Hindawi
Journal of Sensors
Volume 2020, Article ID 8873079, 11 pages
https://doi.org/10.1155/2020/8873079

Research Article

Spatiotemporal Fusion of Remote Sensing Image Based on Deep Learning

Xiaofei Wang and Xiaoyi Wang

School of Electronic Engineering, Heilongjiang University, Harbin 150080, China

Correspondence should be addressed to Xiaofei Wang; nk_wxf@hlju.edu.cn

Received 23 April 2020; Revised 21 May 2020; Accepted 17 June 2020; Published 29 June 2020

Academic Editor: Sang-Hoon Hong

High spatial and temporal resolution remote sensing data play an important role in monitoring the rapid change of the earth surface. However, there is an irreconcilable contradiction between the spatial and temporal resolutions of the remote sensing image acquired from a same sensor. The spatiotemporal fusion technology for remote sensing data is an effective way to solve the contradiction. In this paper, we will study the spatiotemporal fusion method based on the convolutional neural network, which can fuse the Landsat data with high spatial but low temporal resolution and MODIS data with low spatial but high temporal resolution, and generate time series data with high spatial resolution. In order to improve the accuracy of spatiotemporal fusion, a residual convolution neural network is proposed. MODIS image is used as the input to predict the residual image between MODIS and Landsat, and the sum of the predicted residual image and MODIS data is used as the predicted Landsat-like image. In this paper, the residual network not only increases the depth of the superresolution network but also avoids the problem of vanishing gradient due to the deep network structure. The experimental results show that the prediction accuracy by our method is greater than that of several mainstream methods.

1. Introduction

Due to the limitation of the hardware technology of the remote sensing satellite and the cost of satellite launching, it is difficult for the same satellite to obtain the remote sensing image with both high spatial and temporal resolutions. Landsat series satellites can obtain multispectral data with a spatial resolution of 30 m. While multispectral images reflect the spectral information of ground features, when performing classification and other processing, unlike hyperspectral, which has rich dimensions, dimensionality reduction processing is required. Although there are many dimensionality reduction methods, it can achieve dimensionality reduction of hyperspectral images [1]. But multispectral image imaging is more convenient, making it widely used in many fields. With this feature, Landsat data has been widely used in the exploration of earth resources, management of agriculture, forestry, animal husbandry, and natural disaster and environmental pollution monitoring [2–4]. However, the 16-day visit circle of the Landsat satellite and the impact of cloud pollution limit their potential use in monitoring and researching the land surface dynamic changes. On the other hand, Moderate-resolution Imaging Spectroradiometer (MODIS) on Terra/Aqua satellite has a visit circle per 1-2 days, which has a high temporal resolution and can be applied in vegetation phenology [5, 6] and other fields. However, the spatial resolution of MODIS data is 250-1000 m, which has a poor representation of the details of the ground objects and is not enough to observe the heterogeneous landscape.

In 1995, Vignolles et al. [7] first proposed to generate high spatiotemporal resolution data by using spatiotemporal fusion technology. Since then, different types of spatiotemporal fusion methods have been emerging. The spatiotemporal fusion technology of remote sensing images is fused with the spatial features of high spatial but low temporal resolution images and the temporal features of low spatial but high temporal resolution images to generate time series images with high spatial resolution. According to the principle, the existing spatiotemporal fusion models can be divided into three types: reconstruction based, spatial unmixing based, and learning based.

The basic principle of reconstruction-based methods is to calculate the reflectance of the center fusion pixel through a weighting function which takes full account of the spectral, temporal, and spatial information in similar pixels. Gao et al. [8] first proposed the spatiotemporal adaptive reflection fusion model (STARFM), which uses a pair of MODIS and ETM+ reflectance images at known time phase and MODIS reflectance images at predicted time phase to generate 30 m spatial resolution image. Hilker et al. [9] proposed a spatiotemporal fusion algorithm for mapping reflectance change (STAARCH) based on tasseled cap change. The algorithm can not only generate 30 m spatial resolution ETM + like images but also detect highly detailed surface classes. However, the fusion accuracy of STARFM and STAARCH are highly related to the surface landscape heterogeneity, resulting in low fusion accuracy for heterogeneous area. David et al. [10] considered the influence of bidirectional reflectance effect, proposing a semiphysical method to generate fused Landsat ETM+ reflectance using MODIS and Landsat ETM+ data. Zhu et al. [11] based on STARFM considering the reflectivity difference between different sensor imaging systems due to different orbital parameters, band bandwidth, spectral response curve and other factors, the transfer coefficient between different sensor differences is increased, and the enhanced STARFM (ESTARFM) model is proposed to improve the fusion accuracy of complex surface area (heterogeneous area) to a certain extent. The model uses two sets of MODIS and ETM+ reflectance images and MODIS reflectance images to generate 30 m spatial resolution ETM+ like image. Wang and Atkinson [12] proposed a spatiotemporal fusion algorithm consisting of three parts: regression model fitting (RM Fitting), spatial filtering (SF), and residual compensation (RC), referred to as Fit-FC; this method only uses a pair of known high-low resolution image pair as input and can better predict the spatial change between images in different periods. Chiman et al. [13] proposed a simple and intuitive method and has two steps. First, a mapping is established between two MODIS images where one is at an earlier time, t_1, and the other one is at the time of prediction, t_p. Second, this mapping is applied to a known Landsat image at t_1 to generate a predicted Landsat image at t_p.

Spatial-temporal fusion methods based on spatial unmixing performs spatial unmixing of pixels in known low-resolution images and applies the classification results to high-resolution images at the unknown time to predict high-resolution images. Zhukov et al. [14] proposed a spatiotemporal fusion method that considers the spatial variability of pixel reflectivity based on the assumption that the pixel reflectivity does not change drastically between neighbor pixels. This method introduces window technology to predict the high-resolution reflectance of each type of feature. This method is not ideal for the farmland area which changes dramatically in a short time. Wu [15] proposed a spatial and temporal data fusion approach (STDFA) based on the assumption that the temporal variation characteristics of the class reflectivity are consistent with the intraclass pixel reflectivity. This method extracts the temporal change information of ground features from time-series low spatial resolution images and performs classification and density segmentation on known two periods of high spatial resolution images to obtain classified images, so as to obtain class average reflectance for image fusion. On the basis of [15], Wu and Huang [16] comprehensively considered the spatial variability and temporal variation of pixel reflectivity and proposed an enhanced STDFA; the method solves the problem of missing remote sensing data. Hazaymeh and Hassan [17] proposed a relatively simple and more efficient algorithm; the Spatiotemporal Image-Fusion Model (STI-FM) applies clustering to the images first, and, for each cluster, performs a separate prediction. Zhu et al. [18] proposed a flexible spatiotemporal data fusion (FSDAF) based on spectral demixing analysis and thin-plate spline interpolation. The algorithm uses less input data, which is suitable for heterogeneous areas and can effectively preserve the low-resolution details of the image during the prediction period.

