remote sensing

MDPI

Article

Spatiotemporal Fusion of Formosat-2 and Landsat-8 Satellite Images: A Comparison of "Super Resolution-Then-Blend" and "Blend-Then-Super Resolution" Approaches

Tee-Ann Teo * and Yu-Ju Fu

Department of Civil Engineering, National Yang Ming Chiao Tung University, No. 1001, Daxue Rd., East District, Hsinchu City 300, Taiwan; yuju.cv06g@nctu.edu.tw
* Correspondence: tateo@mail.nctu.edu.tw

Abstract: The spatiotemporal fusion technique has the advantages of generating time-series images with high-spatial and high-temporal resolution from coarse-resolution to fine-resolution images. A hybrid fusion method that integrates image blending (i.e., spatial and temporal adaptive reflectance fusion model, STARFM) and super-resolution (i.e., very deep super resolution, VDSR) techniques for the spatiotemporal fusion of 8 m Formosat-2 and 30 m Landsat-8 satellite images is proposed. Two different fusion approaches, namely Blend-then-Super-Resolution and Super-Resolution (SR)-then-Blend, were developed to improve the results of spatiotemporal fusion. The SR-then-Blend approach performs SR before image blending. The SR refines the image resampling stage on generating the same pixel-size of coarse- and fine-resolution images. The Blend-then-SR approach is aimed at refining the spatial details after image blending. Several quality indices were used to analyze the quality of the different fusion approaches. Experimental results showed that the performance of the hybrid method is slightly better than the traditional approach. Images obtained using SR-then-Blend are more similar to the real observed images compared with images acquired using Blend-then-SR. The overall mean bias of SR-then-Blend was 4% lower than Blend-then-SR, and nearly 3% improvement for overall standard deviation in SR-B. The VDSR technique reduces the systematic deviation in spectral band between Formosat-2 and Landsat-8 satellite images. The integration of STARFM and the VDSR model is useful for improving the quality of spatiotemporal fusion.

Keywords: time-series satellite images; image fusion; deep learning; STARFM; VDSR

check for updates

Citation: Teo, T.-A.; Fu, Y.-J. Spatiotemporal Fusion of Formosat-2 and Landsat-8 Satellite Images: A Comparison of "Super Resolution-Then-Blend" and "Blend-Then-Super Resolution" Approaches. *Remote Sens.* **2021**, *13*, 606. https://doi.org/10.3390/rs13040606

Academic Editors: Yanchen Bo, Jie Cheng and Xin Tao

Received: 7 January 2021
Accepted: 5 February 2021
Published: 8 February 2021

1. Introduction

1.1. Motivation

Time-series satellite images are the integration of multitemporal images over a region, and they can be used to analyze spatial temporal variations of the Earth's surface. Owing to the availability of remote sensing open data, we have more data sources to construct time-series satellite images. Examples of such data sources are time-series Landsat satellite images provided by the National Aeronautics and Space Administration and time-series Sentinel satellite images of the European Space Agency (Paris, France). Furthermore, commercial satellites, such as satellite constellations of Planet Labs, can also provide high-temporal-resolution time-series satellite images. The increase in the number of available time-series satellite images has also led to the emergence of more diversified applications for the images, such as vegetation phenology detection [1], water resource management [2], rice crop estimation [3], land cover change detection [4], and regional air quality [5]. In particular, time-series satellite image analysis plays an important role in the application of satellite imagery.

Spatiotemporal fusion methodology has the capability to generate both high-spatial and high-temporal resolution images by employing different sensors. It can improve our ability and flexibility to construct time-series satellite images. As shown in Figure 1, the present study collected satellite images with low cloud cover, taken in 2014 and 2015 over Kaohsiung City, Taiwan. We calculated the normalized difference vegetation index (NDVI) of the center pixel of the images (Figure 1a). We found that if we collected only the time-series data of Landsat-8 (LS-8) NDVI (Figure 1b), there was only one cloud-free dataset between days of year (DOYs) 100 and 200 in these two years, and the dataset was insufficient to determine land cover changes. In fact, the agricultural monitoring usually requires more dateset than city growth. Therefore, additional time-series data from Formosat-2 (FS-2) NDVI (Figure 1c) were added to increase the number of datasets, which led to the number of available datasets from DOYs 100 to 200 increasing to six (Figure 1d). This example shows the importance of integrating different sensors to construct time-series satellite images. However, these two sensors have different spatial resolutions (i.e., 30 m for LS-8 and 8 m for FS-2) and hence, their images cannot be integrated directly. Therefore, spatiotemporal image fusion is required to integrate the FS-2 and LS-8 images.

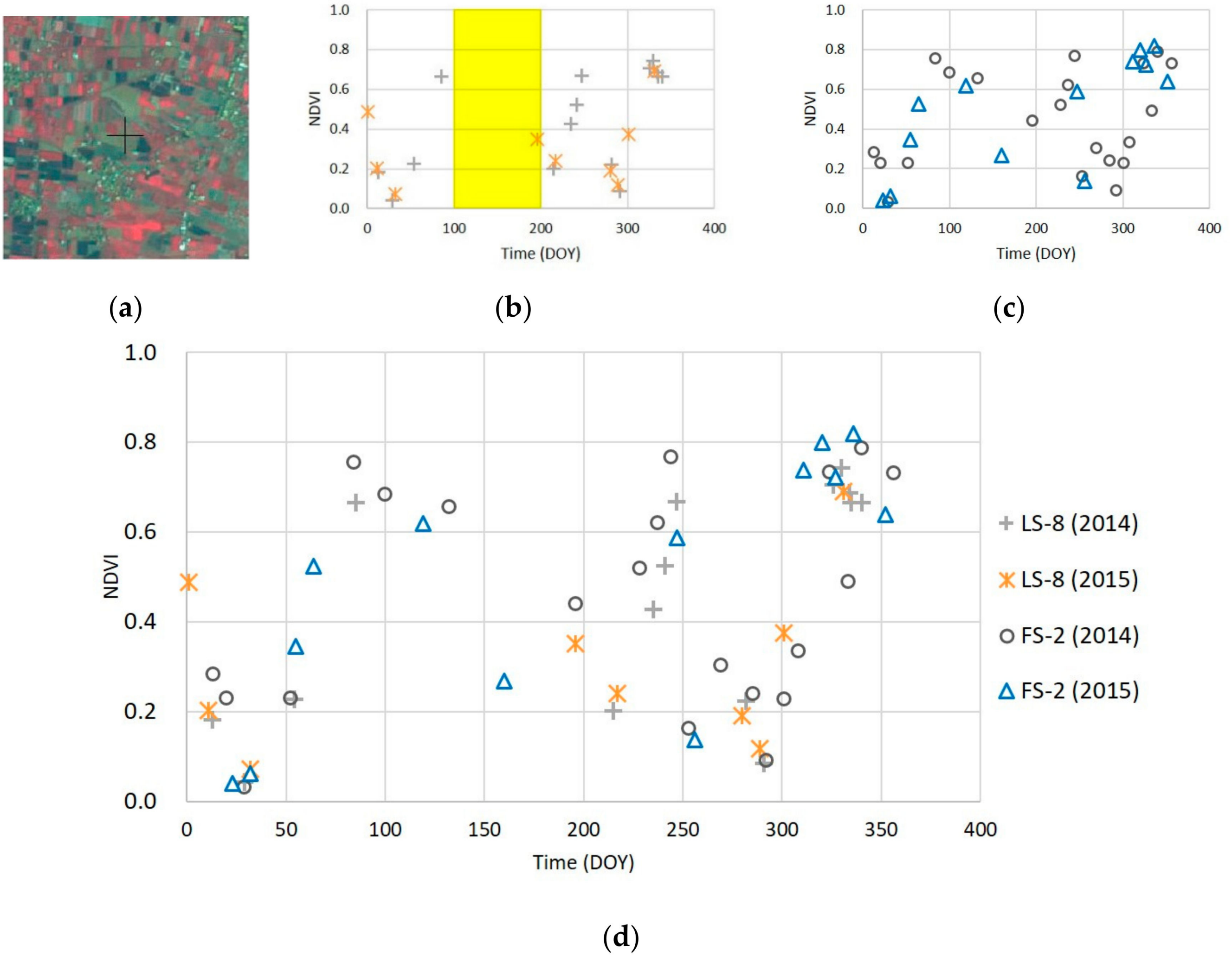


Figure 1. Time-series normalized difference vegetation index (NDVI) data for 2014 and 2015: (**a**) location of the image center for NDVI calculation, (**b**) time-series Landsat-8 NDVI, (**c**) time-series Formosat-2 NDVI, and (**d**) combined time-series NDVI.

In recent years, deep learning has been widely used in the field of image processing. It is a feature-learning method that combines many simple modules to approximate a high-level complex system. With the combination of a large number of simple modules, very complex functions can be learned [6] for modeling complex scenery. Li et al. [7] discussed the fusion of infrared and visible images by using a convolutional neural network (CNN). An advantage of CNNs is their ability to handle complex nonlinear mapping functions and feature extraction at different scales. The result [7] revealed that CNN shows better performance in improving image quality in the image fusion process. Thus, CNNs have high potential for use in spatiotemporal fusion technology for remote sensing images. Most previous studies have discussed feature extraction from CNNs [8,9]. However, relatively few studies have discussed CNN algorithms for the spatiotemporal fusion of satellite images. Hence, the present study performed a detailed investigation of CNN-based spatiotemporal fusion.

