Spatiotemporal Fusion of Land Surface Temperature Based on a convolutional Neural Network

Zhixiang Yin, Penghai Wu, Giles M. Foody, Yanlan Wu, Zihan Liu, Yun Du and Feng Ling

***Abstract*—Due to the trade-off between spatial and temporal resolutions commonly encountered in remote sensing, no single satellite sensor can provide fine spatial resolution land surface temperature (LST) products with frequent coverage. This situation greatly limits applications that require LST data with fine spatiotemporal resolution. Here, a deep learning based SpatioTemporal Temperature Fusion Network (STTFN) method for the generation of fine spatiotemporal resolution LST products is proposed. In STTFN, a multi-scale fusion convolutional neural network is employed to build the complex nonlinear relationship between input and output LSTs. Thus, unlike other LST spatiotemporal fusion approaches, STTFN is able to form the potentially complicated relationships through the use of training data without manually designed mathematical rules making it is more flexible and intelligent than other methods. Additionally, two target fine spatial resolution LST images are predicted and then integrated by a SpatioTemporal-Consistency (STC)-Weighting function to take advantage of spatiotemporal consistency of LST data. A set of analyses using two real LST data sets obtained from Landsat and Moderate Resolution Imaging Spectroradiometer (MODIS), were undertaken to evaluate the ability of STTFN to generate fine spatiotemporal resolution LST products. The results show that, compared to three classic fusion methods (the Enhanced Spatial and Temporal Adaptive Reflectance Fusion Model (ESTARFM), the Spatiotemporal Integrated Temperature Fusion Model (STITFM) and the Two-Stream Convolutional Neural Network for Spatiotemporal Image Fusion (StfNet)), the proposed network produced the most accurate outputs (Average RMSE<1.40˚C and average SSIM> 0.971).**

***Index Terms*—Spatiotemporal fusion, Land surface temperature, Deep learning, Complex nonlinear relationship, Spatiotemporal-Consistency (STC)-Weighting.**

This work was supported by the National Natural Science Foundation of China (grant No. 41501376), the Strategic Priority Research Program of Chinese Academy of Sciences (grant No. XDA 2003030201), the Innovation Group Project of Hubei Natural Science Foundation (grant No. 2019CFA019) and the open fund for Discipline Construction, Institute of Physical Science, and Information Technology at Anhui University. (*Corresponding authors: Penghai Wu, e-mail: wuph@ahu.edu.cn and Feng Ling, e-mail: lingf@asch.whigg.ac.cn*)

Z. Yin, Y. Du and F. Ling are with the Key Laboratory for Environment and Disaster Monitoring and Evaluation of Hubei province, Institute of Geodesy and Geophysics, Chinese Academy of Sciences, Wuhan 430079, China; Z. Yin is also with the University of Chinese Academy of Sciences, Beijing 100049, China, and with the Anhui Province Key Laboratory of Wetland Ecosystem Protection and Restoration, Anhui University, Hefei 230601, China.

P. Wu and Y. Wu are with the Anhui Province Key Laboratory of Wetland Ecosystem Protection and Restoration, Anhui University, Hefei 230601, China, and with the Institute of Physical Science and Information Technology, Anhui University, Hefei, Anhui 230601, China.

G. M. Foody is with School of Geography, University of Nottingham, University Park, Nottingham NG7 2RD, U.K.

Z. Liu is with Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, International Institute for Earth System Science, Nanjing University, Nanjing, Jiangsu 210023, China.

I. Introduction

SATELLITE-derived land surface temperature (LST) is of considerable importance to a diverse array of studies including environmental and climatic change [1, 2], land and water energy exchange with the atmosphere [3, 4] and ecological processes [5, 6]. In such studies, LST products with fine spatiotemporal resolution are often desired [7]. For instance, as a key variable in studies of urban heat island (UHI), LST must be in consistent with simulated hourly energy consumption data from urban buildings to evaluate the UHI impact on energy use [8]. However, due to technological and budget limitations, currently available satellite LST products do not have both fine spatial and temporal resolutions. Specifically, LST products with fine spatial resolution inevitably have coarse temporal resolution, and *vice versa* [9]. This situation arises mainly from the trade-off between spatial and temporal resolutions of satellite sensors and has greatly limited the potential of satellite LST products in various applications.

A variety of methods have been proposed for the generation of fine spatiotemporal resolution LST products [9-15]. Due to the advantages arising from the use of neighboring spatiotemporal change information concurrently, spatiotemporal fusion methods have been widely adopted [11, 16, 17]. Spatiotemporal fusion methods were originally used for the generation of fine spatiotemporal resolution reflectance imagery using images acquired from different satellite sensors with complementary spatial and temporal characteristics [18-20]. Example methods include the Multisensor Multiresolution Technique (MMT) [21], the spatial and temporal adaptive reflectance fusion model (STARFM) [22], an Enhanced version of STARFM (ESTARFM) [23] and the Flexible Spatiotemporal DAta Fusion (FSDAF) [24]. Given that LST and reflectance are both continuous land surface variables, some research has adopted such spatiotemporal fusion methods to estimate fine spatiotemporal LST products directly. For example, Liu *et al.* [25] generated a series of ASTER-like LST products using STARFM for community health research and Ma *et al.* [26] fused MODIS and Landsat LST data using ESTARFM to generate a Landsat-like LST product, which was required to estimate surface evaporation. Nevertheless, some characteristics of LST are different to reflectance. For example, reflectance often varies slowly with time while LST can change rapidly [27]. Consequently,

spatiotemporal fusion methods appropriate for the generation of reflectance imagery may not always be appropriate for the generation of LST products .

Various means have been proposed to enhance the ability of spatiotemporal fusion methods to generate fine spatial resolution LST products. They include considering the correlation between pixels in LST products, introducing temporal change models of LST into the analysis and increasing the number of satellite sensors. For example, Huang *et al.* [25] improved the weight function in STARFM with the aid of bilateral filtering to generate fine spatiotemporal resolution LST products of urban regions, and Wu *et al.* [29] used a variation-based constrained model to produce accurate Landsat-like LST products. By adding the annual temperature cycle (ATC) to the ESTARFM, Weng *et al.* [30] put forward the Spatio-temporal Adaptive Data Fusion Algorithm for Temperature mapping (SADFAT) to predict thermal radiance and LST data. To produce diurnal LST products at a Landsat-scale, Wu *et al.* [11] designed a Spatio-Temporal Integrated Temperature Fusion Model (STITFM) for the estimation of fine temporal and spatial resolution LST products, by integrating data from arbitrary satellite sensors including multi-scale polar-orbiting and geostationary satellites. Recently, several hybrid methods have been presented to take full advantage of different techniques and thus further enhance the spatiotemporal fusion. For instance, Quan *et al.* [16] designed a unified framework to BLEnd Spatiotemporal Temperatures (BLEST) derived from Landsat, MODIS, and geostationary satellites based on the integration of temporal interpolation, spatial downscaling and weight function-based fusion. Alternatively to enhance the spatial resolution of temporally dense LST data series, Xia *et al.* [17] proposed a weighted combination of kernel-driven and fusion-based methods (CKFM), which can inherit the advantages and overcome the shortcomings of each method simultaneously.

Although great progress has been made, spatiotemporal LST fusion remains a challenge because the relations between input and output LST may not be appropriately specified. For example, in some conventional LST fusion techniques [11, 28, 29], it is assumed that the output LST product could be expressed by a linear combination of inputs which may not be appropriate if there are nonlinear temporal changes of LST. Although some research has addressed the effect of temporal variations of LST and employed some mathematical or physical models [16, 30], these theoretical models are not always suitable because LST can be a highly changeable variable. For instance, in BLEST and SADFAT, ATC and diurnal temperature cycle (DTC) models based on prior knowledge are applied to describe the annual and diurnal change of LST. While the ATC and DTC models can to some extent reflect temporal variation of LSTs, they may perform poorly when abrupt changes in LST occur and when some latent relations between input and output LST products do not conform to the prior knowledge [31]. Furthermore, the methods are based on original low-level features such as image texture and location [32]. Therefore, they may not differentiate different objects and thus apply the same model parameters to various objects, while LST of these objects may change differently.