In remote sensing image processing, learning-based methods are more commonly used for classification of ground features [19]. In recent years, spatiotemporal fusion methods based on learning have been widely concerned. In 2012, Huang and Song [20] first introduced sparse representation technology into the process of spatiotemporal fusion and proposed a sparse representation based on a spatiotemporal reflectance fusion model (SPSTFM), which uses the MODIS and ETM+ images of the front and back phases at the predicted time phase. First, use high- and low-resolution difference images to train a couple dictionary representing high- and low-resolution features, and then use a low-resolution image to predict high-resolution image. Song and Huang [21] proposed a sparse representation of spatiotemporal reflectance fusion model using only a pair of known high- and low-resolution image pair, which first enhanced MODIS image by sparse representation to obtain a transition image, then predicted image is generated by combining known high-resolution image with transition image through high-pass modulation. The model reduces the number of known image pair that needs to be inputted, so that the algorithm can be applied in the case of lack of data and has more general applicability. Spatiotemporal fusion method based on feature learning considers the spatial information of changing image. However, there are some limitations in previous methods based on sparse representation. First, the image features need to be designed, which brings complexity and instability to performance. Secondly, the method does not consider the large amount of actual remote sensing data but only develops and validates the algorithm for small-scale research areas.

The convolutional neural network (CNN) [22] model has a simple structure and can be used to solve the problems of target recognition [23] and image classification [24] in computer vision. In recent years, CNN has also been used in the field of superresolution. As the pioneer CNN model for SR, superresolution convolutional neural network (SRCNN) [25] predicts the nonlinear LR-HR mapping via a fully convolutional network and significantly outperforms classical non-DL methods. In the field of remote sensing, Song et al. [26] proposed a five-layer convolutional neural network

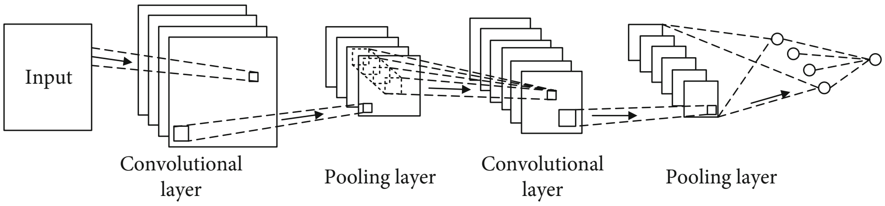

Figure 1: Flowchart of CNN network structure.

(CNN) spatiotemporal fusion model. This model is similar to [21] and is a two-stage model. It learns the CNN nonlinear mapping between MODIS and Landsat images and combines high-pass modulation with a weighting strategy to predict Landsat-like images. Liu et al. [27] proposed a two-stream convolutional neural network *Stf*Net, which not only considered the temporal dependence of remote sensing images but also introduced temporal constraint, the network takes a coarse difference image with the neighboring fine image as inputs and the corresponding fine difference image as output, the method can restore spatial details greater. At present, there are two main problems faced by learning-based spatiotemporal fusion methods: first, the deep-seated network can improve the prediction accuracy; however, the deep-seated network will lead vanishing gradient or convergence difficulty and second, it is difficult to obtain two pairs of suitable prior image pairs as the input of network training. For example, *Stf*Net is a fusion method using two pairs of prior images as input. Considering the above two points, we propose a spatiotemporal fusion model based on residual convolution neural network. The model can only uses a pair of prior images as train input. The MODIS image is very similar to the predicted Landsat image. In other words, the low-frequency information of low-resolution image is similar to that of high-resolution image. In fact, the low-resolution image and the high-resolution image only lack the residual of the high-frequency part. If only train the high-frequency residual between the high resolution and the low resolution, which does not need to spend too much time in the low-frequency part. And can deepen the network structure to avoid problems such as gradient disappearance. For this reason, we introduce the idea of ResNet [28] and set up a spatiotemporal fusion framework of remote sensing image suitable for a small-sample training set for CNN. Considering the time dependence between the image sequences, we use the MODIS-Landsat image pairs of the front and back phases of the prediction image to construct the prediction network, respectively. The experimental results show that compared with benchmark methods, the spectral color and spatial details of our method are closer to the real Landsat image.

The rest of this paper is divided into three sections. In Section 2, the principle of residual CNN is introduced. Section 3 provides the experimental verification process and results. Section 4 gives the conclusion.

2. Methods

In this paper, we use CNN and ResNet to construct a dual stream network to predict Landsat-like images. The principles involved are briefly introduced as follows.

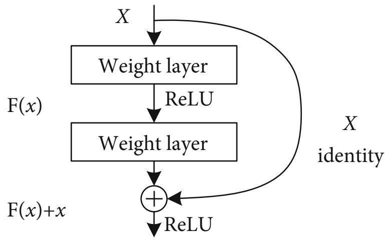

Figure 2: Residual learning unit.

2.1. CNN. Convolutional neural network (CNN) is one of the most representative network models in the deep learning method [29]. With the continuous development of deep learning techniques in recent years, it has achieved very good results in the field of image processing. Compared with the traditional data processing algorithm, CNN avoids the complicated preprocessing work such as manually extracting data from the data, so that it can be directly used in the original data.

CNN is a nonfully connected multilayer neural network, as shown in Figure 1. The main structure consists of a convolutional layer, pooling layer, activation layer, and fully connected layer [30]. The convolutional layer, pooling layer, and activation layer are the feature extraction layers of CNN, which are used to extract the signal features. The fully connected layer is the CNN classifier. Since this paper mainly uses the deep convolution network to extract the spatial characteristics of the remote sensing image, the feature extraction layer of deep convolutional neural networks is analyzed.

2.2. Residual Learning. If the input of a neural network is x, and the expected output is $H(x)$; $H()$ is the expected mapping. If we want to learn such a model, the training difficulty will be greater; if we have learned the more saturated accuracy (or when we find that the error in the lower layer becomes larger), then the next learning goal will be transformed into the learning of identity mapping, that is, to make the input x an approximate output $H(x)$, which is in order to keep in the later hierarchy without causing a drop in accuracy.

