Spatiotemporal Fusion With Only Two Remote Sensing Images as Input

Jingan Wu, Qing Cheng, Huifang Li, *Member, IEEE*, Shuang Li, Xiaobin Guan, and Huanfeng Shen, *Senior Member, IEEE*

***Abstract*—Spatiotemporal data fusion is an effective way of generating a dense time series with a high spatial resolution. Traditionally, the spatiotemporal fusion models, especially the popular ones such as the spatial and temporal adaptive reflectance fusion model, require at least three images as input, i.e., a coarse-resolution image on the target date and a pair of fine- and coarse-resolution images on the reference date. However, this cannot always be satisfied, as the high-quality coarse-resolution image on the reference date may be unavailable in some application scenarios. This led to efforts to achieve data fusion only using the other two images as input. In this article, we proposed an effective strategy that can be combined with any spatiotemporal fusion model to accomplish the fusion with simplified input. To confirm the validity of the method, we comprehensively compared the fusion performances under the two input modalities. In total, 38 tests were conducted with Moderate Resolution Imaging Spectroradiometer (MODIS), Landsat, and Sentinel-2 land surface reflectance products. Results suggest that by applying the proposed method, the fusion performance with only two input images is comparable or even superior to that with three input images. This article challenges the stereotype that spatiotemporal data fusion strictly needs at least three input images. The proposed method extends the application scenarios of spatiotemporal fusion, and creates opportunities to fuse sensors with barely overlapping temporal coverages, such as the Landsat 8 Operational Land Imager and the Sentinel-2 MultiSpectral Instrument.**

***Index Terms*—Landsat 8, Moderate Resolution Imaging Spectroradiometer (MODIS), Sentinel-2, simplified input modality, spatiotemporal data fusion, two input images.**

I. Introduction

DUE to the tradeoff between the swath width and revisit frequency, space-borne remote sensors have to emphasize either the spatial resolution or the temporal resolution, but not both at the same time. The satellite imagery, as a consequence, cannot record the Earth's surface information simultaneously at a fine spatial resolution and a dense temporal frequency [1]–[3]. Acquisitions from medium-resolution sensors, such as the Landsat 8 Operational Land Imager (OLI) and the Terra Advanced Spaceborne Thermal Emission and Reflection Radiometer, have a sub-100-m spatial resolution, but their repeat cycles normally last longer than 10 days. In contrast, although low-resolution sensors, such as the Terra/Aqua Moderate Resolution Imaging Spectroradiometer (MODIS) and the NOAA Advanced Very High-Resolution Radiometer, deliver imagery at a daily basis, the spatial resolution of over hundreds or even thousands of meters cannot fully reflect the spatial details, especially over heterogeneous landscapes. To overcome the limitation, the concept of spatiotemporal data fusion has been put forward [4], [5]. This technique fuses satellite imagery from two sensors with similar spectral band specification, and the synthetic time series simultaneously keeps 1) finer spatial resolution of the two sensors and 2) integrated temporal resolution from the two sensors. It shows great potential to meet the increasing demand for observing and monitoring the Earth's surface at fine spatial and temporal scales [6]–[9].

Traditionally, spatiotemporal data fusion combines observations from two sensors with complementary spatial and temporal resolutions, one with fine spatial resolution but sparse revisit frequency (e.g., 16-day 30-m Landsat OLI/ETM+ imagery) and the other with dense frequency but coarse resolution (e.g., daily 500-m Terra/Aqua MODIS imagery), so as to integrate the advantageous resolution from the two sensors and synthesize the fine-resolution dense time series (e.g., daily 30-m Landsat-like imagery). Great achievements have been made in developing fusion methods over the past decade. Generally, the current models can be categorized into four groups: 1) weight-function-based (or filter-based) methods [10], [11]; 2) unmixing-based methods [12], [13]; 3) Bayesian-based methods [14], [15]; and 4) learning-based methods [16], [17]. The weight-function-based group applies a linear model to describe the relationship between multisource observations over pure coarse-resolution pixels, and further uses a weighting strategy to enhance the prediction over mixed pixels [18], [19]. Among this group, the spatial and temporal adaptive reflectance fusion model (STARFM) [5] is the most popular approach, and extended versions based on STARFM include the enhanced STARFM [20], the spatial and temporal nonlocal filter-based fusion model (STNLFFM) [21], and the fit-FC method [22]. The unmixing-based group uses the spatial unmixing technique for fusion, in which the

Manuscript received June 30, 2020; revised September 3, 2020 and September 24, 2020; accepted September 28, 2020. Date of publication October 1, 2020; date of current version October 22, 2020. This work was supported by the National Key R&D Program of China under Grant 2018YFB2100501. *(Corresponding author: Huanfeng Shen.)*

Jingan Wu, Huifang Li, and Xiaobin Guan are with the School of Resource and Environmental Sciences, Wuhan University, Wuhan 430072, China (e-mail: wujg@whu.edu.cn; huifangli@whu.edu.cn; guanxb@whu.edu.cn).

Qing Cheng is with the School of Computer Science, China University of Geosciences, Wuhan 430074, China (e-mail: qingcheng@whu.edu.cn).

Shuang Li is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430072, China (e-mail: sli@whu.edu.cn).

Huanfeng Shen is with the School of Resource and Environmental Sciences and the Collaborative Innovation Center for Geospatial Technology, Wuhan University, Wuhan 430072, China (e-mail: shenhf@whu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2020.3028116

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

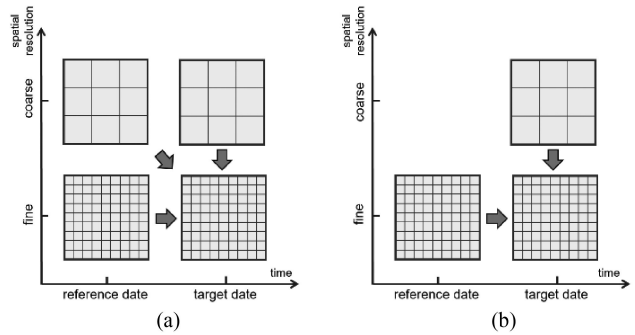


Fig. 1. Spatiotemporal fusion under the two input modalities. (a) Normal input modality. (b) Simplified input modality.

fine-resolution end members are estimated by unmixing the coarse-resolution pixels using the class fractions interpreted from the reference image [23]. Following the framework presented by Zhukov *et al.* [4], the improved models include the spatial and temporal reflectance unmixing model [12] and the flexible spatiotemporal data fusion model (FSDAF) [13]. In the Bayesian-based group, spatiotemporal fusion is considered as a maximum a posteriori (MAP) problem, and the fusion results are produced by maximizing the conditional probability relative to the input and output. For example, as an ideal mathematical framework, the total variation model has been employed for data fusion [10], [15]. The learning-based group builds the relationship between the input and output by utilizing the recent advances in machine learning, including sparse representation [24], [25], extreme learning machine [26], random forest [27], and convolutional neural network [28], [29]. In addition to the above four mainstream groups, several studies conduct the spatiotemporal fusion task by exploring other techniques such as wavelet transformation [30]. It is also worth noting that some researches exploit multisensor satellite imagery to generate homogeneous time series [31], [32] or map land cover types [33], [34]. Although the solutions differ from the models described earlier regarding the research task (e.g., land cover mapping [33]) and the input requirement (e.g., strictly requiring sensors to have similar spatial resolutions [31]), they are associated to the spatiotemporal fusion and represent a much broader research concept. Overall, owing to the ability to integrate the spatial and temporal resolutions from multiple sensors, the spatiotemporal data fusion models have been applied to solve a variety of missions, including monitoring crop growth conditions [35], [6], estimating carbon storage [36], [37], characterizing the urban thermal environment [38], [39], and mapping flood and wildfire events [40], [41].

It has become an implicit consensus that at least three images are required as input for spatiotemporal data fusion [see Fig. 1(a)], i.e., a coarse-resolution image on the target date and a pair of fine- and coarse-resolution images on the reference date, so as to synthesize the fine-resolution image on the target date. Nevertheless, a high-quality coarse-resolution image on the reference date cannot be collected in some application scenarios. As a result, the current fusion models, especially the popular ones, cannot accomplish the fusion task. On one hand, data from two sensors with barely overlapping temporal coverage cannot be fused by these models. For example, ideally, imagery from the Sentinel-2 MultiSpectral Instrument (MSI) and the Landsat 8 OLI can be synergistically used to produce 10-m time series at a nominal frequency of 2–3 days [42], [43]. But, the Landsat mission has a long revisit cycle of 16 days, and it, in most cases, cannot offer temporally matching counterparts to the collected reference Sentinel-2 images. Although an extended timespan between the reference and target dates helps to identify the matching image pairs, it has been reported to result in degraded fusion performance [21]. On the other hand, there are chances that although the coarse-resolution image on the reference date can be collected, it may suffer from degraded quality. For instance, in the MODIS–Landsat fusion applications, the collected MODIS image may have cloud contamination on the reference date [44], as clouds possibly move during the half-an-hour interval between MODIS and Landsat acquisitions. Also, the strong angular effect can be another factor to degrade the quality of MODIS images [45]. In such cases, using the low-quality observations as an input component would incorporate significant errors in the fused results. To solve the problem, a few studies have attempted to accomplish the spatiotemporal fusion with only two images as input, as illustrated in Fig. 1(b). For example, Fung *et al.* [46] proposed the Hopfield neural network spatiotemporal data fusion model, which supports the input of two images; and Wang *et al.* [42] and Shao *et al.* [47] specialized in fusing Landsat 8 and Sentinel-2 observations and developed the fusion models specifically for these two sensors.

For convenience, in this article, the fusion process with three input images [see Fig. 1(a)] is called "normal input modality" or "normal version," while the process with only two input images [see Fig. 1(b)] is called "simplified input modality" or "simplified version." In this article, we have proposed an effective strategy that can enable the spatiotemporal fusion under simplified input modality. To be specific, by combining with any spatiotemporal fusion model (e.g., STARFM), the proposed strategy can accomplish the fusion task in the scenarios where only two input images are available. Unlike the previous studies targeting specific sensor combinations [47], the strategy works with great universality. Besides, given that previous studies have never compared the fusion performance under the two input modalities, we have comprehensively tested and analyzed this point based on the proposed strategy in this article. In the rest of this article, we introduce the proposed method in Section II, describe the experimental setup and test dataset in Section III, present the experimental results and discussions in Section IV, and summarize the findings and contributions of this study in Section V.