Inspired by the powerful nonlinear representation ability of convolutional neural network (CNN) [33], several CNN-based spatiotemporal reflectance fusion methods have recently been proposed. A five-layer CNN was, for example, put forward by Song *et al.* [34] to express the nonlinear relationships between reflectances estimated from MODIS and Landsat data. Alternatively, Tan *et al.* [35] designed a Deep Convolutional Spatiotemporal Fusion Network (DCSTFN) to obtain high spatiotemporal resolution images directly. Additionally, Liu *et al.* [36] presented a Two-Stream Convolutional Neural Network for Spatiotemporal Image Fusion (StfNet) that accommodated temporal dependency. CNN-based fusion methods utilize end-to-end CNN to automatically form the nonlinear relationship between observed pairs of coarse and fine spatial resolution image and use this to predict the target fine spatial resolution image. Because high-level features (e.g., edges of objects and their mutual relations) [37] which contain abundant semantic information can be extracted from the input data through CNN, the established relations between input and output are more practical, thus more favorable results can be acquired. But there are also some limitations when they are adopted for the generation of LST products. For example, the five-layer CNN and DCSTFN both lose some significant fine spatial information in the prediction and StfNet is a shallow network that has limited ability to form the potentially complex nonlinear relationship between the input and output LSTs. Therefore, for the spatiotemporal fusion of LST products, a CNN with enhanced capacity for nonlinear representation would be of considerable value.

This paper proposes a novel SpatioTemporal Temperature Fusion Network (STTFN) method for the accurate generation of fine spatiotemporal resolution LST products. For enhanced modelling of the potentially complex nonlinear relationships between input and output LSTs, STTFN uses a multi-scale fusion CNN. The multi-scale fusion CNN contains three parts: 1) super-resolution of temporal change between the input target and neighboring coarse spatial resolution LST images, 2) high-level feature abstraction of neighboring fine spatial resolution data, and 3) integration of the extracted multi-scale features. Furthermore, to take advantage of the temporal consistency commonly observed in a series of LST images, the STTFN uses two fine-coarse spatial resolution LST image pairs observed before and after the target date to train two multi-scale fusion CNNs, and then combine their predictions using a Spatiotemporal-Consistency (STC)-Weighting function. Considering the special characteristic of temperature, there are four differences of STTFN compared with existing deep learning-based spatiotemporal fusion method: 1) in the super-resolution process, fine spatial resolution image was used to provide fine spatial texture information, 2) residual learning modules were applied to preserve lost features in the convolution and fuse features at different scales, 3) in the training process, Huber loss function was used for decreasing the negative influence of noises and outliers, and 4) in the combination, STC-weighting strategy which considers spatiotemporal consistency was employed to derive the final result. Experiment results showed that the proposed method was effective and performed better than the comparator methods in generating fine spatial resolution LSTs.

The rest of this paper is organized as follows. The proposed STTFN fusion method is presented in Section II. Then, the performance of the proposed method is validated in Section III, which is followed by a discussion in Section IV and the conclusion in Section V.

II. METHODOLOGY

Landsat and MODIS LST data were used as fine and coarse spatial resolution data, respectively. The Landsat LST data has a pixel size of 30 m, and that of MODIS LST data is 1000 m. The aim was to obtain a pair of fine and coarse LST images that pre-dating the target date, for which a coarse MODIS image was available, together with another pair of fine and coarse image that post-dating the target date. Then, from one MODIS LST image at the target date and two pairs of Landsat and MODIS LST images at the neighboring dates, a Landsat-like LST image was predicted. For simplicity, the target date is denoted as t_2, and the dates pre- and post-dating t_2 are denoted as t_1 and t_3, respectively. Accordingly, the MODIS LST image at t_2 is denoted as M_2, and the Landsat-MODIS LST image pairs at dates t_1 and t_3 are denoted as L_1 and M_1 and L_3 and M_3, respectively. As depicted in Fig.1, the proposed STTFN generates a fine spatial resolution LST image via a three-stage process: 1) forward and backward model training, 2) forward and backward prediction and 3) combination of predicted Landsat-like LST images to yield a final fine spatial resolution predicted image for t_2.

The input fine and coarse spatial resolution LST images are first pre-processed. Then, a forward multi-scale fusion CNN, which blends L_1, M_1, and M_3 to predict $L_3^{1\to3}$, is gradually learned and optimized; the superscript 1→3 indicates forward modelling based on the image pairs at t_1 and t_3. Additionally, an optimally trained backward multi-scale fusion CNN can be obtained by using L_3, M_3 and M_1 to predict $L_1^{3\to1}$, the superscript highlights backward modelling. In the prediction stage, two predicted fine spatial resolution LST images, which are expressed as $L_2^{1\to2}$ and $L_2^{3\to2}$, are generated by the optimal forward and backward multi-scale fusion CNNs, respectively. Then $L_2^{1\to2}$ and $L_2^{3\to2}$ are combined via a STC-Weighting function. The following sub-sections explain the proposed STTFN more fully.

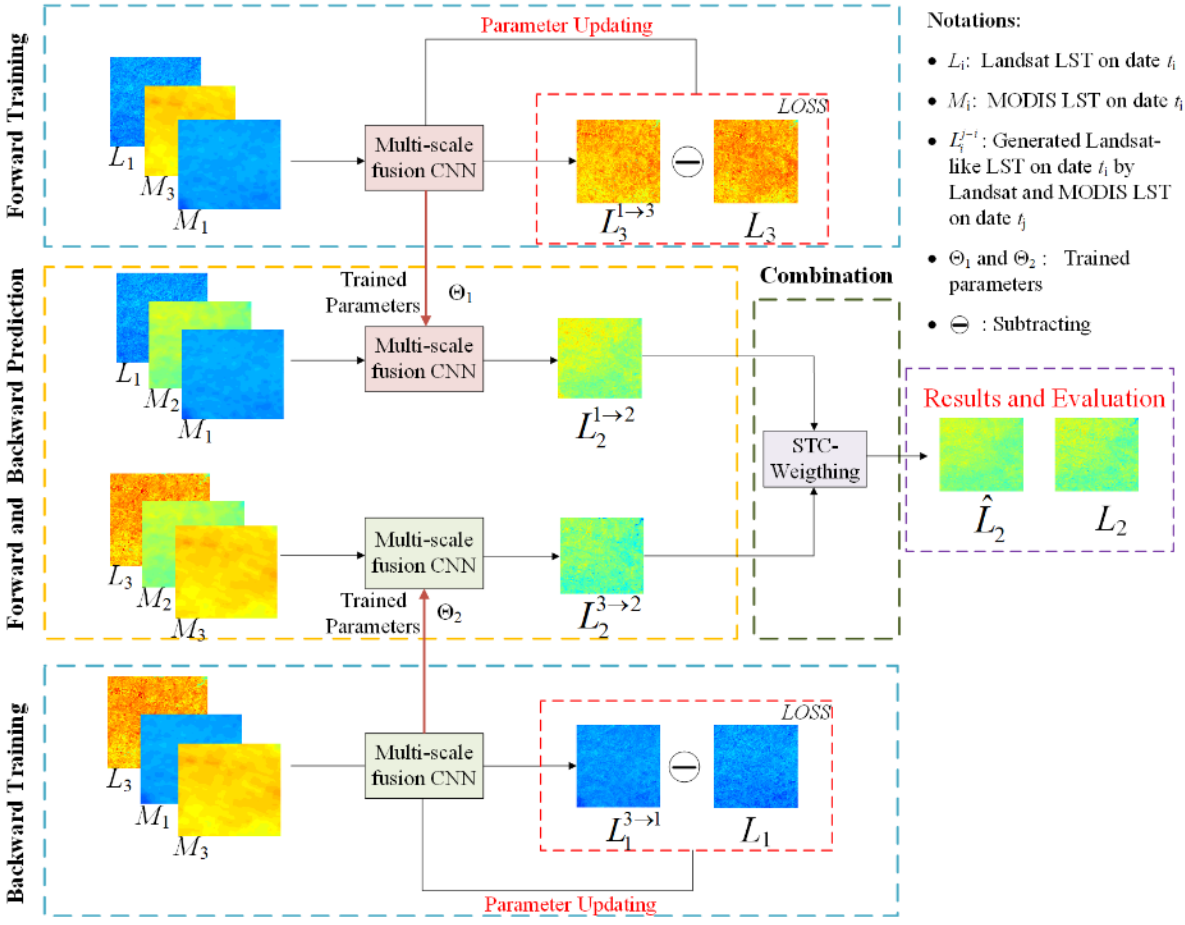

Fig.1. Flowchart of the generation of fine spatial resolution LST image with STTFN.

A. Layers used in the multi-scale fusion CNN

The most commonly used layer in the multi-scale fusion CNN is a convolution layer, which is used to extract and fuse high level feature maps. The inputs to the *i*-th convolution layer are the feature maps (comprised by extracted latent features of input LSTs) Y_{i-1} from the previous layer. To realize the feature extraction and fusion tasks, a convolution layer applies *k* filters

of size $r \times r \times c$ (where c is the number of channels) sliding through the input feature maps with a step size s, which is denoted here as "conv $r \times r$, k". Zero values were padded at the edges of the feature maps to ensure that the output size equaled to that of the input. The output feature maps Y_i were calculated as $Y_i = W_i * Y_{i-1} + b_i$, where W_i denotes the filter parameters and b_i is the bias parameter. The parameters are updated and optimized in the training stage. To improve the speed and performance of the network, a batch normalization layer [38] can be added after each convolution layer. Finally, behind the batch normalization layer, a rectified linear unit (ReLU) activation layer (max (0, ·)) [39] was adopted to ensure that the output is a nonlinear combination of the inputs. Table I lists the detailed composition of different layers of the multi-scale fusion CNN.

