StfNet: A Two-Stream Convolutional Neural Network for Spatiotemporal Image Fusion

Xun Liu, *Student Member, IEEE*, Chenwei Deng, *Senior Member, IEEE*, Jocelyn Chanussot, *Fellow, IEEE*, Danfeng Hong, *Student Member, IEEE*, and Baojun Zhao

***Abstract*— Spatiotemporal image fusion is considered as a promising way to provide Earth observations with both high spatial resolution and frequent coverage, and recently, learning-based solutions have been receiving broad attention. However, these algorithms treating spatiotemporal fusion as a single image super-resolution problem, generally suffers from the significant spatial information loss in coarse images, due to the large upscaling factors in real applications. To address this issue, in this paper, we exploit temporal information in fine image sequences and solve the spatiotemporal fusion problem with a two-stream convolutional neural network called *StfNet*. The novelty of this paper is twofold. First, considering the temporal dependence among image sequences, we incorporate the fine image acquired at the neighboring date to super-resolve the coarse image at the prediction date. In this way, our network predicts a fine image not only from the structural similarity between coarse and fine image pairs but also by exploiting abundant texture information in the available neighboring fine images. Second, instead of estimating each output fine image independently, we consider the temporal relations among time-series images and formulate a temporal constraint. This temporal constraint aiming to guarantee the uniqueness of the fusion result and encourages temporal consistent predictions in learning and thus leads to more realistic final results. We evaluate the performance of the *StfNet* using two actual data sets of Landsat-Moderate Resolution Imaging Spectroradiometer (MODIS) acquisitions, and both visual and quantitative evaluations demonstrate that our algorithm achieves state-of-the-art performance.**

***Index Terms*— Convolutional neural network, spatiotemporal image fusion, super-resolution, temporal consistency, temporal dependence (TD).**

I. Introduction

HIGH spatial resolution remote sensing images with a dense time series play a significant role in studying high-frequency land surface dynamics in heterogeneous landscapes [1]–[4], such as monitoring vegetation seasonality [5], mapping real-time urban hazards [6], and detecting land cover changes [7]. However, due to the hardware limitations and budget constraints, there still exists a "spatial-temporal contradiction" problem in current remote sensing imaging systems, and so far, it is difficult for a single satellite sensor to produce Earth observations with both fine spatial and temporal resolutions [8]–[10]. For instance, the reflectance images acquired from Landsat series, ALOS, GF-1, and GF-2 satellites, are with fine spatial resolutions from 3 to 30 m [11]; however, long revisit cycles of these satellites (16 days for Landsat, 46 days for ALOS, and 5–69 days for GF-1 and GF-2) with frequent cloud contamination and complex topographic effects have severely limited their use in detecting rapid surface changes. Conversely, the Moderate Resolution Imaging Spectroradiometer (MODIS) on the Terra/Aqua, WiFS on IRS-P3, and NOAA Advanced Very High-Resolution Radiometer (AVHRR) revisit the same location on Earth per day and collect daily frequent reflectance images [12], but their low spatial resolutions (250–1000 m) are always not sufficient for quantitative monitoring of land cover changes, especially in the heterogeneous areas.

In order to tackle this problem, spatiotemporal image fusion has emerged in the past decade [13], [14]. These techniques leverage the complementary characteristics of two types of satellite images and blend them to generate high spatial resolution data with frequent coverage, thereby enhancing the capability for monitoring land surface dynamics. To date, spatiotemporal fusion has received significant attention and been widely used in many remote sensing fields such as land cover classification [15], urban flood mapping [16], and heat island monitoring [17].

A. Related Works

Generally speaking, current spatiotemporal image fusion methods can be classified into three groups: reconstruction-based, unmixing-based, and learning-based, respectively.

Among reconstruction-based fusion algorithms, the spatial and temporal adaptive reflectance fusion model (STARFM) is the one developed first [18]. Within STARFM, reflectance changes for pure pixels are supposed to be consistent between coarse and fine images, and accordingly, daily high spatial resolution surface reflectance images are reconstructed by combining neighboring pixels with a weighted-sum strategy. Considering the complex heterogeneous regions, Zhu *et al.* [19] improved STARFM and developed an enhanced STARFM (ESTARFM) involving different conversion coefficients for homogeneous and heterogeneous areas

Manuscript received January 10, 2019; revised February 20, 2019; accepted March 15, 2019. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 91438203. (*Corresponding author: Chenwei Deng.*)

X. Liu, C. Deng, and B. Zhao are with the Beijing Key Laboratory of Embedded Real-Time Information Processing Technology, School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China (e-mail: cwdeng@bit.edu.cn).

J. Chanussot is with Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France.

D. Hong is with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), 82234 Weßling, Germany, and also with Signal Processing in Earth Observation (SiPEO), Technical University of Munich (TUM), 80333 Munich, Germany.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2019.2907310

to modify the weights of neighboring pixels. Shen *et al.* [20] claimed that sensor observation differences may exist among varied land cover types and employed a prior high spatial resolution classification map for reconstruction. Wang *et al.* [21] also developed an operational spatiotemporal fusion framework and integrated the ideas of bidirectional reflectance distribution function correction, automatic coregistration, and automatic selection of input data pairs to improve fusion accuracy. However, these algorithms assuming that land cover type remains unchanged between the known and prediction dates, rarely consider landscape disturbances. Consequently, the aforementioned works may lead to some good results in phenology changing areas, but they may not be effective for the prediction of type changing areas since these land cover type changes can be hardly estimated from similar pixels in the input data.

With respect to unmixing-based models, the central idea is to predict an unknown fine image through spectral unmixing of the available coarse one. Zhukov *et al.* [22] first proposed an unmixing-based multisensor multiresolution fusion framework to integrate satellite images with different spatial resolution and acquired at different times. The fusion procedure generally has the following two steps: 1) spectral unmixing of the input coarse images and 2) fine image generation by replacing spectral information in input fine images according to the unmixing results. Wu *et al.* [23] modified multisensor multiresolution technique and developed a Spatial Temporal Data Fusion Approach (STDFA) by considering the nonlinear temporal change similarities and spatial variations in spectral unmixing. Amorós-López *et al.* [24] suggested that the solved endmembers reflectance should be similar to the imposed class spectra in a sliding window. Thus, a penalty term measuring the spectral distance between the solved and predefined endmembers and is introduced into the cost function during the unmixing process. Recently, Zhu *et al.* [25] also integrated ideas from unmixing-based methods with spatial interpolation and STARFM into one framework and proposed a flexible spatiotemporal data fusion algorithm, named FSDAF. However, these unmixing-based algorithms generally suffer from a wrong estimation of endmember numbers, endmember spectral variability in multitemporal observations, and the spectral mixing nonlinearities [26]. In addition, they still face the same difficulty as reconstruction-based models in estimating pixels in land type changing areas.

In contrast to the two aforementioned groups, learning-based spatiotemporal fusion (LBF) models do not need to specify temporal changing types but cast the prediction of high spatial resolution images as a supervised single-image super-resolution problem. Following the concept of example-based super-resolution [27], LBF aims to establish a complex mapping between the coarse and fine image pairs based on their spatial structural similarity and then predicts the unknown fine images using the corresponding coarse ones. In [28]–[30], dictionary-pair training [31] to coarse and fine image pairs was applied and their patches onto a sparse feature domain with the enforcement of a linear mapping between the coefficients were projected. Liu *et al.* [32] advocated that the coefficients should be similar among the neighboring fine images and proposed a local regularized sparse representation to alleviate the instability problem in fine image prediction. In [33], extreme learning machine (ELM), a single hidden layer feed-forward neural network, is employed to learn a mapping between raw pixels of fine and coarse images directly, and characterized by fast speed of ELM, the fine image can be predicted with much less computational complexity. Inspired by [34], state of the art in single-image super-resolution, Song *et al.* [35] adopted the deep convolutional neural networks to capture large scale spatial information in coarse images and exploited it for the prediction. Moreover, regression tree [36], random forest [37], and artificial neural networks [38] have also been explored in spatiotemporal fusion.

B. Motivation

From the above-mentioned analysis, learning-based models have been extensively studied in recent years and the related super-resolution techniques significantly advanced LBF.

However, it should be mentioned that spatiotemporal image fusion is still a different task compared with the classical natural image super-resolution, and treating spatiotemporal fusion as a single natural image super-resolution problem also faces some great challenges.

First, in spatiotemporal fusion, the magnification factor (usually ranging from 8 to 16) is much larger than that in super-resolution (usually ranging from 2 to 4). In that case, texture details have been severely blurred and distorted in coarse images and limited prior structural information could be utilized for fine image prediction. Moreover, it is known that remote sensing images, compared with natural images, contain more complex heterogeneous areas with abundant texture details, which further increase the difficulty in fine image prediction. Hence, learning a mapping function only from the spatial structure similarity is a severely ill-posed inverse problem and the relationships between fine and coarse images in the previously learned model may not be effective. Consequently, we cannot predict the fine image accurately only from the corresponding coarse one, as shown in Fig. 1.