As shown in the residual network structure diagram in Figure 2, input x is directly transferred to the output as the initial result through "shortcut connections," and the output result is $H(x) = F(x) + x$. When $F(x) = 0$, then $H(x) = x$, which is the constant mapping mentioned above. Therefore, ResNet is equivalent to changing the learning goal, not a complete output of learning, but the difference between the goal value $H(x)$ and x, that is, the so-called residual $F(x) =$

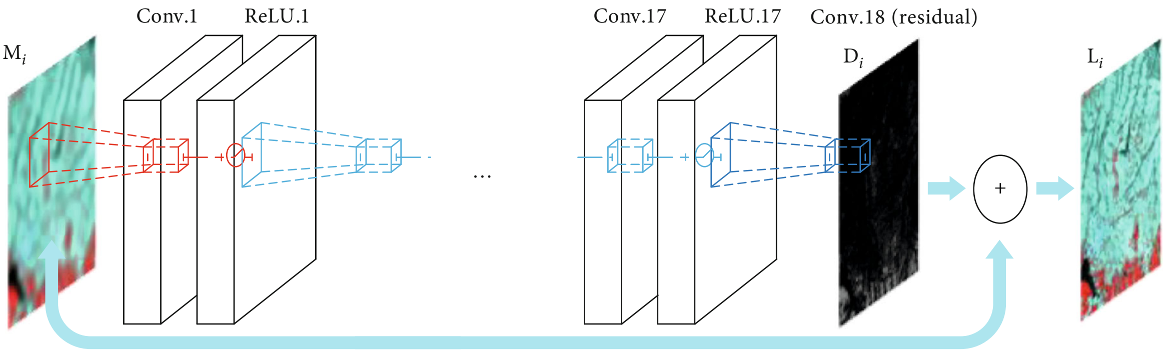

Figure 3: Flowchart illustrating the proposed scheme of two-stream residual learning in CNN.

$H(x) - x$. Therefore, the later training goal is to approach the residual result to 0, so that with the deepening of the network, the accuracy does not decline.

2.3. Spatiotemporal Fusion Using Residual Learning in CNN. In this paper, Landsat image is regarded as high spatial but low temporal resolution data; MODIS image is regarded as high temporal but low spatial resolution data. We express the Landsat image and MODIS image at t_i as $\mathbf{L}_i$ and $\mathbf{M}_i$, respectively. If there are two pairs of prior images, the two-stream residual CNN network uses the known Landsat-MODIS image pair at t_1 and t_3, and MODIS image at t_2 to predict Landsat-like image.

2.3.1. Training Stage. In the training stage, in order to build an nonlinear mapping model between MODIS and Landsat-MODIS residual images, we first up sample the spatial resolution of $\mathbf{M}_i$ to the same size as $\mathbf{L}_i$. Then, the Landsat and MODIS images at the same time are differenced to obtain a residual image $\mathbf{D}_i$. Thus, we expect to learn a mapping function $f(x)$ which approximates $\mathbf{D}_i$. Pixel value in $\mathbf{D}_i$ are likely to be zero or small. We want to predict this residual image. The loss function now becomes $1/2\|y - f(x)\|^2$, where $f(x)$ is the network prediction. We divide the high- and low-resolution images corresponding on the same time into overlapping image patches. Define the set of high- and low-resolution samples as $\mathbf{X}$ and $\mathbf{Y}$, where the corresponding samples are $\mathbf{x}$ and $\mathbf{y}$. The overlapping segmentation is performed here to increase the number of training samples. After predicting the residual image, the ground truth Landsat image is obtained by the sum of the input MODIS image and the predicted residual image.

In the network, the loss layer has three inputs: residual estimation, input MODIS image, and Landsat image. The loss is calculated as the Euclidean distance between the reconstructed image and the real Landsat image. In order to achieve the purpose of high-precision spatiotemporal fusion, we use a very deep convolutional network. We use 18 layers where layers except the first and the last are of the same type: 64 filters of the size $3 \times 3 \times 64$, where a filter operates on 3×3 spatial region across 64 channels (feature maps). The first layer operates on the input image. The last layer, used for image reconstruction, consists of a single filter of size $3 \times 3 \times 64$. The process structure is shown in Figure 3.

Training was performed by using back-propagation-based minibatch gradient descent to optimize regression targets. We set the momentum parameter to 0.9. Training is regularized by weight loss (l_2 penalty multiplied by 0.0001).

2.3.2. Prediction Stage. There are two pairs of prior Landsat-MODIS images and the MODIS image on prediction date, we aim to fuse them to predict the Landsat-like image on prediction date. Denote the prior dates as t_1 and t_3, the prediction date as t_2, we predict $\mathbf{L}_2$ based on the residual learning CNN. $\mathbf{M}_i$, $\mathbf{L}_i$, and $\mathbf{D}_i$ are divided into patches, and their corresponding image patches are $\mathbf{m}_i^k$,$\mathbf{l}_i^k$, and $\mathbf{d}_{1,}^k$ respectively. Taking $\mathbf{m}_1^k$ as the input of CNN, the label is $\mathbf{l}_i^k$, and the sum of the residual image $\mathbf{d}_1^k$ and $\mathbf{m}_1^k$ is used as the prediction. In this paper, the number of network layers is set to 18. In the process of reconstruction, input $\mathbf{m}_2^k$ into the trained network and get the predicted $\mathbf{l}_2^{1k}$. Similarly, $\mathbf{l}_2^{3k}$ can be predicted by Landsat-MODIS image pair at t_3. Considering the temporal correlation between the image at the predicted time and the reference image, we use the corresponding temporal weight when reconstructing each image patch. Finally, the high spatial resolution image patch at the predicted time is obtained:

$$\mathbf{l}_2^k = \omega_1^k * \mathbf{l}_2^{1k} + \omega_3^k * \mathbf{l}_2^{3k}, \qquad (1)$$

where $\mathbf{l}_2^{1k}$ and $\mathbf{l}_2^{3k}$ are the kth predicted patch using $\mathbf{L}_1$ and $\mathbf{L}_3$ as the reference image, respectively, ω_1^k and ω_3^k are the corresponding weight, and determined as follows:

$$\omega_i^k = \frac{1/v_i^k}{\left(1/v_1^k\right) + \left(1/v_3^k\right)} \,(i = 1, 3). \qquad (2)$$

The local weight is calculated by the sum U of normalized difference vegetation index (NDVI) [31] and normalized difference built-up index (NDBI) [32], where v_i^k is used to measure the change degree between MODIS images at two times, and it is the absolute average change of U in $\mathbf{m}_{ij}^k$, where $\mathbf{m}_{ij}^k$ represents the MODIS image change at different times. After each image patch is reconstructed one by one, it is restored to the whole image. In order to ensure the continuity of the reconstructed image, there is an overlap between adjacent patches, and the pixel value of the overlapped part of the