1.2. Previous Studies

Spatiotemporal fusion technology fuses images with high temporal and low spatial resolutions and images with low temporal and high spatial resolutions to produce time-series satellite imagery. An example is the spatiotemporal fusion of MODIS and Landsat images using the spatial and temporal adaptive reflectance fusion model (STARFM) [10]. Furthermore, spatiotemporal fusion technology can be applied to combine images recorded by satellites with similar spatial resolutions, such as the fusion of LS-8 and Sentinel-2 satellite images [11]. The advantage is that time-series satellite images with a consistent spatial resolution can be generated from LS-8 and Sentinel-2 satellite images. Multisensor image fusion simulates high-spatial-resolution time-series images for periods for which only low-spatial-resolution images are available. This image fusion is a key technology for generating time-series satellite images from images acquired by different sensors at different times [12].

From the perspective of time-series data technique, spatiotemporal image fusion can be classified into five categories: unmixing-based, weight-function-based, Bayesian-based, learning-based, and hybrid methods [13,14].

The weight-function-based method has been widely used in image fusion applications [15]. Gao et al. [10] proposed the STARFM, which is the first and the most popular weighted fusion model. It estimates the reflectance of a predicted image by weighing temporal, spectral, and spatial information. It assumes that spatial and temporal changes in a high-spatial-resolution image are the same as those in a low-spatial-resolution image, and consequently, high-spatial-resolution images can be estimated from low-spatial-resolution images. However, its basic assumption renders it unfit for heterogeneous regions. Many improved fusion models have been proposed, such as spatial and temporal adaptive algorithm for mapping reflectance change [16], enhanced STARFM (ESTARFM) [17], and spatial and temporal nonlocal filter-based fusion model (STNLFFM) [15].

The learning-based method uses a machine learning algorithm to establish a nonlinear relationship between the observed and the estimated images. It predicts the high-spatial-resolution image of an observed low-spatial-resolution image. While this method is usually applied to image super-resolution (SR), it is also used for image fusion. Concepts such as dictionary pair learning [18], sparse representation [19], artificial neural network (ANN) [20], deep convolutional neural network [21], and nonlinear mapping CNN combined with SR CNN [22] are applied to determine the nonlinear conversion relationship between low-spatial- and high-spatial-resolution images.

A hybrid method involves the integration of two or more fusion methods. Examples of such methods are flexible spatiotemporal data fusion [13], spatiotemporal remotely sensed images and land cover map fusion model [23], combination of the spatial and temporal reflectance unmixing model (STRUM) [24] with an unmixing-based and weight-function-based method to fuse images, and the unmixing-based Bayesian model [25] integrated with Bayesian-based and unmixing-based methods.

Deep learning is a technique based on a traditional ANN, and it involves the use of multilayer networks to process complex scenery. It uses linear and nonlinear transformations in multilayers to automatically establish a relationship between input and output data. A CNN [26] is a typical deep learning network architecture for image data. It has been proven to be a useful model for performing a wide range of imaging and visual tasks [27]. The CNN technique can be applied to different image processing tasks, for example, image classification and identification, object detection [28,29], image noise reduction [30,31], image resolution improvement [32,33], removal of compression artifacts [34], image color fusion [7], and radar image simulation from optical image [35].

Song et al. [22] pointed out that the capabilities of CNNs originate from three factors: (1) the deep network architecture of a CNN is effective in extracting large-scale image features, (2) efficient and rapid training methods, such as the rectified linear unit (ReLU) [36], batch normalization (BN) [37], and residual learning [38], that have been proposed, and (3) the emergence and popularization of graphic processing units (GPU) dedicated to graphics processing and with powerful parallel computing capability can help speed up training. Zhang et al. [31] proposed a denoising CNN (DnCNN) in which a CNN is used to reduce image noise. DnCNN accelerates the training process and improves image noise removal capability through residual learning, BN, and ReLU. Their experimental results indicated that the characteristics of a CNN deep network model may be useful for effectively predicting and reducing image noise, and also for improving image quality.

1.3. Need for Further Study and Research Purpose

Image SR can be used to enhance the spatial resolution of images with high temporal frequency but low spatial resolution [15]. Dong et al. [32] proposed super-resolution CNN (SRCNN), in which deep learning is introduced in the image SR method. Kim et al. [33] also presented a single-image SR method called very deep SR (VDSR). VDSR also employed a CNN network, but compared with the SRCNN, the neural network of VDSR is deeper and more information can be used to reconstruct the image. The most significant difference between the SRCNN and VDSR is that the training model of the SRCNN learns high-resolution images directly from low-resolution images, whereas the training model of VDSR learns residual images between high- and low-resolution images. Furthermore, to speed up training and the convergence rate, VDSR uses extremely high learning rates; its initial learning rate is 10^4 times higher than that of the SRCNN since it employs residual learning and adjustable gradient clipping. VDSR performs zero padding before convolutions in the training process to maintain the size of feature maps and output images constant; thus, pixels near the image boundary can be correctly predicted.

The spatiotemporal image fusion technique, which is a hybrid method, improves the quality of fusion images by combining the advantages of different approaches. Gevaert and García-Haro [24] proposed STRUM, which integrates unmixing-based and weight-function-based methods. Currently, most of the hybrid methods combine unmixing-based and weight-function-based methods. Drawing inspiration from this fact, this study proposes a hybrid fusion method based on the integration of weight-function-based (i.e., STARFM) and learning-based (i.e., VDSR) methods. The STARFM approach determines weights on the basis of physical parameters (i.e., spectral, temporal, and spatial variations), while VDSR learns weights of neutral networks from the data by itself. VDSR was originally developed to improve the results of image interpolation [33]. In the preprocessing of STARFM, image interpolation is a key process for interpolating a low-resolution image to have the same grid size as a high-resolution image. Since VDSR is capable of improving the results of image interpolation, there appears to be scope for developing a VDSR-assisted STARFM for improving the quality of image fusion.

Jarihani et al. [39] compared the results of index-then-blend and blend-then-index approaches for deriving the vegetation index by the image blending method. The former approach produced higher accuracy. In a hybrid approach, such as one combining STARFM and VDSR, the spatiotemporal image fusion performance should be evaluated for different combinations of processes (i.e., SR-then-Blend and Blend-then-SR). In SR-then-Blend, the role of VDSR is a preprocessing for STARFM. By contrast, in Blend-then-SR, the role of VDSR is a post-processing of STARFM.

1.4. Objectives

This study aims to determine the benefits of combining STARFM and deep learning for spatiotemporal image fusion. We evaluated the performance of the hybrid method for different combinations of processes. Four models that could be used for the spatiotemporal fusion of remote sensing images were compared, namely, STARFM, VDSR, and two hybrid models. The first hybrid model employs STARFM before VDSR (i.e., Blend-then-SR, hereafter abbreviated B-SR), while the second hybrid model employs VDSR before STARFM (i.e., SR-then-Blend, hereafter SR-B). In B-SR, the images are fused by a physical model (i.e., STARFM) and the remaining residuals are then compensated by using an in-depth learning approach (i.e., VDSR). In SR-B, the overall high-frequency details are injected from high-resolution images to low-resolution images by VDSR, and then a physical model (STARFM) is used for image fusion. The VDSR fine tunes the results of cubic interpolation in the preprocessing of STARFM.

Most spatiotemporal image fusion uses images with fixed look-angle satellite, for example, LS-5, LS-7, LS-8 and Sentinel-2 satellites. The FS-2 has the body-pointing capability of 30 degrees in roll and pitch directions, respectively. This body rotation capability is able to collect off-nadir images and to improve the temporal resolution. This study demonstrates the possibility of fusing the fixed look-angle Landsat-8 satellite and body-rotation Formosat-2 satellite in spatiotemporal fusion. The input high- and low-resolution satellite images were 8 m FS-2 and 30 m LS-8 images, while the output fused images were time-series 8 m fused FS-2 images. Finally, the results were evaluated by performing quantitative and qualitative analyses.

2. Material and Methods

2.1. Study Area and Dataset

The test area covered a rural area located in Kao-Hsiung in southwestern Taiwan. The latitude and longitude of image center were $22°52'42''$ N and $120°29'55''$ W, respectively. The area mainly comprised agricultural land, forest land, and building areas. Test images were obtained by the FS-2 and LS-8 satellites, and this study collected 8 m multispectral FS-2 images and 30 m multispectral LS-8 images from January 2014 to January 2016. The total overlapping area between FS-2 and LS-8 images was about 68 km^2. This study employed blue, green, red, and NIR bands of FS-2 and LS-8 images for image fusion. The product level of FS-2 used in this study was Level-2A with systematic correction, and the product level of LS-8 was Level-1 Precision and Terrain (L1TP) with geometric correction. The FS-2 and LS-8 images were precisely co-registered for fusion after preprocessing. Table 1 compares the spectral bandwidth of the corresponding bands of FS-2 and LS-8. FS-2 and LS-8 have corresponding bandwidths in the four bands, but the bandwidth of LS-8 is slightly narrower than that of FS-2. During the period from January 2014 to January 2016, only five pairs of images were recorded on the same day (Table 2 and Figure 2). In the training stage, four pairs of images were used as the training dataset, and one pair of images was used as the independent verification dataset. The quantity and quality of training data were key to the success of image fusion. Therefore, to effectively use image data and generate a deep learning model, in the training stage, we excluded cloud areas (from the LS-8 BQA band) in the training images.

Table 1. Comparison of the spectral bandwidth for FS-2 and LS-8 images.