To address this issue, in this paper, we develop a two-stream convolutional neural network tailored to spatiotemporal image fusion. Unlike the traditional learning-based methods that estimate a fine image only from spatial prior knowledge in the corresponding coarse one, temporal information in fine image sequences is exploited in our model and act as strong priors for alleviating the ill-posedness of spatiotemporal fusion problem. As a result, more plausible patterns could be generated in the results, as shown in Fig. 1(c). We refer to the proposed learning architecture as spatiotemporal fusion network or *StfNet*.

The novelty of this work is twofold.

1) Unlike the traditional learning-based methods that estimate a fine image only from the corresponding coarse one, the proposed *StfNet* incorporates the neighboring fine image and exploits temporal dependence (TD) to predict the unknown fine difference images. In this way, our mapping model does not need to recover the fine difference image only from the severely blurred coarse

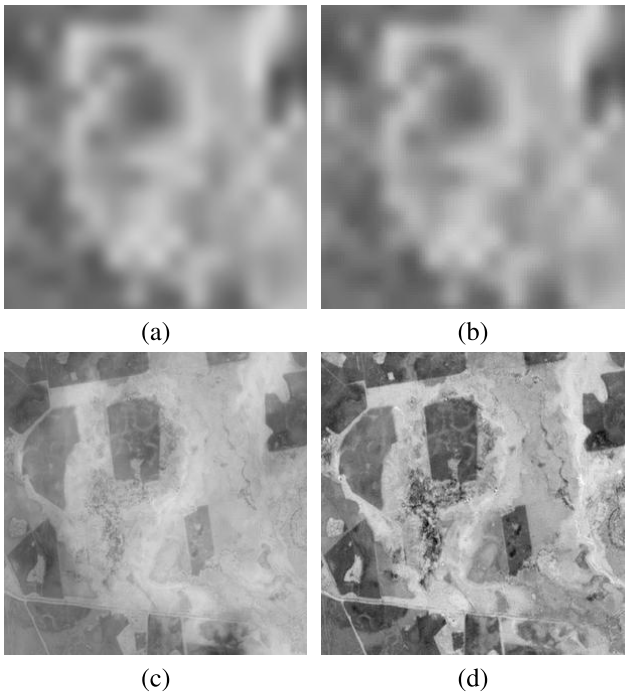


Fig. 1. Example results of the fine image prediction. (a) Coarse image. (b) Fine image predicted by a traditional learning-based model [28]. (c) Fine image predicted by our *StfNet*. (d) Actual fine image.

one. Rather, it has access to the neighboring fine image with abundant texture details and learns to transfer these components to the fine difference image, which results in more plausible image patterns.

2) Instead of estimating each output fine image independently, our network models the temporal relations among time-series images in the form of a temporal constraint during the process of network learning. This temporal constraint aiming to guarantee the uniqueness of the predicted final target fine image from forward and backward dates and encourages temporal consistent predictions and leads to a more accurate final fusion result.

C. Paper Outline

The rest of this paper is organized as follows. In Section II, we formulate the spatiotemporal fusion problem and give detailed descriptions and analysis of the proposed *StfNet*. Section III presents the implementation details of *StfNet* and compares the performance of our algorithm with other relevant spatiotemporal fusion methods on actual Landsat-MODIS images. Finally, the conclusion is drawn in Section IV.

II. Methodology

The high-level outline of the proposed spatiotemporal fusion algorithm has been presented in Fig. 2. In this paper, spatiotemporal fusion aims to predict a target fine image F_2 given in the corresponding coarse one C_2 at date t_2, as well as two coarse and fine image pairs (F_1 and C_1, F_3 and C_3) at neighboring dates t_1 and t_3. Accordingly, we obtain high spatial resolution images with a dense time series.

To alleviate the spatial information loss problem in coarse images, the proposed *StfNet* first incorporates temporal information in image time series, i.e., TD and temporal consistency, to model the mapping between fine and coarse difference image pairs. Then, with the learned mapping, we predict the unknown fine difference images F_{12} and F_{23} from the corresponding coarse ones C_{12} and C_{23}, with the neighboring fine images F_1 and F_3, respectively. Finally, the target image F_2 could be reconstructed.

A. StfNet Architecture

In spatiotemporal fusion, learning a mapping only from the spatial similarity between coarse and fine image pair is a severely ill-posed inverse problem, since the magnification factors are always large. To this end, we incorporate TD and temporal consistency to model the mapping between fine and coarse image pairs and propose a two-stream convolutional neural network architecture, i.e., *StfNet*.

The high-level idea is represented in the sequence of potential network architectures as shown in Fig. 3. We show the basic LBF model in Fig. 3(a), which attempts to learn a mapping between coarse and fine difference image pairs. In Fig. 3(b), we introduce TD into the basic learning-based model and exploit the neighboring fine images for the difference image prediction. We present the proposed *StfNet* in Fig. 3(c), in which two kinds of temporal information in image time series, i.e., TD and temporal consistency, are both incorporated in learning and leveraged for the mapping model. In this way, the ill-posed inverse problem, i.e., recovering details from a largely down-sampled coarse image, could be well alleviated, and one can expect to obtain a more accurate estimation of the missing fine images.

1) Temporal Dependence: In image time series, temporal changes (i.e., difference images) are always correlated with the original image contents in image patterns, and we refer these correlations between different images and neighboring fine images as TD. Unlike the existing learning-based methods estimating a fine image only from the corresponding coarse one, TD is incorporated in our network architecture and serves as a generic prior for the prediction.

Having coarse and fine images C_i and F_i with the difference images C_{ij} and F_{ij}, a straightforward strategy in current learning-based models is leveraging a learning architecture (e.g., sparse representation, ELM) to build a nonlinear mapping relationship $\mathcal{M}$ between the available F_{13} and C_{13}

$$\Phi = \arg\min_{\Phi} \quad \mathcal{L}(\mathcal{M}(\mathrm{C}_{13}; \Phi), L_{13}) \tag{1}$$

where Φ is the parameter of mapping $\mathcal{M}$ and $\mathcal{L}$ is the defined loss function.

To exploit the TD, in our model, the neighboring fine images F_1 and F_3 are incorporated and the mapping function could be learned by

$$\Phi_0 = \arg\min_{\Phi_0} \quad \mathcal{L}(\mathcal{M}_0(\mathrm{C}_{13}, \mathrm{F}_1; \Phi), L_{13}) \tag{2}$$

$$\Phi_1 = \arg\min_{\Phi_1} \quad \mathcal{L}(\mathcal{M}_1(\mathrm{C}_{13}, \mathrm{F}_3; \Phi), L_{13}) \tag{3}$$

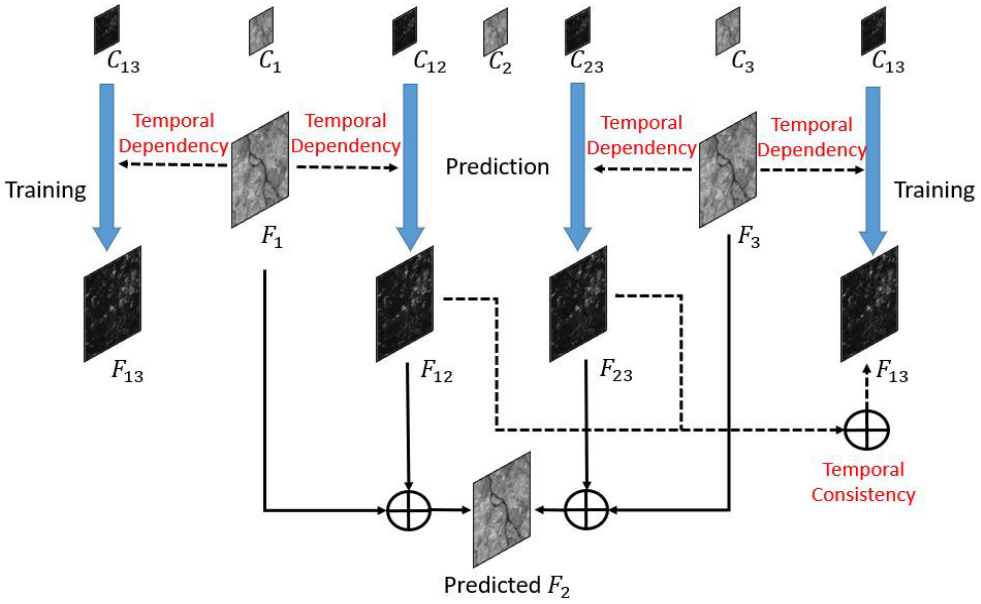


Fig. 2. High-level outline of the proposed LBF model. Here, C_i and F_i represent coarse and fines images acquired at date i, and C_{ij} and F_{ij} denote coarse and fine difference images between t_i and t_j, respectively.