TABLE I
THE COMPOSITION OF THE DIFFERENT LAYERS OF THE MULTI-SCALE FUSION CNN

Extraction-Net		Super-Resolution-Net		Integration-Net	
level	layer	level	layer	level	layer
E1	conv 3 × 3, 32, Batch Normalization, ReLU	SR1	conv 3 × 3, 32	F1	conv 3 × 3, 32, Batch Normalization, ReLU
E2	conv 3 × 3, 32	SR2	conv 1 × 1, 64, Batch Normalization, ReLU	F2	conv 3 × 3, 16, Batch Normalization, ReLU
		SR3	conv 1 × 1, 25, Batch Normalization, ReLU	F3	conv 3 × 3, 1
		SR4	conv 1 × 1, 32, Batch Normalization, ReLU		
		SR5	conv 3 × 3, 32, Batch Normalization, ReLU		
		SR6	conv 3 × 3, 32		

B. Architecture of the multi-scale fusion CNN

The architecture of multi-scale fusion CNN is illustrated using the prediction of $L_2^{1\to 2}$ from L_1, M_1 and M_2 as example (Fig.2). The multi-scale fusion CNN is a fully convolution network, which enables an end-to-end mapping from three input LST images to an output LST image. It first extracts high-level features of L_1 and super resolves the change LST image between M_2 and M_1 at the same time, then fuses and retrieves the extracted feature maps to the fine spatial resolution result. Therefore, the whole architecture of the multi-scale fusion CNN contains three major parts, termed here as the Extraction-Net, Super-Resolution-Net and Integration-Net. The first part of the multi-scale fusion CNN is the Extraction-Net, which is a two-layer convolution network and employed to extract the high-level features of L_1. The derived high-level features can provide fine abundant spatial pattern information for the prediction.

The second part is the Super-Resolution-Net. To obtain the high-level features with abundant spatial pattern information on the temporal changes between M_2 and M_1, the change image is first concatenated with L_1 and then put into the super-resolution module (Fig.2). The focus in the fusion process is on locating of temporal change between the target and neighboring dates. Unlike DCSTFN [35], multi-scale fusion CNN acquires the change image first and then uses it as the input to Super-Resolution-Net. This makes the subsequent network pay more attention to the locations of change and simultaneously reduces its computational burden. By using a super-resolution module, the coarse spatial resolution pixels of the temporal change LST image are disaggregated into fine spatial resolution pixels, which can offer more spatial pattern information. Note that, concatenation of L_1 in the super-resolution is of primary importance, since it provides fine scale information for the disaggregation [40]. To implement the super-resolution, a modified version of the state-of-the-art super-resolution model "Wide Activation for Efficient and Accurate Image Super-Resolution (WDSR)" was used [41], referred to here a slim-WDSR (Fig.2). The core of slim-WDSR are five convolution layers. Through these layers, the slim-WDSR can extract high-level features of the coarse spatial resolution LST image and map them to the features of the fine spatial resolution LST image. However, in the feature extraction and mapping process, some low-level features may be lost [42]. To address this deficiency, two local residual learnings were applied in the slim-WDSR since it can retain shallow features and prevent the problem of misconvergence by adding low-level features to the extracted high-level features [43].

The third part is the Integration-Net, which is made up of three stacked convolutional layers. The generated feature maps are integrated and used to estimate the target LST image gradually through the Integration-Net. Specifically, the feature maps from the Extraction-Net and Super-Resolution-Net are first blended:

$$\text{Features}\left(L_2^{1\to 2}\right) = \text{Features}\left(M_2 - M_1\right) + \text{Features}\left(L_1\right) \quad (1)$$

The blended feature maps are composed of various layers of feature maps, in which different layers will contain different information crucial for estimating the fine spatial resolution LST image. For example, layer one may contain edge information of the involved objects, while layer two is likely to be made up of location information. Therefore, the Integration-Net is used to integrate the various types of information and convert the blended feature maps to the target fine LST image. As aforementioned, low-level features may partly disappear because of convolutions, thus the initial fused image is added to the end of the whole net, which is termed global residual learning. By using local and global residual learning, low- and high-level features at different scales of the input can be fully utilized.

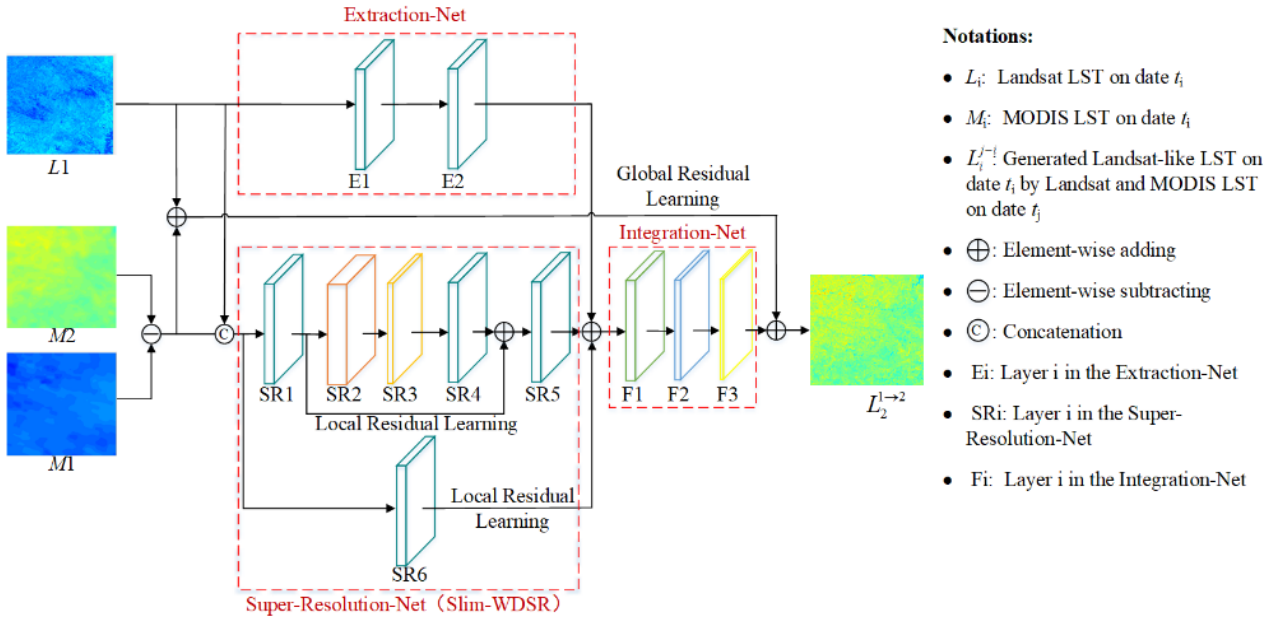

Fig.2. The architecture of the multi-scale fusion CNN (showing the prediction of $L_2^{1\to2}$ from L_1, M_1 and M_2 as an example, M_1 and M_2 were upsampled by bilinear interpolation).

C. *Training and prediction*

Two multi-scale CNNs termed forward multi-scale CNN and backward multi-scale CNN are trained. In other studies [34-36], mean square error (MSE) was used as the loss function in the training. Here, to reduce the effects of noise and outliers, e.g., caused by unidentified cloud in input images, Huber Loss [44] was adopted as the loss function , which is more robust to outliers than MSE loss. The loss of pixel i is denoted as:

$$L(w_t)=\begin{cases}\frac{1}{2}\left\|F(M_{1i},L_{1i},M_{3i};w_t)-L_{3i}\right\|^2 & \left|F(M_{1i},L_{1i},M_{3i};w_t)-L_{3i}\right|\le\delta \\ \delta\times\left|F(M_{1i},L_{1i},M_{3i};w_t)-L_{3i}\right|-\frac{1}{2}\delta^2 & otherwise\end{cases} \quad t=1,2 \tag{2}$$

where δ is a hyper-parameter to determine which formula to be used, here δ was empirically set to 1.0. $F(\cdot)$ denotes the multi-scale fusion network, N is the number of pixels in the input LST imagery, and w_1 and w_2 denote the forward and backward network parameters to be updated during the training process, respectively. In the training stage, the two networks are optimized separately.

For optimization, the weights of the proposed network were initialized to small random values, which were drawn from a Gaussian distribution with zero mean and standard deviation of 1×10^{-3}. The adaptive moment estimation (Adam) with standard back propagation [45] was applied to minimize the loss and update the network weights until convergence (losses of STTFN did not change significantly), a value of β_1=0.9 and β_2=0.999 were set for Adam. The learning rate α was initialized as 1×10^{-4} and is multiplied by a decaying factor 0.1 every 10 epochs to shrink the searching range of the parameters.