II. Method

A. Basic Idea of the Proposed Method

The previous spatiotemporal fusion models (e.g., STARFM) are developed based on the normal input modality, i.e., requiring model input of a coarse-resolution image on the target date and a pair of fine- and coarse-resolution images on the reference

TABLE I
BAND SPECIFICATIONS OF TERRA/AQUA MODIS, LANDSAT 8 OLI, AND SENTINEL-2 MSI IN THE SIX BANDS ACROSS THE VISUAL, NEAR-INFRARED, AND SHORT-WAVE INFRARED SPECTRUM

Band	Wavelength (nm)		
	Aqua/Terra MODIS	Landsat 8 OLI	Sentinel-2 MSI
Blue	459–479	452–512	458–523
Green	545–565	533–590	543–578
Red	620–670	636–673	650–680
NIR	841–876	851–879	855–875
SWIR-1	1628–1652	1566–1651	1565–1655
SWIR-2	2105–2155	2107–2294	2100–2280

date. In this article, we aim to carry out the spatiotemporal fusion under the simplified input modality, i.e., using a coarse-resolution image on the target date and a fine-resolution image on the reference date as input. Given the technical sophistication and wide acceptance of the previous models, it would be an interesting idea to generate a simulated coarse resolution image on the reference date on our own, so as to use the previous models for the fusion task. Fortunately, the multisensor observations used in spatiotemporal fusion usually present high radiometric consistency, providing the basis for implementing the above idea. As described in [5], [20], the satellites imagery from multiple optical sensors used for data fusion are highly consistent and comparable. For instance, the three data sources in this article (Aqua/Terra MODIS, Landsat 8 OLI, and Sentinel-2 MSI) show a band-to-band correspondence in the six bands across the visible, near-infrared, and short-wave infrared spectrum, as reported in Table I. The similar bandwidth specifications further lead to a high level of radiometric consistency among the observations, as confirmed by previous studies [48], [49]. As a result, if the geolocation errors and slight radiometric differences are ignored, the coarse-resolution image can be approximately considered as a degraded observation of the fine-resolution image on the same date. In this case, if the high-quality coarse-resolution image on the reference date cannot be collected, we can generate a simulated image by imposing an image degradation process on the corresponding fine-resolution image. The simulated image is then fed into any existing fusion model (e.g., STARFM), along with the two input images, and the spatiotemporal fusion task can be completed.

B. Description of the Proposed Method

Before describing the details of the proposed method, some notations and definitions are given for convenience. F and C denote the fine-resolution and coarse-resolution observations, respectively; and t_k and t_p are the reference date and the target date, respectively. Under the normal input modality, in order to produce the fine-resolution image F_{t_p} on the target date, the coarse-resolution image C_{t_p} on the target date and the fine- and coarse- resolution image pair F_{t_k} and C_{t_k} on the reference date are needed as input. Given the fact that the coarse-resolution image C_{t_k} on the reference date is potentially unavailable in some application scenarios, the proposed method is aimed at implementing the fusion under the simplified input modality, i.e., with the two images F_{t_k} and C_{t_p} as input. To enable the simplified version of spatiotemporal fusion, a simulated image C'_{t_k} is derived as the analog of the potentially unavailable image C_{t_k}. As mentioned before, due to the similar band specifications between the two sensors to be fused, the multisensor observations show high overall consistency. Therefore, we can generate a simulated coarse-resolution image on the reference date C'_{t_k} by imposing an image degradation process on the fine-resolution image F_{t_k}. The image degradation model involves complicated steps to downsample the fine-resolution image to the coarse spatial resolution. In this article, we consider the two parts: pixel aggregation and image blurring. The pixel aggregation step is performed on the fine-resolution image F_{t_k}, so as to coordinate its spatial resolution to the coarse resolution. In the implementation, the fine-resolution pixels within the extent of each coarse-resolution pixel are averaged as the simulated coarse-resolution pixel [48]. Mathematically, the pixel aggregation step can be formulated as

$$C'_{t_k}\ (x,y) = \frac{1}{m}\ \times \sum_{i=1}^{m} F_{t_k}\left(x_i, y_i\right) \tag{1}$$

where (x, y) is the coordinate index of a coarse-resolution pixel, and (x_i, y_i) is the index of the ith fine-resolution pixel within the extent of the coarse-resolution pixel. m is the number of fine-resolution pixels within the coarse-resolution pixel. For the special cases that spatial resolution gap between the two sensors is not an integer, the fine-resolution pixels within a coarse-resolution pixel do not equally contribute to the aggregated coarse-resolution pixel. Instead, they are weighted according to their overlapping area in the extent of the coarse-resolution pixel. The image blurring step is also considered in the method. Note that although incorporating this step helps to simulate the image degradation process in the real scenarios, it may cause the over-smoothness of the simulated coarse-resolution image. Therefore, the inclusion of this step depends on experimental results. In this article, we applied a 3×3 Gaussian kernel with a standard deviation of 1.0 to blur the aggregated image, and test results suggested the image blurring step improved the fusion performance for MODIS–Landsat fusion, but decreased the performance for Landsat–Sentinel fusion. This is probably due to the remarkably different zoomed-in scale factor of spatial resolution (about 16 times vs. 3 times) in the two cases. In the following experiment, the image blurring step is employed for the MODIS–Landsat tests, but not for the Landsat–Sentinel tests.

Although the radiometric consistency is overall high, the spectral bandwidth shows slight differences between sensors, inevitably resulting in small observation bias between the multisensor observations. The modeling of some fusion methods relates to the observation bias. For example, STARFM assumes the small bias to be stable on the reference and target dates [5], and FSDAF assumes the bias to be reduced by applying a radiometric normalization procedure [13]. In this case, if we directly use the simulated image for fusion, theoretically, the small observation bias would exist on the target date, but not on the reference date. In order to satisfy the above assumptions relating to the observation bias, the coarse-resolution image on the target date should be adjusted to eliminate the observation

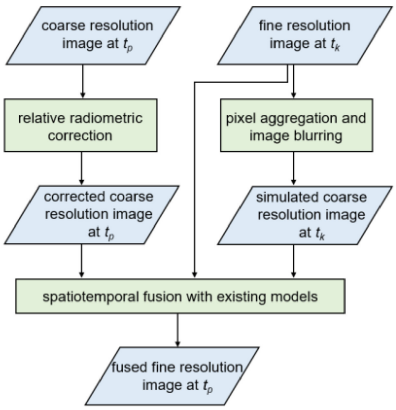


Fig. 2. Flowchart of the developed method.

bias. Therefore, a relative radiometric correction procedure is embedded in the proposed method. The procedure works in a local way. As reported in [48], [49], for a given pixel location, a linear relationship can be assumed between the different data sources:

$$C'_{t_p}(x,y) = a \times C_{tp}(x,y) + b \tag{2}$$

where C'_{t_p} represents the corrected coarse-resolution image. a and b are the slope and intercept coefficients, respectively, which characterize the systematic transformation between the two data sources. The two coefficients can be obtained by linearly regressing a collection of coincident observations from the two sensors. Note that the fine-resolution imagery has to be resized to the coarse resolution to eliminate the impact of the different spatial resolutions. In this article, for the MODIS–Landsat fusion, as the coefficients may vary slightly from locations, we used all the collected coincident image pairs over the study region to estimate the coefficients and applied a 5×5 window for each location to ensure the local adjustment. For the Landsat–Sentinel fusion, the barely overlapping acquisition dates of the two satellites led to a small number of coincident observations over the study region, which significantly reduced the robustness of the estimated coefficients. Thus, the coefficients in Zhang *et al.* [49] based on a continental-scale assessment were employed. After performing the radiometric correction, the fine-resolution image F_{t_k}, the simulated coarse-resolution image C'_{t_k}, and the corrected coarse-resolution image C'_{t_p} were fed into the spatiotemporal fusion model to produce the synthetic image on the target date. The flowchart of the developed method is shown in Fig. 2.

C. Three Spatiotemporal Fusion Models

In practical applications, the proposed strategy has to be combined with an existing spatiotemporal fusion model to accomplish the fusion task. The three popular fusion models, namely, STARFM, STNLFFM, and FSDAF, were employed in the research. This section briefly introduces them. For more details, please refer to Gao *et al.* [5], Cheng *et al.* [21], and Zhu *et al.* [13].

1) STARFM: In STARFM, it is assumed that the small multisensor observation bias remains stable over a pure coarse-resolution pixel during the acquisition interval. Accordingly, an unknown fine-resolution pixel on the target date can be estimated as the sum of the corresponding coarse-resolution pixels on the same date and the observation bias on the reference date. To enhance the prediction accuracy over mixed pixels, the neighboring similar pixels of the same land-cover type are combined based on a weighting scheme. The STARFM model is mathematically depicted as

$$F_{t_p}(x,y) = \sum_{j=1}^{n} w(x_j,y_j) \times \left[C_{t_p}(x_j,y_j) + F_{t_k}(x_j,y_j) - C_{t_k}(x_j,y_j)\right] \tag{3}$$

where (x,y) and (x_j,y_j) are the coordinate indices of the central pixel and the jth neighboring similar pixel around the central one. n is the number of similar pixels for prediction. $w(x_j,y_j)$ is the weight of the jth similar pixel, which is determined by the spectral similarity, the temporal difference, and the spatial distance between the jth similar pixel and the central pixel.

2) STNLFFM: STNLFFM is an improved version of STARFM. It follows the basic framework of STARFM, but presents differences in the fundamental assumption, the weight assignment, and the procedure of identifying similar pixels. In STNLFFM, it is assumed that the temporal change of land surface between the multitemporal observations is constant for the different sensors, and thus the model derived from the coarse-resolution sensor can be applied to the fine-resolution one. The similar pixel information is also integrated to ease the influence of the spatial resolution gap. Mathematically, the STNLFFM model is formulated as

$$F_{t_p}(x,y) = \sum_{j=1}^{n} w(x_j,y_j) \times \left[g(x_j,y_j) \times F_{t_k}(x_j,y_j) + h(x_j,y_j)\right] \tag{4}$$

where (x,y) and (x_j,y_j) are in the same definition as in STARFM. $g(x_j,y_j)$ and $h(x_j,y_j)$ are the linear coefficients characterizing the temporal change of the jth similar pixel, and they can be estimated by linearly regressing the similar pixels recorded in the two coarse-resolution images. $w(x_j,y_j)$ is the weight of the jth similar pixel.