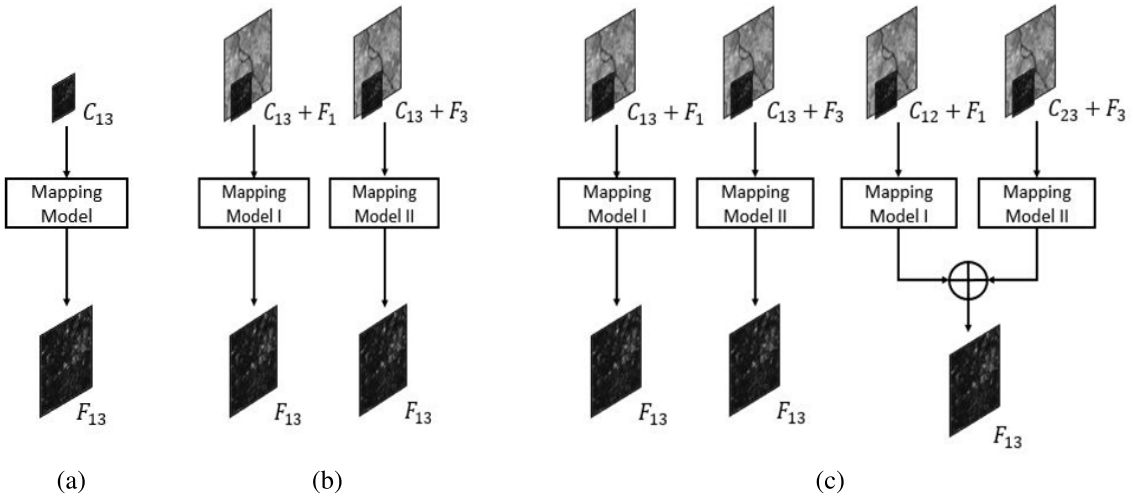


Fig. 3. Illustration of three learning model structures for spatiotemporal fusion problem. (a) Traditional LBF models. (b) LBF model with TD. (c) Our *StfNet* architecture with both TD and temporal consistency.

where $\mathcal{M}_0$ and $\mathcal{M}_1$ are the mappings to exploit fine images acquired at forward and backward dates, respectively.

Compared with the traditional learning-based methods, our algorithm not only learns a structure prior of the fine difference image from the available coarse and fine difference image pair but also models the correlation between a difference image with the neighboring fine images. In this way, our mapping model does not need to recover the fine difference image only from the severely blurred coarse one. Rather, it has access to the neighboring fine image with abundant texture details and learns to transfer these components to the fine difference image, which results in more plausible image patterns.

2) Temporal Consistency: In LBF, two fine difference images F_{12} and F_{23} are predicted from the corresponding coarse ones C_{12} and C_{23} and then reconstruct the target fine image F_2 from the neighboring fine images, respectively. Instead of estimating these output fine difference images at forward and backward imaging dates independently, we introduce temporal relations among time image series and formulate a temporal constraint in the learning.

Without loss of generality, given coarse and fine three-image sequences, the difference images F_{ij} can be calculated as

$$\begin{aligned} F_{12} &= F_2 - F_1 \\ F_{23} &= F_3 - F_2 \\ F_{13} &= F_3 - F_1. \end{aligned} \tag{4}$$

Then, a temporal constraint among the fine image sequence could be formulated as follows:

$$F_{13} = F_{12} + F_{23}. \tag{5}$$

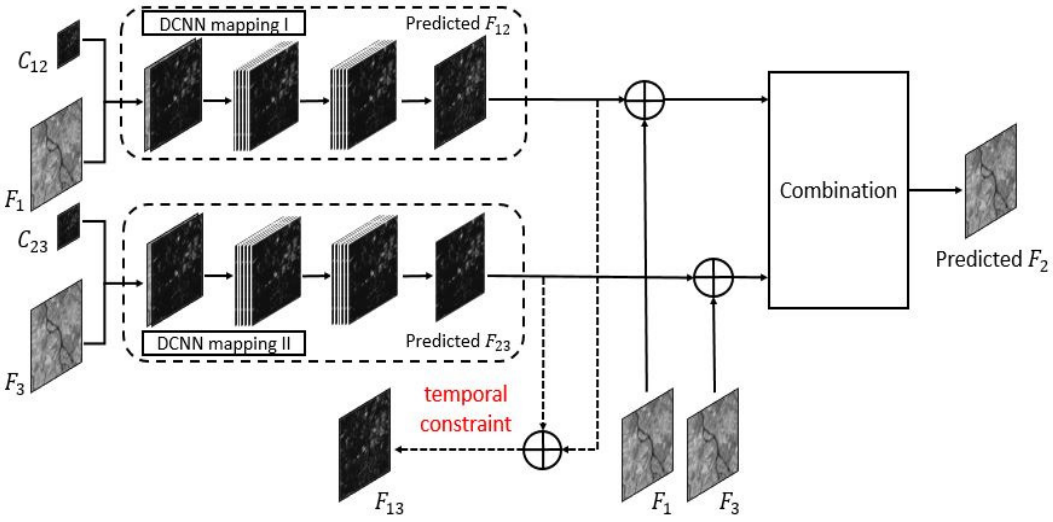


Fig. 4. Flowchart of the proposed two-stream convolutional neural network *StfNet* for prediction and target fine image reconstruction.

Temporal consistency can also be interpreted to guarantee the uniqueness of the final fusion result F_2. That is to say, we have two estimated target fine images F_2 derived by F_1 and F_3, and it is necessary to prevent these two F_2 predictions being greatly different from each other in any way. Temporal consistency applies such constraint in the learning and infers the mapping models from "unlabeled" data using their hidden temporal relations. Consequently, temporal consistent predictions are encouraged and more accurate fusion result could be achieved.

3) Network Architecture: To build the mapping from a coarse difference image with the neighboring fine image to the fine difference image, we adopt a three-layer convolutional neural network architecture [39]–[41] as our network model $\mathcal{M}_0$ and $\mathcal{M}_1$, inspired by the previous work on super-resolution [27].

Specifically, this learning architecture is composed of three convolutional layers followed by rectified linear unit (ReLU) activations in both the input and the hidden layers, and linear activation in the output layer. For each layer, a number of convolutional filters are first applied to represent the input x_l as a set of feature maps, as follows:

$$z_l = w_l * x_l + b_l \tag{6}$$

where z_l is the convolutional feature maps of input, w_l and b_l denote the filters and biases, respectively, and "*" is the convolutional operation. After filtering, in the input and hidden layer, a pointwise nonlinear function, named ReLUs, is then adopted to speed up the convergence of the network, as follows:

$$y_l = \max(0, z_l) \tag{7}$$

where y_l is the output feature maps. These feature maps comprising of nonlinear combination of multiple input layers and enable the coupling between the coarse difference image and the neighboring fine image. In this way, our mapping model captures complementary information in different input layers and transfers information from the neighboring fine image to the coarse difference image for prediction.

B. Network Training

Training our proposed *StfNet* architecture requires the parameter estimation for two convolutional neural networks M_0 and M_1. Following traditional LBF, we assume that the relationships between coarse and fine image pairs are invariant in the same period and select the available coarse and fine difference image pairs (C_{13} and F_{13}) from t_1 to t_3 as the training data set. To benefit from the TD, the neighboring fine images (F_1 and F_3) at t_1 and t_3 are also used as inputs. Under the temporal constraint, we can obtain the objective function of the proposed network architecture as follows:

$$\{\Phi_0, \Phi_1\} = \arg\min_{\Phi_0,\Phi_1} \{\mathcal{L}_R + \lambda \mathcal{L}_T\}. \tag{8}$$

Here, Φ_0 and Φ_1 denote the network parameters of two convolutional neural network mapping models M_0 and M_1, respectively; $\mathcal{L}_R$ is the reconstruction loss with TD and calculated by

$$\mathcal{L}_R = \mathcal{L}(\mathcal{M}_0(C_{13}, F_1; \Phi_0), F_{13}) + \mathcal{L}(\mathcal{M}_1(C_{13}, F_3; \Phi_1), F_{13}) \tag{9}$$

where $\mathcal{L}$ is the mean square error (MSE)-based loss function. The second penalty term $\mathcal{L}_T$ is the temporal loss from temporal consistency and defined as

$$\mathcal{L}_T = \mathcal{L}(\widetilde{F}_{13}, F_{13}) \tag{10}$$

where

$$\widetilde{F}_{13} = \mathcal{M}_0(C_{12}, F_1; \Phi_0) + \mathcal{M}_1(C_{23}, F_3; \Phi_1) \tag{11}$$

and λ is the weighting parameter.

