})\|_F^2\\ &= \|\Delta\widehat{\boldsymbol{F}}-\boldsymbol{G}(\Delta\boldsymbol{C})\|_F^2.\end{aligned}\tag{20}$$

After comparing (19) and (20), we can conclude that the STFDCNN (which uses $\Delta\boldsymbol{C}$ to approximate $\Delta\boldsymbol{F}$) can be regarded as a traditional method (i.e., it assumes that change information can be straightforwardly transferred from one sensor to another). Therefore, a comparison between the losses of BiaSTF and STFDCNN can be considered as a comparison between the sensor-bias derived STF model and the traditional model. In the following, we make a detailed comparison between the losses of the proposed BiaSTF and traditional methods. The final loss function of BiaSTF, obtained by plugging (18) into (17), becomes

$$\begin{aligned}\mathrm{E}(\|\widehat{\boldsymbol{B}}-\boldsymbol{B}\|_F^2) &= \mathrm{E}(\|\widehat{\boldsymbol{B}}-(\Delta\boldsymbol{F}-\boldsymbol{G}(\Delta\boldsymbol{C}))\|_F^2)\\ &= \mathrm{E}(\|\widehat{\boldsymbol{B}}+\boldsymbol{G}(\Delta\boldsymbol{C})-\Delta\boldsymbol{F}\|_F^2).\end{aligned}$$

With respect to STFDCNN, taking (20) and (19) into account, we can model its loss as follows:

$$\mathrm{E}(\|\widehat{\boldsymbol{R}}-\boldsymbol{R}\|_F^2)=\mathrm{E}(\|\boldsymbol{G}(\Delta\boldsymbol{C})-\Delta\widehat{\boldsymbol{F}}\|_F^2).\tag{21}$$

Following [36], it is easy to infer that[2)]

$$\mathrm{E}(\|\widehat{\boldsymbol{B}}+\boldsymbol{G}(\Delta\boldsymbol{C})-\Delta\boldsymbol{F}\|_F^2)<\mathrm{E}(\|\boldsymbol{G}(\Delta\boldsymbol{C})-\Delta\widehat{\boldsymbol{F}}\|_F^2).$$

Therefore, we can conclude that the proposed BiaSTF has a better expected loss than the one of STFDCNN. For illustrative purposes, Figure 4 plots the training residual of the proposed BiaSTF and the traditional STFDCNN on two different datasets, i.e., CIA and LGC datasets, that will be used for

2) For a detailed proof, we refer to the work in [36].

Table 1 Structure of the CNN architecture used by the proposed BiaSTF

Layer	Filter size	Stride	Activation function
Conv1	$7\times7\times n_1$	(1, 1)	ReLU
Conv2	$5\times5\times n_2$	(1, 1)	ReLU
Conv3	$3\times3\times n_3$	(1, 1)	ReLU
Conv4	$3\times3\times1$	(1, 1)	–

detailed evaluation in Section 3. It can be observed that the final residual of BiaSTF is much smaller than those of STFDCNN, which confirms the introspections in our theoretical analysis. Furthermore, as shown in Figure 4, in the beginning of BiaSTF the residual behaves the same as in STFDCNN. This is expected, as the first subproblem (16) is exactly the same as in STFDCNN. However, in the later stage of BiaSTF the residual further degrades to a much lower level by taking advantage from the bias learning in the second subproblem (18)—STFDCNN converged in the first stage. Therefore, we can conclude that our proposed approach is expected to achieve better STF reconstruction.

3 Experimental results

In this section, we evaluate the proposed BiaSTF approach by using two dtasets consisting of pairs of Landsat-MODIS images. For comparative purposes, three methods, including ESTARFM (a classical weighted function-based method) [9], FSDAF (a representative unmixing-based method) [4], and STFDCNN (a CNN-based learning method) [29], are also evaluated. For the competitors, the parameters are set according to the original contributions to ensure their optimal performance. We would like to emphasize that these three methods are based on the traditional sptio-temporal fusion model, in which only the changes are used for the reconstruction of the final image. Our proposed method uses (in addition to the changes) the bias of the two sensors in the reconstruction process. Finally, for our proposed BiaSTF, the weights in the final fusion model in (2) are determined by evaluating the similarity between the predictions of $\widehat{\boldsymbol{F}}^{(k-1,k)}$ and $\widehat{\boldsymbol{F}}^{(k+1,k)}$, and the coarse images $\boldsymbol{C}^{(k)}$.

3.1 CNN setting

For the proposed BiaSTF, we use four convolutional layers to construct the network architecture, where the outputs of the first three layers are activated by using the ReLU funcion. The detailed settings are illustrated in Table 1, where n_1, n_2, and n_3 represent the number of kernels for the convolutional layers: 1, 2, and 3, respectively. Specifically, in this study we set n_1, n_2, and n_3 to 128 owing to the complicated spatial and temporal changes of the studied sites. Taking into account the limitations of the memory of the GPU used in experiments, the training images are tailored into patches with size of 128 $\times$ 128 for learning purposes. The final implementation is carried out using Tensorflow and Keras on a GTX 1080Ti GPU.