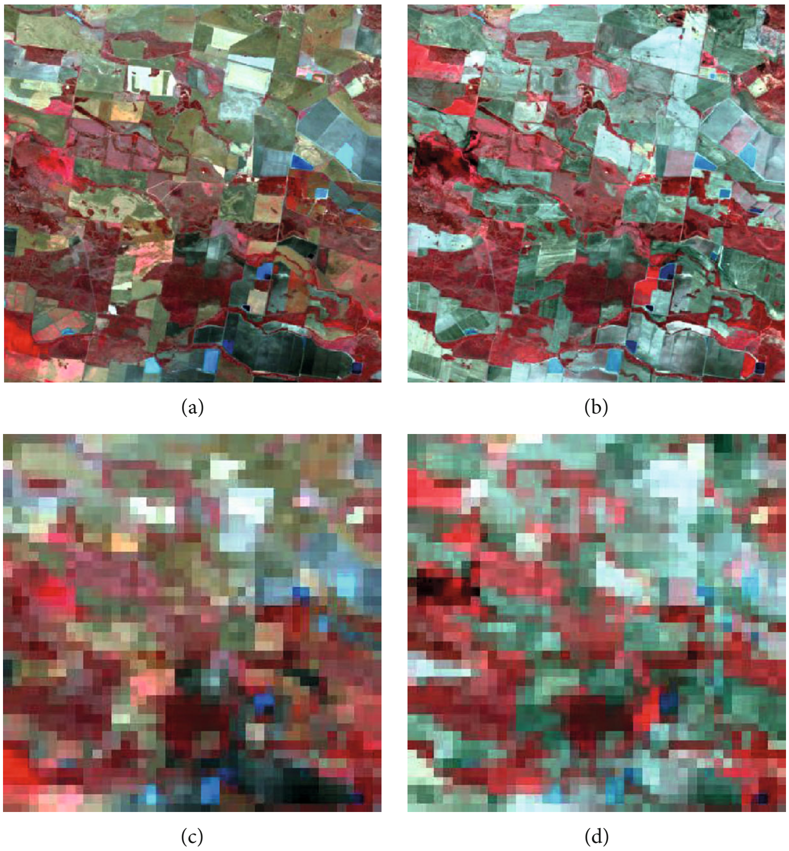

Figure 4: 30 m Landsat 8 and 500 m MODIS images for the first dataset (green, red, and NIR bands as RGB). (a) and (b) are 30 m Landsat images on 2 July 2013 and 17 August 2013, respectively, (c) and (d) are the corresponding 500 m MODIS images for (a) and (b).

image patch is taken as the mean value when the whole image is restored.

3. Experiments

Two datasets were used in the experiments. The first dataset contains two pairs of MODIS-Landsat images, and the second dataset contains three pairs of MODIS-Landsat images. Both areas are located in Coleambally, New South Wales, Australia. MODIS data uses the surface reflectance of MOD09A1 (500 m) and MOD09Q1 (250 m) for 8-day synthetic products. We up sampled all MODIS in the dataset to the same resolution as the Landsat image of the corresponding date. Compared with natural images, remote sensing images have a large size and rich details. Therefore, the remote sensing images are overlapped and divided into patches to obtain the training set. In the paper, the images of the two areas are overlapped divided into 33×33 image patches. The above image patch set is used as the train set and prediction set. We compare our method with the mainstream and advanced methods (including STARFM, FSDAF, Fit-FC, STDFA, STI-FM, HCM, ESTARFM, SPSTFM, and *Stf*Net), which will be described in detail in this section.

3.1. Experiment on the First Dataset. In order to verify the applicability of our proposed spatial-temporal fusion method based on residual convolutional neural network for one prior Landsat-MODIS image pair, we use a single-stream network to verify and compare the same data as the input of STARFM, FSDAF, Fit-FC, STDFA, STI-FM, and HCM.

In this experiment, two pairs of Landsat and MODIS surface reflectance images covering a 20 km × 20 km area in Coleambally are used. The two pair images were acquired on 2 July 2013 and 17 August 2013. Figure 4 shows the 30 m Landsat images (upper row) and 500 m MODIS images (lower row) using green-red-NIR as RGB composite image. Then, we use the bicubic interpolation method to downscale the 500 m MODIS image into 30 m. Our experimental task used the pair of Landsat-MODIS images on 2 July 2013 and the MDOIS image on 17 August 2013 to predict the 30 m Landsat-like image on 17 August 2013. At the same time, STARFM, FSDAF, Fit-FC, STDFA, STI-FM, and HCM are tested with the same input in this experiment, and the true 30 m Landsat image acquired on 17 August 2013 is used as the reference to evaluate the accuracy of fusion results.

Figure 5 shows the fusion results by four methods (STARFM, Fit-FC, FSDAF, STDFA, STI-FM, HCM, and our method). Obviously, the prediction accuracy by our method is greater. For example, the highlighted areas in the bottom left part of subarea S, for Fit-FC, STDFA, FSDAF, STI-FM, HCM, and STARFM, some dark green pixels are incorrectly predicted as purple pixels. In addition, the highlighted areas in the bottom right part of the subarea, Fit-FC, STDFA, FSDAF, STI-FM, HCM, and STARFM