3) FSDAF: FSDAF combines the unmixing-based framework and the weighted-function-based framework to capture both the gradual and abrupt land-cover type changes. It generates the fused results based on the three steps as follows:

1) The temporal changes recorded by the coarse-resolution pixels are unmixed to obtain the fine-resolution temporal change of each class.
2) The residuals in the first step are distributed to the fine resolution for improving the fusion accuracy.
3) The neighboring similar pixels are integrated to eliminate the block artifacts and enhance the prediction robustness.

The FSDAF model is mathematically described as

$$F_{t_p}(x,y) = F_{t_k}(x,y) + \sum_{j=1}^{n} w(x_j,y_j) \times [\Delta F(c) + r(x_j,y_j)] \quad (5)$$

where (x,y) and (x_j,y_j) are in the same definition as in STARFM. $\Delta F(c)$ is the temporal change of the land-cover class c, noting that the pixel at (x_j,y_j) belongs to the class c. $r(x_j,y_j)$ is the residual distributed to the jth similar pixel. $w(x_j,y_j)$ is the weight of the jth similar pixel, which is measured based on the spatial distance between the target pixel and the similar pixel.

III. Experimental Setup and Test Data

A. Experimental Setup

To validate the effectiveness of the proposed strategy, we compare the fusion performances of the same model under the two input modalities. If the simplified version achieves comparable or even superior performance to the normal version, the validity of the method can be confirmed. To be specific, in each test, we collect the four images, i.e., the two image pairs on the reference date and the target date, respectively. The coarse-resolution image on the target date and the image pair on the reference date is used as model input: the normal version uses all the three images as input, while the simplified version uses only two input images. The fused results under two input modalities are evaluated against the observed fine-resolution image on the target date to reveal the fusion performance.

In the following experiments, the two kinds of scenarios are investigated. In the first scenario (i.e., scenario 1), the high-quality coarse-resolution image can be collected on the reference date. In the second scenario (i.e., scenario 2), we can collect the coarse-resolution image on the reference date, but the collected image is of low quality, i.e., suffering from cloud contamination or significant inconsistency as compared to the coincident fine-resolution image due to the mismatched temporal coverage. The difference between the two scenarios is that the collected coarse-resolution images on the reference dates are of good quality in scenario 1, while they are of low quality in scenario 2. In this article, we conducted 38 tests in total, among which 26 tests (i.e., 24 MODIS–Landsat tests and 2 Landsat–Sentinel tests) are grouped into scenario 1, and 12 tests (i.e., 2 MODIS–Landsat tests and 10 Landsat–Sentinel tests) are grouped into scenario 2.

B. Data Description

Two fusion choices were investigated in this article: MODIS–Landsat fusion and Landsat–Sentinel fusion. Detailed descriptions of the experimental materials are given as follows.

1) Fusion of Terra MODIS and Landsat 5 TM Observations: The time-series dataset provided by Emelyanova *et al.* [50], which has been widely used in data fusion research [13], [16], was employed in this article. The study area is located in the Lower Gwydir Catchment in northern New South Wales, Australia. The dataset is made up of 14 pairs of coincident MODIS–Landsat images, and the details on the data locations and acquisition dates are given in Table II. All the Landsat images were acquired by the Thematic Mapper (TM) instrument onboard Landsat 5 and were atmospherically corrected by the algorithm developed by Li *et al.* [51]. The MODIS images were collected from the MOD09GA Collection 6 daily reflectance products and were reprojected and resampled to the projection and spatial resolution of the Landsat observations. Two subset regions were used for the experiments, with each region covering 1000 × 1000 pixels at the spatial resolution of the Landsat observations. In each region, 14 MODIS–Landsat image pairs were ordered chronologically, and the two adjacent pairs formed a dataset for the follow-up tests, with the former date as the reference and the latter date as the target. As a result, 26 tests in total were performed in the two regions. By visually checking the test materials, we found that in most tests, the MODIS images on the reference dates showed good quality, but the MODIS image on August 6, 2004, was partly contaminated by cloud cover (the Landsat image on the same date is cloud-free). According to the criterion to categorize the two scenarios, the two tests with August 6, 2004, as the reference date were grouped into scenario 2, while the rest 24 tests were grouped into scenario 1.

TABLE II
Spatial Location and Acquisition Dates of the Test Materials

	Data location	Acquisition dates
MODIS-Landsat fusion	MODIS H/V: 31/11 Landsat path/row: 91/80	16 Apr 2004; 02 May 2004; 05 July 2004; 06 Aug 2004; 22 Aug 2004; 25 Oct 2004; 26 Nov 2004; 12 Dec 2004; 28 Dec 2004; 13 Jan 2005; 29 Jan 2005; 14 Feb 2005; 02 Mar 2005; 03 Apr 2005
Landsat-Sentinel fusion	Landsat path/row: 123/37 Sentinel-2 tile number: T50SKB	24 July 2017 (L: 26 July 2017); 15 Sep 2017 (L: 12 Sep 2017) ; 30 Oct 2017 (L: 30 Oct 2017); 19 Dec 2017 (L: Dec 17 2017)

2) Fusion of Landsat 8 OLI and Sentinel-2 MSI Observations: The study site is located at Zhumadian, Henan province, China. Due to the 16-day revisit cycle of the Landsat 8 satellite, normally, we cannot find the matching counterparts on the same date as the collected Sentinel-2 MSI observations. For example, only one matching pair without cloud cover was collected in 2017 if the strict condition was applied. In order to obtain more materials, the image pairs acquired with an interval of fewer than 3 days were considered acceptable. Four cloud-free image pairs in total were collected for 2017, and the acquisition dates of the test images are shown in Table II. The Landsat observations were collected as surface reflectance products. The Sentinel-2 observations were downloaded as Level 1C product, and they were atmospherically corrected by the Sen2Cor plugin developed by the European Space Agency to derive surface reflectance products. As in the previous case, two subset regions were used for the tests, with each region covering 1000 × 1000 pixels at a 10-m spatial resolution. In each region, the four collected image pairs were ordered chronologically, and

any two pairs were combined as a test dataset, with the former date as the reference and the latter date as the target. In total, we have performed 12 tests over the two regions. A detailed visual examination indicated that the image pairs from July and September presented significant differences characterized as phenological or land-cover changes occurring during the 3-day interval, and thus the Landsat images from July and September should be considered as low-quality matching counterparts to the Sentinel-2 images. According to the criterion to categorize the two scenarios, the 10 tests using the image pairs in July and September as reference were grouped into scenario 2, while the rest two tests were group into scenario 1.

It should be noted that the SWIR bands of Sentinel-2 images are at the 20-m spatial resolution, and ideally, they should be downscaled to 10 m before performing the spatiotemporal fusion. But given the fact that applying the downscaling method would introduce additional uncertainty, we directly fused and evaluated the SWIR bands at the 20-m spatial resolution in this article. This helps to directly reveal the fusion performance without incorporating effects from the downscaling step.

C. Model Parameter Setup and Computational Time

Three spatiotemporal fusion models, namely, STARFM, STNLFFM, and FSDAF, were employed in the experimental analysis, due to their wide acceptance and program availability. In this article, the model parameters were optimized through trial-and-error tests. For the MODIS–Landsat fusion, the window sizes in STARFM, STNLFFM, and FSDAF were set to 21, 41, and 41, respectively. The cluster number in STARFM and FSDAF was set to 7 based on a visual interpretation. For the Landsat–Sentinel fusion, due to the close spatial resolutions, the window sizes were set small, as 7, 13, and 13 for the three models, respectively. The cluster number in STARFM and FSDAF was set to 8.

The computational time varies among the three models, since they are implemented with different programming environment. But notice that, for the same model, 1) the simplified version usually requires about two more seconds than the normal version, due to the additional processing steps of image degradation and radiometric correction (e.g., for a MODIS–Landsat fusion test based on STNLFFM, normal version vs. simplified version: 51 vs. 53 s); and 2) the computational efficiency is higher for the sensor combination choice with close spatial resolutions than that with considerably contrasting resolutions, because a smaller window size can be set for the sensor combination choice with closer spatial resolutions (e.g., for normal version of STNLFFM, a MODIS–Landsat fusion test vs. a Landsat-Sentinel fusion test: 51 vs. 7 s).

D. Quality Evaluation of the Fused Results

The fused results are evaluated against the collected fine-resolution images on the target dates visually and quantitatively. The visual comparison checks the consistency between the fused results and the observed images. The quantitative assessment is performed by adopting the five widely used metrics: the mean absolute error (MAE), the root-mean-square error (RMSE), the correlation coefficient (CC), the spectral angle mapper (SAM), and the structural similarity index (SSIM). Among the five measures, MAE and RMSE measure the overall radiometric difference between the simulated image and the observed image, CC shows their degree of correlation, SAM assesses the spectral distortion of the fused result from the spectral fidelity aspect, and SSIM shows the similarity of the spatial structures between the fusion result and the observed image. The ideal values of MAE, RMSE, SAM, CC, and SSIM are 0, 0, 0, 1, and 1, respectively.