3.2 Sites and datasets

Two datasets are used in our experiments, which are publicly available from [37]. They are collected by Landsat (with 25 m spatial resolution and six spectral bands: 1–5 and 7) and MODIS (500 m spatial resolution and the corresponding spectral bands: 1–5 and 7), respectively. The band orders of the two sensors have been adjusted to match each other.

• The area covered by the first dataset is the coleambally irrigation area (CIA) in New South Wales, Australia. In total, 15 cloud-free Landsat-MODIS image pairs are used, all with size of 1500 $\times$ 1500 pixels and acquired between October 2001 and May 2002, which comprise a growing period for the crops. Notice that, CIA is a heterogenous area with multiple phenological changes [37]. As a result, it becomes a widely used benchmark for the evaluation of STF methods. For the two CNN-based learning methods (STFDCNN and the proposed BiaSTF), the image pairs 1–7 and 15 are used for training, while the remaining pairs (8–14) are used in the prediction stage. For illustrative purposes, Figure 5 shows the 1st,

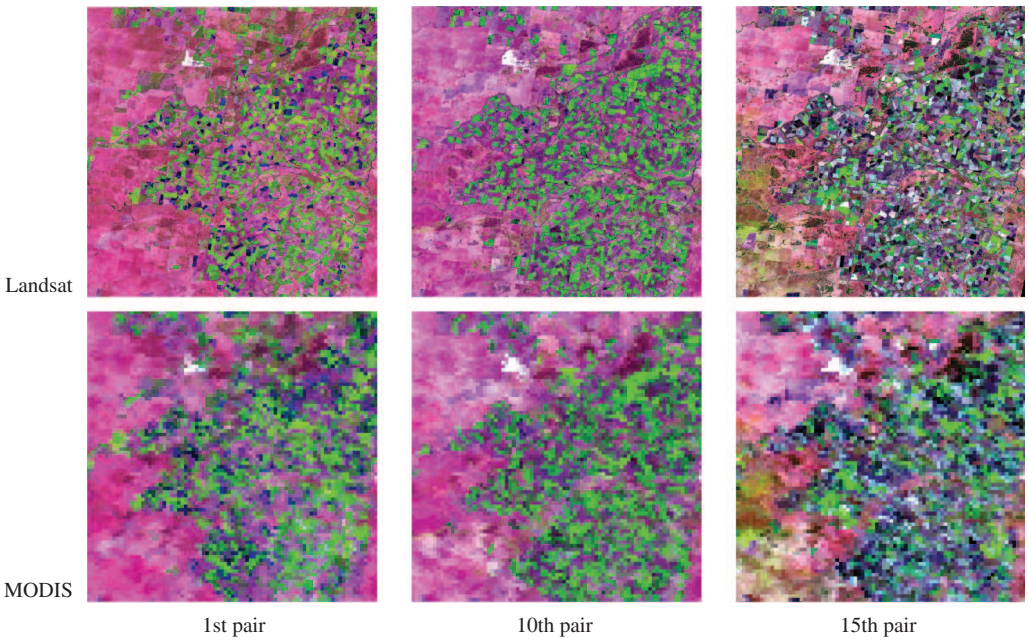

Figure 5 (Color online) Examples from CIA dataset, from which we can observe that there are significant phenological changes.

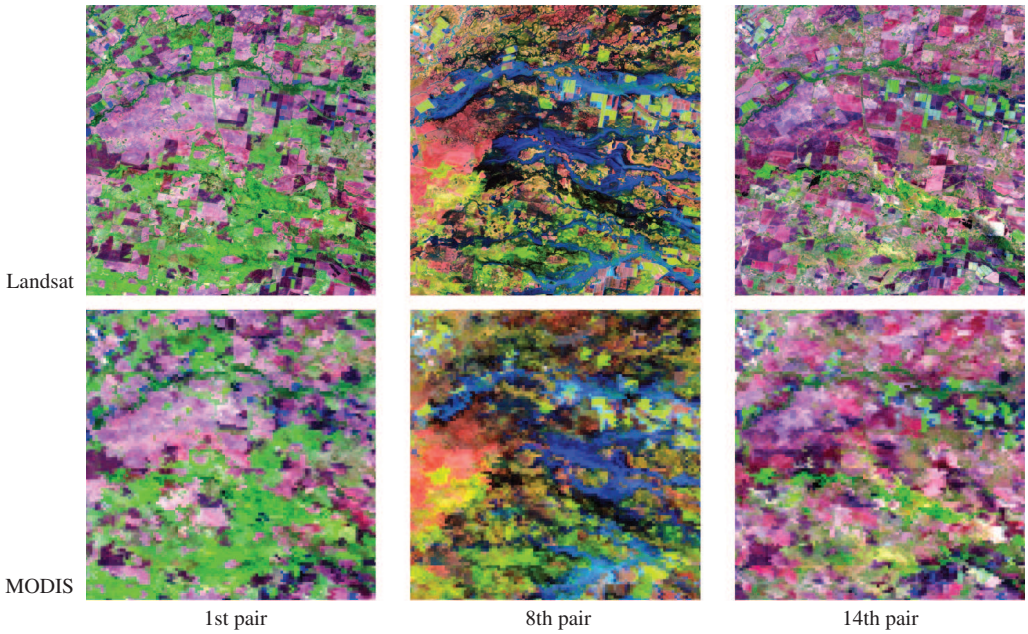

Figure 6 (Color online) Examples from LGC dataset, in which significant land-cover type changes can be observed.