Bands	FS-2		LS-8	
	Spectral Bands	**Bandwidth (μm)**	**Spectral Bands**	**Bandwidth (μm)**
Blue	Band 1	0.45~0.52	Band 2	0.45~0.51
Green	Band 2	0.52~0.60	Band 3	0.53~0.59
Red	Band 3	0.63~0.69	Band 4	0.64~0.67
Near Infrared	Band 4	0.76~0.90	Band 5	0.85~0.88

Table 2. Data acquisition date for training and independent verification.

Sensor	Training Dataset	Accuracy Analysis
FS-2, LS-8	2014/01/13, 2014/01/29 2015/02/01, 2016/01/26	2014/12/06

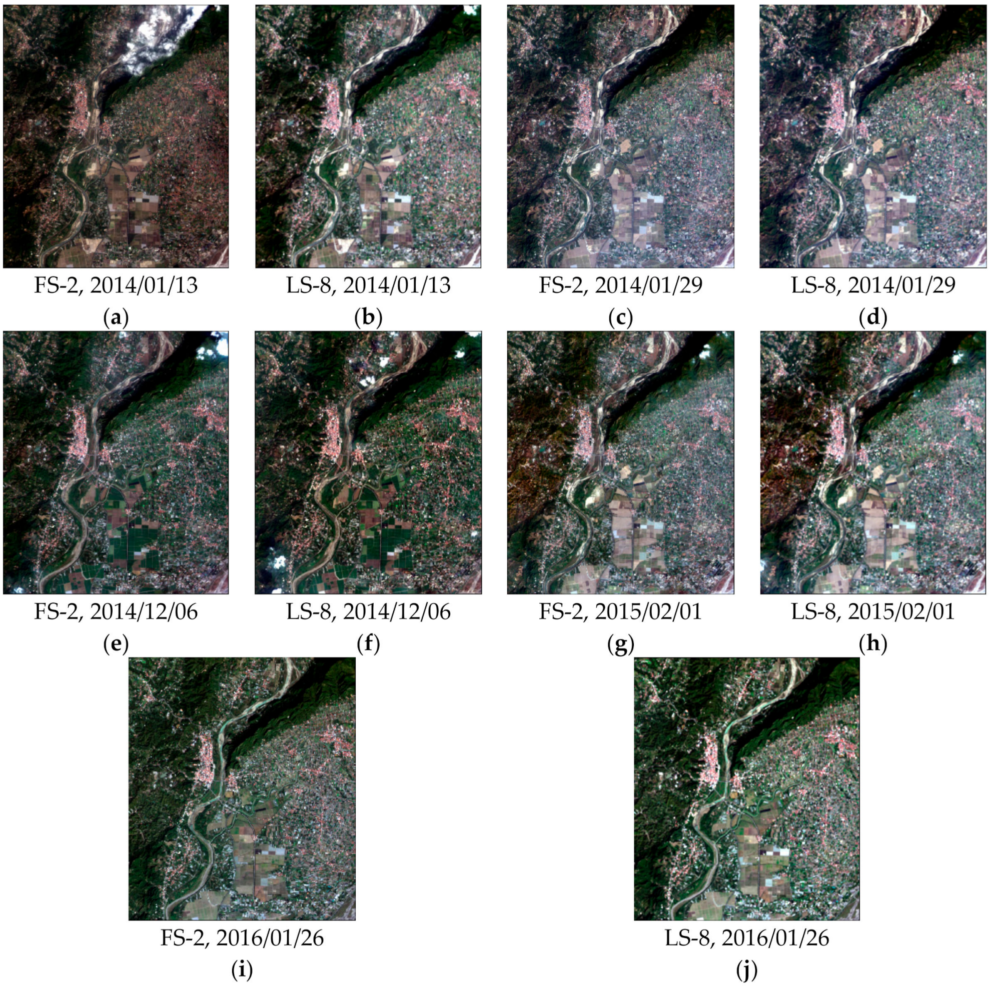
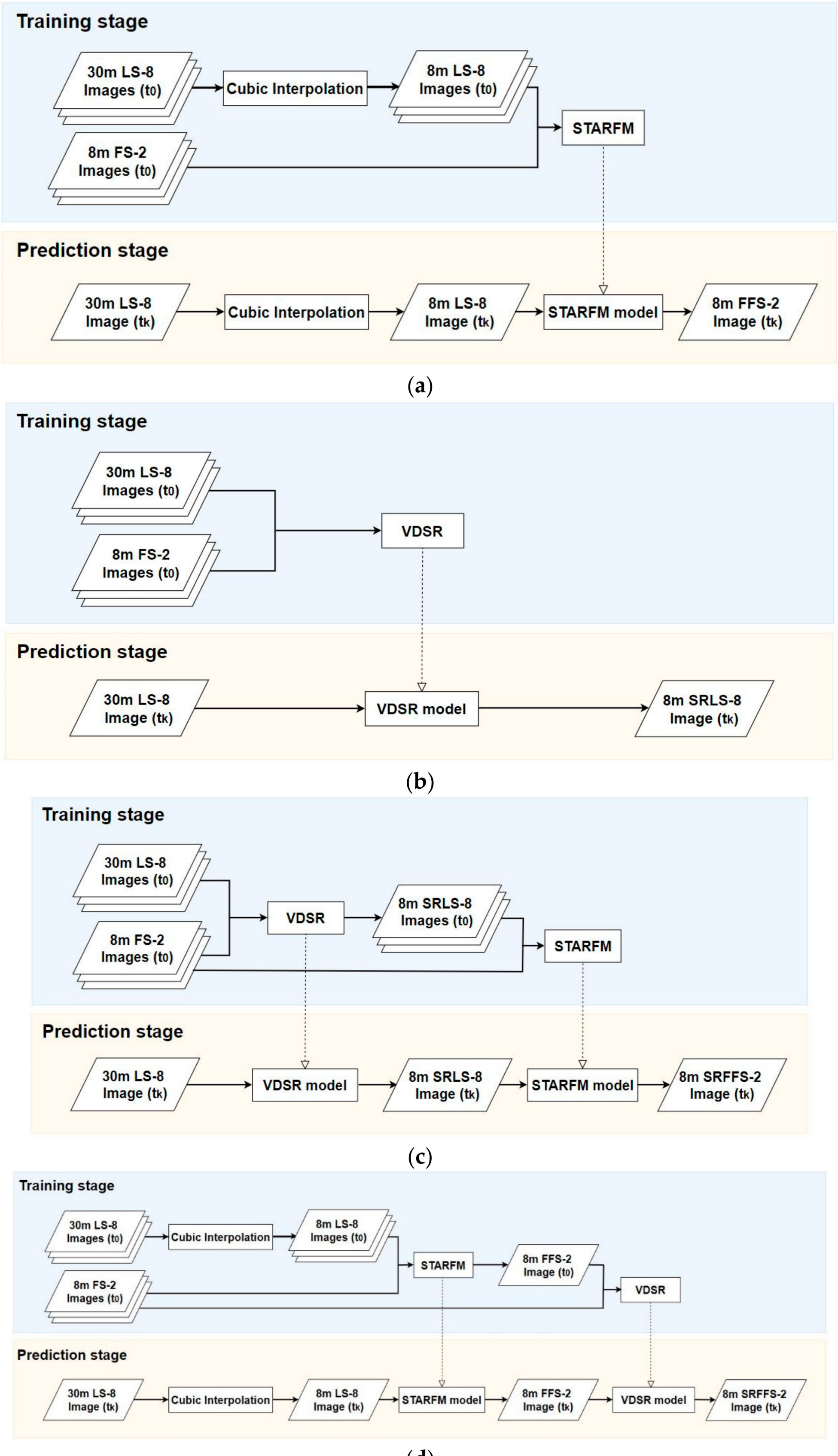

FS-2, 2014/01/13
(**a**)
LS-8, 2014/01/13
(**b**)
FS-2, 2014/01/29
(**c**)
LS-8, 2014/01/29
(**d**)
FS-2, 2014/12/06
(**e**)
LS-8, 2014/12/06
(**f**)
FS-2, 2015/02/01
(**g**)
LS-8, 2015/02/01
(**h**)
FS-2, 2016/01/26
(**i**)
LS-8, 2016/01/26
(**j**)

Figure 2. Image pairs of FS-2 and LS-8 images.

2.2. *Methodologies*

This study aimed to develop a hybrid spatiotemporal image fusion approach. The proposed scheme comprised five parts: (1) preprocessing of input FS-2 and LS-8 images, (2) image blending by STARFM (Figure 3a), (3) image SR by VDSR (Figure 3b), (4) development of hybrid spatiotemporal image fusion approach (Figure 3c,d), and (5) accuracy analysis. In the data preprocessing, satellite images from two different sensors were preprocessed to obtain data with consistent geometric and radiometric characteristics. The STARFM generates high-spatial- and high-temporal-resolution images by blending images with high temporal and low spatial resolution with images with low temporal and high spatial resolution, and VDSR compensates high-frequency details for the low-resolution images to construct an SR image. Two different combinations (i.e., B-SR vs. SR-B) were compared. Finally, several quantitative evaluation indicators were used to assess the quality of the fused image.



(**a**)

(**b**)

(**c**)

(**d**)

Figure 3. Workflow of proposed methods: (**a**) workflow of traditional STARFM; (**b**) workflow of VDSR; (**c**) workflow of SR-B; (**d**) workflow of B-SR.

2.2.1. Data Preprocessing

If different satellite images are to be compared spatially and temporally during spatiotemporal image fusion, the images to be fused should have the same radiometric response, ground sampling distance, image size, and coordinate system. Therefore, the images should be preprocessed by using radiometric correction, image co-registration, image clipping, and image resampling.