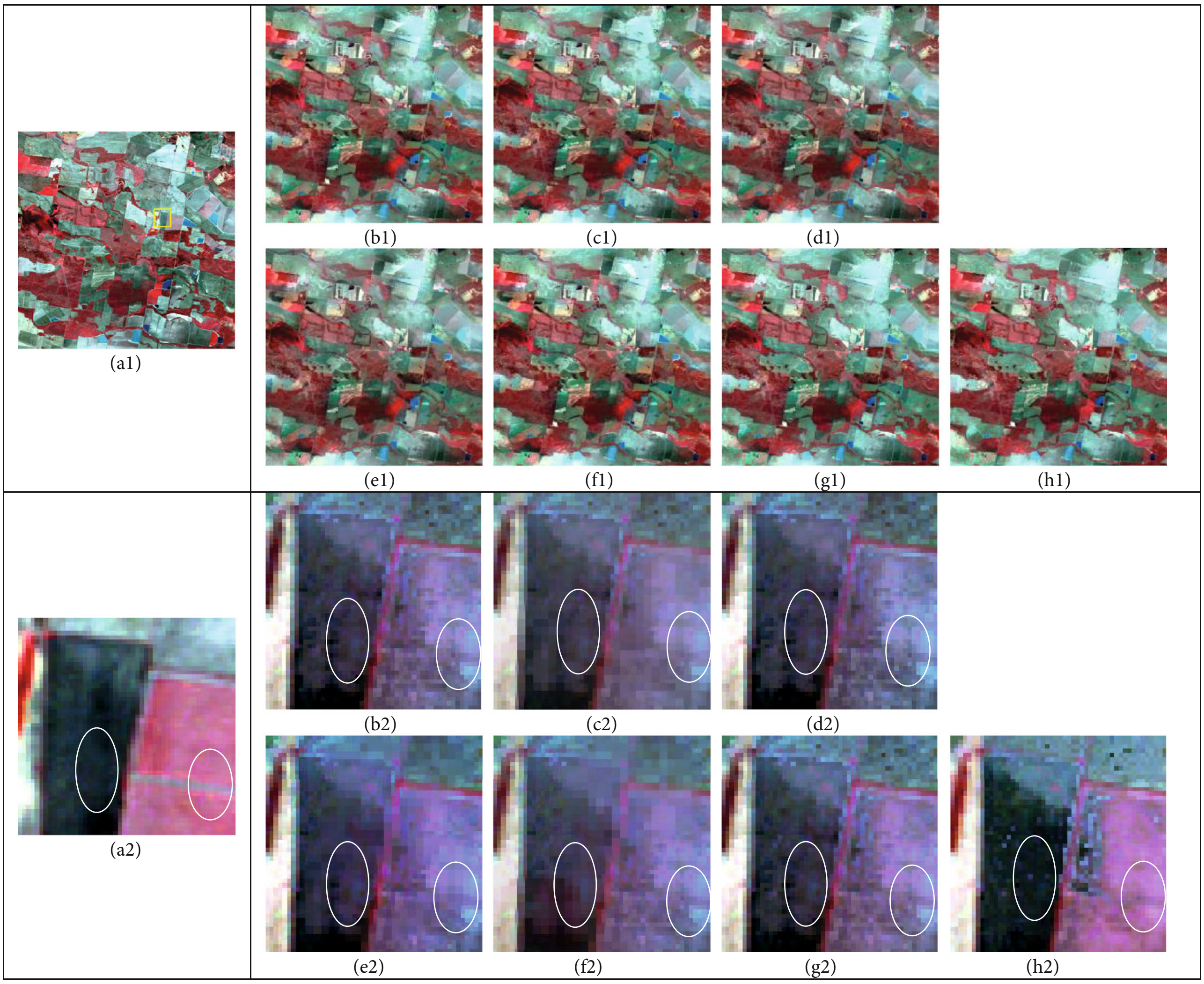

Figure 5: 30 m Landsat 8 results for the first dataset (green, red, and NIR bands as RGB). (a1) is the 30 m true Landsat 8 image on 17 August, 2013. (b1) is the 30 m Fit-FC-derived Landsat 8 images on 17 August, 2013. (c1) is the 30 m STDFA-derived Landsat 8 images on 17 August, 2013. (d1) is the 30 m FSDAF-derived Landsat 8 images on 17 August, 2013. (e1) is the 30 m STI-FM-derived Landsat 8 images on 17 August, 2013. (f1) is the 30 m HCM-derived Landsat 8 images on 17 August, 2013. (g1) is is the 30 m STARFM-derived Landsat 8 images on 17 August, 2013. (h1) is the 30 m our method-derived Landsat 8 images on 17 August, 2013. Row 1 shows the results for the whole area, and Row 2 is the results for a heterogeneous subareas (S) marked in (a1).

incorrectly predicted some red pixels as purple and blue pixels. However, our method is closer to the reference image. The main reason is that the Fit-FC method directly applies the known linear coefficients of the low-resolution image to fit the high-resolution image on the prediction period. Therefore, when the spatial resolution difference between the high- and low-resolution images is large, there will be obvious “block effect,” for example, the spatial resolution difference between Landsat image and MODIS image is nearly 17 times. STDFA assumes that the temporal variation characteristics of the same surface coverage class in coarse pixels are consistent, but there may be inconsistencies in practical applications, so the fusion result is affected. The accuracy of the FSDAF spatiotemporal fusion algorithm is low, which is mainly caused by two aspects: The prediction accuracy of FSDAF is worse, which is mainly caused by two aspects: at first, the known high-resolution data needs to be classified, and the classification accuracy by unsupervised classification method (such as the K-means method) will have a certain impact on the results; at second, when the spatial resolution difference between high- and low-resolution data is large, the endmember (that is, high-resolution pixel) represented area will be more refined. When the number of categories is small, the fusion result will be relatively smooth, and when the number of categories is large, the fitting accuracy will also be reduced (such as in low-resolution pixels, if the richness of a certain category is low, the total prediction error will be increase). STI-FM is susceptible to interference from outliers, so when the spatial characteristics change significantly, the prediction effect is not good. The method of using gradation mapping is greatly affected by the heterogeneous region, so HCM failed to show the best performance in this experiment. STARFM considers the similarity of neighboring pixels, so the prediction accuracy is relatively stable. However, the

Table 1: Quantitative assessment for the Coleambally dataset.

	Bands	Fit-FC	STDFA	FSDAF	STI-FM	HCM	STARFM	Our method
RMSE	Blue	0.0088	0.0089	0.0091	0.0098	0.0094	0.0086	0.0081
	Green	0.0113	0.0115	0.0117	0.0124	0.0120	0.0110	0.0105
	Red	0.0150	0.0152	0.0154	0.0164	0.0159	0.0146	0.0139
	NIR	0.0240	0.0244	0.0248	0.0268	0.0258	0.0233	0.0219
	SWIR1	0.0328	0.0331	0.0339	0.0363	0.0351	0.0318	0.0299
	SWIR2	0.0308	0.0317	0.0319	0.0344	0.0331	0.0299	0.0279
	Mean	0.0205	0.0208	0.0211	0.0227	0.0219	0.0199	0.0187
CC	Blue	0.8761	0.8773	0.8674	0.8463	0.8574	0.8833	0.8967
	Green	0.8786	0.8928	0.8705	0.8510	0.8612	0.8853	0.8976
	Red	0.8925	0.8931	0.8856	0.8687	0.8775	0.8982	0.9085
	NIR	0.7713	0.7722	0.7550	0.7171	0.7367	0.7850	0.8098
	SWIR1	0.8327	0.8331	0.8207	0.7925	0.8072	0.8429	0.8627
	SWIR2	0.8393	0.8398	0.8269	0.7978	0.8130	0.8497	0.8697
	Mean	0.8484	0.8489	0.8377	0.8122	0.8255	0.8574	0.8742
UIQI	Blue	0.8653	0.8662	0.8564	0.8350	0.8462	0.8728	0.8867
	Green	0.8680	0.8687	0.8597	0.8400	0.8503	0.8750	0.8879
	Red	0.8835	0.8841	0.8763	0.8591	0.8681	0.8895	0.9008
	NIR	0.7644	0.7652	0.7488	0.7125	0.7313	0.7775	0.8011
	SWIR1	0.8251	0.8259	0.8132	0.7855	0.7999	0.8352	0.8557
	SWIR2	0.8313	0.8319	0.8190	0.7903	0.8053	0.8418	0.8622
	Mean	0.8396	0.8403	0.8289	0.8037	0.8168	0.8486	0.8657

premise of STARFM is that the spectrum of similar pixels in the neighborhood is constant and there is no land cover change during the observation period, which makes the model susceptible to environmental and phenological changes, resulting in large prediction errors, especially for heterogeneous areas. Our method uses deep convolutional neural networks to more effectively extract the features of low-resolution images and residual images and constructs a mapping relationship between low-resolution images and residual images through a residual learning network. This mapping relationship is nonlinear mapping and is more in line with the change of ground features. In addition, the number of layers in the network is deepened through residual learning, which strengthens the robustness of the network. Therefore, the experimental results based on our method have better visual effects.