10th and 15th pairs from the CIA dataset. It can be seen that there are significant phenological changes among these three pairs.

• The area covered by the second dataset is the lower Gwydir catchment (LGC), also located New South Wales, Australia. In total, 14 cloud-free Landsat-MODIS image pairs are used, all with size of 2000 × 2000 pixels and acquired between April 2004 and April 2005. LGC suffers a serious flood in mid-December 2004. Specifically, the 8th pair is often used to test the ability of STF methods to characterize land-cover type changes. Similar to our experiments for the CIA dataset, we choose 8 image pairs for training purposes. Owing to the fact that the 8th image pair records the flood, we specifically choose the pairs 1–6, 13 and 14 for training, and the remaining pairs (7–12) for the prediction stage. For illustrative

purposes, Figure 6 shows the 1st, 8th and 14th pairs from the LGC dataset. It is obvious that there are significant land-cover type changes among these three pairs.

3.3 Evaluation metrics

Four metrics are considered for quantitative evaluation as the following.

• The root mean square error (RMSE), which measures the reflectance difference between the predicted image and the real image [4]. The lower the RMSE is, the better the performance is.

• The correlation coefficient (CC), which evaluates the linear relationship between the predicted and real reflectance [4]. The higher the CC is, the better the performance is.

• The erreur relative global adimensionnelle de synthese (ERGAS), which indicates the overall spectral similarity of two images [38]. The lower the ERGAS is, the better the performance is.

• The structure similarity (SSIM), which evaluates overall structure similarities between the predicted and real images [39]. SSIM is a index measuring the spatial similarity. As a result, higher SSIM values indicate lower spatial distortion.

• The spectral angle mapper (SAM), which measures the spectral angle differences between the predicted and real images. SAM is a index measuring the spectral similarity. As a result, lower SAM values indicate lower spectral distortion.

3.4 Results and analysis

3.4.1 *Experiment 1: overall quantitative evaluation*

Table 2 summarizes the conducted quantitative evaluation experiments for the two considered datasets, where the RMSE, CC, and SSIM refer to the mean values obtained for all six bands. It can be observed that the proposed approach obtains the best results in most cases.

It should be noted that, for the SSIM metric (indicating the spatial similarity) and the SAM metric (indicating the spectral similarity), which respectively evaluate the spatial and spectral distortions, the proposed BiaSTF exhibits significant advantages over the other tested methods. Therefore, we can conclude that the proposed BiaSTF can significantly reduce the distortion in comparison with the other methods, which results from the model difference, where a bias of sensors is included in our newly proposed model. In other words, by taking advantage from the bias modeling, the proposed BiaSTF can quantitatively achieve better reconstruction from both the spatial and spectral viewpoints.

3.4.2 *Experiment 2: quantitative and qualitative analysis*

In this experiment, we perform a detailed quantitative and qualitative analysis of the obtained predictions, with the following main observations.

• For the CIA dataset, we particularly address the prediction results for the 10th pair, which suffers from severe phenological changes. Table 3 shows the results of the obtained quantitative assessment. It is remarkable that the proposed approach achieves better results for all the considered (six) bands. Concerning the SSIM and SAM metrics, the proposed BiaSTF exhibits significant advantages over the other tested methods. As already mentioned, this indicates that the proposed BiaSTF exhibits smaller spatial and spectral distortions in the STF process. In the following, we provide a qualitative illustration of the obtained results. Figure 7 shows the obtained predictions. As it can be observed from the zoomed regions, the reconstructed color (lower-leftmost plot) obtained by the proposed BiaSTF is more similar to the ground-truth, which indicates that our BiaSTF exhibits less spectral distortion. On the other hand, as shown in the upper-leftmost plot, our BiaSTF can preserve better the spatial details, meaning that it exhibits less spatial distortion. Therefore, from both the quantitative and qualitative viewpoints, the proposed BiaSTF achieves better fusion performance, with less spatial and spectral distortions, when compared with methods based on the traditional model.

• For the LGC dataset, we specifically choose the 8th pair for detailed analysis because this pair contains significant land-cover changes, which cause its spatial structure to be quite different from that

Table 2 Quantitative assessment of the fusion results obtained for the two considered datasets