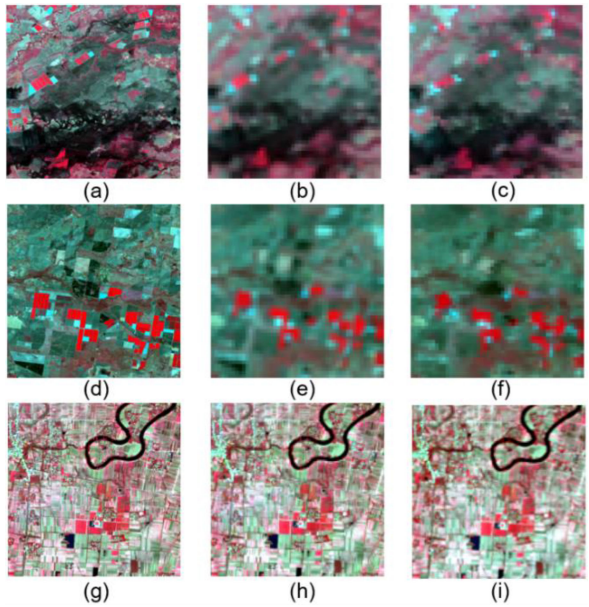


Fig. 3. Observed and simulated coarse-resolution images (NIR, red, and green bands as RGB). (a)–(c) Observed Landsat image, the simulated MODIS image, and the observed MODIS image on December 28, 2004, respectively. (d)–(f) Images on March 2, 2005, displayed in the same order as the previous case. (g)–(i) Observed Sentinel image, the simulated Landsat image, and the observed Landsat image on October 30, 2017, respectively.

IV. Experimental Results and Discussion

A. Comparison Between the Simulated and Observed Coarse-Resolution Images

Are the simulated coarse-resolution images on the reference dates highly consistent with the corresponding observed images? This is the basis of determining whether or not the proposed strategy can be applied. The collected images, including the MODIS–Landsat pairs and the Landsat–Sentinel pairs, were used for the validation. By applying the proposed strategy, we can generate a simulated coarse-resolution image, and the simulated images are then evaluated against the observed ones. As described in Section III-B, the observed coarse-resolution images in some datasets suffered from cloud contamination or land surface changes, so they were unsuitable as the evaluation standard and excluded from the evaluation. Fig. 3 illustrates the results of three tests, including two MODIS–Landsat pair tests and one Landsat–Sentinel pair test. The acquisition dates of the three image pairs are December 28, 2004; March 2, 2005; and October 30, 2017, respectively. The results indicate that, overall,

TABLE III
QUANTITATIVE ASSESSMENT OF THE MODIS–LANDSAT TEST AND THE LANDSAT–SENTINEL TEST

Sensors	Band	MAE	RMSE	CC	Intercept	Slope	p-value
Terra MODIS/ Landsat 5 TM	Blue	0.0103	0.0128	0.8776	1.0023	0.0085	<0.0001
	Green	0.0090	0.0127	0.8711	1.0016	0.0036	<0.0001
	Red	0.0117	0.0163	0.8904	0.9830	0.0065	<0.0001
	NIR	0.0187	0.0288	0.9256	1.0382	−0.0039	<0.0001
	SWIR-1	0.0214	0.0293	0.9069	1.0023	0.0037	<0.0001
	SWIR-2	0.0205	0.0292	0.8937	0.9427	0.0136	<0.0001
Landsat 8 OLI/ Sentinel-2 MSI	Blue	0.0113	0.0145	0.7353	0.8651	0.0186	<0.0001
	Green	0.0079	0.0106	0.8629	0.9900	0.0052	<0.0001
	Red	0.0094	0.0126	0.9058	1.0739	−0.0036	<0.0001
	NIR	0.0096	0.0141	0.9305	1.0264	−0.0044	<0.0001
	SWIR-1	0.0155	0.0215	0.9470	1.0319	0.0032	<0.0001
	SWIR-2	0.0231	0.0285	0.9345	1.0710	0.0074	<0.0001

TABLE IV
QUANTITATIVE EVALUATION OF THE THREE TESTS IN SCENARIO 1

		MAE	RMSE	SAM	CC	SSIM
Fig. 4	Normal	**0.0167**	**0.0228**	4.7066	**0.8743**	**0.8891**
	Simplified	0.0176	0.0240	**4.3794**	0.8714	0.8795
Fig. 5	Normal	**0.0187**	**0.0259**	4.8921	0.7694	**0.8797**
	Simplified	0.0195	0.0264	**4.6644**	**0.7834**	0.8754
Fig. 6	Normal	0.0167	0.0224	4.3465	**0.8244**	0.9152
	Simplified	**0.0163**	**0.0223**	**4.2107**	0.8090	**0.9241**

TABLE V
AVERAGED QUANTITATIVE RESULTS AMONG THE 24 MODIS–LANDSAT FUSION TESTS IN SCENARIO 1

		MAE	RMSE	SAM	CC	SSIM
STARFM	Normal	0.0219	0.0298	6.9325	0.7982	0.8538
	Simplified	**0.0207**	**0.0278**	**6.4411**	**0.8131**	**0.8575**
STNLFFM	Normal	0.0196	0.0265	5.8736	0.8258	**0.8726**
	Simplified	**0.0193**	**0.0260**	**5.8232**	**0.8293**	0.8695
FSDAF	Normal	0.0206	0.0278	6.2536	0.8114	0.8682
	Simplified	**0.0197**	**0.0263**	**5.9745**	**0.8231**	**0.8715**

TABLE VI
QUANTITATIVE EVALUATION OF THE THREE TESTS IN SCENARIO 2

		MAE	RMSE	SAM	CC	SSIM
Fig. 8	Normal	0.0204	0.0267	5.8807	0.9103	0.8723
	Simplified	**0.0120**	**0.0158**	**3.9609**	**0.9558**	**0.9706**
Fig. 9	Normal	0.0309	0.0423	4.8884	0.7174	0.8440
	Simplified	**0.0230**	**0.0312**	**3.8206**	**0.7528**	**0.8986**
Fig. 10	Normal	0.0291	0.0379	7.6075	0.6456	0.8442
	Simplified	**0.0186**	**0.0248**	**4.4018**	**0.7568**	**0.9152**

TABLE VII
AVERAGED RESULTS OF THE QUANTITATIVE EVALUATION AMONG THE 10 LANDSAT–SENTINEL FUSION TESTS IN SCENARIO 2

		MAE	RMSE	SAM	CC	SSIM
STARFM	Normal	0.0286	0.0380	7.3798	0.6929	0.8421
	Simplified	**0.0214**	**0.0293**	**5.7017**	**0.7160**	**0.8940**
STNLFFM	Normal	0.0280	0.0376	7.0010	0.6721	0.8231
	Simplified	**0.0213**	**0.0291**	**5.5060**	**0.7024**	**0.8726**
FSDAF	Normal	0.0280	0.0372	6.9877	0.6994	0.8523
	Simplified	**0.0209**	**0.0284**	**5.5477**	**0.7318**	**0.9033**

the simulated coarse-resolution images [see Fig. 3(b), (e), and (h)] present close radiometric characteristics and spatial patterns to the observed coarse-resolution images [see Fig. 3(c), (f), and (i)], confirming their high similarity. It should be noted that, although the observed coarse-resolution images are employed as the evaluation standard, they are not perfect standards as there exist slight intrinsic spectral differences due to bandwidth differences between two sensors. Therefore, the simulated images present similar spectral visual effects to the observed one, but not completely the same.

In addition, the simulated images were quantitatively assessed against the observed ones by the measures of MAE, RMSE, and CC. The linear fit between the simulated images and the observed images was also developed for all the included tests in the two sensor combination choices by using ordinary least-squares regression. The evaluation results are listed in Table III. In the MODIS–Landsat case, the simulated MODIS images are highly correlated with the observed MODIS images, with CC values ranging from 0.8711 to 0.9256 in the six bands. The linear fit between the simulated and observed MODIS images falls around the ideal 1:1 line, and the regression in six bands is highly significant (p-value < 0.0001), revealing the goodness of fit. In the Landsat–Sentinel case, all bands except for the blue one show high consistency. The blue band presents a certain degree of inconsistency, with a lower CC value of 0.7353 and a departure of the linear fit from the ideal 1:1 line. Possibly, this is caused by the bits of thin clouds in the scenes, and the blue band is more vulnerable to the cloud effect than the other bands due to its shorter wavelength. Overall, the proposed strategy can produce a simulated coarse-resolution image that is consistent and comparable to the observed one, providing the chances to adopt the simplified versions of spatiotemporal fusion.

B. Evaluation in Scenario 1

This section is aimed at comparing the spatiotemporal fusion performance under the two input modalities in scenario 1, in which the high-quality coarse-resolution image can be collected on the reference date. We have conducted 26 tests in total, including 24 MODIS–Landsat tests and two Landsat–Sentinel tests. In each test, we generated the fused results under the two input modalities, respectively. As the three models were employed, in each test, we generated six fused results totally (i.e., three results under the normal input modality and three results under the simplified input modality).

To visually assess the fusion performance, the results of three tests are shown in Figs. 4–6, including two MODIS–Landsat

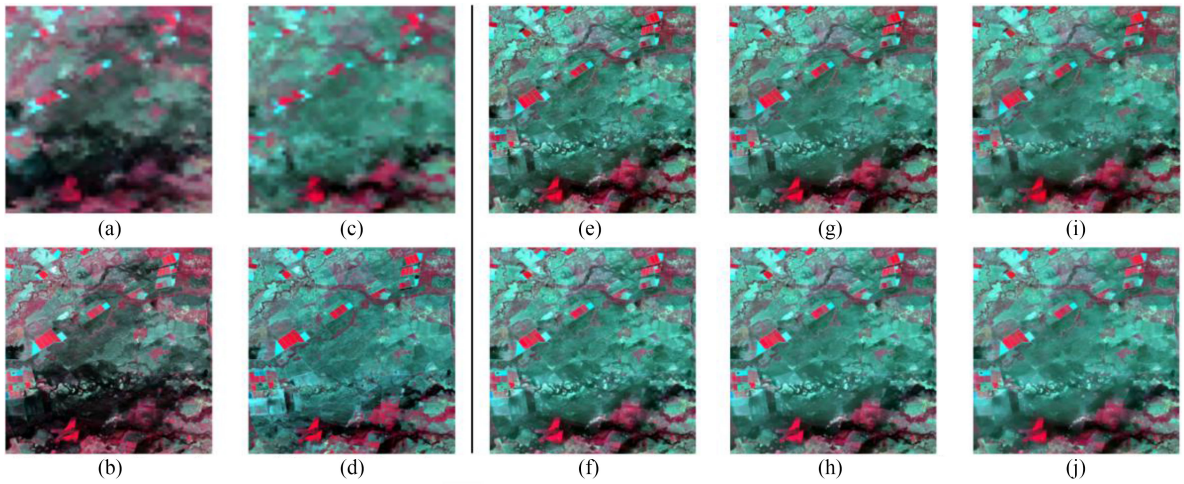


Fig. 4. Test data and fused results in the MODIS–Landsat fusion test (NIR, red, and green bands as RGB). The target date and the reference date are January 13, 2005, and December 28, 2004, respectively. (a)–(d) Observed MODIS and Landsat images on the reference date and the target date, with (d) used as the evaluation standard. (e)–(j) Fused results under the two input modalities. (a) MODIS (reference). (b) Landsat (reference). (c) MODIS (target). (d) Landsat (target). (e) STARFM (normal). (f) STARFM (simplified). (g) STNLFFM (normal). (h) STNLFFM (simplified). (i) FSDAF (normal). (j) FSDAF (simplified).