In the prediction stage, two fine spatial resolution images $L_2^{1\to2}$ and $L_2^{3\to2}$ at the target date, t_2, are produced from the two trained networks. $L_2^{1\to2}$ is generated from M_1, L_1 and M_2 with the trained forward multi-scale fusion CNN (Fig.2), while $L_2^{3\to2}$ is obtained from M_3, L_3, and M_2 via the backward multi-scale fusion CNN. These predicted images are denoted as follows:

$$\begin{cases}L_2^{1\to2}=F(M_1,L_1,M_2;w_1)\\ L_2^{3\to2}=F(M_3,L_3,M_2;w_2)\end{cases} \tag{3}$$

Then, the two predicted fine spatial resolution LST images are combined to produce the final predicted LST image for the target date, $\hat{L}_2$:

$$\hat{L}_2=p_1\times L_2^{1\to2}+p_3\times L_2^{3\to2} \tag{4}$$

where p_1 and p_3 are the weighting parameters for $L_2^{1\to2}$ and $L_2^{3\to2}$, respectively. To determine p_1 and p_3 in the combination, the spatial consistency between the two predicted fine spatial resolution images and the corresponding coarse spatial resolution LST image was considered and a novel weighting strategy termed here STC-Weighting was proposed. It was different from previous fusion methods [34, 36], in which the temporal change between the coarse spatial resolution images at the neighboring dates (M_1 and M_3) and target date (M_2) was used to calculate the weights (termed here Temporal Consistency (TC)-Weighting), which can be denoted as:

$$p_i=\frac{\frac{1}{\left|M_i-M_2\right|}}{\frac{1}{\left|M_1-M_2\right|}+\frac{1}{\left|M_3-M_2\right|}} \quad i=1,3 \tag{5}$$

In STC-Weighting, weight magnitude depends on the difference between the predicted fine spatial resolution images and the coarse spatial resolution image at the target date, i.e., a smaller difference results in a higher weight. Specifically, the differences between M_2 and $L_2^{1\to2}$, M_2 and $L_2^{3\to2}$ are used to calculate the weighting parameters:

$$p_i=\frac{\frac{1}{\left|L_2^{i\to2}-M_2\right|}}{\frac{1}{\left|L_2^{1\to2}-M_2\right|}+\frac{1}{\left|L_2^{3\to2}-M_2\right|}} \quad i=1,3 \tag{6}$$

III. EXPERIMENTAL PARTS

A. Study areas and data

Two study areas were used here. The first study area was located in Évora, Portugal (Fig.3(a)). This was a large region of expansive plains covered mainly by Holm and Cork Oaks trees. The second study area was in the Shengjin Lake nature reserve, China (Fig.3(b)). The land cover mosaic in this study area was more complex than in Évora and, in addition, the lake's water level fluctuated greatly between seasons and hence its extent was variable [46].

The MODIS/Terra LST and Emissivity Daily L3 Global 1-km SIN Grid product (MOD11A1, Collection 6) were obtained from the website of The Level 1 and Atmosphere Archive and Distribution System Distributed Active Archive Center (https://ladsweb.modaps.eosdis.nasa.gov/), the RMSEs of the MOD11A1 product are within 2 ℃ for most landcovers [47, 48]. Landsat LST data were derived from imagery acquired in Landsat TM/ETM+ thermal bands provided at 30 m pixel size, which were downloaded from the United States Geological Survey (USGS) Earth Explorer (http://earthexplorer.usgs.gov/). Data gaps caused by the ETM+ SLC-off problem were filled by a linear interpolation algorithm. Landsat LSTs were estimated with a single channel algorithm, whose accuracy is reportedly within 1.5 ℃ [49].

For Évora, 18 cloud-free Landsat-MODIS LST image pairs between January 2010 and October 2011 were obtained (Table II). These image pairs each covered an area of 39 km×39 km and, hence, each Landsat LST products comprised 1300 pixels × 1300 pixels while the MODIS LST product comprised 39 pixels × 39 pixels. The data acquisition dates included almost every month of the year and, therefore, captured phenological change.

For the Shengjin Lake nature reserve study area, 5 cloud-free Landsat-MODIS LST image pairs, which covered an area of 24 km × 24 km, were acquired. The Landsat LST product comprised 800 pixels × 800 pixels and the MODIS LST product comprised 24 pixels× 24 pixels. The rise and fall of the lake water level during the period of study caused changes in the areal extent of the water body.

In the training process, according to the structure of the multi-scale fusion CNN, the input coarse resolution images are not the original MODIS data but the interpolated image that match the size of Landsat data, and bilinear interpolation was used to interpolate the MODIS data. The original input Landsat and interpolated MODIS LST images were cropped into image patches of 40 pixels × 40 pixels. Considering the different image sizes and landcover types and via a series of experiments, the cropping strides were set to 20 and 15 for Évora and Shengjin Lake, respectively. Accordingly, a total of 4096 and 2704 image patches from each original input LST image were used to train the network for Évora and Shengjin Lake, respectively. These image patches were randomly chose in the training to ensure convergence and generalization ability of network.

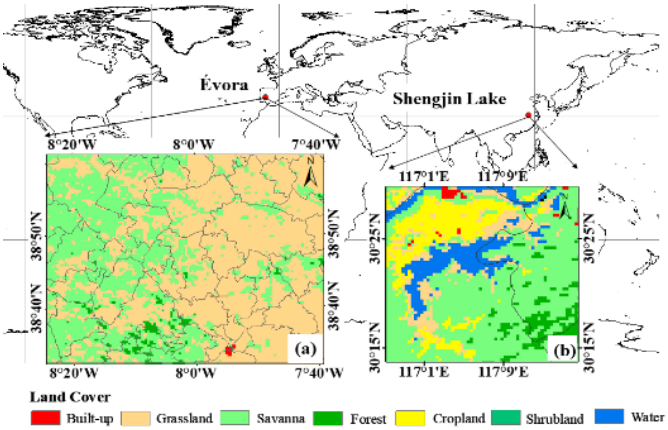

Fig.3. Study areas. (a) The land cover map of Évora area; (b) the land cover map of Shengjin Lake nature reserve area. Both of them are derived from the MODIS yearly land cover product in 2011 and 2016 respectively under the International Geosphere–Biosphere program (IGBP) [50] .

TABLE II
THE LST DATA USED FOR THE TWO STUDY AREAS

Évora				Shengjin Lake nature reserve			
Pair Number	Date	Landsat sensor and overpass time	MODIS overpass time	Pair Number	Date	Landsat sensor and overpass time	MODIS overpass time
1	28 Jan 2010	TM, 10:59	10:40	1	6 Nov 2016	ETM+, 10:47	11:03
2	13 Feb 2010	TM,10:59	10:57	2	8 Dec 2016	ETM+, 10:46	11:05
3	9 Mar 2010	ETM+, 11:00	11:09	3	26 Feb 2017	ETM+, 10:46	11:09
4	10 Apr 2010	ETM+, 11:00	11:08	4	14 Mar 2017	ETM+, 10:46	11:08
5	20 May 2010	TM, 10:59	10:30	5	15 Apr 2017	ETM+, 10:46	11:09
6	21 Jun 2010	TM, 10:59	10:31				
7	24 Aug 2010	TM, 10:59	10:30				
8	9 Sep 2010	TM,10:59	10:29				
9	11 Oct 2010	TM, 10:58	10:37				
10	4 Nov 2010	ETM+, 11:01	11:09				
11	20 Mar 2011	TM, 10:58	10:31				
12	5 Apr 2011	TM, 10:58	10:41				
13	15 May 2011	ETM+, 11:02	11:11				
14	24 Jun 2011	TM, 10:58	10:30				
15	26 Jul 2011	TM, 10:57	10:30				
16	27 Aug 2011	TM, 10:57	10:33				
17	12 Sep 2011	TM, 10:57	10:30				
18	6 Oct 2011	ETM+, 11:02	11:09				

B. *Comparator methods*

For an informative assessment, three traditional fusion methods were used for comparison, the ESTARFM [23] , the STITFM [11] and the StfNet [36]. These methods all assume images from different satellite sensors observed at the same date are comparable and correlated. ESTARFM is a widely adopted spatiotemporal fusion method which uses images from two satellite sensors as input [23], while STITFM is a modified version of ESTARFM accounting for input images from arbitrary satellite sensors [11]. Both methods obtain spatial variations from the input fine spatial resolution images and acquire temporal changes from the input fine temporal resolution images through linear weight functions [51]. StfNet applies a two-stream CNN to build the nonlinear relations between input and output images and predicts a fine spatiotemporal resolution image by incorporating temporal information to fine spatial resolution image series [36]. Here, the inputs of these methods are a coarse spatial resolution image at the target date and two pairs of coarse and fine images pre- and post-dating the target date.