For optimization, we adopt the stochastic gradient descent (SGD) with the standard back propagation to minimize the reconstruction and temporal loss jointly. In particular, the weights in our *StfNet* are updated as

$$\Delta_{i+1} = m \cdot \Delta_i + \eta \cdot \frac{\partial(L_R + \lambda L_T)}{\partial W_i^{n_l}} \tag{12}$$

$$W_{i+1}^{n_l} = W_i^{n_l} + \Delta_{i+1} \tag{13}$$

where $n \in 0, 1$ and $l \in 1, 2, 3$ are the indices of two mapping models M_0 and M_1, and the layers, i is the iteration number, m is the momentum, η is the learning rate, and $(\partial(L_R + \lambda L_T)/\partial W_i^{nl})$ is the derivative.

C. Prediction and Target Image Reconstruction

Having the trained two-stream convolutional neural network, we can predict two fine difference images F_{12} and F_{23} and, then, reconstruct the target unknown fine image F_2. The flowchart of the prediction and target image reconstruction is presented in Fig. 4.

Considering TD, we predict two fine difference images F_{12} and F_{23} from the corresponding coarse ones C_{12} and C_{23}, as well as the neighboring fine images F_1 and F_3

$$F_{12} = \mathcal{M}_0(C_{12}, F_1; \Phi_0) \tag{14}$$

$$F_{23} = \mathcal{M}_1(C_{23}, F_3; \Phi_1). \tag{15}$$

Then, we reconstruct the target fine image F_2 by an adaptive local weighting strategy

$$F_2 = \alpha * (F_1 + F_{12}) + (1 - \alpha) * (F_3 - F_{23}) \tag{16}$$

where α and $1 - \alpha$ are the weighting parameters for the predicted image F_2 from F_1 and F_3, respectively.

To decide the weighting parameter in reconstruction, we believe that a more similar coarse image leads to a more reliable fine image prediction. That is, if there are less changes between two coarse images C_2 and C_k ($k = 1$ or 3), it is possible that the target fine image F_2 is more similar to the neighboring fine image F_k, and the result reconstructed from F_k should be more accurate. Therefore, we employ the absolute differences between coarse images to measure temporal change degrees and calculate the reconstruction weighting parameters as

$$\alpha = \begin{cases} 1 & \text{if } \upsilon_{c_{23}} - \upsilon_{c_{12}} > \delta \\ 0 & \text{if } \upsilon_{c_{12}} - \upsilon_{c_{23}} > \delta \\ \dfrac{1/\upsilon_{c_{12}}}{1/\upsilon_{c_{12}} + 1/\upsilon_{c_{23}}} & \text{else} \end{cases} \tag{17}$$

where $\upsilon_{c_{12}}$ and $\upsilon_{c_{12}}$ represent the absolute average changes of C_{12} and C_{23}, respectively; δ is a changing threshold and empirically set to 0.2 in our experiments.

III. Experimental Results and Analysis

In this section, we first give the data sets used in the experiments and describe the implementation details of our proposed *StfNet*. Then, the experimental results are presented. Furthermore, we report more discussions about our network, including the validity of TD and temporal consistency, the network convergence, and the computational efficiency.

A. Data Sets

To evaluate the performance of different fusion algorithms, we prepared Landsat ETM+ (30 m) and MODIS (250–500 m) surface reflectance images as fine and coarse images, respectively, and performed the experiments on two actual data sets.

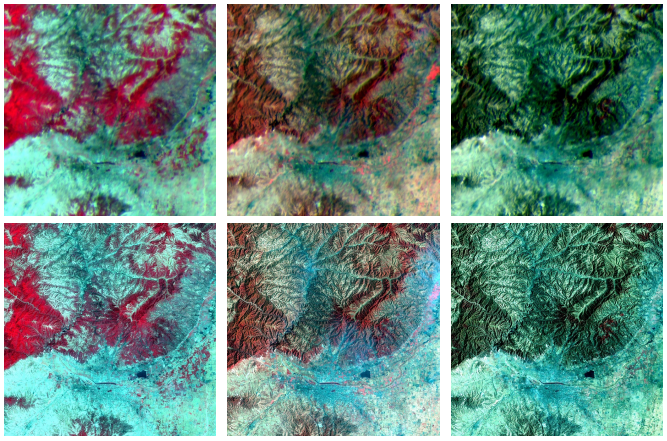


Fig. 5. Composite surface reflectance of (Top row) MODIS and (Bottom row) Landsat data acquired at different dates for *Taiyuan* data set.

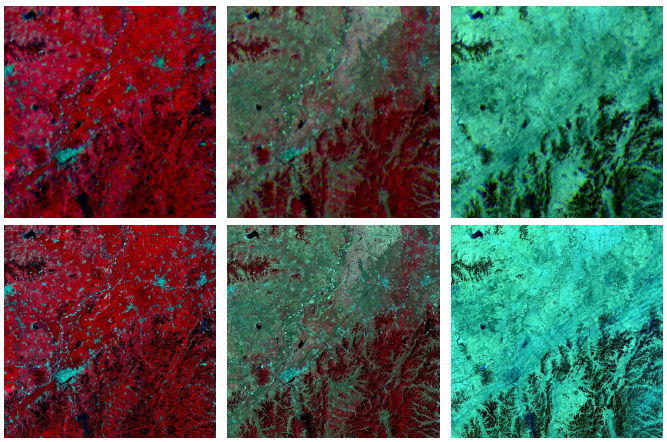


Fig. 6. Composite surface reflectance of (Top row) MODIS and (Bottom row) Landsat data acquired at different dates for *Shenyang* data set.

The first data set is from the city of Taiyuan (37°84′ N, 112°51′ W) located in the middle of China and captured on June 5, 2002, October 11, 2002, and November 12, 2002, as shown in Fig. 5. The second data set is from the city of Shenyang (41°76′ N, 123°30′ W) located in the northeast of China and captured on August 1, 2001, September 28, 2001, and November 15, 2001, as shown in Fig. 6. These tested data are 6 bands surface reflectance images with a size of 2000 × 2000 pixels and cover complex study areas with a size of 60 km × 60 km. For each data set, the Landsat reflectance images at the second date are the target images to be reconstructed and serve as the reference data for evaluation.

Herein, the Landsat surface reflectance images are available in the United States Geological Survey (http://earthexplorer.usgs.gov/) and have been radiometrically and geometrically corrected using Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS). The MODIS surface reflectance images have been downloaded from the Land Processes Distributed Active Archive Center (http://lpdaac.usgs.gov/lpdaac/), and then reprojected and resampled with the MODIS Reprojection Tool.

B. Implementation Details

In our proposed *StfNet*, we employ the three-layer convolutional neural network as the basic mapping model, which is a comparatively shallow network and quite suitable for the spatiotemporal fusion task in which the training data are limited. We also prefer a lightweight structure network and the filter numbers are set as 32 and 16 in the first and second layer, respectively. Moreover, we set the convolutional filter size as $f_1 = 9$, $f_2 = 5$, and $f_3 = 5$ for the three layers, to ensure a large receptive field and capture sufficient spatial information in the coarse images.

For the training phase, given the coarse difference image C_{13} with neighboring fine images F_1 and F_3 as inputs, and the corresponding fine difference image F_{13} as output, we crop these images to patches of size $50 \times 50 \times 6$ pixels and thus have 1600 nonoverlapped patch samples for training the network. We do not employ any augmentation strategy to augment the training samples, since the *StfNet* is a relatively shallow neural network and has a small number of parameters. These samples are expected to capture sufficient variability of the estimated images at the same scene, and the *StfNet* could learn an effective mapping from them.

For optimization, our network weights are initialized to small random values from a Gaussian distribution with zero mean and standard deviation of 0.001. The mini-batch size is set as 64 to fit into the GPU memory. The weight decay is set as 10^{-6} and the momentum is set as 0.9. We initialize the learning rate as 5×10^{-4} with a division by 10 every 10^5 iterations and iterate the model for 3×10^5 times to ensure convergence. The learning model is implemented with *Caffe* package and runs on an NVIDIA Titan Xp GPU with 12 GB of RAM.

Regarding the prediction phase, since our network is fully convolutional, it can process multispectral images of arbitrary size theoretically. However, limited by the memory of GPU, in practice, we tailor input images into tiles of size 250×250 pixels for prediction and adjacent tiles have an overlap to avoid boundary artifacts.

C. Comparison and Evaluation

Several algorithms have been employed for comparison in this paper, including the traditional reconstruction-based model, STARFM [18]; a flexible unmixing-based spatiotemporal fusion approach, FSDAF [25]; and the competitive-learning-based algorithms, error-bound-regularized semi-coupled dictionary learning-based fusion model (EBSCDL) [30], ELM-based fusion model (ELM-FM) [33] and the recently developed spatiotemporal fusion model based on convolutional neural networks (STFCNNs) [35]. To ensure a fair comparison, all these algorithms adopt the default parameters given by the authors in our experiments.