The FS-2 images used in this study were Level-2A images, which are ortho ready standard images. First, an additional reference FS-2 orthoimage was selected as a base map. All the FS-2 Level-2A images were orthorectified to the base map [40]. Next, a radiometric correction [41] was applied to convert the digital numbers in the FS-2 images to top-of-atmosphere reflectance using the physical parameters in the Dimap image description file.

The LS-8 images used in this study were L1TP terrain and precision corrected images. We found a systematic offset between LS-8 and the corrected FS-2 images. Therefore, we used cloud-free LS-8 and corrected FS-2 images to perform frequency domain image matching [42] and to determine the systematic bias between these two images. The systematic bias was then applied to the upper-right coordinates of the LS-8 images. A radiometric correction determined by using the physical parameters in the MTL image description file was also applied to the LS-8 images [43]. As the spatial resolution of the images to be fused should be the same, the corrected LS-8 images were further resampled into 8 m/pixel using cubic interpolation.

2.2.2. Method 1: Image Blending Using Spatiotemporal Image Fusion Method

In this part, we employed the STARFM developed by Gao et al. [10] as the spatiotemporal image fusion model. The STARFM is a physical model that fuses images with high temporal and low spatial resolution and images with low temporal and high spatial resolution to generate fused images with high temporal and high spatial resolution. The images with high temporal and low spatial resolution provided temporal information, while images with low temporal and high spatial resolution provided spatial information.

This study used 8 m FS-2 images as images with low temporal and high spatial resolution and 30 m LS-8 images as images with high temporal and low spatial resolution images for image fusion. As the FS-2 and LS-8 images were preprocessed to have the same geometrical and radiometric characteristics, these two datasets could be compared with each other directly. The reflectance of the FS-2 high spatial resolution pixel corresponding to the LS-8 low spatial resolution homogeneous pixel on date t_0 can be expressed as Equation (1), while reflectance of FS-2 and LS-8 image on date t_k can be defined as shown in Equation (2):

$$F\left(x_i, y_j, t_0\right) = L\left(x_i, y_j, t_0\right) + \varepsilon_0, \tag{1}$$

$$F\left(x_i, y_j, t_k\right) = L\left(x_i, y_j, t_k\right) + \varepsilon_k, \tag{2}$$

Here, (x_i, y_j) is a given pixel location for both FS-2 and LS-8 images, F is FS-2 data (high-spatial-resolution image), L is LS-8 data (low-spatial-resolution image), t_0 and t_k are the acquisition dates for both FS-2 and LS-8 images (the observation date and prediction date, respectively), and ε_0 and ε_k represent the difference between the FS-2 and LS-8 reflectance values at t_0 and t_k, respectively.

If it is assumed that the land cover and the systematic error of the pixel (x_i, y_j) did not change at t_0 and t_k, that is, the difference in the spectral reflectance between different dates is similar, then $\varepsilon_0 = \varepsilon_k$. Thus, Equations (1) and (2) can be used to obtain Equation (3). However, the relationship between LS-8 and FS-2 images is highly complex because of the following reasons: (1) LS-8 observations might not be homogeneous pixels and may contain mixed land cover types when considered at the FS-2 spatial resolution. (2) There is a high chance that the land cover type will change during the period from the observation date (t_0) to the prediction date (t_k). Furthermore, the transformation of the land cover status and the bidirectional reflectance distribution function would also change the reflectance during the interval from the observation date (t_0) to the prediction date (t_k).

Therefore, the linear equation (i.e., Equation (3)) is not sufficient, and the fusion model must consider a weighting function. Consequently, STARFM utilized a moving window to obtain neighboring pixels with pixels's spectrally similar during the fusion process and then used the weighting function to estimate the center pixel of the image on the prediction date (i.e., Equation (4)):

$$F\left(x_i, y_j, t_k\right) - F\left(x_i, y_j, t_0\right) = L\left(x_i, y_j, t_k\right) - L\left(x_i, y_j, t_0\right), \tag{3}$$

$$F(x_{m/2}, y_{m/2}, t_k) = \Sigma_{\mathrm{i}=1}^{m} \Sigma_{j=1}^{m} \Sigma_{k=1}^{n} W_{ijk} \times \left(L\left(x_i, y_j, t_k\right) + F\left(x_i, y_j, t_0\right) - L\left(x_i, y_j, t_0\right)\right), \tag{4}$$

where m is the size of the moving window, $\left(x_{m/2}, y_{m/2}\right)$ is the central pixel of the moving window, and W_{ijk} is the combined weights for a neighboring pixel, including spectral, temporal, and spatial distance variations.

The combined weight (W_{ijk}) (i.e., Equation (5)) determines the contribution of each neighboring pixel to predict the reflectance of the central pixel, which depends on the variation of the images in terms of the spectral, temporal, and spatial distances (i.e., Equation (6)). The spectral variation (S_{ijk}) is the spectral difference between the FS-2 and LS-8 reflectances on the same date (i.e., Equation (7)). The smaller the difference, the greater the similarity between the reflectances of the FS-2 image and the averaged surrounding pixels. Thus, S_{ijk} will be assigned a higher weight. The temporal variation (T_{ijk}) is the difference in time between the input training and the predicted LS-8 images (i.e., Equation (8)). A smaller value indicates that the land cover does not change significantly during the period from t_0 to t_k. T_{ijk} will also be assigned a higher weight. The spatial variation (D_{ijk}) is the relative spatial distance between the central pixel of the moving window and the surrounding spectrally similar candidate pixels on date t_0 (i.e., Equation (9)). The candidate pixels near the central pixel have a higher weight. In Equation (9), A is a constant parameter used to define the relative importance of the spatial distance to the difference between spectral and temporal distances. The larger the A value, the smaller the weight of D_{ijk}:

$$W_{ijk} = \left({}^{1}/_{C_{ijk}}\right) / \Sigma_{\mathrm{i}=1}^{m} \Sigma_{j=1}^{m} \Sigma_{k=1}^{n} \left({}^{1}/_{C_{ijk}}\right), \tag{5}$$

$$C_{ijk} = S_{ijk} \times T_{ijk} \times D_{ijk}, \tag{6}$$

$$S_{ijk} = \left|F\left(x_i, y_j, t_0\right) - L\left(x_i, y_j, t_0\right)\right|, \tag{7}$$

$$T_{ijk} = \left|L\left(x_i, y_j, t_0\right) - L\left(x_i, y_j, t_k\right)\right|, \tag{8}$$

$$D_{ijk} = 1.0 + \sqrt{\left(x_{m/2} - x_i\right)^2 + \left(y_{m/2} - y_j\right)^2} / A, \tag{9}$$

The input data required for applying the STARFM should include at least one pair of high- and low-spatial-resolution images obtained on the same date, and a low-spatial-resolution image on the prediction date. The output data is a high-spatial-resolution fused image on the prediction date. Generally, the STARFM includes training and prediction stages. The training stage is aimed at determining the combined weights from FS-2 and LS-8 images captured on the same date. The first step uses high-spatial-resolution images to find candidate pixels that are spectrally similar to the central pixel in the moving window, and the second step filters out inappropriate candidate pixels on the basis of the uncertainties in the spectral information of the FS-2 and LS-8 images. The third step assigns weights according to the pixels' variation in terms of the spectral, temporal, and spatial information. The higher the weight, the more the contribution of the central pixel reflectance to the prediction. In the prediction stage, the predicted reflectance for LS-8 is estimated from the pretrained weights. A more detailed description of the STARFM with regard to the determination of weights can be found in the paper of Gao et al. [10].

2.2.3. Method 2: Image SR Using VDSR

Kim et al. [33] proposed VDSR, which involves the use of deep learning. This method uses a very deep convolutional network inspired by the visual geometry group network (VGGNet). VDSR determines the details of images to solve single-image SR (SISR) problems, that is, to reconstruct a higher-resolution image from a low-resolution image. The reconstruction method employs a residual-learning CNN to train and predict the residual map between the low- and high-resolution images. Subsequently, the residual image is compensated back to the low-resolution image to obtain the corresponding high-resolution image.

The first layer of VDSR's network structure (Figure 4) is the input layer, which is a receptive field of size $(2D + 1) \times (2D + 1)$; D represents the total number of convolutional layers in the network. The VDSR network used in this research had twenty convolutional layers; thus, the size of the receptive field was 41 × 41. The middle layer consists of a repetitive cascade of 19 pairs of convolutional layers and ReLU layers. Each convolutional layer includes 64 filters of size 3 × 3 × 64. Zero padding is performed before each convolution operation to ensure that all feature maps are of the same size. This is done to maintain the size of the output image identical to that of the input image. The last layer is a convolutional layer composed of a single filter of size 3 × 3 × 64, which is the residual image used for image reconstruction.

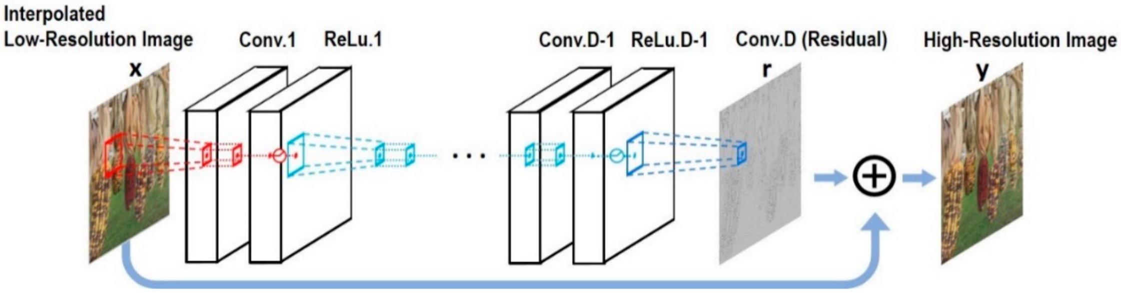


Figure 4. Network structure of very deep super-resolution (VDSR) [33].