Regarding quantitative evaluation, the absolute error (AE), root mean square error (RMSE) and structural similarity (SSIM) were calculated using the real fine resolution image at the target date as reference. AE denotes the difference between pixels of the real and predicted images, RMSE represents the difference between the real and predicted images, while SSIM reflects the correlation of spatial details between the real and predicted images. The most accurate prediction will have a low AE and RMSE as well as a high SSIM.

C. *Validity of STC-Weighting and Huber Loss*

The performance of STC-Weighting and TC-Weighting based combination was compared (Fig.4). It shows that STC-Weighting and TC-Weighting yielded similar results in most dates, while TC-Weighting had a poor performance on 11 October 2010. It may because that spatial consistency was not considered in TC-Weighting. Three subsets of Landsat LST images on 09 September 2010, 11 October 2010, and 4 November 2010 were shown in Fig.5. We can see that, there were many anomalous values in the gap-filled Landsat ETM+ LC-off LST image on 4 November 2010, resulting in large difference of spatial patterns between Landsat LSTs on 11 October 2010 and 4 November 2010. However, as indicated in Table III, the difference of MODIS LSTs between the target date (11 October 2010) and the date post-dating the target date (4 November 2010) is smaller than that between the target date (11 October 2010) and the date pre-dating the target date (9 September 2010). Therefore, in TC-Weighting based combination, the backward prediction based on the image pair on 4 November 2010 would obtain a higher weight due to the smaller difference of MODIS LSTs with the target date. While the error of the backward prediction is much larger than the forward prediction based on image pair from 9 September 2010 (Fig.6) due to spatial inconsistency, and resulted in large errors of the TC-Weighting based combination. In contrast, the STC-Weighting based combination considered the spatiotemporal consistency between the predicted results and the actual MODIS data at the target date and, thus, yielded a more robust result. Additionally, final results from the combinations of forward and backward predictions through STC-Weighting are better than the forward and backward predictions (Fig.7) due to the consideration of spatiotemporal consistency.

TABLE III
AVERAGE VALUE OF MODIS AND LANDSAT LSTS ON TARGET DATE (11 OCTOBER 2010), NEIGHBORING DATES (09 SEPTEMBER 2010 AND 4 NOVEMBER 2010)

Date	09 Sep 2010	11 Oct 2010	4 Nov 2010
MODIS	34.27	21.61	23.03
Landsat	34.91	23.02	23.12

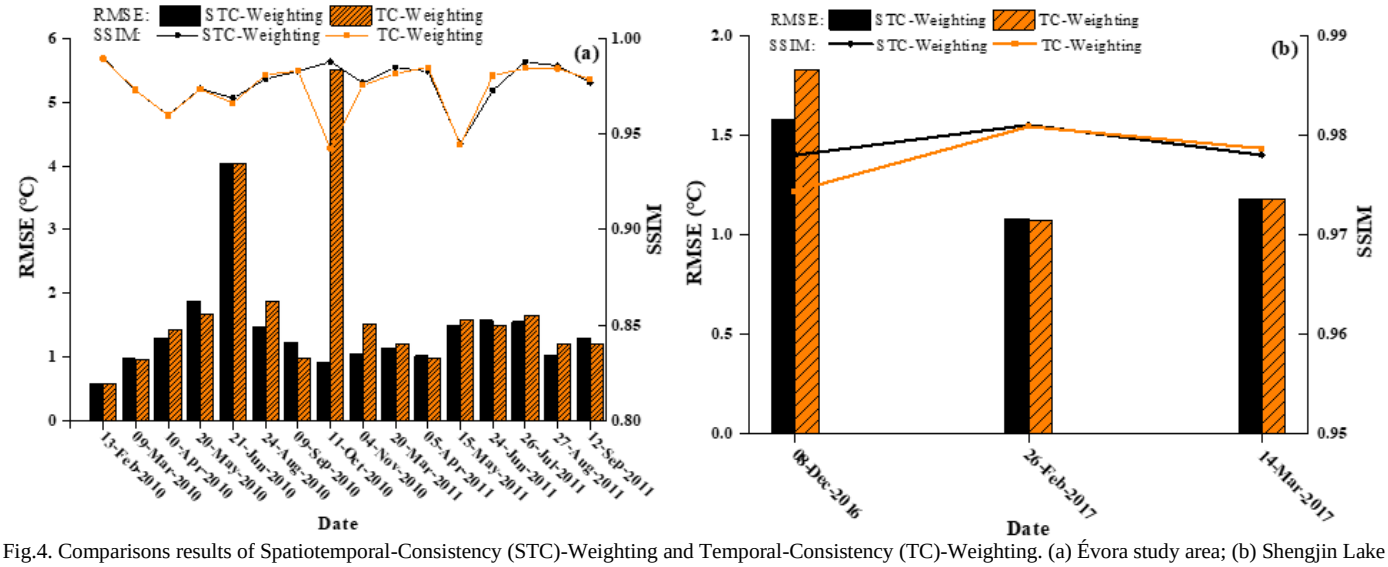

Fig.4. Comparisons results of Spatiotemporal-Consistency (STC)-Weighting and Temporal-Consistency (TC)-Weighting. (a) Évora study area; (b) Shengjin Lake study area.

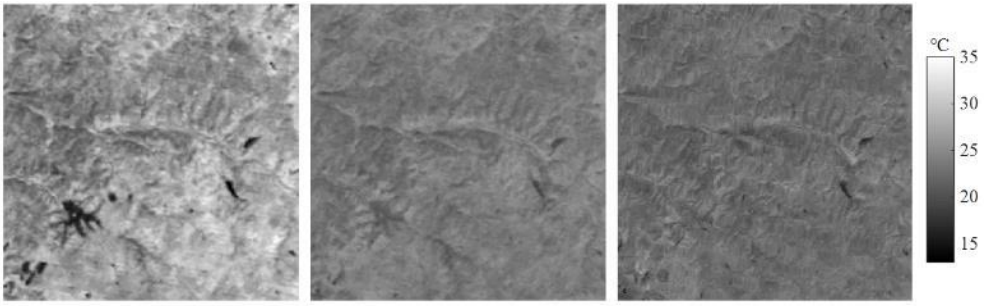

Fig.5. Subsets of Landsat LST images on 09 September 2010, 11 October 2010, and 4 November 2010.

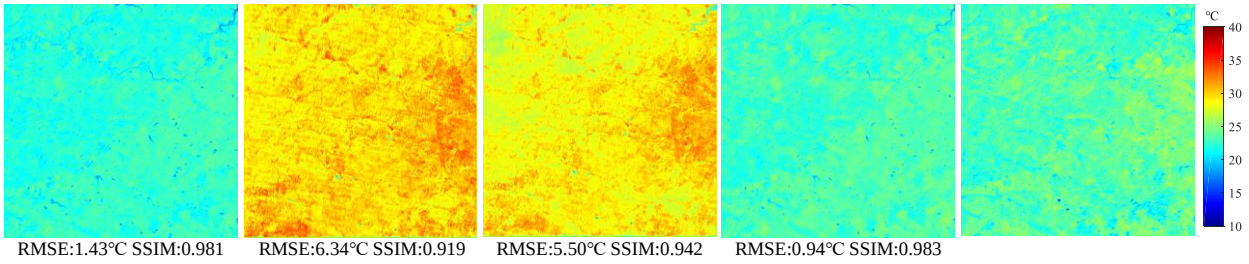

Fig.6. Forward and backward predictions compared with TC-Weighting and STC-Weighting based result and actual LST on 11 Oct 2010. (a) Forward prediction; (b) Backward prediction; (c) TC-Weighting based result; (d) STC-Weighting based result; (e) Actual Landsat LST.

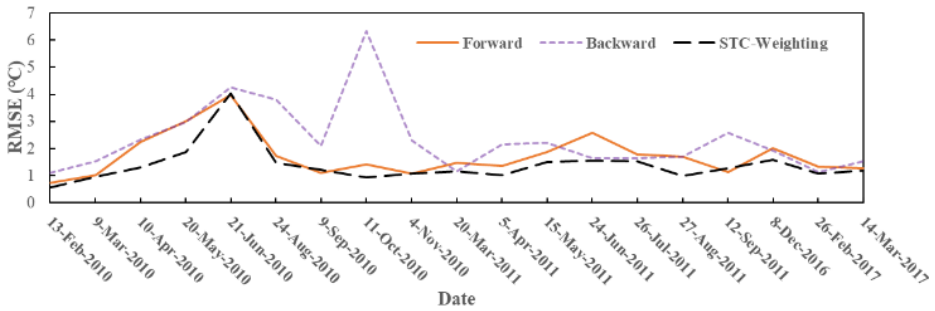

Fig.7. RMSEs of forward and backward predictions compared with STC-Weighting results on all dates

Huber Loss was adopted as the loss function to optimize the network is of considerable importance for decreasing the negative influences of the abnormal values in the input LST images. The lower RMSEs of predictions in Évora area with Huber loss compared to MSE loss (Fig.8) indicates that more favorable results can be derived by Huber Loss.