Regarding the evaluation, the availability of original Landsat image F allows us to evaluate the fusion result $\tilde{F}$ with a full referenced manner, and three widely used metrics, root-mean-square error (RMSE), correlation coefficient (CC), and structural similarity (SSIM), are adopted for quantitative assessment.

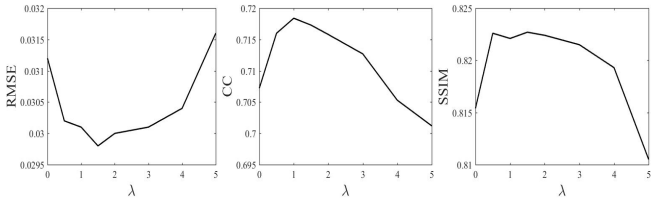


Fig. 7. *StfNet* performance of the different weighting parameter λ in terms of three objective metrics.

D. Parameter Analysis

In this section, we analyze the influence of the weighting parameter λ. We use a separate data set from northeastern New South Wales, Australia, and three Landsat and MODIS reflectance image pairs acquired on May 2, 2004, December 12, 2004, and March 2, 2005 are involved. The influence of parameter λ has been depicted in Fig. 7 and the fusion performance is evaluated by the average index values for six bands.

From Fig. 7, we can see that as λ increases, the metrics CC and SSIM get larger while RMSE becomes smaller, which means that the temporal loss encourages more temporal consistent predictions and leads to more accurate fusion results. However, when λ is too large, the important reconstruction loss may be ignored and this will result in bad fusion performances for all of the evaluation metrics. Therefore, λ should have an appropriate value and is set at 1 in our experiments.

E. Experimental Results

1) Objective Evaluation: Table I shows the quantitative outcomes of different methods on the *Taiyuan* data set. We can see that STARFM and FSDAF generate the worst results in all metrics and their performances are quite unstable for different bands. For instance, FSDAF gets a comparable performance as other methods on band 1 while it generates the results that are quite far from the ground truth in band 5. Regarding learning-based algorithms, EBSCDL, ELM-FM, and STFCNN show much better performances in this data set since they do not need to specify temporal reflectance changing types but predict them in a general super-resolution framework. As reported in Table I, our proposed algorithm consistently outperforms the other approaches in terms of RMSE, CC, and SSIM. These results indicate that our model achieves the closest result to the ground truth (the smallest RMSE and CC) and presents most structural details in actual Landsat images (the largest SSIM).

The objective performances of *Shenyang* data set for different fusion methods are presented in Table II. We observe that STARFM again provides the worst performances in all metrics since it can only handle the phenology changes in the homogenous areas. However, it can be seen that the tested data set consists of large heterogeneous regions with abundant texture details and thus STARFM fails. The unmixing-based

TABLE I
QUANTITATIVE ASSESSMENT OF DIFFERENT SPATIOTEMPORAL FUSION METHODS FOR TAIYUAN DATA SET

Index	Band	STARFM [18]	FSDAF [25]	EBSCDL [30]	ELM-FM [33]	STFCNN [35]	**Proposed**
RMSE	Band 1	0.0166	0.0151	0.0155	0.0152	0.0159	**0.0149**
	Band 2	0.0141	0.0147	0.0136	0.0129	0.0134	**0.0122**
	Band 3	0.0148	0.0169	0.0142	0.0137	0.0149	**0.0121**
	Band 4	0.0260	0.0293	0.0146	0.0144	0.0153	**0.0142**
	Band 5	0.0272	0.0316	0.0233	0.0201	0.0193	**0.0179**
	Band 6	0.0253	0.0298	0.0217	0.0212	0.0203	**0.0191**
	Average	0.0207	0.0229	0.0171	0.0162	0.0165	**0.0151**
CC	Band 1	0.8073	0.8299	0.8155	0.8231	0.8002	**0.8298**
	Band 2	0.8719	0.8428	0.8860	0.8967	0.9004	**0.9070**
	Band 3	0.8924	0.8518	0.9153	0.9203	0.9070	**0.9312**
	Band 4	0.6964	0.6507	0.9199	0.9223	0.9216	**0.9258**
	Band 5	0.8597	0.7966	0.9296	0.9379	0.9372	**0.9432**
	Band 7	0.8562	0.8038	0.9149	0.9202	0.9201	**0.9283**
	Average	0.8307	0.7893	0.8969	0.9034	0.8978	**0.9109**
SSIM	Band 1	0.8996	0.8665	0.8972	0.9064	0.8852	**0.9240**
	Band 2	0.9023	0.8252	0.8840	0.8964	0.8952	**0.9163**
	Band 3	0.8820	0.8021	0.8741	0.8813	0.8552	**0.9027**
	Band 4	0.6833	0.6206	0.8887	0.8977	0.8747	**0.9042**
	Band 5	0.8369	0.7089	0.8676	0.8848	0.8935	**0.9004**
	Band 7	0.8177	0.7390	0.8636	0.8678	0.8756	**0.8859**
	Average	0.8370	0.7604	0.8792	0.8891	0.8802	**0.9056**

TABLE II
QUANTITATIVE ASSESSMENT OF DIFFERENT SPATIOTEMPORAL FUSION METHODS FOR SHENYANG DATA SET

Index	Band	STARFM [18]	FSDAF [25]	EBSCDL [30]	ELM-FM [33]	STFCNN [35]	**Proposed**
RMSE	Band 1	0.0122	0.0112	0.0113	0.0111	0.0115	**0.0110**
	Band 2	0.0105	0.0109	0.0091	0.0090	0.0087	**0.0085**
	Band 3	0.0159	0.0160	0.0133	0.0133	0.0145	**0.0125**
	Band 4	0.0270	0.0309	0.0232	0.0229	0.0223	**0.0213**
	Band 5	0.0311	0.0267	0.0231	**0.0229**	0.0260	0.0232
	Band 7	0.0307	0.0269	0.0273	0.0264	0.0233	**0.0217**
	Average	0.0208	0.0199	0.0179	0.0176	0.0177	**0.0165**
CC	Band 1	0.8055	0.8461	0.8533	0.8579	0.8537	**0.8636**
	Band 2	0.8306	0.8321	0.8939	0.8901	0.8986	**0.9016**
	Band 3	0.8289	0.8328	0.8965	0.8981	0.8821	**0.9032**
	Band 4	0.7244	0.6912	0.8123	0.8206	0.8309	**0.8370**
	Band 5	0.8037	0.8593	**0.9047**	0.8991	0.8927	0.8963
	Band 7	0.7434	0.8195	0.8239	0.8309	0.8412	**0.8582**
	Average	0.7894	0.8135	0.8641	0.8661	0.8665	**0.8773**
SSIM	Band 1	0.9047	0.9162	0.9077	0.9137	0.8834	**0.9223**
	Band 2	0.8727	0.8645	0.9094	0.9075	0.9125	**0.9281**
	Band 3	0.8002	0.8214	0.8645	0.8638	0.8640	**0.8832**
	Band 4	0.6565	0.6736	0.7518	0.7579	0.8028	**0.8070**
	Band 5	0.7079	0.7782	0.8246	0.8222	0.8204	**0.8254**
	Band 7	0.6365	0.7665	0.8116	0.8099	0.8107	**0.8133**
	Average	0.7631	0.8034	0.8449	0.8469	0.8490	**0.8632**

approach, FSDAF, also fails since it generally suffers from the wrong estimation of endmember numbers, endmember spectral variability in multitemporal observations, and the spectral mixing nonlinearities in the spectral unmixing process. Two learning-based models, SPSTFM and ELM-FM, perform better on the data set, and the recently developed STFCNN model also generates the competitive results. However, our *StfNet* benefitting from the powerful temporal information in fine image sequences, gets always the best performances for all the metrics in all bands except in the case of band 5 on RMSE and CC indices where ELM-FM and EBSCDL only show slightly better performance.

2) Subjective Evaluation: Apart from the objective evaluation, the subjective results are also demonstrated, and for a better visual inspection, a close-up view is presented in the right-bottom of each subpicture.