	CIA dataset					LGC dataset				
	Pair	ESTARFM	FSADF	STFDCNN	BiaSTF	Pair	ESTARFM	FSADF	STFDCNN	BiaSTF
RMSE	8th	0.0301	0.0331	0.0256	**0.0227**	7th	0.0265	0.0247	0.0378	**0.0240**
	9th	0.0249	0.0282	0.0263	**0.0235**	8th	0.0386	0.0374	0.0346	**0.0334**
	10th	0.0263	0.0243	0.0274	**0.0230**	9th	0.0382	0.0383	0.0275	**0.0239**
	11th	0.0265	0.0278	0.0285	**0.0244**	10th	0.0227	0.0237	0.0252	**0.0211**
	12th	0.0241	0.0269	0.0251	**0.0215**	11th	0.0272	0.0283	0.0242	**0.0240**
	13th	0.0213	0.0231	0.0226	**0.0208**	12th	**0.0165**	0.0230	0.0268	0.0167
	14th	0.0229	0.0251	0.0229	**0.0203**	13th	0.0156	0.0255	0.0245	**0.0154**
CC	8th	0.8740	0.8514	0.9154	**0.9312**	7th	0.7295	0.7530	0.6456	**0.7981**
	9th	0.9113	0.8784	0.8990	**0.9186**	8th	0.6871	0.7078	0.7166	**0.7454**
	10th	0.9075	0.9206	0.8939	**0.9267**	9th	0.7444	0.6938	0.8372	**0.8644**
	11th	0.8806	0.8734	0.8621	**0.8993**	10th	0.8975	0.8722	0.8879	**0.9048**
	12th	0.8371	0.7936	0.8302	**0.8644**	11th	0.8900	0.8809	0.9061	**0.9075**
	13th	0.8643	0.8505	0.8462	**0.8670**	12th	0.9356	0.8878	0.8476	**0.9391**
	14th	0.8602	0.8223	0.8481	**0.8794**	13th	**0.9326**	0.8560	0.8313	0.9309
ERGAS	8th	0.8382	0.9239	0.7531	**0.6445**	7th	0.8418	0.8005	1.2632	**0.7925**
	9th	0.8121	0.9214	0.8663	**0.7740**	8th	2.3058	2.0301	2.0952	**2.0083**
	10th	0.8838	0.8097	0.8981	**0.7720**	9th	1.7146	1.7339	1.2056	**1.0770**
	11th	0.8827	0.9474	0.9694	**0.8131**	10th	0.8059	0.8612	0.8949	**0.7528**
	12th	0.8843	0.9616	0.9172	**0.7901**	11th	0.9556	0.9861	0.8486	**0.8433**
	13th	**0.8601**	0.9027	0.9579	0.8943	12th	0.5657	0.7771	0.9810	**0.5639**
	14th	0.8503	0.9607	0.8462	**0.7540**	13th	0.5015	0.7823	0.8909	**0.5000**
SSIM	8th	0.8921	0.8715	0.9288	**0.9412**	7th	0.8244	0.8403	0.7250	**0.8658**
	9th	0.9277	0.8958	0.9174	**0.9323**	8th	0.7481	0.7523	0.7813	**0.8022**
	10th	0.9246	0.9353	0.9164	**0.9407**	9th	0.7969	0.7811	0.8853	**0.9074**
	11th	0.9111	0.9049	0.8972	**0.9233**	10th	0.9224	0.9078	0.9094	**0.9302**
	12th	0.8977	0.8697	0.8923	**0.9174**	11th	0.9081	0.8996	0.9233	**0.9250**
	13th	0.9146	0.9048	0.9034	**0.9165**	12th	0.9549	0.9189	0.8836	**0.9572**
	14th	0.9080	0.8838	0.9034	**0.9224**	13th	**0.9538**	0.8861	0.8766	0.9532
SAM	8th	0.0728	0.0834	0.0742	**0.0619**	7th	0.0889	0.0799	0.1317	**0.0738**
	9th	**0.0716**	0.0874	0.0874	0.0730	8th	0.2781	0.2433	0.2213	**0.2156**
	10th	0.0783	0.0740	0.0914	**0.0705**	9th	0.1227	0.1785	0.1241	**0.1185**
	11th	0.0823	0.0890	0.1004	**0.0794**	10th	0.0743	0.0935	0.0749	**0.0651**
	12th	0.0770	0.0914	0.0893	**0.0705**	11th	0.0719	0.0805	0.0701	**0.0689**
	13th	0.0737	0.0774	0.0824	**0.0680**	12th	0.0534	0.0681	0.0877	**0.0504**
	14th	0.0601	0.0723	0.0762	**0.0590**	13th	0.0502	0.0801	0.0819	**0.0463**

Table 3 Quantitative assessment of the fusion results obtained for the 10th pair of the CIA dataset

	ESTARFM			FSDAF			STFDCNN			BiaSTF		
	RMSE	CC	SSIM	RMSE	CC	SSIM	RMSE	CC	SSIM	RMSE	CC	SSIM
Band1	0.0134	0.9002	0.9379	0.0115	0.9168	0.9509	0.0126	0.8726	0.9317	**0.0111**	**0.9232**	**0.9543**
Band2	0.0138	0.8984	0.9351	0.0130	0.9094	0.9419	0.0138	0.8864	0.9288	**0.0122**	**0.9124**	**0.9446**
Band3	0.0211	0.9111	0.9247	0.0199	0.9216	0.9339	0.0222	0.8976	0.9137	**0.0192**	**0.9242**	**0.9358**
Band4	0.0348	0.8929	0.9009	0.0307	0.9160	0.9197	0.0378	0.8789	0.8878	**0.0286**	**0.9280**	**0.9322**
Band5	0.0382	0.9211	0.9248	0.0357	0.9319	0.9352	0.0389	0.9188	0.9229	**0.0340**	**0.9365**	**0.9394**
Band6	0.0368	0.9214	0.9242	0.0353	0.9283	0.9303	0.0395	0.9091	0.9138	**0.0331**	**0.9360**	**0.9381**
ERGAS		0.8838			0.8097			0.8981			**0.7720**	
SAM		0.0783			0.0740			0.0914			**0.0705**	

of its temporally adjacent image pairs. Table 4 shows the obtained results, while Figure 8 also shows the obtained predictions. Similar to the observations in our previous experiment, the proposed approach

Figure 7 (Color online) Prediction results obtained for the 10th pair of the CIA dataset.