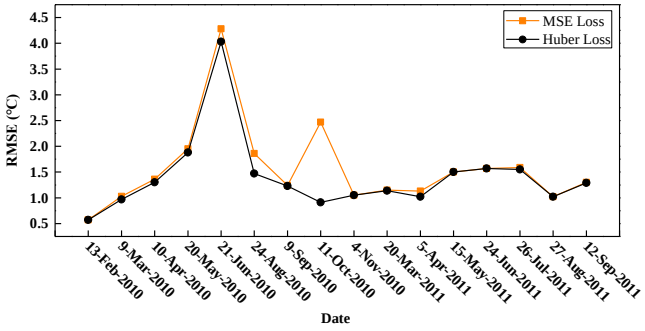

Fig.8. RMSEs of predictions by using MSE loss and Huber Loss.

D. Evaluations Based on Actual Data

In the prediction stage, all Landsat-MODIS pairs were arranged in numeric order as defined in Table II. A Landsat-like LST image on the target date was predicted using the corresponding MODIS LST image and two Landsat-MODIS LST image pairs. The acquisition dates of the two Landsat-MODIS LST image pairs were those that immediately pre- and post-dated the target date. For example, the Landsat-MODIS LST image pair numbers 1 and 3, and the MODIS LST image number 2 are used to predict a Landsat-like LST image for the date of MODIS image number 2. Therefore, excluding the first and last LST image pairs, 16 Landsat-like LST images were predicted for Évora area and 3 Landsat-like LST images were predicted for Shengjin Lake reserve area. Because the multi-scale fusion CNN in the STTFN is a fully convolutional network, it can theoretically process images with arbitrary size. Therefore, original LST images without cropping were entered into the network to yield the predicted image. The accuracy of the prediction was assessed relative to the actual Landsat image for the target date.

For the Évora study area, the quantitative evaluations for the fusion results of ESTARFM, STITFM, StfNet and STTFN are summarized in Table IV. It is evident that, the proposed STTFN produced the highest SSIMs on the prediction dates except for 9 March 2010 and 24 Aug 2010. Additionally, with the exception of three dates (9 March 2010, 21 June 2010, and 20 Mar 2011), the RMSEs for the predictions from STTFN were all lower than those from ESTARFM, STITFM, and StfNet. Finally, the STTFN produced the most accurate predictions in terms of the average RMSE and SSIM.

TABLE IV
QUANTITATIVE EVALUATION FOR THE ÉVORA STUDY AREA. (MOST ACCURATE RESULT HIGHLIGHTED IN BOLD)

Date	ESTARFM		STITFM		StfNet		STTFN	
	RMSE(˚C)	SSIM	RMSE(˚C)	SSIM	RMSE(˚C)	SSIM	RMSE(˚C)	SSIM
13 Feb 2010	2.06	0.932	1.98	0.931	1.30	0.980	**0.57**	**0.987**
09 Mar 2010	**0.95**	**0.967**	1.37	0.950	1.95	0.956	0.97	0.966
10 Apr 2010	2.14	0.944	2.86	0.903	1.49	0.945	**1.31**	**0.949**
20 May 2010	1.90	0.963	3.68	0.887	2.44	0.957	**1.88**	**0.966**
21 Jun 2010	4.37	0.945	**2.63**	0.936	4.03	0.950	4.03	**0.962**
24 Aug 2010	2.69	0.968	5.33	0.913	1.82	**0.976**	**1.46**	0.973
09 Sep 2010	1.31	0.977	3.14	0.958	1.59	0.974	**1.21**	**0.978**
11 Oct 2010	2.79	0.955	2.82	0.915	1.07	0.980	**0.94**	**0.983**
04 Nov 2010	1.15	0.970	1.90	0.950	1.07	0.970	**1.07**	**0.971**
20 Mar 2011	**1.14**	0.977	1.38	0.946	1.22	0.979	1.15	**0.981**
05 Apr 2011	1.15	0.980	1.88	0.948	1.33	0.977	**1.03**	**0.983**
15 May 2011	1.65	0.929	2.46	0.895	1.72	0.927	**1.52**	**0.931**
24 Jun 2011	2.59	0.955	2.75	0.872	2.98	0.953	**1.55**	**0.967**
26 Jul 2011	1.80	0.978	2.65	0.933	2.31	0.982	**1.53**	**0.985**
27 Aug 2011	1.23	0.980	2.45	0.942	1.77	0.978	**1.00**	**0.982**
12 Sep 2011	1.53	0.962	2.94	0.937	2.31	0.969	**1.29**	**0.972**
Average	1.90	0.961	2.64	0.927	1.90	0.966	**1.40**	**0.971**

For 21 June 2010, the RMSE values computed for the prediction from ESTARFM, StfNet and STTFN were large (all > 4.0 ˚C), much larger than those on other dates, while the RMSE for STITFM was relatively small. This might because of the large discrepancy between MODIS and Landsat LST on 21 June 2010 (Fig.9), and the overall differences between Landsat and MODIS images on 20 May 2010 and 24 August 2010 were much smaller than that on 21 June 2010. This caused the relationships formed in ESTARFM, StfNet and STTFN between neighboring dates to be inappropriate for predicting the fine spatial resolution image on the target date. For STITFM, the fine spatial resolution image prediction is based mainly on adding the differences between Landsat and MODIS images on neighboring dates to the MODIS image on the target date [11] rather than forming relations between the images and, thus, produced a relatively accurate result.

Table V summarises the quantitative assessment of the output from different algorithms applied to the Shengjin Lake area. Here, the predictions from STTFN had the highest SSIMs on all dates. The RMSEs of STTFN on 08 December 2016 and 26 February 2017 are lower than for the other methods. The only date for which the prediction with STTFN did not yield the lowest RMSE was 14 Mar 2017 for which ESTARFM has a very slightly lower RMSE. The average results of the four methods also indicate the STTFN generally produced the most accurate predictions.

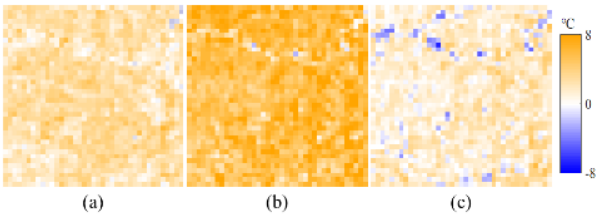

Fig.9. Differences of aggregated Landsat (pixel size of 1000m) and MODIS LST images on (a) 20 May 2010, (b) 21 June 2010 and (c) 24 August 2010.

TABLE V
QUANTITATIVE EVALUATION FOR THE SHENGJIN LAKE STUDY AREA. (MOST ACCURATE RESULT HIGHLIGHTED IN BOLD)

Date	ESTARFM		STITFM		StfNet		STTFN	
	RMSE(˚C)	SSIM	RMSE(˚C)	SSIM	RMSE(˚C)	SSIM	RMSE(˚C)	SSIM
08 Dec 2016	1.81	0.943	3.07	0.920	3.84	0.952	**1.59**	**0.975**
26 Feb 2017	1.58	0.952	2.03	0.949	2.46	0.967	**1.08**	**0.977**
14 Mar 2017	**1.09**	0.971	3.21	0.911	2.42	0.961	1.18	**0.973**
Average	1.49	0.955	2.77	0.927	2.90	0.960	**1.28**	**0.975**

The AE distribution of image pixels for the Évora study area (Fig.10 (a)) and Shengjin Lake study area (Fig.10 (b)) on all prediction dates also shows that STTFN has a better performance in both study areas (almost 60% AE < 1.0 ˚C). ESTARFM has a comparable AE distribution with STTFN for Shengjin Lake study area (Fig.10 (b)), while it performs not well in Évora study area (less than 40% AE < 1.0 ˚C) (Fig.10 (a)). ESTARFM, STITFM and StfNet have similar distribution of AE in Évora study area, but they perform differently in Shengjin Lake area.

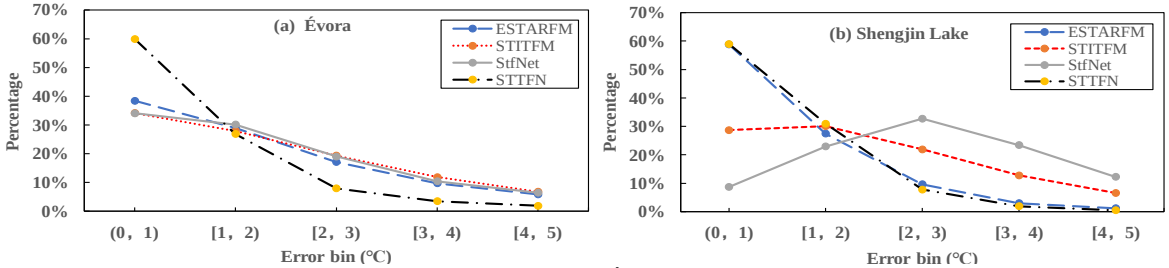

Fig.10. AE distribution of the four methods. (a) Évora study area; (b) Shengjin Lake study area.