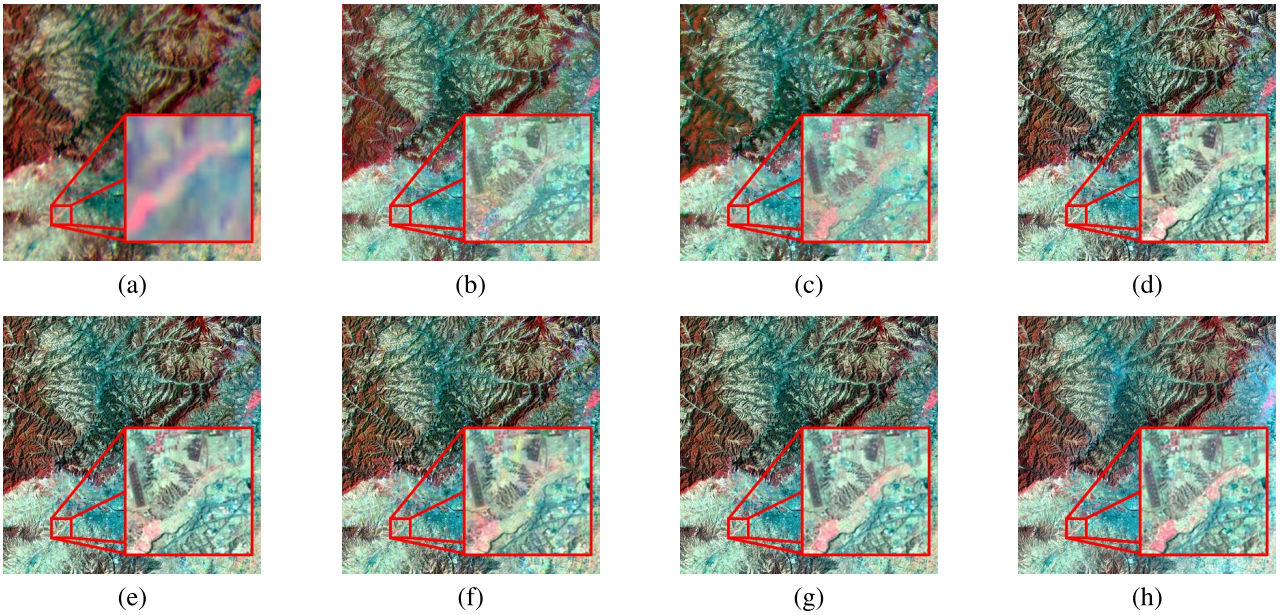


Fig. 8. Prediction results for the target Landsat image in *Taiyuan* data set. (a) Observed MODIS image. (b) Predicted by STARFM [18]. (c) Predicted by FSDAF [25]. (d) Predicted by EBSCDL [30]. (e) Predicted by ELM-FM [33]. (f) Predicted by STFCNN [35]. (g) Predicted by the proposed model. (h) Observed Landsat image.

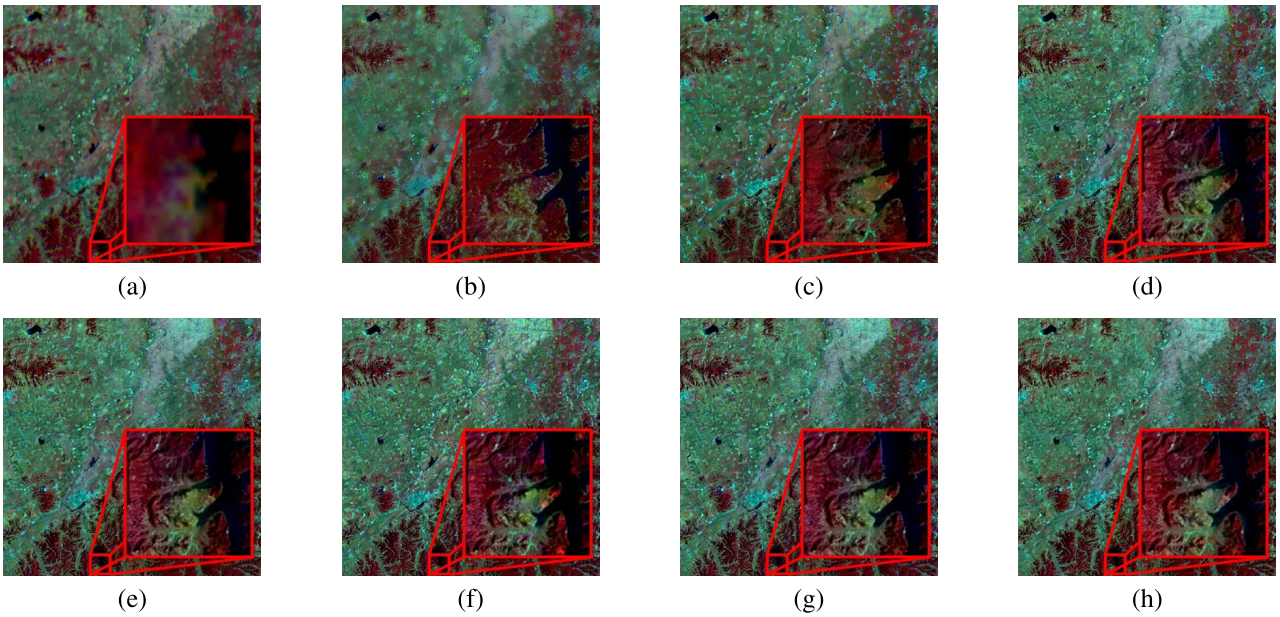


Fig. 9. Prediction results for the target Landsat image in *Shenyang* data set. (a) Observed MODIS image. (b) Predicted by STARFM [18]. (c) Predicted by FSDAF [25]. (d) Predicted by EBSCDL [30]. (e) Predicted by ELM-FM [33]. (f) Predicted by STFCNN [35]. (g) Predicted by the proposed model. (h) Observed Landsat image.

From the results of the *Taiyuan* data set (Fig. 8), it can be seen that the result from STARFM has been distorted, especially in areas with abundant texture details. Moreover, in the middle of the close-up view, the temporal dynamic area that should be red but has been totally missed due to the fact that STARFM assumes the land cover type will not change in the image sequences. As for the result from FSDAF, some of the details have been largely blurred and thus cannot be visible. Learning-based models lead to more accurate results in the temporal dynamic area and ELM-FM and STFCNN shows relatively better performances compared to EBSCDL since more spatial information is incorporated in the prediction. Our proposed model achieves the most convincing results and the fused image is the closest to the actual Landsat image.

The subjective results of *Shenyang* data set have been presented in Fig. 9. We can see that all these algorithms are generally able to estimate the fine target image. However, it is still observed that STARFM has introduced serious artifacts and the landscape cannot be recognized in the close-up view. For FSDAF, the fused image has been largely blurred and some

TABLE III
OBJECTIVE PERFORMANCE OF DIFFERENT FUSION STRATEGIES IN LBF

Data Set	Index	LBF	LF with TD	Our *StfNet*
Data Set 1	RMSE	0.0159	0.0155	**0.0151**
	CC	0.9041	0.9086	**0.9109**
	SSIM	0.8936	0.9002	**0.9056**
Data Set 2	RMSE	0.0173	0.0170	**0.0165**
	CC	0.8680	0.8721	**0.8773**
	SSIM	0.8549	0.8598	**0.8632**

details have been lost especially for the heterogeneous regions. The results from two learning-based models, EBSCDL and ELM-FM, achieve better performance but there still exist spectral distortions in the middle regions. Regarding STFCNN, since the high-pass components injection step in the fusion, the details seem to be clearer, but some visible artifacts are also introduced in the result. Fig. 9(h) shows the results from *StfNet* and we can see that our proposed algorithm generates the most visually similar result without any obvious artifacts in the image.

F. Validity of Temporal Dependence and Temporal Consistency

To further illustrate the advantages of TD and temporal consistency in our proposed *StfNet*, we conduct additional comparisons among three learning architectures (see Fig. 3), including the basic LBF without TD and temporal consistency, LBF only with TD, and LBF both with TD and temporal consistency (i.e., our *StfNet*). For a fair comparison, all the experimental environments and settings remain the same.

Table III lists the quantitative performances of the three aforementioned models, and the average index values for six bands are presented. We observe that the adoption of TD reduces RMSE and increases CC and SSIM values, compared to LBF on both data sets. This is due to the fact that LBF with TD has access to the neighboring fine image and exploits the correlated spatial information for the fine image prediction. Our proposed *StfNet* with both TD and temporal consistency, encouraging more temporal consistent predictions, further outperforms LBF with TD and generates more accurate fusion results.