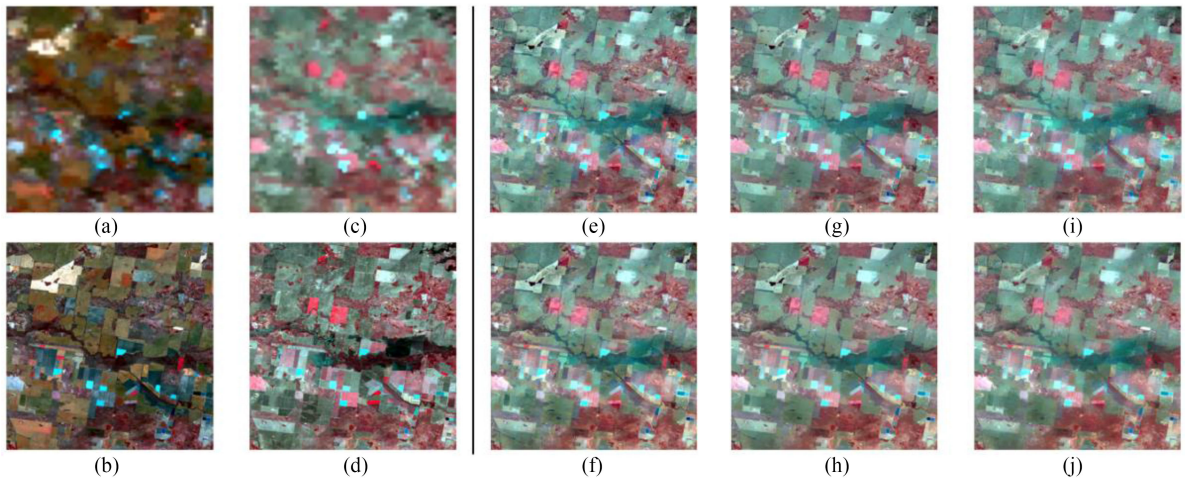


Fig. 5. Test data and fused results in the MODIS–Landsat fusion test (NIR, red, and green bands as RGB). The target date and the reference date are November 26, 2004 and October 25, 2004, respectively. (a)–(d) Observed MODIS and Landsat images on the reference date and the target date, with (d) used as the evaluation standard. (e)–(j) Fused results under the two input modalities. (a) MODIS (reference). (b) Landsat (reference). (c) MODIS (target). (d) Landsat (terget). (e) STARFM (normal). (f) STARFM (simplified). (g) STNLFFM (normal). (h) STNLFFM (simplified). (i) FSDAF (normal). (j) FSDAF (simplified).

tests and one Landsat–Sentinel test. The target dates in the three tests are January 13, 2005 (using December 28, 2004, as the reference date), November 26, 2004 (using October 25, 2004, as the reference date), and December 19, 2017 (using October 30, 2017, as the reference date). The visual comparison in Fig. 4 reveals that as compared with the observed Landsat image in Fig. 4(d), the fused results in Fig. 4(e)–(j) are generally consistent, but with slight spectral distortion. The fused images under the simplified input modality in Fig. 4(f), (h), and (j) closely resemble those under the normal input modality in Fig. 4(e), (g), and (i), indicating that the simplified version of spatiotemporal fusion can generate comparable results to the normal version. The visual comparison in Figs. 5 and 6 reveals the same finding that in scenario 1, the fused images under the simplified input modality are comparable to the results under the normal input modality. By taking STNLFFM as an example, Table IV reports the quantitative results of the three tests. In the former two tests, the normal version slightly outperforms the simplified version, while in the latter test, the simplified version performs slightly better. But, in general, the differences between the two input modalities are insignificant, as the values of quantitative measures vary in a very small extent.

The quantitative evaluation was performed on the 24 MODIS–Landsat fusion tests to comprehensively assess the fusion performance in scenario 1. By taking the STNLFFM model as an example, Fig. 7 shows the MAE values and the SSIM values of the fused images among the 24 tests. It can be found that, in more than half of these tests, the normal version of spatiotemporal fusion obtains lower MAE values and higher SSIM values, as compared with the simplified version, but, on average, the

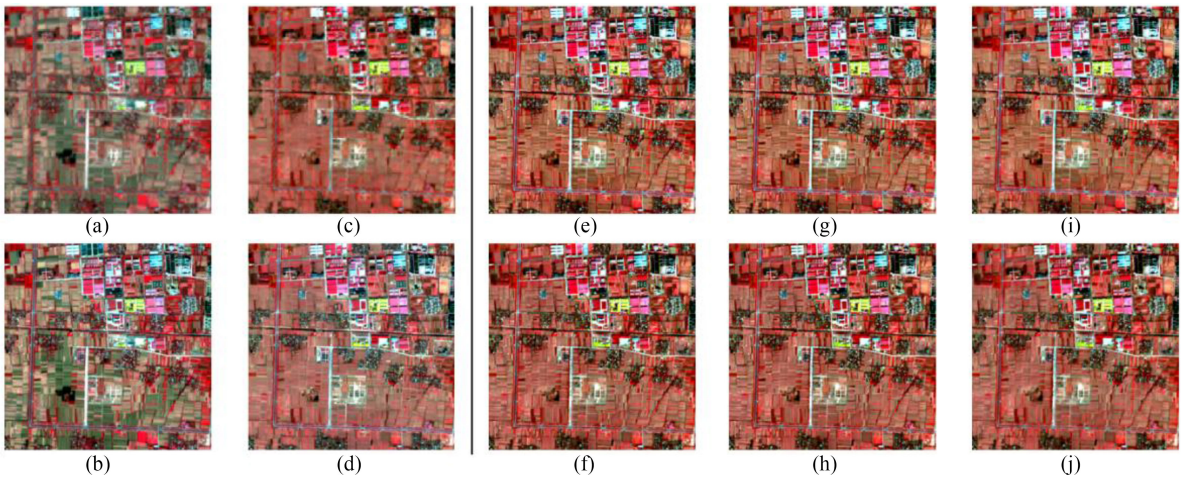


Fig. 6. Test data and fused results in the Landsat–Sentinel fusion test (NIR, red, and green bands as RGB). The target date and the reference date are December 19, 2017 and October 30, 2017, respectively. (a)–(d) Observed Landsat and Sentinel-2 images on the reference date and the target date, with (d) used as the evaluation standard. (e)–(j) Fused results under the two input modalities. (a) Landsat (reference). (b) Sentinel (reference). (c) Lansat (target). (d) Sentinel (target). (e) STARFM (normal). (f) STARFM (simplified). (g) STNLFFM (normal). (h) STNLFFM (simplified). (i) FSDAF (normal). (j) FSDAF (simplified).

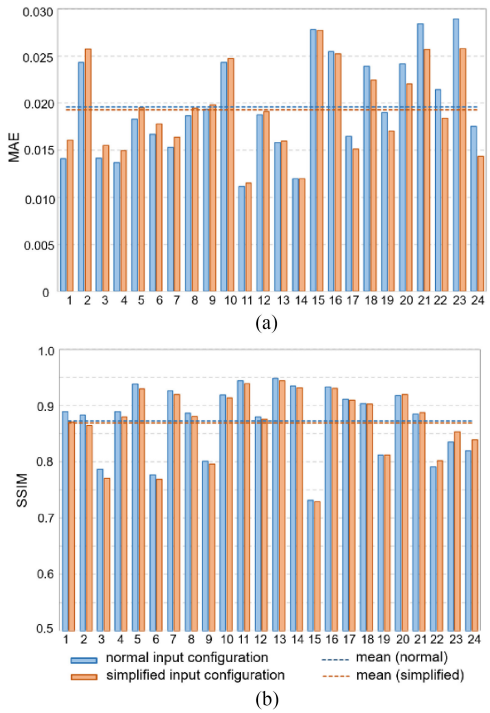


Fig. 7. Quantitative descriptions in terms of MAE and SSIM in the 24 MODIS–Landsat fusion tests. The STNLFFM model is taken as an example. The dashed line shows the mean value among the 24 tests. The order of the 24 tests in the horizontal axis is reorganized: the closer to the left part, the more the normal version outperforms the simplified version, and vice versa.

simplified versions show comparable performance to the normal versions (MAE: 0.0196 vs. 0.0193; SSIM: 0.8726 vs. 0.8695). Table V lists the averaged quantitative descriptions among the 24 tests. Overall, a similar tendency is shown in the three models, i.e., although less input data is used, the simplified version of spatiotemporal fusion obtains even slightly better performance than the normal version. For example, by adopting the simplified version, SAM is decreased by 0.0503–0.4914, and CC is increased by 0.0035–0.0149. As a result, it can be concluded that by using the proposed strategy, the fusion performance with only two input images is comparable to or even slightly superior to that with three input images.

C. Evaluation in Scenario 2

This section is aimed at comparing the fusion performance under the two input modalities in scenario 2, in which although we can get the coarse-resolution image on the reference date, the collected data is of low quality. In total, the 12 tests have been conducted, including two Landsat–MODIS tests and 10 Landsat–Sentinel tests. As in the previous section, we generated the fused results under the two input modalities and then assessed the quality of fused results visually and quantitatively.