Table 1 lists the objective evaluation results of four fusion methods and uses three common fusion evaluation methods of remote sensing image, including root mean square error (RMSE) [33], correlation coefficient (CC) [34], and universal image quality index (UIQI) [35]. The ideal values for RMSE, CC, and UIQI are 0, 1, and 1, respectively. From Table 1, we can see that for the six bands of all fusion results, the fusion results of our method have smaller RMSE and larger CC and UIQI. Our method is compared with other six methods (STARFM, Fit-FC, FSDAF, STDFA, STI-FM, and HCM); the gain of the mean CC is 0.0259, 0.0365 0.0168, 0.0253, 0.0620, and 0.0487, and the gain of the mean UIQI is 0.0261, 0.0368, 0.0175, 0.0254, 0.0620, and 0.0489, respectively. The mean RMSE is reduced by 0.0018, 0.0024, 0.0012, 0.0021, 0.0040, and 0.0032, respectively. In addition, the fusion result based on our method is better than STATFM, and STARFM is better than Fit-FC, the rest of the sequence is STDFA>FSDAF>HCM>STI-FM. The main reason is when the spatial resolution difference between high- and low-resolution images is large, Fit-FC directly applies the fitting coefficients of low-resolution images into high-resolution images, which causes large errors; FSDAF also has similar fitting errors. STDFA assumes that the temporal variation characteristics of the same surface coverage class in coarse pixels are consistent, but there may be inconsistencies in practical applications, so the fusion result is affected. Although STARFM considers the similarity of neighboring pixels, the reconstruction method of each pixel cannot consider the continuity of the image. STI-FM is susceptible to interference from outliers, so when the spatial characteristics change significantly, the prediction effect is not good. HCM using gradation mapping is greatly affected by the heterogeneous region. Our method can better restore the continuity of the image by reconstructing the image patch.

3.2. Experiment on the Second Dataset. In this experiment, three pairs of Landsat-MODIS images covering 30 km × 30 km area of Coleambally are used to verify the applicability of our method for two pairs of prior images. The three pairs of images were acquired on 6 April, 2012, 12 May, 2012, and 20 July, 2012, respectively. Figure 6 shows the 30 m Landsat

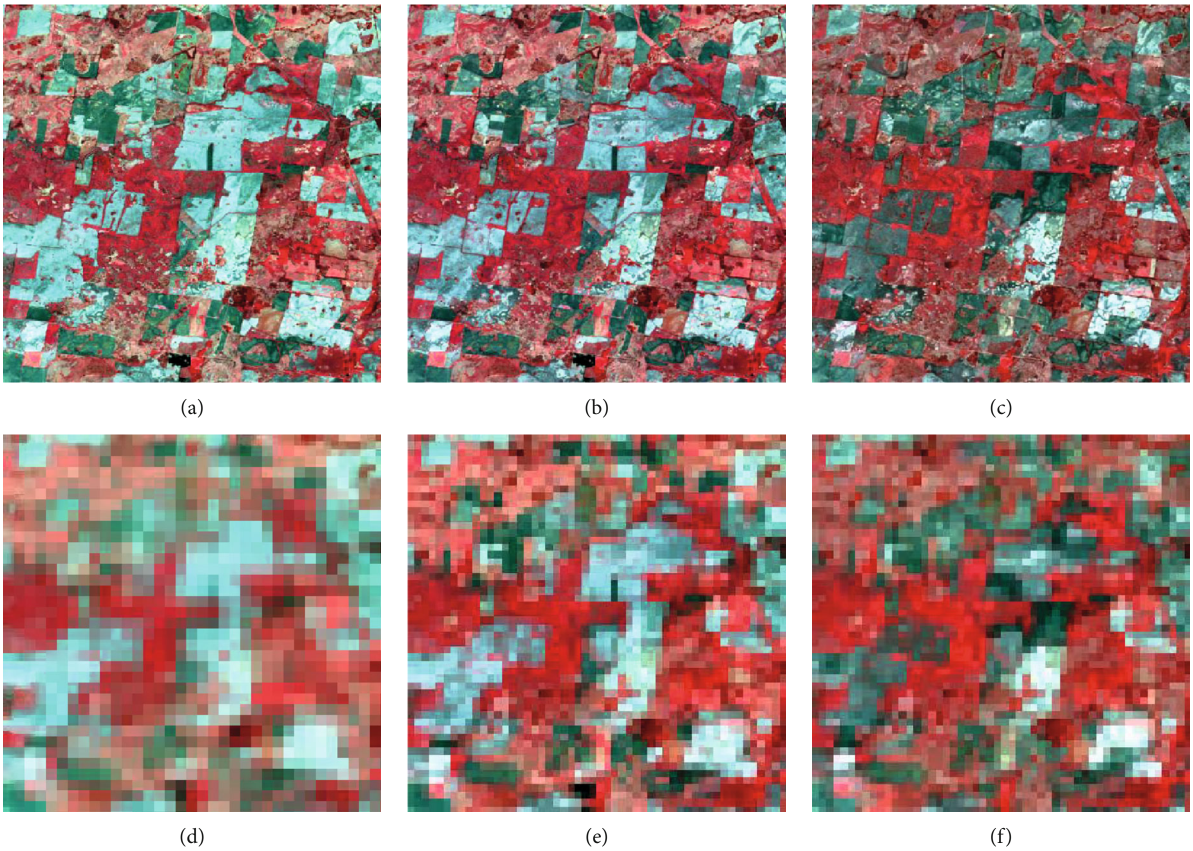

Figure 6: 30 m Landsat and 500 m MODIS images for the second dataset (green, red, and NIR bands as RGB). (a), (b), and (c) are 30 m Landsat images on 6 April, 2012, 12 May, 2012, and 20 July, 2012, respectively, and (d–f) are the corresponding 500 m MODIS images for (a–c).

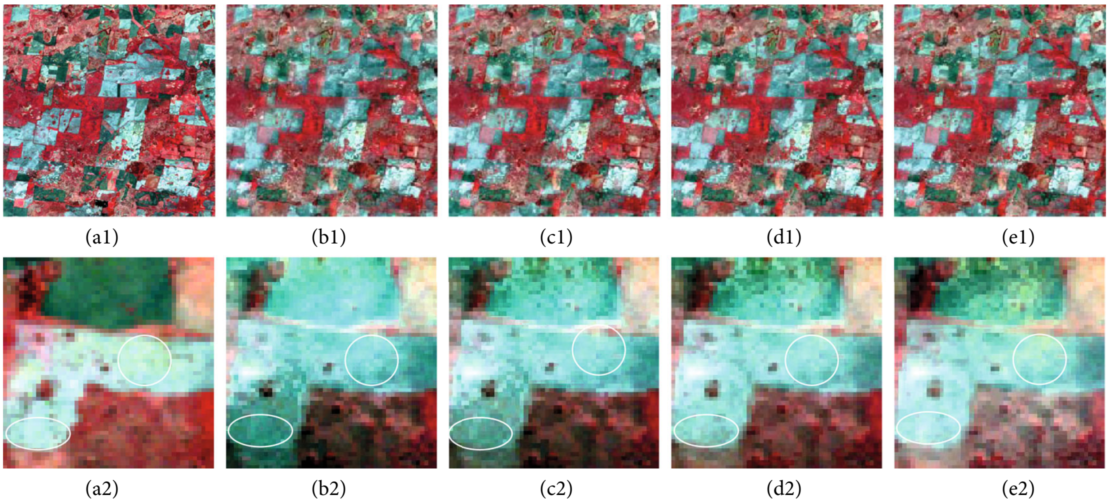

Figure 7: 30 m Landsat results for the second dataset (green, red, and NIR bands as RGB). (a1) is the 30 m true Landsat image on July 11, 2012. (b1) is the 30 m ESTARFM-derived Landsat images on 12 May, 2012. (c1) is the 30 m SPSTFM-derived Landsat images on 12 May, 2012. (d1) is the 30 m *Stf*Net-derived Landsat images on 12 May, 2012. (e1) is the 30 m our method-derived Landsat images on 12 May, 2012. Row 1 shows the results for the whole area, and Row 2 is the results for the heterogeneous sub-areas (S1) marked in (a1).

image (upper row) and 500 m MODIS image (lower row) using green-red-NIR as RGB composite image. This experiment is to verify the accuracy of our methods based on two pairs of prior images, we use the two image pairs on 6 April, 2012 and 20 July, 2012, and MODIS image on 12 May, 2012 to predict the Landsat-like image on 12 May, 2012.