We also take a patch contains 200×200 pixels from *Taiyuan* data set band 3 as an example and visualize the fine difference image prediction results of three learning architectures. From Fig. 10(a), we can see that LBF fails to recover the texture details in the fine difference image and the result is still severely blurred. This was expected since the magnification factor between coarse and fine image pairs is always too large and limited structural information could be exploited in the prediction. In contrast, LBF with TD significantly improves the fusion results with more pleasing texture details, as shown in Fig. 10(b). In Fig. 10(c), our *StfNet* powered by TD and temporal consistency produces even finer details under the temporal constraint and achieves the most visually similar result compared with the reference image. Based on

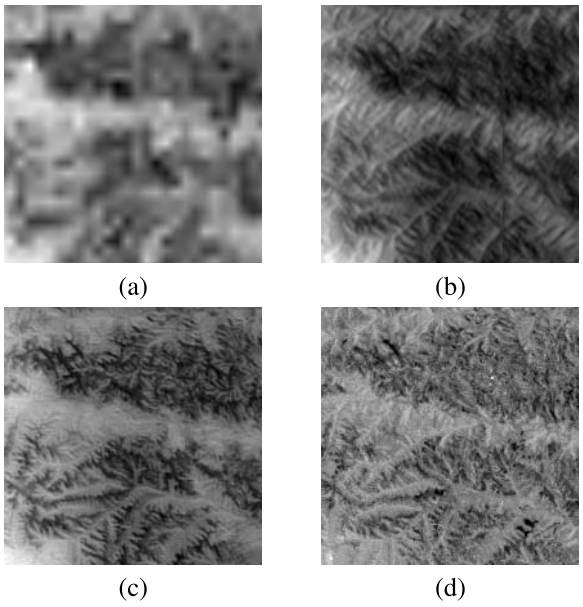


Fig. 10. Fine difference image prediction results of different fusion strategies. (a) LBF without TD and temporal consistency. (b) LBF with only TD. (c) Our *StfNet* with both TD and temporal consistency. (d) Reference fine difference image.

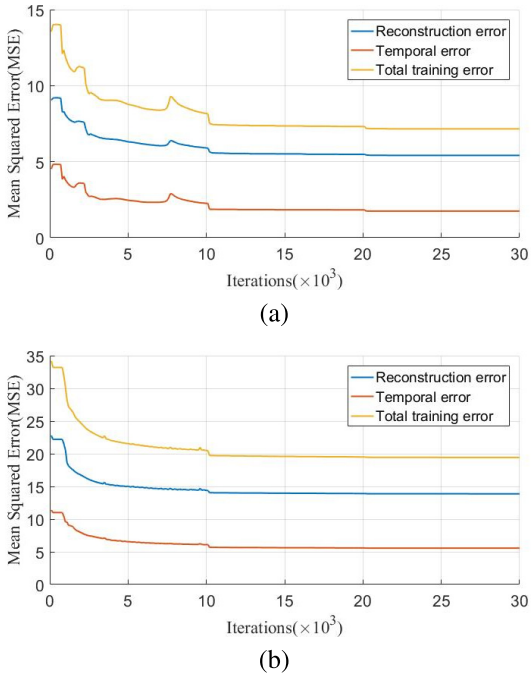


Fig. 11. Learning curves of our *StfNet* on two data sets. (a) *Taiyuan* data set. (b) *Shenyang* data set.

the comparisons mentioned earlier, it can be concluded that TD and temporal consistency in our *StfNet* both contribute to the fine image prediction and lead to more realistic results in spatiotemporal fusion.

G. Convergence Analysis

In this section, we also analyze the convergence of the proposed technique and Fig. 11 presents the reconstruction

TABLE IV
COMPUTATION COSTS OF DIFFERENT SPATIOTEMPORAL FUSION METHODS (IN SECONDS)

	STARFM [18]	FSDAF [25]	EBSCDL [30]	ELM-FM [33]	STFCNN (GPU) [35]	*StfNet* (GPU)
Training Time (s)	-	-	25948	37	9291	6907
Prediction Time (s)	329	264	2735	8	14	13
Total Time (s)	329	264	28683	45	9305	6920

error, temporal error, and total training error on both data sets in the training stage. We can see that, during the optimization, our network converges smoothly on both data sets and the error varies significantly at the beginning of the training process. In Fig. 11(a), the network has some minor fluctuations but then the error reduces steadily and stabilizes. These learning curves show that our network can learn an effective mapping between coarse and fine difference image pairs and predict the fine difference images.

H. Computational Efficiency

In this section, the computational efficiency of different spatiotemporal fusion methods is reported. We conduct a comparison on the red band image of *Taiyuan* data set of size 2000×2000 pixels, and to ensure a fair comparison, all the experiments are performed on a Linux workstation equipped with an Intel Xeon E5-2670 processor with 2.30 GHz and 128-GB RAM and a NVIDIA Geoforce GTX Titan Xp GPU with 12 GB of RAM. We download the source codes of STARFM and FSDAF online. For learning-based algorithms, EBSCDL and ELM-FM are coded on MATLAB, and *StfNet* and STFCNN are implemented with the support of *Caffe* and MATLAB. It should be noted that the computation of our *StfNet* and STFCNN is performed by *Caffe* with GPU acceleration, whereas the other algorithms are not available with GPU acceleration and, therefore, measured without GPU support.

The computational time of different spatiotemporal fusion algorithms is recorded in Table IV. We can see that the reconstruction and unmixing-based spatiotemporal fusion algorithms (STARFM and FSDAF) show fast running speed since no training process is needed in fusion. For learning-based methods, ELM-FM is also efficient since the random generation of input weights, but EBSCDL and STFCNN are quite computationally demanding and take time. Despite GPU support, our *StfNet* also has a slower running speed than STARFM, FSDAF, and ELM-FM but shows faster running speed compared with STFCNN since we only need to train a relatively slighter neural network.

IV. CONCLUSION

In this paper, we proposed a two-stream convolutional neural network (*StfNet*) by incorporating temporal information in fine image sequences (i.e., TD and temporal consistency) for spatiotemporal image fusion. Our network takes a coarse difference image with the neighboring fine image as inputs and the corresponding fine difference image as output. In this way, the fine images are predicted not only from the structural similarity between coarse and fine image pairs but also by exploiting texture information in temporally neighboring images. Moreover, considering the temporal relations among time-series images, a temporal constraint was formulated in our model and encouraged more temporal consistent results. The experiments have been conducted on actual Landsat and MODIS images and both objective and subjective verifications demonstrated that our model achieved the best performances compared with several state-of-the-art algorithms.

ACKNOWLEDGMENT

The authors would like to thank the editors and anonymous reviewers for their insightful comments and suggestions that lead to this improved version and clearer presentation of the technical content.

REFERENCES

[1] C. Lin, Y. Li, Z. Yuan, A. K. H. Lau, C. Li, and J. C. Fung, "Using satellite remote sensing data to estimate the high-resolution distribution of ground-level $pm_{2.5}$," *Remote Sens. Environ.*, vol. 156, pp. 117–128, Jan. 2015.

[2] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[3] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: A survey," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 84–100, Jan. 2018.

[4] L. Zhang, Q. Zhang, B. Du, X. Huang, Y. Y. Tang, and D. Tao, "Simultaneous spectral-spatial feature selection and extraction for hyperspectral images," *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 16–28, Jan. 2018.

[5] Q. Yu, P. Gong, N. Clinton, G. Biging, M. Kelly, and D. Schirokauer, "Object-based detailed vegetation classification with airborne high spatial resolution remote sensing imagery," *Photogramm. Eng. Remote Sens.*, vol. 72, no. 7, pp. 799–811, 2006.

[6] M. A. White and R. R. Nemani, "Real-time monitoring and short-term forecasting of land surface phenology," *Remote Sens. Environ.*, vol. 104, no. 1, pp. 43–49, 2006.

[7] M. C. Hansen and T. R. Loveland, "A review of large area monitoring of land cover change using Landsat data," *Remote Sens. Environ.*, vol. 122, pp. 66–74, Jul. 2012.

[8] C. M. Gevaert and F. J. García-Haro, "A comparison of STARFM and an unmixing-based algorithm for Landsat and MODIS data fusion," *Remote Sens. Environ.*, vol. 156, pp. 34–44, Jan. 2015.

[9] Y. Qu, H. Qi, and C. Kwan, "Unsupervised sparse Dirichlet-net for hyperspectral image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 2511–2520.

[10] 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 J. Photogramm. Remote Sens.*, vol. 147, pp. 193–205, Jan. 2019.

[11] J. G. Masek *et al.*, "A Landsat surface reflectance dataset for North America, 1990-2000," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 1, pp. 68–72, Jan. 2006.

[12] C. O. Justice *et al.*, "The Moderate Resolution Imaging Spectroradiometer (MODIS): Land remote sensing for global change research," *IEEE Trans. Geosci. Remote Sens.*, vol. 36, no. 4, pp. 1228–1249, Jul. 1998.

[13] B. Chen, B. Huang, and B. Xu, "Comparison of spatiotemporal fusion models: A review," *Remote Sens.*, vol. 7, no. 2, pp. 1798–1835, 2015.

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

[15] C. Senf, P. J. Leitão, D. Pflugmacher, S. van der Linden, and P. Hostert, "Mapping land cover in complex mediterranean landscapes using Landsat: Improved classification accuracies from integrating multi-seasonal and synthetic imagery," *Remote Sens. Environ.*, vol. 156, pp. 527–536, Jan. 2015.

[16] F. Zhang, X. Zhu, and D. Liu, "Blending MODIS and Landsat images for urban flood mapping," *Int. J. Remote Sens.*, vol. 35, no. 9, pp. 3237–3253, 2014.