Table 4 Quantitative assessment of the fusion results obtained for the 8th pair of the LGC dataset

	ESTARFM			FSDAF			STFDCNN			BiaSTF		
	RMSE	CC	SSIM	RMSE	CC	SSIM	RMSE	CC	SSIM	RMSE	CC	SSIM
Band1	0.0161	0.6905	0.8597	0.0160	0.6820	0.8577	0.0166	0.6697	0.8513	**0.0157**	**0.7126**	**0.8673**
Band2	0.0228	0.6928	0.8059	0.0226	0.6921	0.8063	0.0234	0.6934	0.8045	**0.0223**	**0.7115**	**0.8168**
Band3	0.0281	0.6955	0.7795	0.0277	0.6904	0.7775	0.0295	0.6952	0.7754	**0.0275**	**0.7169**	**0.7938**
Band4	0.0481	0.7201	0.7487	0.0428	0.7947	0.7923	0.0376	0.8273	0.8329	**0.0373**	**0.8374**	**0.8471**
Band5	0.0660	0.6725	0.6595	0.0647	0.7008	0.6603	0.0558	0.7491	0.7264	**0.0556**	**0.7656**	**0.7543**
Band6	0.0505	0.6515	0.6356	0.0506	0.6868	0.6200	0.0451	0.6654	0.6975	**0.0417**	**0.7317**	**0.7348**
ERGAS		2.3058			2.0301			2.0952			**2.0083**	
SAM		0.2781			0.2433			0.2213			**0.2156**	

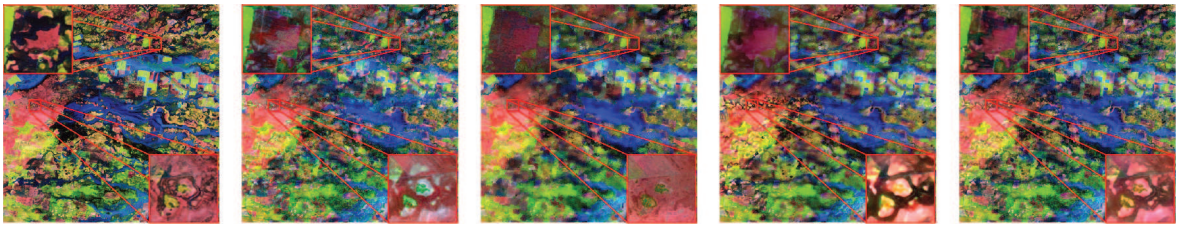

Figure 8 (Color online) Prediction results obtained for the 8th pair of the LGC dataset.

obtains the best results from a quantitative standpoint (as shown by Table 4). On the other hand, as demonstrated in Figure 8, the proposed BiaSTF achieves better performance than the one obtained by the other methods from a qualitative perspective also. Actually, as mentioned before, this image is quite challenging from the viewpoint of reconstruction. We can observe that, from Figure 8, the image obtained by the proposed BiaSTF exhibits much better similarity in terms of color (i.e., spectral similarity) and spatial structure (i.e., spatial similarity) with regards to the ground-truth than those obtained by ESTARFM, FSDAD and STFDCNN, considering the two zoomed regions as illustrative examples (it is clear that the reconstruction obtained by BiaSTF exhibits much less distortion when compared with the other three methods). Therefore, from this experiment we can conclude that the proposed BiaSTF exhibits better STF performance when compared to the methods based on the traditional model.

3.4.3 *Experiment 3: qualitative evaluation of the bias*

As we described in Section 2, traditional methods only use changes from different acquisition times to reconstruct the missing images, while the proposed BiaSTF exploits both the changes and the bias between different sensors. In the former two experiments, we have provided quantitative results. In this experiment, we perform a qualitative evaluation of the impact of changes and bias on the proposed method. Figures 9 and 10 show the obtained changes maps, bias maps and bias errors (calculated as the difference between the real bias maps and the obtained bias maps) for the 10th pair of the CIA dataset and the 8th pair of the LCG dataset, respectively. After a qualitative evaluation of the results in Figures 9 and 10, the following observations can be highlighted.