Figure 7 shows the 30 m prediction results on August 12, 2012 based on the four methods (ESTARFM, SPSTFM, *Stf*Net, and our method). It is worth noting that the ESTARFM result is the worst, *Stf*Net is better than SPSTFM, and our method is better than *Stf*Net. For example, the highlighted areas in the bottom left of subarea S, ESTARFM, and SPSTFM incorrectly predicted some light green pixels as dark green pixels. Although the prediction by *Stf*Net is similar to the reference image, but there are some yellow pixels which had been incorrectly predicted to be blue pixels. However, for our method, the prediction is closer to the true reference image. Compared with the three benchmark methods, our method provides excellent performance. The main reason is ESTARFM assumes that during the observation period, the conversion coefficients between high- and low-resolution images remain unchanged, but in actual conditions, land types and coverage will change, so this assumption is not applicable in the areas with significant changes. SPSTFM utilizes sparse representation and dictionary learning approaches in the signal domain to increase prediction accuracy for land cover change and heterogeneous region. Although the network structure, compared with our method, SPSTFM only applicable for small-scale regions and cannot extract sufficient image features. Although *Stf*Net can produce more accurate prediction results by deep network, but the data contained in the training process is too large and the network is hard to convergent, which also has a certain impact on the prediction accuracy. Our method using residual learning network can only learn the difference information between high- and low-resolution images. Since the low-frequency information between high- and low-resolution images is similar, if we directly learn mapping relationship between high- and low-resolution images, it will increase the amount of calculation, which also introduces errors. Through residual learning, not only the nonlinear mapping relationship between high- and low-frequency information can be directly learned, but also the network layers can be deepened, which enhances the accuracy and stability of the network structure.

Table 2 shows the comparison results in RMSE, CC, and UIQI. From Table 2, we can see that for six bands, the fusion results by our method can obtain smaller average RMSE and larger CC and UIQI. It is easy to find that our method is better than *Stf*Net, the *Stf*Net is better than SPSTFM, and ESTARFM is the worst among the four approaches. Specifically, the CC gains of our method over ESTARFM, SPSTFM, and *Stf*Net are 0.0508, 0.0257, and 0.0134, and the UIQI gains are 0.0467, 0.0238, and 0.0126, respectively. The main reason is that ESTARFM assumes that the conversion coefficients remain unchanged during the observation period, but there are land cover types in this area, such as subregion S, so the conversion coefficients are not consistent, so the prediction results are greatly biased. SPSTFM takes the image patch as the reconstruction unit and considers the continuity between adjacent pixels, so it has strong robustness in dealing with complex surface changes. However, due to the instability of forcing the same sparse coefficient of high- and low-resolution dictionaries to construct the mapping relationship, the performance in this experiment is worse than our method. *StfNet* has a deep network layers; however, it is difficult to converge due to directly training the mapping relationship between high- and low-resolution images, which also leads to network instability. Our method through residual network not only improves the stability of the network but also enhances the accuracy of the fusion results.

Table 2: Quantitative assessment for the second dataset.

	Bands	ESTARFM	SPSTFM	*Stf*Net	Our method
RMSE	Blue	0.0118	0.0112	0.0109	0.0104
	Green	0.0139	0.0132	0.0128	0.0122
	Red	0.0217	0.0208	0.0203	0.0196
	NIR	0.0361	0.0332	0.0317	0.0301
	SWIR1	0.0306	0.0288	0.0279	0.0268
	SWIR2	0.0369	0.0356	0.0348	0.0340
	Mean	0.0252	0.0238	0.0231	0.0222
CC	Blue	0.8512	0.8727	0.8837	0.8964
	Green	0.8520	0.8729	0.8833	0.8951
	Red	0.8537	0.8756	0.8864	0.8990
	NIR	0.7131	0.7601	0.7824	0.8055
	SWIR1	0.8519	0.8735	0.8843	0.8955
	SWIR2	0.8725	0.8900	0.8987	0.9077
	Mean	0.8324	0.8575	0.8698	0.8832
UIQI	Blue	0.8425	0.8622	0.8723	0.8847
	Green	0.8431	0.8624	0.8719	0.8834
	Red	0.8297	0.8483	0.8574	0.8688
	NIR	0.7040	0.7500	0.7714	0.7940
	SWIR1	0.8455	0.8655	0.8753	0.8858
	SWIR2	0.8529	0.8667	0.8735	0.8810
	Mean	0.8196	0.8425	0.8537	0.8663

4. Conclusion

In this paper, we propose a residual convolution neural network to predict Landsat-like image, and the method can be applied to the case where there is only a pair of prior images. This method mainly includes two steps: firstly, use the known MODIS-Landsat image pair to train the residual convolutional neural network and secondly, input MODIS image at predicted phase to reconstruct Landsat-like image. Compared with the several benchmark algorithms (STARFM, FSDAF, Fit-FC, ESTARFM, SPSTFM, and *Sft*Net), our method has the advantages of learning algorithm, which takes the image patch as the reconstruction unit and considers the continuity between adjacent pixels. Training the residual to construct the depth network not only enhances the stability of the network but also improves the prediction accuracy.

The spatiotemporal fusion methods based on learning have greater prediction accuracy for heterogeneous regions. In this paper, we use a multilayer convolution neural network to extract spatial features. In the future work, we will try to design more effective methods to extract spatial features to improve the recognition ability of change information. In recent years, deep learning has received extensive attention.

Deep learning needs a lot of data to train model. Because of the characteristics of large amount of data and rich information in remote sensing data, we can use the “big data” characteristics of remote sensing data to train more effective mapping relationship between MODIS and Landsat images by deep learning training, so as to improve the prediction accuracy. In addition, although the spatiotemporal fusion models based on learning have outstanding performance, but the calculation time is longer, which is also a “common failure” based on the learning method. Therefore, our future work will follow the idea of improving the accuracy of fusion results and reducing computational complexity.

Data Availability

Data is not available for the following reasons: In this paper, we received remote sensing data from Institute of Remote Sensing Applications Chinese Academy of Sciences and conducted an experiment, but without the consent of Institute of Remote Sensing Applications Chinese Academy of Sciences, the author cannot judge whether data is available or not.