The results in Tables IV and V indicate that the predictions of the proposed STTFN were closest to the actual data and retained the most structural similarities. Fig.10 shows that STTFN not only yielded the most accurate results but had the stablest performance. This is because the proposed STTFN established a realistic nonlinear relationship between input and output LSTs, whereas ESTARFM and STITFM are limited to linear relationships. Though StfNet builds the nonlinear relationship between input and output LSTs, it is a shallow network, which is unable to adequately represent complex nonlinear relationships. Additionally, CNN is able to extract the high-level features which contain abundant semantic information (e.g., edges of objects in images and their mutual relations), thus, the average SSIMs of STTFN and StfNet, which include CNN, in the two study areas are higher than the other two methods. To illustrate the quality of the spatial representation of LST in the various fusion results, predictions from the four methods and the actual Landsat LST images, and the error maps between four predictions and the corresponding actual LST images on 24 Aug 2010 (T_a and d_a) and 24 Jun 2011 (T_b and d_b) for Évora and on 26 February 2017 (T_c and d_c) for Shengjin Lake reserve are depicted in Fig.11, respectively. It is evident that on 24 August 2010 and 24 June 2011 large parts of the predicted LST images obtained from STITFM and ESTARFM are higher than the actual values, especially for STITFM (Fig.11). The predictions from ESTARFM are visually close to the actual LST image for 26 February 2017, but there are some considerably higher values as well (red rectangle in Fig.11.) The predicted image from StfNet for 24 August 2010 was visually good but too smooth, while that for 24 June 2011 showed some over-prediction. The smoothness problem also occurred for the 26 February 2017, for which large areas show predicted values lower than the actual values. Overall, the predicted LST images generated by STTFN are most consistent with the actual LST images, which agrees with the SSIMs in Tables IV and V.

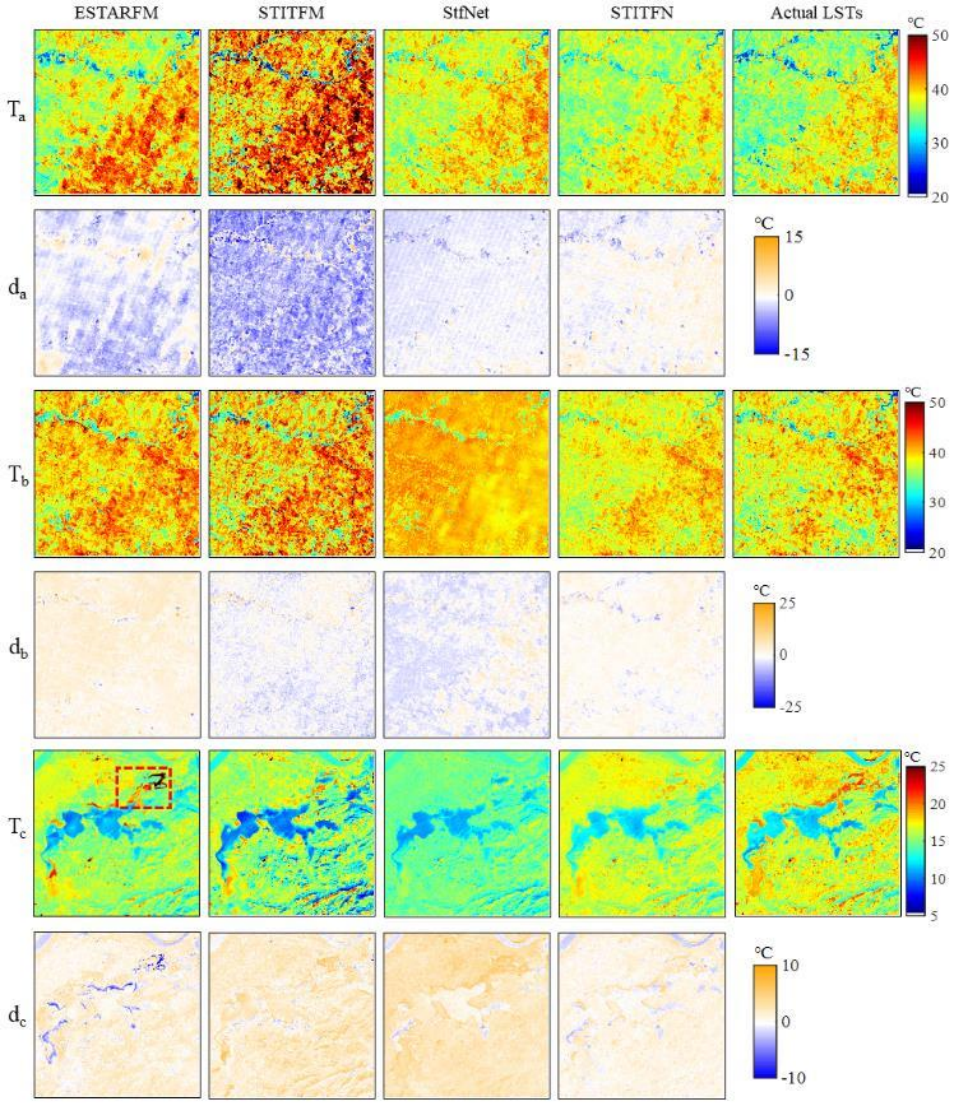

Fig.11. Examples of the fusion results and the actual Landsat LSTs and the error maps (d_a, d_b and d_c) of the predictions on 24 Aug 2010 (T_a) and 24 Jun 2011(T_b) for Évora and on 26 February 2017 (T_c) for Shengjin Lake reserve.

For ESTARFM and STITFM, each predicted pixel is a weighted combination of the inputs. Therefore, if there are any abnormal values in the input LST image, such as local inconsistency in the gap-filled Landsat ETM+ SLC-off images, the output of the fusion product will contain corresponding abnormal values. For example, Fig.12 shows the input LST

images and the corresponding results of the four methods for 13 February 2010. Some values in the left-bottom bold rectangle part of the MODIS LST image (Fig.12 (f)) are clearly not consistent with the Landsat LST image on 9 March 2010 (Fig.12 (c)). This resulted in the generation of several corresponding abnormal pixels in the predicted LST images of ESTARFM and STITFM, as shown in Fig.12 (g) and (h). With STTFN, the predicted image is the combination of the two predicted outcomes from the learned forward and backward multi-scale fusion CNNs. Therefore, if the abnormal values occur in only one pair of neighboring images, this pair will be given low weight in the abnormal regions and hence has little negative influence. The core procedure of StfNet is similar to STTFN, so it also has few abnormal values in the corresponding region. Similarly, when anomalous values occur in the gap-filled Landsat ETM+ SLC-off LST image, the results of ESTARFM and STITFM are impacted greatly, e.g., the road in the red bold box disappeared (Fig.13 (a) and (b)), while the effect on predictions from StfNet and STTFN is relatively small (Fig.13).

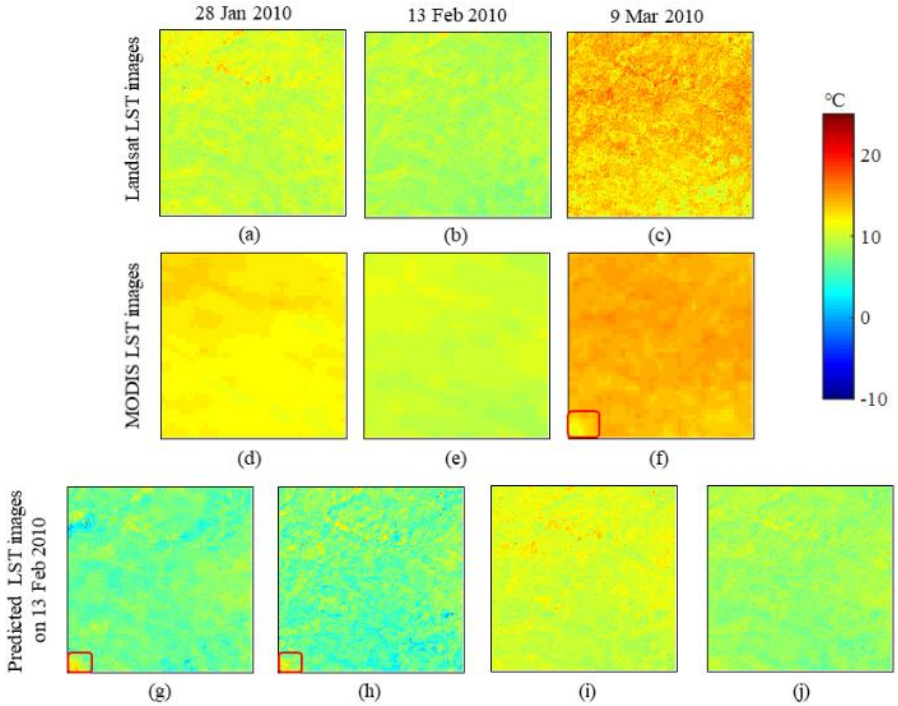

Fig.12. Input LST images and predictions from the four methods for 13 February 2010. (a) Landsat LST image on 28 January 2010; (b) Landsat LST image on 13 February 2010; (c) Landsat LST image on 9 March 2010; (d) MODIS LST image on 28 January 2010; (e) MODIS LST image on 13 February 2010; (f) MODIS LST imagery on 9 March 2010; (g) the predicted LST image from ESTARFM; (h) the predicted LST image from STITFM; (i) the predicted LST image from StfNet; (j) the predicted LST image from STTFN.