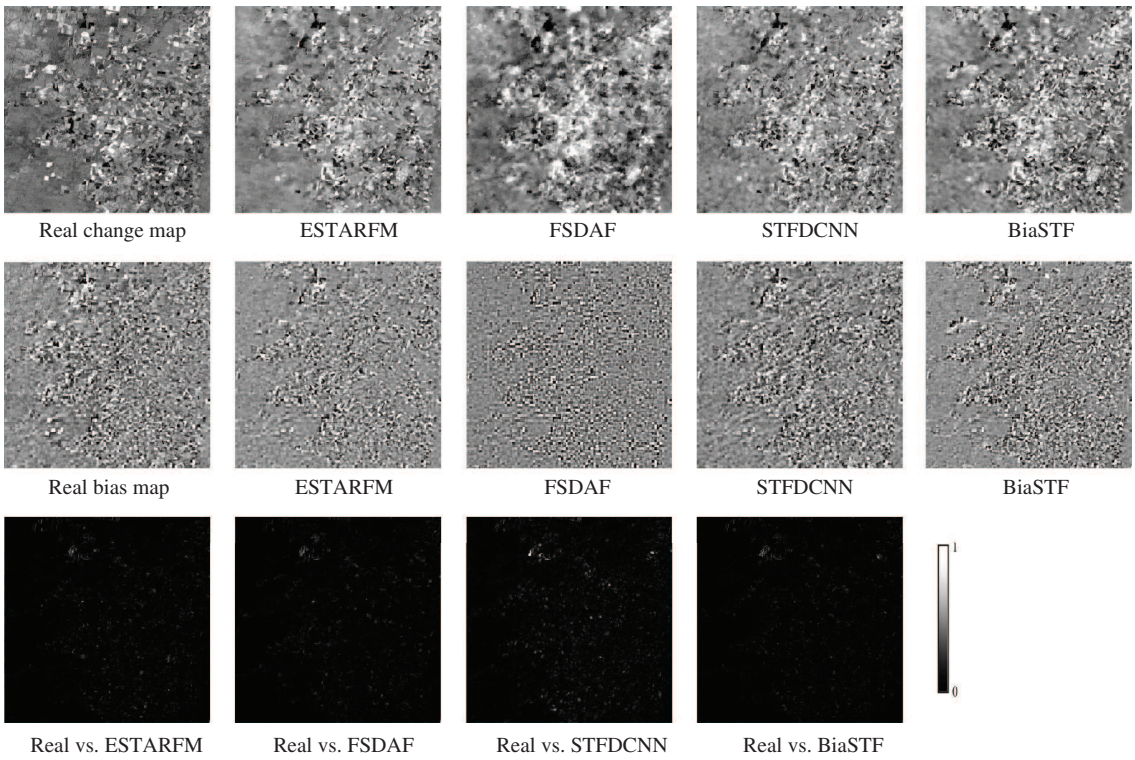

Figure 9 Change maps (first row), bias maps (second row), and bias square error maps (third row) obtained by different methods for the CIA dataset, using the 10th pair.

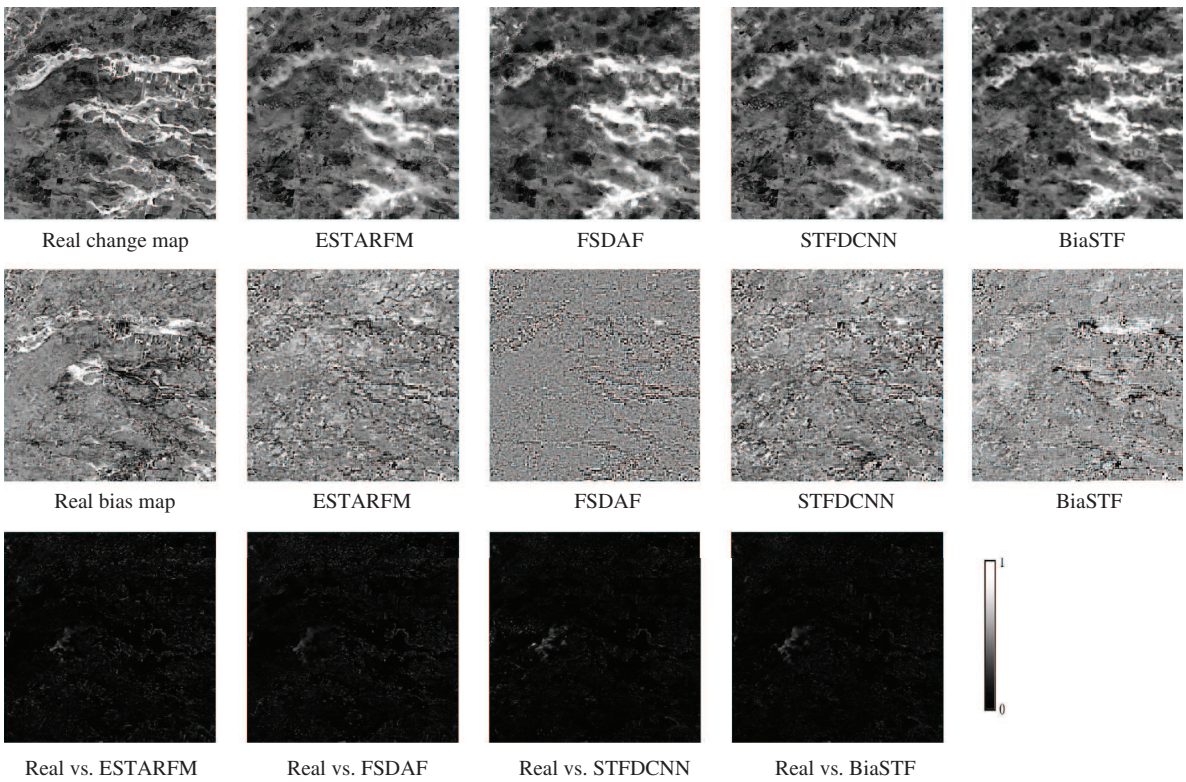

Figure 10 Change maps (first row), bias maps (second row) and bias square error maps (third row), obtained by different methods for the LGC dataset, using the 8th pair.