Figs. 8–10 display the data in the three tests, including one MODIS–Landsat test and two Landsat–Sentinel tests. In Fig. 8, the observed MODIS image on the reference date is partly contaminated by clouds, and thus it is considered low-quality input. In the two Landsat–Sentinel tests shown in Figs. 9 and 10, due to the 16-day revisit cycle of the Landsat mission, we cannot have the Landsat observations acquired exactly on the reference dates. Instead, the Landsat images on the neighboring dates are collected. However, the collected Landsat images present distinct inconsistency to the corresponding Sentinel-2 images on the reference dates [e.g., see the regions marked in yellow in Fig. 9(a) and (b)], which is characterized as the land surface change due to the mismatched acquisition dates, and thus, they are also considered low-quality input. As the low-quality coarse-resolution image on the reference date is incorporated as the input component, the normal version of spatiotemporal fusion yields fused results with significant errors. For instance, as compared with the observed Sentinel-2 image in Fig. 9(d), the fused images under the normal input modality show distinct spectral distortion. Conversely, the simplified version of

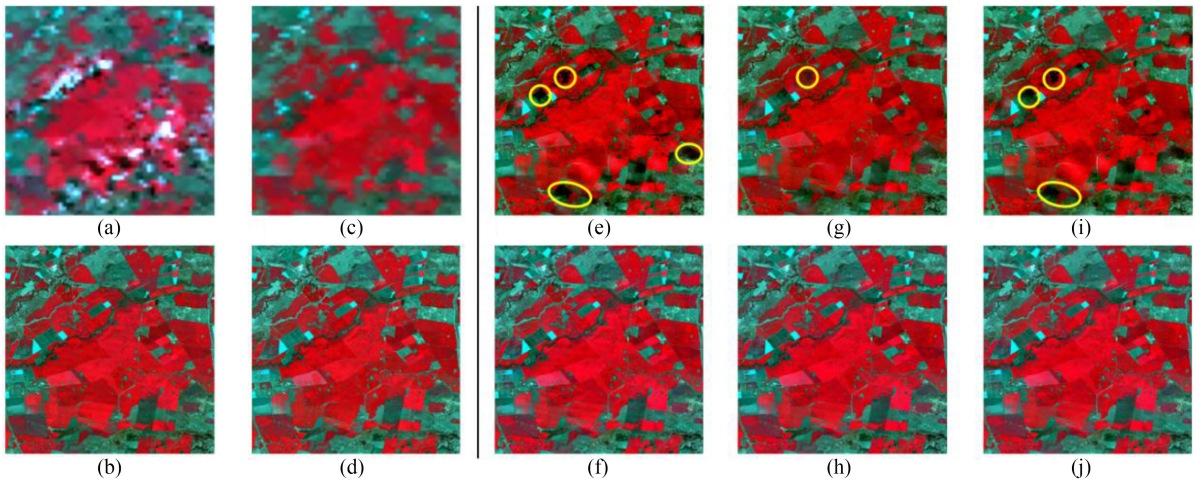


Fig. 8. Test data and fused results in the MODIS–Landsat fusion test (NIR, red, and green bands as RGB). The target date and the reference date are August 22, 2004 and August 6, 2004, respectively. (a)–(d) Observed MODIS and Landsat images on the reference date and the target date, with (d) used as the evaluation standard. (e)–(j) Fused results under the two input modalities. (a) MODIS (reference). (b) Landsat (reference). (c) MODIS (target). (d) Landsat (target). (e) STARFM (normal). (f) STARFM (simplified). (g) STNLFFM (normal). (h) STNLFFM (simplified). (i) FSDAF (normal). (j) FSDAF (simplified).

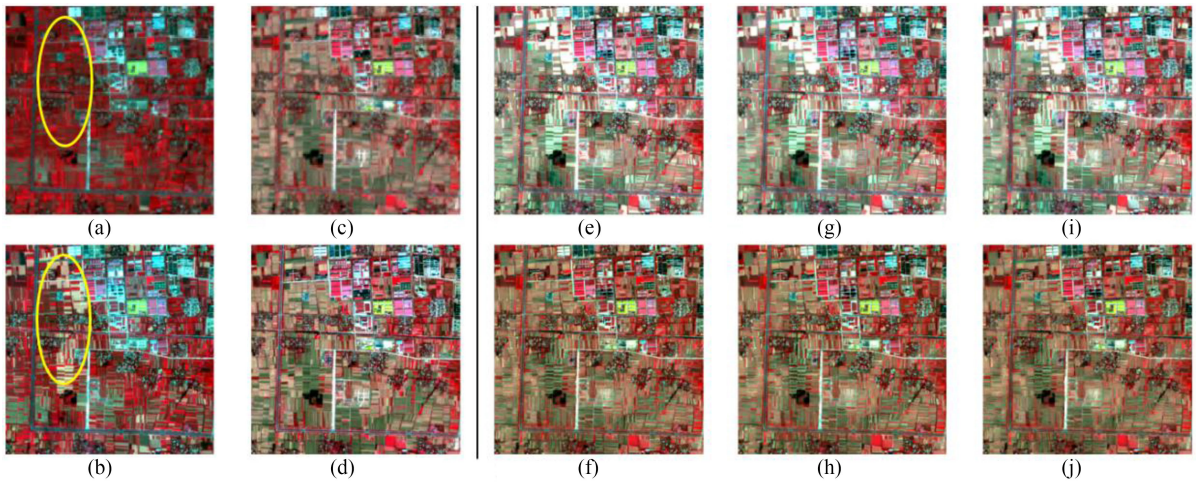


Fig. 9. Test data and fused results in the Landsat-Sentinel fusion test (NIR, red, and green bands as RGB). The target date and the reference date are October 10, 2017, and September 15, 2017, respectively. (a) Observed Landsat image on September 15, 2017. (b) Observed Sentinel image on September 12, 2017. (c)–(d) Observed Landsat and Sentinel images on October 10, 2017, with (d) used as the evaluation standard. (e)–(j) Fused results under the two input modalities. (a) Landsat (reference). (b) Sentinel (reference). (c) Landsat (target). (d) Sentinel (target). (e) STARFM (normal). (f) STARFM (simplified).(g) STNLFFM (normal). (h) STNLFFM (simplified). (i) FSDAF (normal). (j) FSDAF (simplified).

spatiotemporal fusion avoids the above problem by using only two input images, and accordingly, the results [see Fig. 9(f), (h), and (j)] generally conform with the observed fine-resolution image on the target date [see Fig. 9(d)]. By taking FSDAF as an example, Table VI reports the quantitative descriptions under the two input modalities in the three tests. It can be found that in the three tests, the simplified version of spatiotemporal fusion outperforms the normal version, with significantly lower MAE values and higher CC values.

The quantitative evaluation was performed for the 10 Landsat–Sentinel fusion tests. By taking the FSDAF model as an example, Fig. 11 shows the MAE values and the SSIM values of the fused images among the 10 tests. It is shown that in almost all the tests, the simplified version obtains higher accuracy than the normal version, with lower MAE values and higher SSIM values, and the average values among the 10 tests report the precision differences are significant between the two cases (MAE: 0.0280 vs. 0.0209; SSIM: 0.8523 vs. 0.9033). Table VII lists the averaged quantitative descriptions among the 10 tests. We can find that the simplified version of spatiotemporal fusion obtains considerably higher accuracy than the normal version. For example, as compared with the normal version, the simplified version decreases SAM by 1.4400–1.6781 and increases CC by 0.0231–0.0324. This is in accordance with the findings from the visual comparison. Therefore, we can summarize that, for the scenario in which we cannot collect the high-quality coarse-resolution image on the reference date, the fused images show considerable differences under the two input

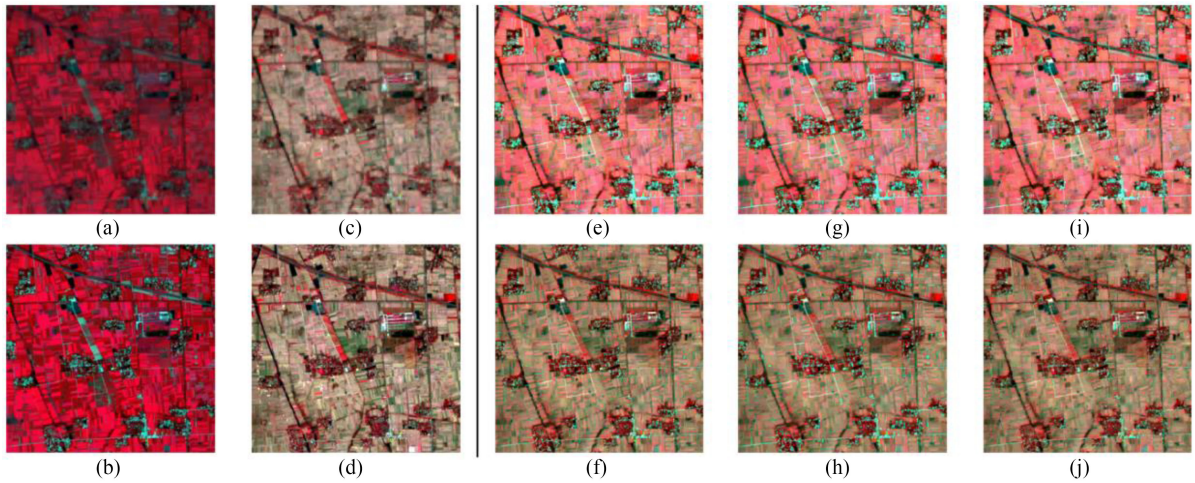


Fig. 10. Test data and fused results in the Landsat–Sentinel fusion test (NIR, red, and green bands as RGB). The target date and the reference date are October 10, 2017 and July 24, 2017, respectively. (a) Observed Landsat image on July 24, 2017. (b) Observed Sentinel image on July 26, 2017. (c)–(d) Observed Landsat and Sentinel images on October 10, 2017, with (d) used as the evaluation standard. (e)–(j) Fused results under the two input modalities. (a) Landsat (reference). (b) Sentinel (reference). (c) Landsat (target). (d) Sentinel (target). (e) STARFM (normal). (f) STARFM (simplified). (g) STLFFM (normal). (h) STNLFFM (simplified). (i) FSDAF (normal). (j) FSDAF (simplified).