During the training process, VDSR learned to predict the difference between input and output images to avoid vanishing gradient and exploding gradient problems and to increase the training model's convergence speed. In this study, the VDSR model was used to estimate the residual between high-resolution and low-resolution images on the same date. The learning process was intended to establish a nonlinear relationship between the low-resolution image and the residual image. Residual learning involved learning the high-frequency variation of the image to improve the spatial details of the image. The input in the training process was a dataset ($\left\{\mathrm{L}^{(\mathrm{i})}, \mathrm{F}^{(\mathrm{i})}\right\}_{\mathrm{i}=1}^{\mathrm{N}}$) composed of multiple pairs of low-resolution images (***L***) and high-resolution images (***F***). The output was the residual image (***r***) (i.e., Equation (10)), that is, the difference between high and low-resolution images (Figure 5). The training model is expressed by Equation (11), where f is the trained deep learning SR model and $\hat{\mathrm{F}}$ is the predicted target image. The loss function is expressed in Equation (12). A more detailed description of VDSR can be found in Kim et al.'s paper [33]:

$$r = F - L, \tag{10}$$

$$\hat{F} = f(L), \tag{11}$$

$$loss = \frac{1}{2}\|r - f(L)^2\|, \tag{12}$$

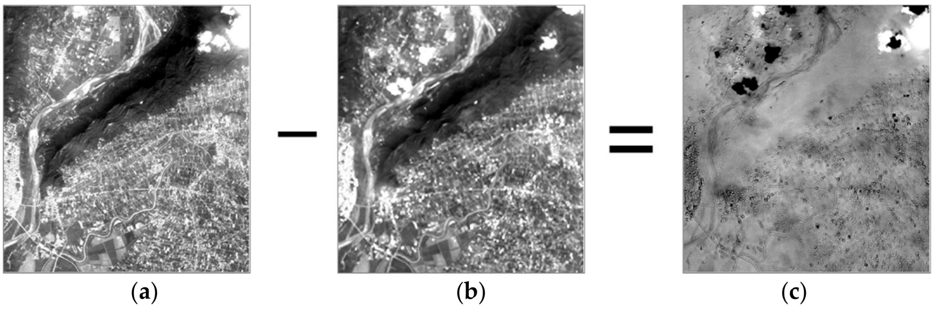


Figure 5. Generation of a residual image for VDSR: (**a**) a high-resolution FS-2 image, (**b**) a low-resolution LS-8 image, and (**c**) the residual image.

The purpose of using VDSR for image fusion was to learn the spatial details of the input image (i.e., residual) by using a CNN between low- and high-resolution images acquired on the same day. Both FS-2 and LS-8 pre-processed images have the same pixel size in the calculation of the residual between FS-2 and LS-8 images. In the training stage, this study used FS-2 and LS-8 images recorded on the same day to calculate the residual image. The input training images are LS-8 and residual images for the same day. VDSR is applied to train a deep neural network with LS-8 and residual images. In the prediction stage, the VDSR's model is used to predict the residual images from time-series LS-8 images. High-frequency details from VDSR's residual image is then injected into the time-series LS-8 images. Finally, the 8 m fused image on the prediction date can be obtained by combining the LS-8 image and predicted residual image.

In the VDSR training model, the optimization method is stochastic gradient descent with momentum. The momentum is set to 0.9. The initial learning rate is set to 0.1, and it is reduced 10 times after every 10 epochs for a total of 100 epochs. The patch size is 41 × 41 pixels, while 256 patches are randomly selected from the image pair. The minibatch size is set to 64. In particular, a multispectral image used in this study comprised four spectral bands, and therefore, VDSR trained each spectral band separately.

2.2.4. Method 3: Hybrid Spatiotemporal Fusion Approach SR-B

This study proposes a hybrid spatiotemporal fusion approach in which the STARFM and VDSR model are combined to produce time-series satellite imagery. The VDSR model was applied to reduce the difference between the two types of satellite imagery or to reduce the difference between the fused image and the observed image. Hence, this study proposes two different input training data sets for VDSR models. The first type of training data were an LS-8 image and a residual image (i.e., difference between LS-8 and FS-2 images), and they were used to learn residuals between two different sensors, recorded on the same day. The second type of training data were a fused image (i.e., results of the STARFM) and a residual image (i.e., difference between fused and FS-2 images), and they were used to learn residuals between the fused image and the observed image.

The first hybrid model was SR-B, which employed VDSR before the STARFM. The first VDSR training model used LS-8 images to learn and to predict the residuals between itself and corresponding FS-2 images, and it subsequently added the predicted residual images to the low-resolution LS-8 images to generate the SR LS-8 images (SRLS-8). The spatial details of LS-8 were enhanced by VDSR. The SRLS-8 image and FS-2 image were blended by the STARFM to produce time-series fusion satellite images. The workflow is shown in Figure 3c. The concept underlying SR-B was to reduce the spatial and spectral differences of high- and low-resolution images before image fusion. In other words, VDSR was intended to preprocess the input data of the STARFM. VDSR was used to improve the results of the image interpolation stage in the STARFM for facilitating a comparison between the traditional STARFM and SR-B (Figure 3a,c).

2.2.5. Method 4: Hybrid Spatiotemporal Fusion Approach B-SR

The second hybrid model was B-SR, which employed the STARFM before VDSR. The STARFM generated a fused FS-2 image directly from LS-8 and FS-2 images. The second VDSR training model then used the fused FS-2 images to learn and predict the residual between itself and the corresponding FS-2 images, after which it added the residual images to the fused FS-2 images to generate SR fused FS-2 images. The workflow is shown as Figure 3d. The concept underlying B-SR was to compensate the residuals between fused FS-2 and the original high-resolution FS-2 image by using deep learning technique. VDSR post-processes the output of STARFM. B-SR was used to compensate the residual between the result of the STARFM and the original FS-2 image for facilitating a comparison between the traditional SR-B and B-SR (Figure 3c,d).

2.3. Accuracy Analysis

The quality assessment includes the quantitative analysis for the entire area and qualitative analysis for the vegetation and building regions. The quantitative analysis involves absolute and relative indexes. The absolute index evaluates the nature of the fused image itself, and therefore, it is calculated using the fused image. The relative index compares the observed and fused images. It uses the observed image as a benchmark to evaluate the correlation between the real observed and synthetized fusion images. The absolute indexes were entropy [44] and the blind/referenceless image spatial quality evaluator (BRISQUE) [45], and the relative indexes were reflectance bias, structural similarity (SSIM) [46], and peak signal-to-noise ratio (PSNR) [47]. Among these five indicators, the reflectance bias was used to evaluate the difference between the observed and fused images in different spectral bands, and the other four indicators were used to assess the visual performance of the fused image:

(1) Reflectance bias: This index is used to evaluate the degree of difference in reflectance among observed and fused images. This study calculated the average and standard deviation (SD) of the reflectance bias between observed images and fused images. The lower difference indicates better result.

(2) SSIM: This index is used to evaluate the similarity of the overall structure between observed and fused images. This index is based on the human visual system to extract structural information for comparing the luminance, contrast, and structure between images. The SSIM ranges from -1 to 1. The larger the value, the higher the similarity between the two images. The expression for the SSIM is presented in Equation (13), where $l(x, y)$ is the luminance comparison function, $c(x, y)$ is the contrast comparison function, $s(x, y)$ is the structure comparison function, μ_x and μ_y are the mean of images x and y, σ_x and σ_y are the SDs of images x and y, σ_x^2 and σ_y^2 are the variances of images x and y, σ_{xy} is the cross-covariance between images x and y, and C_1, C_2, and C_3 are constants used to maintain the stability of $l(x, y)$, $c(x, y)$, and $s(x, y)$, respectively.

$$SSIM = l(x,y)\cdot c(x,y)\cdot s(x,y) = \left(\frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}\right)\cdot\left(\frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}\right)\cdot\left(\frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}\right), \quad (13)$$

(3) PSNR: This index is used to assess the degree of distortion of the fused image. This study used the observed image as the reference undistorted image. The ratio of the maximum value of an image signal to the noise in an image was used as the evaluation index. The larger the value of this index, the higher degree of undistortion between the two images. The PSNR is given by Equation (14), where $\boldsymbol{x}$ and $\boldsymbol{y}$ are the observed image and fused image, respectively, $\boldsymbol{n}$ is the image bit depth, and MSE is the mean square error between the observed image and the fused image. In the

absence of noise, the observed image and the fused image are identical, and the MSE is equal to 0; therefore, the PSNR is infinite.

$$PSNR(x,y) = 10 \times \log_{10}\left(\frac{(2^{\mathrm{n}}-1)^2}{MSE}\right)(\mathrm{d}B), \quad (14)$$

(4) Entropy: The entropy is used to assess the amount of information contained in an image. Generally, a clear image provides more detailed information than a blurred image. Hence, the greater the entropy of a fused image, the greater the amount of information contained in the fused image. The equation of entropy is presented in Equation (15), where n is the total number of grayscale levels, N_i is the number of pixel i in the image, and N_s N_s is the total number of pixels in the image:

$$ENTROPY = -\sum_{i=0}^{n-1}\left(\frac{N_i}{N_s}\right)\log_2\left(\frac{N_i}{N_s}\right), \quad (15)$$

(5) BRISQUE: The quality of a fused image is evaluated according to the natural characteristics of the fused image, and it is the reference value of the image quality obtained from the characteristics of natural statistics of the image. The scene statistics of locally normalized luminance coefficients are used to quantify the distortion in the image and assess the quality of the image. BRISQUE ranges from 0 to 100, with the value of 0 representing an undistorted image. This implies that a smaller value indicates lower distortion and better image quality. Details on BRISQUE can be found in Mittal et al.'s [45] paper.