• First of all, concerning the changes, it is clear that the change maps obtained by the BiaSTF are more similar to the real ones than those obtained by the other tested methods. This is a very interesting observation because, theoretically, both learning-based methods (i.e., STFDCNN and BiaSTF) would be expected to achieve the same (or very similar) change maps. Indeed, the maps produced by these two methods are very similar. However, the maps obtained by BiaSTF are slightly better than those of STFDCNN. This is mainly owing to the fact that we are incorporating a two-step learning in proposed BiaSTF, so that the inclusion of the learning step for the bias reinforces and strenghtens the learning of the changes.

• Furthermore, we can clearly observe that there is a bias between the two considered sensors. For illustrative purposes, we not only compute the bias between the two sensor for our approach, but also for the other tested methods. It is clear that the bias maps of our BiaSTF are more similar to the real ones, while those obtained from the reconstructions are quite different from the real ones. This is expected, as there is no consideration of the bias in the traditional model. In other words, with the inclusion of the bias, the proposed BiaSTF is able to learn a model close to real scenarios.

• Finally, we can also see that the bias error maps (calculated as the difference between the real and obtained bias maps) obtained by the proposed BiaSTF are better than those obtained using the other methods. This confirms that the reconstructions achieved by BiaSTF exhibit less error when compared to those obtained by the other tested method, as already indicated by our quantitative experiments.

4 Conclusion

In this study, we have presented the BiaSTF for remotely sensed images, where the bias is learnt by a CNN-based method. An essential consideration exploited by our approach is that the sensor bias also should be transferred from one sensor to another, apart from the direct changes. Another important contribution of this study is the design of a new CNN-based method to efficiently learn the bias. Our quantitative and qualitative experiments, conducted on two different datasets, reveal that the proposed BiaSTF exhibits very good performance in STF when compared with other traditional methods. This is because the proposed BiaSTF exploits both the changes and the bias between different sensors, while traditional methods only use changes from different acquisition times to reconstruct the missing images. A final relevant contribution of the proposed approach is that, with the inclusion of the bias in the model, the spectral and spatial distortions are significantly reduced, resulting in remarkable performance in terms of STF. In the future, we will develop more advanced deep learning-based strategies to learn the bias and changes more efficiently and effectively.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant Nos. 61771496, 61571195, 61901208), National Key Research and Development Program of China (Grant No. 2017YFB0502900), Guangdong Provincial Natural Science Foundation (Grant Nos. 2016A030313254, 2017A030313382), Science and Technology Project of Jiangxi Provincial Department of Education (Grant No. GJJ180962), and Natural Science Foundation of Jiangxi China (Grant No. 20192BAB217003). The authors would like to thank the contributors for sharing their codes for the algorithms of ESTARFM, FSDAF and STFDCNN.

References

1 Johnson M D, Hsieh W W, Cannon A J, et al. Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. Agric For Meteor, 2016, 218–219: 74–84

2 Shen M, Tang Y, Chen J, et al. Influences of temperature and precipitation before the growing season on spring phenology in grasslands of the central and eastern Qinghai-Tibetan Plateau. Agric For Meteor, 2011, 151: 1711–1722

3 Li X C, Zhou Y Y, Asrar G R, et al. Response of vegetation phenology to urbanization in the conterminous United States. Glob Change Biol, 2017, 23: 2818–2830

4 Zhu X L, Helmer E H, Gao F, et al. A flexible spatiotemporal method for fusing satellite images with different resolutions. Remote Sens Environ, 2016, 172: 165–177

5 Zhu X L, Cai F, Tian J, et al. Spatiotemporal fusion of multisource remote sensing data: literature survey, taxonomy, principles, applications, and future directions. Remote Sens, 2018, 10: 527

6 Zhang H K, Huang B, Zhang M, et al. A generalization of spatial and temporal fusion methods for remotely sensed surface parameters. Int J Remote Sens, 2015, 36: 4411–4445

7 Gao F, Masek J G, Schwaller M R, et al. On the blending of the landsat and MODIS surface reflectance: predicting daily landsat surface reflectance. IEEE Trans Geosci Remote Sens, 2006, 44: 2207–2218

8 Hilker T, Wulder M A, Coops N C, 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, 2009, 113: 1613–1627

9 Zhu X L, Chen J, Gao F, et al. An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions. Remote Sens Environ, 2010, 114: 2610–2623

10 Weng Q H, Fu P, Gao F. Generating daily land surface temperature at Landsat resolution by fusing Landsat and MODIS data. Remote Sens Environ, 2014, 145: 55–67

11 Huang B, Wang J, Song H H, et al. Generating high spatiotemporal resolution land surface temperature for urban heat island monitoring. IEEE Geosci Remote Sens Lett, 2013, 10: 1011–1015