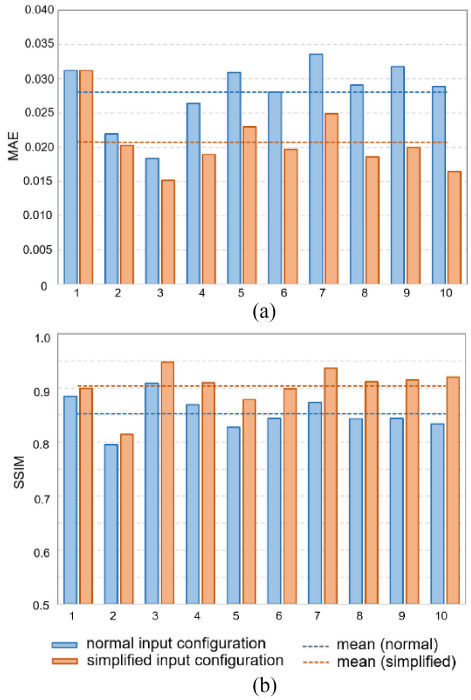


Fig. 11. Quantitative descriptions in terms of MAE and SSIM in the 10 Landsat–Sentinel fusion tests. The FSDAF model is taken as an example. The dashed line shows the mean value of the 10 tests. The order of the 10 tests in the horizontal axis is reorganized: the closer to the right part in the horizontal axis, the more the simplified version outperforms the normal one, and vice versa.

modalities, and the results derived under the simplified case are significantly superior to those under the normal case.

D. Discussion

1) Sources of Spectral Errors in the Fused Images: The spectral errors in the fused images are a key issue in the spatiotemporal fusion, and this problem exists for both the normal version and the simplified version. As we described before, even though the multiple sensors show similar band specifications, the multisource data inevitably have observation differences and the differences would slightly vary with acquisition dates due to factors such as illumination and viewing conditions. Hence, for the normal version of spatiotemporal fusion, the basic assumptions that the observation differences remain stable would not be completely satisfied, resulting in the spectral errors in the fused results. As for the simplified version, two sources potentially lead to spectral errors. First, the degradation model from a fine spatial resolution to a coarse spatial resolution is very complex in real scenarios. Although we have considered the image downsampling and blurring steps, the estimated image may still contain slight spectral errors as compared with the coarse-resolution image observed in the same scene. Second, we introduce a radiometric correction step to improve the consistency between the data sources. This step helps to characterize the systematic transformation, but as the observation differences slightly vary with acquisition dates, there would exist random uncertainty for the corrected observations for a given date. These two sources of spectral errors would be accumulated in the fused results under the simplified input modality.

It is noteworthy that even the proposed strategy cannot completely eliminate spectral errors, according to the experimental results in Section IV-B, it is already enough to obtain comparable or even slightly superior performance as compared with the normal version of spatiotemporal fusion.

2) Potential Applications of the Proposed Method: The proposed method can be combined with the existing spatiotemporal fusion models, especially the popular ones such as STARFM, to accomplish the fusion for the scenarios in which only two input images are available. The applications of the method are twofold. Most importantly, it can be used to fuse data from sensors with barely overlapping temporal coverage, such as Landsat 8 OLI

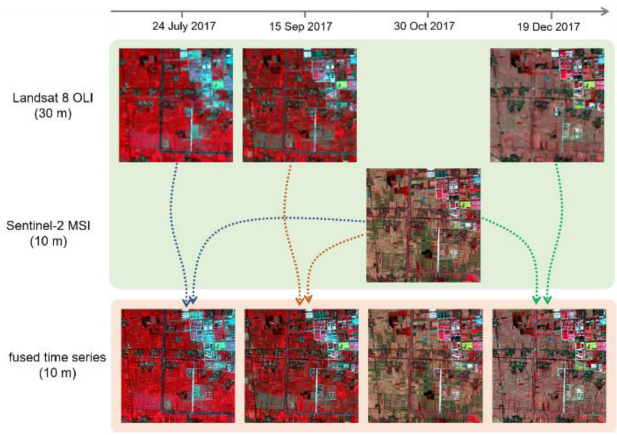


Fig. 12. Generation of 10-m time series by fusing observations from Landsat 8 OLI and Sentinel-2 MSI (NIR, red, and green bands as RGB).

and Sentinel-2 MSI. In this case, few coincident observations can be identified as reference image pairs, on account of the 16-day revisit cycle of the Landsat mission and the negative effect of cloud cover. By exploiting the proposed method, we can fuse the data from such two sensors under the simplified input modality. Besides, the proposed method can benefit some special cases. For example, although a coarse-resolution image can be acquired on the reference date, it may suffer from cloud covers or strong angular effects. In these cases, the proposed method avoids the problem of inputting the degraded image and can be expected to achieve satisfactory fusion performance.

It should be emphasized that the prerequisite for adopting the proposed method for fusion is that high consistency is revealed between the multisensor observations. The land surface reflectance products were investigated in this article. As the MODIS, Landsat, and Sentinel-2 land surface reflectance products are highly comparable and consistent, the proposed method can be applied. Comprehensive validation of the data consistency should be considered in the first place before extending the proposed approach to other sensors and quantitative products.

3) Generation of Time-Series Data: The spatiotemporal fusion technique is aimed at generating a fine-resolution dense time series by fusing a coarse-resolution dense time series with a fine-resolution sparse time series. To demonstrate the ability of spatiotemporal fusion for generating the time series, we provide an example of fusing observations from Landsat 8 OLI and Sentinel-2 MSI in Fig. 12. The 10-m Sentinel-2 image on October 30 and the three 30-m Landsat 8 images on July 24, September 15, and December 19, are collected as input. The three Landsat images are fused with the Sentinel image on October 30, respectively, by using the proposed strategy and FSDAF, and produced the synthetic 10-m images. The derived time series provide denser observations than a single data source, representing the enhanced ability to capture the temporal change of land surface. Besides, the 10-m spatial resolution of the synthetic time series reveals a strong capacity to characterize the spatial details of Earth's surface, especially over heterogeneous landscapes.

V. Conclusion

Traditionally, spatiotemporal data fusion requires at least three input images, i.e., a coarse-resolution image on the target date and a pair of fine- and coarse-resolution images on the reference date. However, new application scenarios call for efforts to conduct spatiotemporal fusion with only two input images, i.e., a coarse-resolution image on the target date and a fine-resolution image on the reference date. In this article, we developed a universal method that can be used to accommodate the existing fusion models requiring three input images for the scenarios in which only two input images are available. Based on the developed method, we comprehensively compared and assessed the spatiotemporal fusion performance under the two input modalities. According to the experimental results, by applying the proposed strategy, the fusion performance with only two input images is comparable or even superior to that with three input images. The findings in this article challenge the stereotype that spatiotemporal fusion strictly requires at least three input images. Furthermore, the proposed method extends the potential applications of the existing fusion models, especially the popular ones, such as STARFM, and allows us to adapt these models for new scenarios. It is also worth noting that most of the existing spatiotemporal fusion models were considered unable to fuse the data from sensors with barely overlapping temporal coverage, such as Landsat 8 OLI and Sentinel-2 MSI, while our study indicates these models can be adapted by the method proposed in this article to accomplish the fusion task.

Acknowledgment

The authors appreciate the Editors and anonymous Reviewers for their valuable suggestions.

References

[1] H. K. Zhang, B. Huang, M. Zhang, K. Cao, and L. Yu, "A generalization of spatial and temporal fusion methods for remotely sensed surface parameters," *Int. J. Remote Sens.*, vol. 36, no. 17, pp. 4411–4445, 2015.

[2] X. Zhu, F. Cai, J. Tian, and T. K.-A. Williams, "Spatiotemporal fusion of multisource remote sensing data: Literature survey, taxonomy, principles, applications, and future directions," *Remote Sens.*, vol. 10, no. 4, 2018, Art. no. 527.

[3] M. Belgiu and A. Stein, "Spatiotemporal image fusion in remote sensing," *Remote Sens.*, vol. 11, no. 7, 2019, Art. no. 818.

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

[5] 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.

[6] F. Gao *et al.*, "Toward mapping crop progress at field scales through fusion of Landsat and MODIS imagery," *Remote Sens. Environ.*, vol. 188, pp. 9–25, 2017.

[7] A. O. Onojeghuo, G. A. Blackburn, J. Huang, D. Kindred, and W. Huang, "Applications of satellite 'hyper-sensing' in Chinese agriculture: Challenges and opportunities," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 64, pp. 62–86, 2018.

[8] P. Ghamisi *et al.*, "Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state-of-the-art," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 1, pp. 6–39, Mar. 2019.

[9] Y. Tang, Q. Wang, K. Zhang, and P. M. Atkinson, "Quantifying the effect of registration error on spatio-temporal fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 487–503, Jan. 2020.