3. Results

In this experiment, the same dataset were used to train four different image fusion methods (i.e., STARFM only, VDSR only, B-SR hybrid method, and SR-B hybrid method). After training the fusion models with the training dataset, a 30 m LS-8 image recorded on 6 December 2014, was used to predict 8 m fused images with the different fusion models. For the evaluation of the accuracy of the models, the FS-2 image acquired on 6 December 2014, was used as an independent check image. The four fused images are shown in Figure 6. The results from the methods were verified by using the five indicators (i.e., entropy, BRISQUE, SSIM, PSNR, and reflectance bias) in the following section.

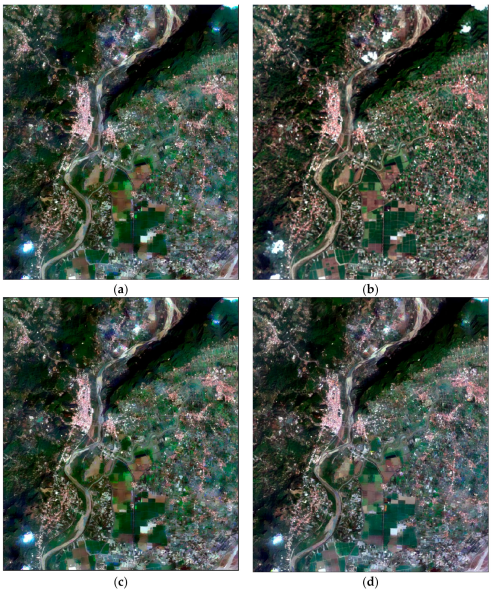


Figure 6. Fused images predicted with a 30 m LS-8 image recorded on 6 December 2014. (**a**) the STARFM, (**b**) VDSR, (**c**) B-SR, and (**d**) SR-B.

4. Discussions

Section 4.1 discusses the quantitative analysis for the entire area, while Section 4.2 focuses on qualitative analysis in the vegetation and building regions.

4.1. Quantitative Analysis

The mean and SD of the reflectance bias for individual bands (i.e., B, G, R, NIR) and all bands (i.e., four bands) are provided for comparing the performance of different methods (Table 3). The reflectance bias of NIR was larger than that of the other three bands. A possible reason is that the variation of the NIR reflectance was significantly larger than that of the other bands (Table 4). The test area was mostly covered by vegetation. Consequently, the NIR reflectance of chlorophyll was larger than the reflectance of the other bands. Gao et al. [10] employed the STARFM to fuse Landsat-7 and MODIS images. The reflectance bias between the real image and the predicted image in the NIR band was also markedly larger than blue, green, and red bands. From the preceding discussion, the NIR band is evidently more difficult to predict using the STARFM, but the VDSR model could improve this problem. Because the VDSR model learned information from across sensors, it not only improved the spatial resolution of the image, but also reduced the systematic deviation between the two satellite images in the spectral bands.

Table 3. Reflectance bias between fused and observed images (unit: $\rho \times$ 10,000).

	Bands	STARFM	VDSR	B-SR	SR-B
Mean (Δ)	Blue	64.428	55.910	64.054	55.426
	Green	69.959	71.482	69.654	64.475
	Red	89.626	92.461	89.323	81.204
	NIR	297.946	254.803	297.837	298.994
	4 Bands	130.489	118.664	130.217	125.025
SD (Δ)	Blue	88.568	81.576	88.534	80.500
	Green	111.092	113.134	111.060	103.288
	Red	145.661	151.123	145.639	137.954
	NIR	341.856	228.877	341.839	344.398
	4 Bands	171.794	143.677	171.768	166.535

Table 4. Comparison of FS-2 and LS-8 image reflectance acquired on 6 December 2014 (unit: $\rho \times$ 10,000).

2014/12/06					
Bands	**Satellites**	**Min**	**Max**	**Mean**	**SD**
Blue	FS-2	1043	2216	1250.568	114.589
	LS-8	1005	5950	1282.070	173.274
Green	FS-2	770	2585	1095.109	155.392
	LS-8	657	6195	1055.731	192.599
Red	FS-2	544	3084	899.848	218.592
	LS-8	312	6743	833.172	255.384
NIR	FS-2	609	3699	2273.934	640.359
	LS-8	384	8992	2686.187	818.275

A comparison of the results of reflectance bias between the STARFM and VDSR showed that the results of VDSR were slightly better than those of the STARFM in the NIR band. The VDSR requires a large number of training datasets in the training stage. This study considered only four image pairs (image size: 1360 × 1580 pixels) to establish the VDSR network model. For a limited number of training images, the VDSR is still better than STARFM. A comparison of the results of reflectance bias between the individual and hybrid methods showed that the mean bias of the hybrid method was smaller and better than that of the STARFM-only method. The hybrid method is slightly better than the traditional method. The hybrid strategy exploited the advantages of the STARFM and VDSR methods to minimize the reflectance bias. It could be inferred that the integration of weight-function-based and learning-based fusion models for image fusion can help increase the similarity between the fused image and the observed image.

This study also compares two different hybrid models. The results of SR-B (SR then blend) were better than those of B-SR (blend then SR). The overall mean bias of SR-B was 4% lower than B-SR, and 3% improvement for overall standard deviation in SR-B. In SR-B, VDSR improved the resolution of the LS-8 image owing to its training with FS-2 images before STARFM was applied. In this way, VDSR provided a better SRLS-8 image than the traditional cubic-interpolated LS-8 image. The SR-B method could obtain more information from FS-2 before the application of STARFM. While VDSR in SR-B learned the difference between the original LS-8 and FS-2 images, VDSR in B-SR learned the difference between an LS-8 image from STARFM and an FS-2 image. Although the result of B-SR was slightly better than that of the traditional STARFM used individually, the error in the LS-8 image from the STARFM could affect VDSR in the B-SR hybrid method. Therefore, the results of SR-B are better than those of B-SR, and it is recommended that SR be performed before blend.

The other four image quality assessment indicators for the different fusion methods were also provided for comparison (Table 5). Both SSIM and PSNR compare the observed and fused images. The SSIM evaluates the similarity between images, while the PSNR examines the distortion between images. The SSIM of VDSR was lower than STARFM, implying that the overall SSIM of the VDSR method was lower than STARFM because of the limited training data set. The SSIMs of the B-SR and SR-B were similar, and the difference was only 0.004. In terms of the degree of reflectance's distortion using the PSNR, the SR-B approach showed better results compared with the other three methods. The PSNR and reflectance bias showed similar behavior because both indicators were based on the residual of reflectance. In summary, both SSIM and PSNR indicated that the SR-B approach combining VDSR and the STARFM could minimize the reflectance differences between observed and fused images. The SR-B is slightly better than B-SR and the PSNR's difference between SR-B and B-SR was 0.531.

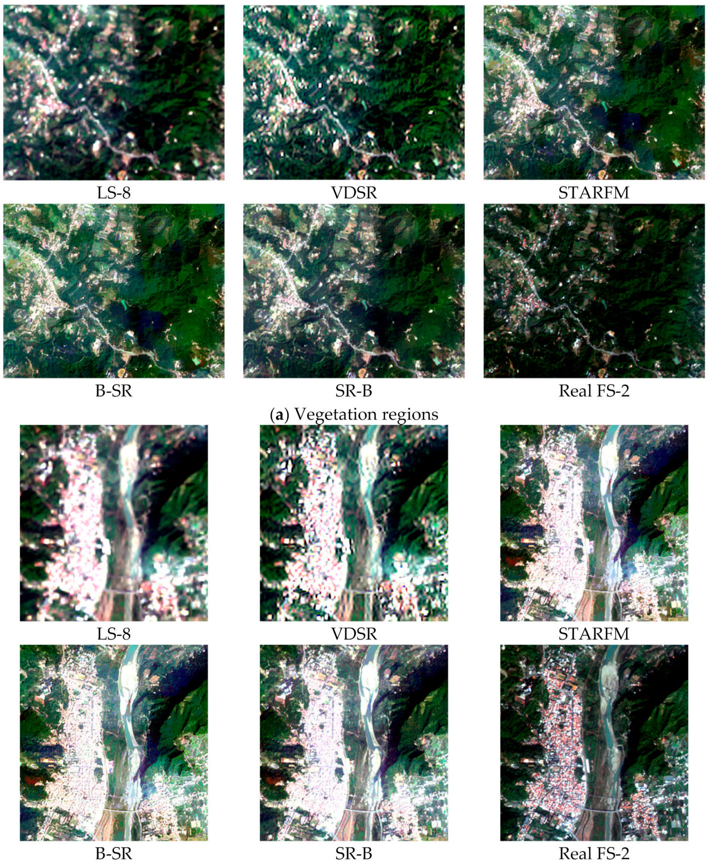
Table 5. Comparison of image quality assessment indicators for different fusion methods.