12 Zhang W, Li A, Jin H, et al. An enhanced spatial and temporal data fusion model for fusing landsat and MODIS surface reflectance to generate high temporal landsat-like data. Remote Sens, 2013, 5: 5346–5368

13 Niu Z. Use of MODIS and Landsat time series data to generate high-resolution temporal synthetic Landsat data using a spatial and temporal reflectance fusion model. J Appl Remote Sens, 2012, 6: 063507

14 Wu M Q, Huang W, Niu Z, et al. Generating daily synthetic landsat imagery by combining Landsat and MODIS data. Sensors, 2015, 15: 24002–24025

15 Gevaert C M, García-Haro F J. A comparison of STARFM and an unmixing-based algorithm for Landsat and MODIS data fusion. Remote Sens Environ, 2015, 156: 34–44

16 Song H H, Huang B. Spatiotemporal satellite image fusion through one-pair image learning. IEEE Trans Geosci Remote Sens, 2013, 51: 1883–1896

17 Huang B, Song H H. Spatiotemporal reflectance fusion via sparse representation. IEEE Trans Geosci Remote Sens, 2012, 50: 3707–3716

18 Wu B, Huang B, Zhang L. An error-bound-regularized sparse coding for spatiotemporal reflectance fusion. IEEE Trans Geosci Remote Sens, 2015, 53: 6791–6803

19 Li D C, Li Y R, Yang W F, et al. An enhanced single-pair learning-based reflectance fusion algorithm with spatiotemporally extended training samples. Remote Sens, 2018, 10: 1207

20 Zhao C Y, Gao X B, Emery W J, et al. An integrated spatio-spectral-temporal sparse representation method for fusing remote-sensing images with different resolutions. IEEE Trans Geosci Remote Sens, 2018, 56: 3358–3370

21 Jiang C, Zhang H Y, Shen H F, et al. Two-step sparse coding for the pan-sharpening of remote sensing images. IEEE J Sel Top Appl Earth Observ Remote Sens, 2014, 7: 1792–1805

22 Boyte S P, Wylie B K, Rigge M B, et al. Fusing MODIS with Landsat 8 data to downscale weekly normalized difference vegetation index estimates for central Great Basin rangelands, USA. GISci Remote Sens, 2018, 55: 376–399

23 Ke Y H, Im J, Park S, et al. Downscaling of MODIS one kilometer evapotranspiration using Landsat-8 data and machine learning approaches. Remote Sens, 2016, 8: 215

24 Liu X, Deng C, Wang S, et al. Fast and accurate spatiotemporal fusion based upon extreme learning machine. IEEE Geosci Remote Sens Lett, 2016, 13: 2039–2043

25 Dong C, Loy C C, He K, et al. Image super-resolution using deep convolutional networks. IEEE Trans Pattern Anal Mach Intell, 2016, 38: 295–307

26 Dong C, Chen C L, He K, et al. Learning a deep convolutional network for image super-resolution. In: Computer Vision—ECCV 2014. Berlin: Springer, 2014

27 Wei Y, Yuan Q, Shen H, et al. Boosting the accuracy of multispectral image pansharpening by learning a deep residual network. IEEE Geosci Remote Sens Lett, 2017, 14: 1795–1799

28 Yuan Q, Wei Y, Meng X, et al. A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening. IEEE J Sel Top Appl Earth Observ Remote Sens, 2018, 11: 978–989

29 Song H H, Liu Q, Wang G, et al. Spatiotemporal satellite image fusion using deep convolutional neural networks. IEEE J Sel Top Appl Earth Observ Remote Sens, 2018, 11: 821–829

30 Nair V, Hinton G E. Rectified linear units improve restricted boltzmann machines. In: Proceedings of International Conference on International Conference on Machine Learning, 2010. 807–814

31 Hu W, Huang Y Y, Li W, et al. Deep convolutional neural networks for hyperspectral image classification. J Sensors, 2015, 2015: 1–12

32 Chen Y, Jiang H, Li C, et al. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. IEEE Trans Geosci Remote Sens, 2016, 54: 6232–6251

33 Jia X Y, Xu X M, Cai C B, et al. Single image super-resolution using multi-scale convolutional neural network. In:

Advances in Multimedia Information Processing—PCM 2017. Berlin: Springer, 2017. 149–157

34 Kingma D P, Ba J. Adam: a method for stochastic optimization. In: Proceedings of the 3rd International Conference for Learning Representations, San Diego, 2015

35 He K, Zhang Z, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 770–778

36 He L, Rao Y Z, Li J, et al. Pansharpening via detail injection based convolutional neural networks. IEEE J Sel Top Appl Earth Observ Remote Sens, 2019, 12: 1188–1204

37 Emelyanova I V, McVicar T R, van Niel T G, et al. Assessing the accuracy of blending Landsat-MODIS surface reflectances in two landscapes with contrasting spatial and temporal dynamics: a framework for algorithm selection. Remote Sens Environ, 2013, 133: 193–209

38 Renza D, Martinez E, Arquero A. A new approach to change detection in multispectral images by means of ERGAS index. IEEE Geosci Remote Sens Lett, 2013, 10: 76–80

39 Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process, 2004, 13: 600–612