[17] B. Huang, J. Wang, H. Song, D. Fu, and K. Wong, "Generating high spatiotemporal resolution land surface temperature for urban heat island monitoring," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 5, pp. 1011–1015, Sep. 2013.

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

[19] X. Zhu, J. Chen, F. Gao, X. Chen, and J. G. Masek, "An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions," *Remote Sens. Environ.*, vol. 114, no. 11, pp. 2610–2623, 2010.

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

[21] P. Wang, F. Gao, and J. G. Masek, "Operational data fusion framework for building frequent landsat-like imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 7353–7365, Nov. 2014.

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

[23] M. Wu, Z. Niu, C. Wang, C. Wu, and L. Wang, "Use of MODIS and Landsat time series data to generate high-resolution temporal synthetic Landsat data using a spatial and temporal reflectance fusion model," *J. Appl. Remote Sens.*, vol. 6, no. 1, 2012, Art. no. 063507.

[24] J. Amorós-López *et al.*, "Multitemporal fusion of Landsat/TM and ENVISAT/MERIS for crop monitoring," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 23, pp. 132–141, Aug. 2013.

[25] X. Zhu, E. H. Helmer, F. Gao, D. Liu, J. Chen, and M. A. Lefsky, "A flexible spatiotemporal method for fusing satellite images with different resolutions," *Remote Sens. Environ.*, vol. 172, pp. 165–177, Jan. 2016.

[26] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

[27] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[28] B. Huang and H. Song, "Spatiotemporal reflectance fusion via sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 10, pp. 3707–3716, Oct. 2012.

[29] H. Song and B. Huang, "Spatiotemporal satellite image fusion through one-pair image learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 4, pp. 1883–1896, Apr. 2013.

[30] B. Wu, B. Huang, and L. Zhang, "An error-bound-regularized sparse coding for spatiotemporal reflectance fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6791–6803, Dec. 2015.

[31] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 184–199.

[32] X. Liu, C. Deng, and B. Zhao, "Spatiotemporal reflectance fusion based on location regularized sparse representation," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 2562–2565.

[33] X. Liu, C. Deng, S. Wang, G.-B. Huang, B. Zhao, and P. Lauren, "Fast and accurate spatiotemporal fusion based upon extreme learning machine," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 2039–2043, Dec. 2016.

[34] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang, "Coupled dictionary training for image super-resolution," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3467–3478, Aug. 2012.

[35] H. Song, Q. Liu, G. Wang, L. Hang, and B. Huang, "Spatiotemporal satellite image fusion using deep convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 3, pp. 821–829, Mar. 2018.

[36] S. P. Boyte, B. K. Wylie, M. B. Rigge, and D. Dahal, "Fusing MODIS with Landsat 8 data to downscale weekly normalized difference vegetation index estimates for central Great Basin rangelands, USA," *GISci. Remote Sens.*, vol. 55, no. 3, pp. 376–399, 2018.

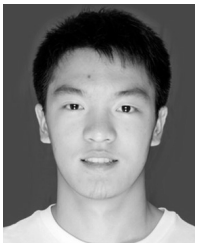
[37] Y. Ke, J. Im, S. Park, and H. Gong, "Downscaling of MODIS one kilometer evapotranspiration using Landsat-8 data and machine learning approaches," *Remote Sens.*, vol. 8, no. 3, p. 215, 2016.

[38] V. Moosavi, A. Talebi, M. H. Mokhtari, S. R. F. Shamsi, and Y. Niazi, "A wavelet-artificial intelligence fusion approach (WAIFA) for blending Landsat and MODIS surface temperature," *Remote Sens. Environ.*, vol. 169, pp. 243–254, Nov. 2015.

[39] J. Han, R. Quan, D. Zhang, and F. Nie, "Robust object co-segmentation using background prior," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1639–1651, Apr. 2018.

[40] J. Han, G. Cheng, Z. Li, and D. Zhang, "A unified metric learning-based framework for co-saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2473–2483, Oct. 2010.

[41] G. Cheng, P. Zhou, and J. Han, "Duplex metric learning for image set classification," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 281–292, Jan. 2018.



Xun Liu (S'16) received the B.Eng. degree from the Beijing Institute of Technology, Beijing, China, in 2014, where he is currently pursuing the Ph.D. degree with the School of Electrical and Information Engineering.

His research interests include remote sensing image fusion, super-resolution, and machine learning.

Chenwei Deng (M'09–SM'15) received the Ph.D. degree in signal and information processing from the Beijing Institute of Technology, Beijing, China, in 2009.

He was a Post-Doctoral Research Fellow with the School of Computer Engineering, Nanyang Technological University, Singapore. Since 2012, he has been an Associate Professor and then a Full Professor with the School of Information and Electronics, Beijing Institute of Technology. He has authored or coauthored more than 50 technical papers in refereed international journals and conferences and co-edited one book. His research interests include video coding, quality assessment, perceptual modeling, feature representation, object recognition, and image fusion.

Jocelyn Chanussot (M'04–SM'04–F'12) received the M.Sc. degree in electrical engineering from the Grenoble Institute of Technology (Grenoble INP), Grenoble, France, in 1995, and the Ph.D. degree from the Universit de Savoie, Annecy, France, in 1998.

In 1999, he was with the Delegation Generale de l'Armement (DGA), French National Defense Department, Geography Imagery Perception Laboratory, Arcueil, France. Since 1999, he has been with Grenoble INP. He has been a Visiting Scholar with Stanford University, Stanford, CA, USA; KTH Royal Institute of Technology, Stockholm, Sweden; and National University of Singapore (NUS), Singapore. Since 2013, he has been an Adjunct Professor with the University of Iceland, Reykjavík, Iceland. From 2015 to 2017, he was a Visiting Professor with the University of California at Los Angeles (UCLA), Los Angeles, CA, USA. He is currently a Professor of signal and image processing with Grenoble INP, where he is conducting his research at the Grenoble Images Speech Signals and Automatics Laboratory (GIPSA-Lab). His research interests include image analysis, multicomponent image processing, nonlinear filtering, data fusion, and machine learning in remote sensing.

Dr. Chanussot is a member of the IEEE Geoscience and Remote Sensing Society AdCom, the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society from 2006 to 2008, and the Institut Universitaire de France from 2012 to 2017. He was the co-recipient of the NORSIG 2006 Best Student Paper Award, the IEEE GRSS 2011 and 2015 Symposium Best Paper Award, the IEEE GRSS 2012 Transactions Prize Paper Award, and the IEEE GRSS 2013 Highest Impact Paper Award. He was the General Chair of the first IEEE GRSS Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote sensing (WHISPERS). He was the Chair from 2009 to 2011 and the Co-Chair from 2005 to 2008 of the GRS Data Fusion Technical Committee. He was the Program Chair of the IEEE International Workshop on Machine Learning for Signal Processing in 2009. He is an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and the IEEE TRANSACTIONS ON IMAGE PROCESSING. He was the Editor-in-Chief of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING from 2011 to 2015. He was a Guest Editor for the PROCEEDINGS OF THE IEEE in 2013 and the *IEEE Signal Processing Magazine* in 2014. He is a 2018 Highly Cited Researcher (Clarivate Analytics). He is the Founding President of IEEE Geoscience and Remote Sensing French Chapter from 2007 to 2010, which received the 2010 IEEE GRS-S Chapter Excellence Award.

Danfeng Hong (S'16) received the B.Sc. degree in computer science and technology from the Neusoft College of Information, Northeastern University, Shenyang, China, in 2012, and the M.Sc. degree in computer vision from Qingdao University, Qingdao, China, in 2015. He is currently pursuing the Ph.D. degree with Signal Processing in Earth Observation, Technical University of Munich (TUM), Munich, Germany.

Since 2015, he has been a Research Associate with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Weßling, Germany. In 2018, he was a Visiting Student with GIPSA-Lab, Grenoble INP, CNRS, University of Grenoble Alpes, Grenoble, France, under the supervision of Prof. J. Chanussot. His research interests include signal/image processing and analysis, pattern recognition, machine/deep learning, and their applications in Earth vision.

Baojun Zhao received the Ph.D. degree in electromagnetic measurement technology and equipment from the Harbin Institute of Technology (HIT), Harbin, China, in 1996.

From 1996 to 1998, he was a Post-Doctoral Fellow with the Beijing Institute of Technology (BIT), Beijing, China, where he is currently a Full Professor. He is also the Vice Director of Laboratory and Equipment Management, Beijing, and the Director of the National Signal Acquisition and Processing Professional Laboratory, Beijing. He has authored or coauthored more than 100 publications and received 5 provincial/ministerial-level scientific and technological progress awards in these fields. His research interests include image/video coding, image recognition, infrared/laser signal processing, and parallel signal processing.