	STARFM	VDSR	B-SR	SR-B
SSIM	0.906	0.894	0.906	0.910
PSNR	34.763	33.933	34.774	35.305
Entropy	2.433	2.389	2.433	3.001
BRISQUE	25.335	50.326	25.178	24.383

From the perspective of the average amount of information, the entropy of SR-B showed more information than that of the other approaches. The entropy of VDSR was the lowest, while that of STARFM and B-SR were similar. The SR-B improve the information content than B-SR and the Entropy's difference between SR-B and B-SR was 0.568. The improvement rate was about 23% as the SR-B was sharpening the input image before image blending. In the evaluation of image quality, BRISQUE was used as a no-reference image quality indicator. The BRISQUE of SR-B was better than that of the other three approaches. The results of SR-B was slightly better than B-SR. Owing to the insufficient amount of training data for the VDSR-only approach, this approach showed lower performance. In summary, the overall performance of the SR-B approach yielded better accuracies than the other methods.

4.2. Qualitative Analysis

To examine the performance of different methods for different land covers, vegetation and building regions were chosen for performing a qualitative analysis. Figure 7 compares the results of different image fusion methods with the FS-2 real observation image recorded on 6 December 2014. In the visual analysis, most fused images were similar to the real FS-2 observation, except the VDSR-only fused image. A comparison of the interpolated LS-8 image and VDSR-only fused image showed that the sharpness of the VDSR result was better than the interpolated LS-8. From the perspective of image interpolation, VDSR refines and provides better results than the traditional cubic interpolation. Therefore, the integration of VDSR and STARFM can improve the fusion quality.



(**a**) Vegetation regions

(**b**) Building regions

Figure 7. Comparison of two different land covers on 6 December 2014: (**a**) vegetation area and (**b**) building area.

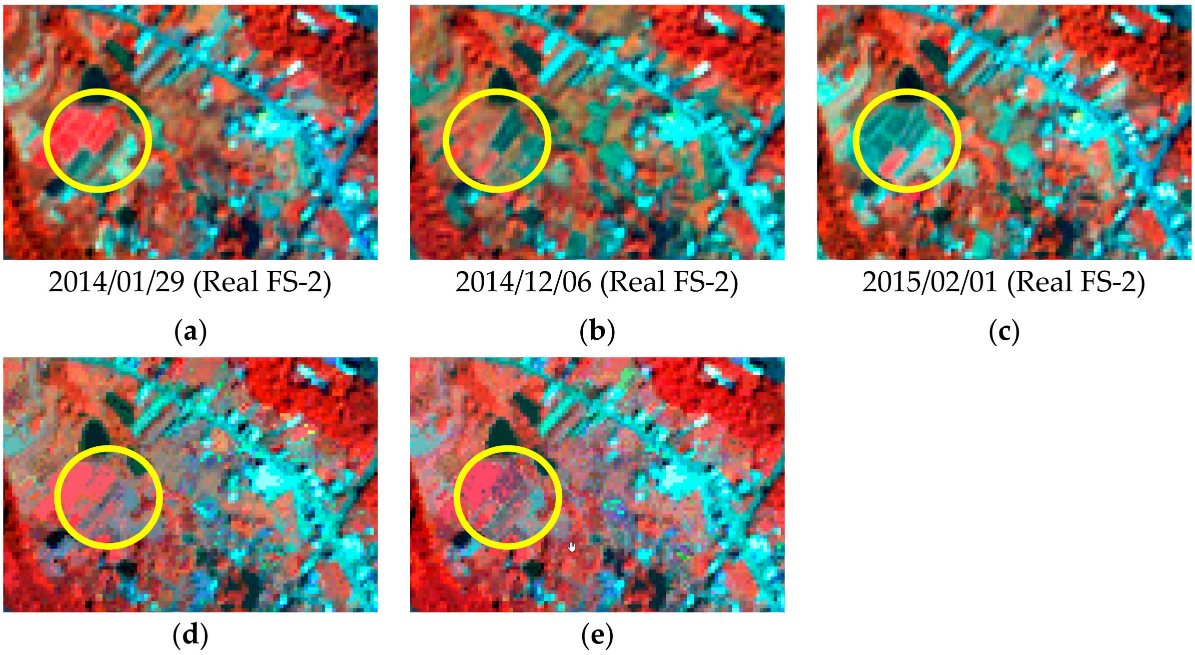
Owing to the spectral variation of the building region being larger than that of the vegetation region, the spectral distortion in the building region was slightly larger than that in the vegetation region. However, the entropy (information content) of the building region was better than that of the vegetation area. In the improvement of entropy for the STARFM and hybrid methods, the building region showed higher improvement than the vegetation area. In other words, the hybrid methods showed better performance in the building region compared with the vegetation region.

From the perspective of visual performance, a detailed comparison of the hybrid methods (i.e., SR-B and B-SR) and the STARFM-only method revealed that the hybrid methods not only showed superior spatial resolution from the 30 m LS-8 image to the 8 m fused image, but also enhanced local details in the mountain's shadow region. In other words, the hybrid methods could improve the spatial information and local spectral information. Table 6 shows that most quality indexes of the hybrid methods were better than those of the STARFM-only method. Furthermore, the fused images from SR-B showed better performance than the other fused images.

Table 6. Reflectance bias between the local areas of fused and observed images (unit: $\rho \times 10{,}000$).

		STARFM	VDSR	B-SR	SR-B
Vegetation regions	Mean (Δ)	95.275	104.924	94.933	98.884
	SD (Δ)	78.705	84.129	78.605	80.996
	SSIM	0.936	0.909	0.936	0.935
	PSNR	40.069	38.907	40.106	40.137
	Entropy	3.612	3.625	3.614	3.645
	BRISQUE	29.959	53.749	30.001	32.744
Building regions	Mean (Δ)	99.675	133.020	99.383	98.708
	SD (Δ)	91.094	115.218	91.037	90.339
	SSIM	0.932	0.876	0.932	0.933
	PSNR	38.601	35.863	38.625	38.796
	Entropy	4.228	4.000	4.230	4.400
	BRISQUE	21.241	55.826	21.256	21.769

Land cover change is a challenging issue in spatiotemporal fusion, and the results of image fusion were usually affected by seasonal crop changing (Figure 8a–c). To compare the SR-B and B-SR methods in detail, the fused image using SR-B (Figure 8e) did not perform very well in the land cover changed area. It was still more similar to the real observation image figure (Figure 8b) compared to the fused image using B-SR (Figure 8d). The possible reason was the VDSR in SR-B compensated the high-frequency of LS-8 image. Consequently, the image blending produces better results. In summary, the performance of SR-B was better than that of B-SR.


2014/01/29 (Real FS-2)

(**a**)

2014/12/06 (Real FS-2)

(**b**)

2015/02/01 (Real FS-2)

(**c**)

(**d**)

(**e**)

Figure 8. Comparison of B-SR and SR-B in land cover changed region. (**a**) Training; (**b**) Testing; (**c**) Training; (**d**) 2014/12/06 B-SR; (**e**) 2014/12/06 SR-B.

5. Conclusions

This study developed a hybrid spatiotemporal image fusion approach involving a deep learning model and a physical model. The deep learning model is the VDSR model, which improves the spatial resolution of low-resolution images, and the physical model is the STARFM model, which considers physical parameters such as pixel distance, spectral, temporal and spatial variations. Two different hybrid fusion strategies (i.e., B-SR vs. SR-B) were developed and discussed on the basis of experiments. In SR-B, SR replaces the image resampling stage used for interpolating the pixel size of a low-resolution image into that of a high-resolution image. This image resampling stage is an essential step for obtaining low- and high-resolution images with the same pixel size. In B-SR, the SR refines the spatial details after the STARFM is applied.

The major contribution of this study is to develop hybrid spatiotemporal image fusion methods (i.e., B-SR and SR-B) using STARFM and VDSR. In the experiment, STARFM, VDSR, B-SR, and SR-B were used to fuse the satellite images of FS-2 and LS-8 to produce spatiotemporal fusion data, and several quality indexes (i.e., reflectance bias, entropy, BRISQUE, SSIM, and PSNR) were used to assess the quality of fused images. The experimental results demonstrated that the combination of the spatiotemporal fusion techniques of the STARFM and the VDSR model based on a CNN architecture helped to increase the similarity between fusion images and observation images. Overall, the quality of the fusion results obtained using SR-B was better than that generated with the other methods. The results showed that running the VDSR model to learn the difference between low-resolution images and high-resolution images before applying STARFM could reduce the variation in spatial and spectral resolution between the fused image and the observed image. Besides, it could also yield a fused image that is better in visual performance and is most similar to the observation image.

The difference between the fusion result of SR-B and the real image was smaller than the result obtained by employing the STARFM alone. In the comparison of SR-B ad B-SR, the overall spectral bias of SR-B was lower than B-SR; moreover, the entropy of SB-R was also higher than B-SR. This demonstrated that the combination of VDSR model based on the CNN architecture and the spatiotemporal fusion techniques of the STARFM can help to increase the similarity between fusion images and observation images. Therefore, this study recommends the use of SR-B rather than B-SR.

The STARFM is not effective for simulating sudden land cover changes in the short term. Although the combination of the STARFM and VDSR models in this study could improve the quality of fusion images, its capability is limited. Therefore, in order to acquire better fusion results, a weight-based fusion model modified from STARFM, such as ESTARFM, may be used in future studies. Furthermore, since only five pairs of LS-8 and FS-2 satellite image pairs were used in this research, there was not much training data that could be used to train the VDSR model, which influenced the learning effectiveness of VDSR. In addition, a long time interval between the image pairs also affects the fusion results of the STARFM. Hence, it is recommended to examine the use of Sentinel-2 images with higher temporal resolution in future studies.