[10] P. Wu, H. Shen, T. Ai, and Y. Liu, "Land-surface temperature retrieval at high spatial and temporal resolutions based on multi-sensor fusion," *Int. J. Digit. Earth*, vol. 6, pp. 113–133, 2013.
[11] Q. Wang, Y. Tang, X. Tong, and P. M. Atkinson, "Virtual image pair-based spatio-temporal fusion," *Remote Sens. Environ.*, vol. 249, 2020, Art. no. 112009.
[12] C. M. Gevaert and F. J. García-Haro, "A comparison of STARFM and an unmixing-based algorithm for Landsat and MODIS data fusion," *Remote Sens. Environ.*, vol. 156, pp. 34–44, 2015.
[13] X. Zhu, E. H. Helmer, F. Gao, D. Liu, J. Chen, and M. A. Lefsky, "A flexible spatiotemporal method for fusing satellite images with different resolutions," *Remote Sens. Environ.*, vol. 172, pp. 165–177, 2016.
[14] L. Liao, J. Song, J. Wang, Z. Xiao, and J. Wang, "Bayesian method for building frequent landsat-like NDVI datasets by integrating MODIS and landsat NDVI," *Remote Sens.*, vol. 8, no. 6, 2016, Art. no. 452.
[15] H. Shen, X. Meng, and L. Zhang, "An integrated framework for the spatio–temporal–spectral fusion of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7135–7148, Dec. 2016.
[16] H. Song, Q. Liu, G. Wang, R. Hang, and B. Huang, "Spatiotemporal satellite image fusion using deep convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 3, pp. 821–829, Mar. 2018.
[17] Z. Tan, L. Di, M. Zhang, L. Guo, and M. Gao, "An enhanced deep convolutional model for spatiotemporal image fusion," *Remote Sens.*, vol. 11, no. 24, 2019, Art. no. 2898.
[18] 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.
[19] Q. Weng, P. Fu, and F. Gao, "Generating daily land surface temperature at Landsat resolution by fusing Landsat and MODIS data," *Remote Sens. Environ.*, vol. 145, pp. 55–67, 2014.
[20] X. Zhu, J. Chen, F. Gao, X. Chen, and J. G. Masek, "An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions," *Remote Sens. Environ.*, vol. 114, no. 11, pp. 2610–2623, 2010.
[21] 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.
[22] Q. Wang and P. M. Atkinson, "Spatio-temporal fusion for daily Sentinel-2 images," *Remote Sens. Environ.*, vol. 204, pp. 31–42, 2018.
[23] A. Minghelli-Roman, M. Mangolini, M. Petit, and L. Polidori, "Spatial resolution improvement of MeRIS images by fusion with TM images," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 7, pp. 1533–1536, Jul. 2001.
[24] 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.
[25] 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.
[26] X. Liu, C. Deng, S. Wang, G. Huang, B. Zhao, and P. Lauren, "Fast and accurate spatiotemporal fusion based upon extreme learning machine," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 2039–2043, Dec. 2016.
[27] Y. Ke, J. Im, S. Park, and H. Gong, "Downscaling of MODIS one kilometer evapotranspiration using Landsat-8 data and machine learning approaches," *Remote Sens.*, vol. 8, no. 3, 2016, Art. no. 215.
[28] X. Liu, C. Deng, J. Chanussot, D. Hong, and B. Zhao, "StfNet: A two-stream convolutional neural network for spatiotemporal image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6552–6564, Sep. 2019.
[29] Y. Li, J. Li, L. He, J. Chen, and A. Plaza, "A new sensor bias-driven spatio-temporal fusion model based on convolutional neural networks," *Sci. China Inf. Sci.*, vol. 63, no. 1674-733X, 2020, Art. no. 140302.
[30] F. W. Acerbi-Junior, J. G. P. W. Clevers, and M. E. Schaepman, "The assessment of multi-sensor image fusion using wavelet transforms for mapping the Brazilian Savanna," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 8, no. 4, pp. 278–288, 2006.
[31] Y. T. Solano-Correa, F. Bovolo, and L. Bruzzone, "Generation of homogeneous VHR time series by nonparametric regression of multisensor bitemporal images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7579–7593, Oct. 2019.
[32] L. G. Denaro and C. Lin, "Hybrid canonical correlation analysis and regression for radiometric normalization of cross-sensor satellite imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 976–986, Feb. 2020.
[33] P. Benedetti, D. Ienco, R. Gaetano, K. Ose, R. G. Pensa, and S. Dupuy, "M^3 fusion: A deep learning architecture for multiscale multimodal multitemporal satellite data fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 12, pp. 4939–4949, Dec. 2018.
[34] H. Sheng, X. Chen, J. Su, R. Rajagopal, and A. Y. Ng, "Effective data fusion with generalized vegetation index: Evidence from land cover segmentation in agriculture," 2020. [Online]. Available: https://doi.org/10.1109/CVPRW50498.2020.00038
[35] J. Amorós-López *et al.*, "Multitemporal fusion of Landsat/TM and ENVISAT/MERIS for crop monitoring," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 23, pp. 132–141, 2013.
[36] T. Dong *et al.*, "Estimating winter wheat biomass by assimilating leaf area index derived from fusion of Landsat-8 and MODIS data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 49, pp. 63–74, 2016.
[37] X. Guan, H. Shen, X. Li, W. Gan, and L. Zhang, "Climate control on net primary productivity in the complicated mountainous area: A case study of yunnan, China," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 12, pp. 4637–4648, Dec. 2018.
[38] H. Liu and Q. Weng, "Enhancing temporal resolution of satellite imagery for public health studies: A case study of West Nile Virus outbreak in Los Angeles in 2007," *Remote Sens. Environ.*, vol. 117, pp. 57–71, 2012.
[39] H. Shen, L. Huang, L. Zhang, P. Wu, and C. Zeng, "Long-term and fine-scale satellite monitoring of the urban heat island effect by the fusion of multi-temporal and multi-sensor remote sensed data: A 26-year case study of the city of Wuhan in China," *Remote Sens. Environ.*, vol. 172, pp. 109–125, 2016.
[40] F. Zhang, X. Zhu, and D. Liu, "Blending MODIS and Landsat images for urban flood mapping," *Int. J. Remote Sens.*, vol. 35, no. 9, pp. 3237–3253, 2014.
[41] R. Ghosh, P. K. Gupta, V. Tolpekin, and S. K. Srivastav, "An enhanced spatiotemporal fusion method—Implications for coal fire monitoring using satellite imagery," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 88, 2020, Art. no. 102056.
[42] Q. Wang *et al.*, "Fusion of Landsat 8 OLI and Sentinel-2 MSI data," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3885–3899, Jul. 2017.
[43] M. Claverie *et al.*, "The harmonized landsat and sentinel-2 surface reflectance data set," *Remote Sens. Environ.*, vol. 219, pp. 145–161, 2018.
[44] H. Shen, J. Wu, Q. Cheng, M. Aihemaiti, C. Zhang, and Z. Li, "A spatiotemporal fusion based cloud removal method for remote sensing images with land cover changes," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 3, pp. 862–874, Mar. 2019.
[45] D. Xie, F. Gao, L. Sun, and M. Anderson, "Improving spatial-temporal data fusion by choosing optimal input image pairs," *Remote Sens.*, vol. 10, no. 7, 2018, Art. no. 1142.
[46] H. C. Fung, S. M. Wong, and W. P. Chan, "Spatio-temporal data fusion for satellite images using hopfield neural network," *Remote Sens.*, vol. 11, no. 18, 2019, Art. no. 2077.
[47] Z. Shao, J. Cai, P. Fu, L. Hu, and T. Liu, "Deep learning-based fusion of Landsat-8 and Sentinel-2 images for a harmonized surface reflectance product," *Remote Sens. Environ.*, vol. 235, 2019, Art. no. 111425.
[48] M. Feng *et al.*, "Global surface reflectance products from Landsat: Assessment using coincident MODIS observations," *Remote Sens. Environ.*, vol. 134, pp. 276–293, 2013.
[49] H. K. Zhang *et al.*, "Characterization of Sentinel-2A and Landsat-8 top of atmosphere, surface, and nadir BRDF adjusted reflectance and NDVI differences," *Remote Sens. Environ.,* vol. 215, pp. 482–494, 2018.
[50] I. V. Emelyanova, T. R. McVicar, T. G. Van Niel, L. T. Li, and A. I. J. M. van Dijk, "Assessing the accuracy of blending Landsat–MODIS surface reflectances in two landscapes with contrasting spatial and temporal dynamics: A framework for algorithm selection," *Remote Sens. Environ.*, vol. 133, pp. 193–209, 2013.
[51] F. Li *et al.*, "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.

Jingan Wu received the B.S. degree in geographic information system from Anhui University, Hefei, China, in 2015. He is currently working toward the Ph.D. degree in cartography and geographic information engineering at the School of Resource and Environmental Sciences, Wuhan University, Wuhan, China.

His research interests include spatiotemporal data fusion and missing information reconstruction of remote sensing images.

Qing Cheng received the B.S. degree in the geographic information system and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2010 and 2015, respectively.

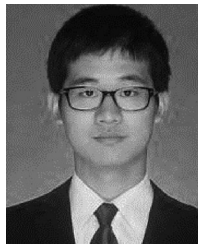
She is currently an Associate Professor with the School of Computer Science, China University of Geosciences, Wuhan, China. Her research interests include remote sensing data reconstruction, data fusion, and urban remote sensing.

Huifang Li (Member, IEEE) received the B.S. degree in the geographic information science from China University of Mining and Technology, Xuzhou, China, in 2008, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2013.

She is currently an Associate Professor with the School of Resource and Environmental Sciences, Wuhan University. Her research interests include radiometric correction of remote sensing images, including cloud correction, shadow correction, and urban thermal environment analysis and alleviation.

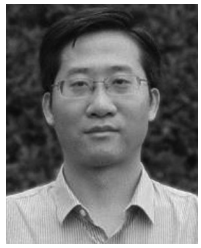
Shuang Li received the Ph.D. degree in cartography and geographic information engineering from Wuhan University, Wuhan, China, in 2010.

She is currently an Associate Professor with the School of Remote Sensing and Information Engineering, Wuhan University. Her research interest includes remote sensing image processing.



Xiaobin Guan received the B.S. and Ph.D. degrees in geographical information system from the School of Resource and Environmental Sciences, Wuhan University, Wuhan, China, in 2013 and 2018, respectively.

He is currently a Postdoctoral Research Assistant with the School of Resource and Environmental Sciences, Wuhan University. His research interests include the processing of multisource remote-sensing images and its application in the terrestrial ecosystem and global change.



Huanfeng Shen (Senior Member, IEEE) received the B.S. degree in surveying and mapping engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2002 and 2007, respectively.

In 2007, he was with the School of Resource and Environmental Sciences (SRES), Wuhan University, where he is currently a Luojia Distinguished Professor and an Associate Dean with SRES. He was or is the PI of two projects supported by the National Key Research and Development Program of China, and six projects supported by the National Natural Science Foundation of China. He has authored more than 100 research papers in peer-reviewed international journals. His research interests include remote sensing image processing, multisource data fusion, and intelligent environmental sensing.

Dr. Shen is a Council Member of China Association of Remote Sensing Application, Education Committee Member of Chinese Society for Geodesy Photogrammetry and Cartography, and Theory Committee Member of Chinese Society for Geospatial Information Society. He is currently a member of the Editorial Board of *Journal of Applied Remote Sensing* and *Geography and Geo-information Science*.