Conflicts of Interest

The authors declare that they have no conflict of interest.

References

[1] B. Rasti, D. Hong, R. Hang et al., “Feature extraction for hyperspectral imagery: the evolution from shallow to deep (Overview and Toolbox),” *IEEE Geoscience and Remote Sensing Magazine*, vol. 8, no. 3, pp. 63–92, 2020.

[2] M. C. Anderson, R. G. Allen, A. Morse, and W. P. Kustas, “Use of Landsat thermal imagery in monitoring evapotranspiration and managing water resources,” *Remote Sensing of Environment*, vol. 122, pp. 50–65, 2012.

[3] F. D. van der Meer, H. M. A. van der Werff, F. J. A. van Ruitenbeek et al., “Multi- and hyperspectral geologic remote sensing: A review,” *International Journal of Applied Earth Observation & Geo information*, vol. 14, no. 1, pp. 112–128, 2012.

[4] D. Hong, N. Yokoya, N. Ge, J. Chanussot, and X. X. Zhu, “Learnable manifold alignment (LeMA): a semi-supervised cross-modality learning framework for land cover and land use classification,” *ISPRS Journal of Photogrammetry & Remote Sensing*, vol. 147, pp. 193–205, 2019.

[5] S. Ganguly, M. A. Friedl, B. Tan, X. Zhang, and M. Verma, “Land surface phenology from MODIS: characterization of the collection 5 global land cover dynamics product,” *Remote Sensing of Environment*, vol. 114, no. 8, pp. 1805–1816, 2010.

[6] X. Zhang, M. A. Friedl, C. B. Schaaf et al., “Monitoring vegetation phenology using MODIS,” *Remote Sensing of Environment*, vol. 84, no. 3, pp. 471–475, 2003.

[7] C. Vignolles, M. Gay, G. Flouzat, and P. Puyou-Lascassies, “Spatiotemporal connection of remote sensing data concerning agricultural production modelisation at a middle scale,” in *1995 International Geoscience and Remote Sensing Symposium, IGARSS '95. Quantitative Remote Sensing for Science and Applications*, Firenze, Italy, Italy, July 1995.

[8] 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 Transactions on Geoscience and Remote Sensing*, vol. 44, no. 8, pp. 2207–2218, 2006.

[9] T. Hilker, M. A. Wulder, N. C. Coops et al., “A new data fusion model for high spatial- and temporal-resolution mapping of forest disturbance based on Landsat and MODIS,” *Remote Sensing of Environment*, vol. 113, no. 8, pp. 1613–1627, 2009.

[10] P. R. David, J. Junchang, L. Philip, S. Crystal, H. Matt, and L. Erik, “Mufti-temporal MODIS-Landsat fusion for relative and prediction of Landsat data,” *Remote Sensing of Environment*, vol. 112, no. 6, pp. 3112–3130, 2008.

[11] 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 Sensing of Environment*, vol. 114, no. 11, pp. 2610–2623, 2010.

[12] Q. Wang and P. M. Atkinson, “Spatio-temporal fusion for daily Sentinel-2 images,” *Remote Sensing of Environment*, vol. 204, pp. 31–42, 2018.

[13] C. Kwan, B. Budavari, F. Gao, and X. Zhu, “A hybrid color mapping approach to fusing MODIS and Landsat images for forward prediction,” *Remote Sensing*, vol. 10, no. 4, pp. 520–529, 2018.

[14] B. Zhukov, D. Oertel, F. Lanzl, and G. Reinhackel, “Unmixing-based multisensor multiresolution image fusion,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 3, pp. 1212–1226, 1999.

[15] M. Wu, “Use of MODIS and Landsat time series data to generate high-resolution temporal synthetic Landsat data using a spatial and temporal reflectance fusion model,” *Journal of Applied Remote Sensing*, vol. 6, no. 13, p. 063507, 2012.

[16] M. Wu, W. Huang, Z. Niu, and C. Wang, “Generating daily synthetic Landsat imagery by combining Landsat and MODIS data,” *Sensors*, vol. 15, no. 9, pp. 24002–24025, 2015.

[17] K. Hazaymeh and Q. K. Hassan, “Spatiotemporal image-fusion model for enhancing the temporal resolution of Landsat-8 surface reflectance images using MODIS images,” *Journal of Applied Remote Sensing*, vol. 9, no. 1, p. 096095, 2015.

[18] 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 Sensing of Environment*, vol. 172, pp. 165–177, 2016.

[19] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, “CoSpace: common subspace learning from hyperspectral-multispectral correspondences,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 4349–4359, 2019.

[20] B. Huang and H. Song, “Spatiotemporal reflectance fusion via sparse representation,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 10, pp. 3707–3716, 2012.

[21] H. Song and B. Huang, “Spatiotemporal satellite image fusion through one-pair image learning,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 4, pp. 1883–1896, 2013.

[22] D. Yu and L. Deng, “Deep learning and its applications to signal and information processing exploratory,” *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 145–154, 2011.

[23] F. H. C. Tivive and A. Bouzerdoum, “A new class of convolutional neural Networks (SICoNNets) and their appl-ication of face detection,” in *Proceedings of the International Joint Conference on Neural Networks, 2003*, pp. 2157–2162, Portland, OR, USA, July 2003.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Image Net classification with deep convolutional neural networks,” in

Proceedings of the 2012 Advances in Neural Information Processing Systems, pp. 1097–1105, Lake Tahoe, Nevada, USA, 2012, Curran Associates, Inc..

[25] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *EEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[26] H. Song, Q. Liu, G. Wang, R. Hang, and B. Huang, "Spatiotemporal satellite image fusion using deep convolutional neural networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 3, pp. 821–829, 2018.

[27] X. Liu, C. Deng, J. Chanussot, D. Hong, and B. Zhao, "StfNet: a two-stream convolutional neural network for spatiotemporal image fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6552–6564, 2019.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 2016.

[29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[30] M. Liang and X. Hu, "Recurrent convolutional neural network for object recognition," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, June 2015.

[31] E. F. Lambin and A. H. Strahler, "Indicators of land-cover change for change-vector analysis in multitemporal space at coarse spatial scales," *International Journal of Remote Sensing*, vol. 15, no. 10, pp. 2099–2119, 2007.

[32] C. He, P. Shi, D. Xie, and Y. Zhao, "Improving the normalized difference built-up index to map urban built-up areas using a semiautomatic segmentation approach," *Remote Sensing Letters*, vol. 1, no. 4, pp. 213–221, 2010.

[33] Z. Zhang and R. S. Blum, "A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application," *Proceedings of the IEEE*, vol. 87, no. 8, pp. 1315–1326, 1999.

[34] K. D. Kim and J. H. Heo, "Comparative study of flood quantiles estimation by nonparametric models," *Journal of Hydrology*, vol. 260, no. 1-4, pp. 176–193, 2002.

[35] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, 2002.