Author Contributions: T.-A.T. provided the overall conception of this research, designs the methodologies and experiments, and wrote the majority of the manuscript; Y.-J.F. contributed to the implementation of proposed algorithms, conducts the experiments and performs the data analyses. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Ministry of Science and Technology of Taiwan, grant number MOST 108-2628-E-009-002-MY3.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would also like to thank the National Space Organization of Taiwan for providing the test datasets for experiments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sakamoto, T.; Yokozawa, M.; Toritani, H.; Shibayama, M.; Ishitsuka, N.; Ohno, H. A Crop phenology detection method using time-series MODIS data. *Remote Sens. Environ.* **2005**, *96*, 366–374. [CrossRef]
2. Lymburner, L.; Botha, E.; Hestir, E.; Anstee, J.; Sagar, S.; Dekker, A.; Malthus, T. Landsat 8: Providing continuity and increased precision for measuring multi-decadal time series of total suspended matter. *Remote Sens. Environ.* **2016**, *185*, 108–118. [CrossRef]
3. Son, N.T.; Chen, C.F.; Chen, C.R.; Sobue, S.I.; Chiang, S.H.; Maung, T.H.; Chang, L.Y. Delineating and predicting changes in rice cropping systems using multi-temporal MODIS data in myanmar. *J. Spat. Sci.* **2017**, *62*, 235–259. [CrossRef]

4. Zeng, Z.; Estes, L.; Ziegler, A.D.; Chen, A.; Searchinger, T.; Hua, F.; Wood, E.F. Highland cropland expansion and forest loss in Southeast Asia in the twenty-first century. *Nat. Geosci.* **2018**, *11*, 556. [CrossRef]
5. Huang, C.-H.; Ho, H.-C.; Lin, T.-H. Improving the image fusion procedure for high spatiotemporal aerosol optical depth retrieval: A case study of urban area in Taiwan. *J. Appl. Remote Sens.* **2018**, *12*, 042605. [CrossRef]
6. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436. [CrossRef] [PubMed]
7. Li, H.; Wu, X.J.; Kittler, J. Infrared and visible image fusion using a deep learning framework. In Proceedings of the International Conference on Pattern Recognition, Beijing, China, 20–24 August 2018; pp. 2705–2710.
8. Chen, Y.; Gan, W.; Jiao, S.; Xu, Y.; Feng, Y. Salient feature selection for CNN-based visual place recognition. *IEICE Trans. Inf. Syst.* **2018**, *101*, 3102–3107. [CrossRef]
9. Ma, W.; Zhang, J.; Wu, Y.; Jiao, L.; Zhu, H.; Zhao, W. A novel two-step registration method for remote sensing images based on deep and local features. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 4834–4843. [CrossRef]
10. Gao, F.; Masek, J.; Schwaller, M.; Hall, F. On the blending of the landsat and MODIS surface reflectance: Predicting daily landsat surface reflectance. *IEEE Trans. Geosci. Remote Sens.* **2006**, *44*, 2207–2218.
11. Storey, J.; Roy, D.P.; Masek, J.; Gascon, F.; Dwyer, J.; Choate, M. A note on the temporary misregistration of Landsat-8 operational land imager (OLI) and sentinel-2 multi-spectral instrument (MSI) imagery. *Remote Sens. Environ.* **2016**, *186*, 121–122. [CrossRef]
12. Weng, Q. *Remote Sensing Time Series Image Processing*, 1st ed.; CRC Press: Boca Raton, FL, USA, 2018; p. 243.
13. Zhu, X.; Helmer, E.H.; Gao, F.; Liu, D.; Chen, J.; Lefsky, M.A. A flexible spatiotemporal method for fusing satellite images with different resolutions. *Remote Sens. Environ.* **2016**, *172*, 165–177. [CrossRef]
14. Zhu, X.; Cai, F.; Tian, J.; Williams, T. Spatiotemporal fusion of multisource remote sensing data: Literature survey, taxonomy, principles, applications, and future directions. *Remote Sens.* **2018**, *10*, 527.
15. Cheng, Q.; Liu, H.; Shen, H.; Wu, P.; Zhang, L. A spatial and temporal nonlocal filter-based data fusion method. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 4476–4488. [CrossRef]
16. Hilker, T.; Wulder, M.A.; Coops, N.C.; Linke, J.; McDermid, G.; Masek, J.G.; White, J.C. 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. [CrossRef]
17. Zhu, X.; Chen, J.; Gao, F.; Chen, X.; Masek, J.G. An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions. *Remote Sens. Environ.* **2010**, *114*, 2610–2623. [CrossRef]
18. Song, H.; Huang, B. Spatiotemporal satellite image fusion through one-pair image learning. *IEEE Trans. Geosci. Remote Sens.* **2012**, *51*, 1883–1896. [CrossRef]
19. Huang, B.; Song, H. Spatiotemporal reflectance fusion via sparse representation. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 3707–3716. [CrossRef]
20. Moosavi, V.; Talebi, A.; Mokhtari, M.H.; Shamsi, S.R.F.; Niazi, Y. A wavelet-artificial intelligence fusion approach (WAIFA) for blending landsat and MODIS surface temperature. *Remote Sens. Environ.* **2015**, *169*, 243–254. [CrossRef]
21. Tan, Z.; Di, L.; Zhang, M.; Guo, L.; Gao, M. An enhanced deep convolutional model for spatiotemporal image fusion. *Remote Sens.* **2019**, *11*, 2898. [CrossRef]
22. Song, H.; Liu, Q.; Wang, G.; Hang, R.; Huang, B. Spatiotemporal satellite image fusion using deep convolutional neural networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 821–829. [CrossRef]
23. Li, X.; Ling, F.; Foody, G.M.; Ge, Y.; Zhang, Y.; Du, Y. Generating a series of fine spatial and temporal resolution land cover maps by fusing coarse spatial resolution remotely sensed images and fine spatial resolution land cover maps. *Remote Sens. Environ.* **2017**, *196*, 293–311. [CrossRef]
24. 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. [CrossRef]
25. Xue, J.; Leung, Y.; Fung, T. An unmixing-based bayesian model for spatio-temporal satellite image fusion in heterogeneous landscapes. *Remote Sens.* **2019**, *11*, 324. [CrossRef]
26. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation applied to handwritten zip code recognition. *Neural Comp.* **1989**, *1*, 541–551. [CrossRef]
27. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [CrossRef]
28. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587. [CrossRef]
29. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
30. Eigen, D.; Krishnan, D.; Fergus, R. Restoring an image taken through a window covered with dirt or rain. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 633–640. [CrossRef]
31. Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; Zhang, L. Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Trans. Image Process.* **2017**, *26*, 3142–3155. [CrossRef] [PubMed]
32. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 295–307. [CrossRef] [PubMed]

33. Kim, J.; Kwon Lee, J.; Mu Lee, K. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 27–30 June 2016; pp. 1646–1654. [CrossRef]
34. Svoboda, P.; Hradis, M.; Barina, D.; Zemcik, P. Compression artifacts removal using convolutional neural networks. *J. WSCG* **2016**, *24*, 63–72.
35. He, W.; Yokoya, N. Multi-temporal sentinel-1 and-2 data fusion for optical image simulation. *ISPRS Int. J. Geoinf.* **2018**, *7*, 389. [CrossRef]
36. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the International Conference on Machine Learning (ICML-10), Haifa, Israel, 21–24 June 2010; pp. 807–814.
37. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 7–9 July 2015; Volume 37, pp. 448–456.
38. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
39. Jarihani, A.A.; McVicar, T.R.; Van Niel, T.G.; Emelyanova, I.V.; Callow, J.N.; Johansen, K. Blending Landsat and MODIS data to generate multispectral indices: A comparison of "Index-then-Blend" and "Blend-then-Index" approaches. *Remote Sens.* **2014**, *6*, 9213–9238. [CrossRef]
40. Teo, T.A.; Shih, T.Y.; Chen, B. Automatic georeferencing framework for time series formosat-2 satellite imagery using open source software. In Proceedings of the Asian Conference on Remote Sensing, New Delhi, India, 23–27 October 2017.
41. McInerney, D.; Kempeneers, P. Orfeo toolbox. In *Open Source Geospatial Tools*; Springer: Cham, Switzerland, 2015; pp. 199–217.
42. Stone, H.S.; Orchard, M.T.; Chang, E.C.; Martucci, S.A. A fast direct fourier-based algorithm for subpixel registration of images. *IEEE Trans. Geosci. Remote Sens.* **2001**, *39*, 2235–2243. [CrossRef]
43. Randrianjatovo, R.N.; Rakotondraompiana, S.; Rakotoniaina, S. Estimation of land surface Temperature over reunion island using the thermal infrared channels of Landsat-8. In Proceedings of the IEEE Canada International Humanitarian Technology Conference (IHTC), Montreal, Canada, 1–4 June 2014.
44. Shannon, C.E. A mathematical theory of communication. *Bell. Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
45. Mittal, A.; Moorthy, A.K.; Bovik, A.C. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.* **2012**, *21*, 4695–4708. [CrossRef] [PubMed]
46. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
47. Hore, A.; Ziou, D. Image quality metrics: PSNR vs. SSIM. In Proceedings of the International Conference on Pattern Recognition (ICPR 2010), Istanbul, Turkey, 23–26 August 2010; pp. 2366–2369.