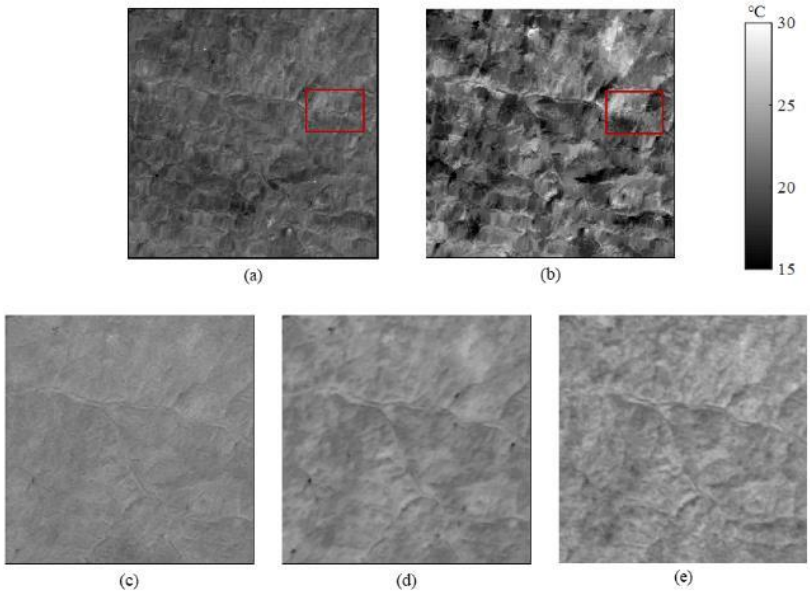

Fig.13. Fusion results of the four methods when anomalous values exist in gap-filled Landsat LST image on 11 October 2010. (a) the predicted LST image from ESTARFM; (b) the predicted LST image from STITFM; (c) the predicted LST image from StfNet; (d) the predicted LST image from STTFN; (e) the actual Landsat LST image. Roads were not totally captured in (a) and (b) (in red box) due to anomalous values in the gap-filled Landsat ETM+ SLC-off LST image (In gray to highlight the spatial and structure information).

E. Computation Efficiency

Using a 1300 pixels ×1300 pixels area extracted from the Landsat LST images as an example, the computational efficiency of the different fusion methods was compared. All experiments are carried out on the same computer equipped with an Intel Core i7-6700 processor with 3.4GHz and 24GB RAM, and a NVIDIA 1080Ti GPU with 11 GB of RAM. StfNet and STTFN were implemented using the TensorFlow framework with GPU acceleration, while ESTARFM and STITFM were tested under CPU mode. The comparison of computation efficiency for different methods is shown in Table VI. It is evident that ESTARFM was the most time-consuming method. The time cost of STTFN was close to STITFM and higher than for StfNet. For STTFN and StfNet, most computational time is spent on training, while StfNet is a shallow network with three layers, which requires less time for training. In general, CNN-based methods are more time-consuming than traditional methods because the training process is slow. However, STTFN employs the Adam algorithm is employed to update the parameters of the network, which has a much faster convergence speed than stochastic gradient descent (SGD) algorithm. Furthermore, batch normalization layers are used to overcome overfitting and speed up the training process. Thus, the total time of STTFN is less than that of ESTARFM and close to that of STITFM.

TABLE VI
COMPUTATION EFFICIENCY COMPARISON (IN SECONDS)

Methods	ESTARFM	STITFM	StfNet	STTFN
Training Time	-	-	78	170
Predicting Time	718	155	6	8
Total Time	718	155	84	178

IV. DISCUSSIONS

A SpatioTemporal Temperature Fusion Network (STTFN) was proposed to predict fine spatiotemporal resolution Landsat-like LST images from MODIS and Landsat observations. The effectiveness of the proposed network was assessed using data from two study areas: Évora, Portugal and Shengjin Lake reserve, China.

The enhanced performance of STTFN relative to the other methods arises primarily from the ability to accommodate nonlinear relations between input and output LST data and a specific weight function-based combination. STTFN is a multi-scale fusion CNN based method that achieves high level feature extraction and fusion at different levels. The slim-WDSR super resolves the temporal change image between two input coarse spatial resolution LST images, and, to provide fine spatial detail, during the super resolution analysis the temporally neighboring fine spatial resolution LST image is concatenated with the change image. This allows to represent fine spatial resolution change information. Some low-level features may be lost in the convolution process of CNN. To solve this problem, STTFN uses two local and one global residual learning processes which can retain shallow features. Therefore, STTFN can derive high- and low-level features at the same time. Finally, a key component of STTFN is the use of the STC-Weighting function. In previous spatiotemporal fusion methods [34, 36], the temporal changes between the coarse spatial resolution LST images at the neighboring and target dates are employed to calculate weight parameters. However, the STC-Weighting function also accommodates spatial consistencies between the two predicted fine spatial resolution images and the corresponding coarse image. Note that, despite the performance of STTFN was validated in two study areas here, it can be applied globally if sufficient training data set is available.

There are, however, some limitations to STTFN. First, the performance of STTFN greatly relies on the training samples, an inherent problem of CNN. Therefore, LST changes which are not contained in the training data may not necessarily be predicted accurately. The incorporation of additional inputs (e.g., *in-situ* LST) in combination with other models (e.g., radiation transfer model) may help to solve this problem. Second, since the overpass time of MODIS and Landsat is similar (as indicated in Table II, time differences between them are all within half an hour), the data was not corrected for possible inconsistencies, but the results from combination of forward and backward predictions would not be ideal if the MODIS and Landsat LST did not match at the target date. At present, time normalization method has been applied to eliminated the time inconsistent of MODIS LST product [52], we will try to integrate the time normalization method into CNN in further study. Additionally, there may be large discrepancies between MODIS and Landsat LSTs for some image pairs because they were retrieved by different algorithms. These discrepancies would have large negative impacts on the fusion results (e.g., the result on 21 June 2010). Therefore, the accuracy of LST product is an important factor for the accuracy of STTFN as well. Third, two pairs of temporally close LST images are required to train the network and predict the result, which is sometimes difficult to obtain due to cloud cover. This problem would be reduced by use of methods to fill missing values in remotely-sensed LST series [53]. Finally, some customized parameters in STTFN (e.g., cropping size and stride of training image patches, learning rate and decay rates of Adam) should be set in advance, they were set via experience and a series of experiments. Final experiment showed STTFN were able to generate promising results for different date sets with these set values, but there may be several better values for these parameters which can derive more accurate results, how to acquire the best values of these parameters is a hot-topic in deep learning and needs further studies.

V. CONCLUSION

The proposed STTFN was used successfully to generate fine spatial resolution LST image. STTFN specified the nonlinear relations between input and output LST based on an integration of features extracted at different levels and fusion through a specially designed convolution neural network. The main novelty of STTFN is that: 1) it more fully uses low- and high-level features of the input LSTs through use of residual learning unit and super-resolution module, which enables spatial information at different scales to be obtained, and by its enhanced capacity to accommodate potentially complicated nonlinear relations between input and output data; 2) it uses Huber Loss as the loss function in training which is robust to

outliers and, hence, yields enhanced outputs; and 3) it preserves the spatiotemporal consistency of LST images through STC-Weighting, which considerably enhances output quality. The proposed method was tested on actual data for two study areas, and compared with three classic fusion methods (ESTARFM, STITFM, and StfNet). The results indicated that STTFN has the ability to produce accurate and stable fine spatiotemporal resolution LST image. Moreover, it also has a good performance in terms of computational efficiency. STTFN is designed to yield more accurate LST products with fine spatiotemporal resolution, and therefore to support monitoring diurnal land-surface and ecological dynamics. Future improvements may include developing strategies to tackle missing value problem, analyzing network parameters, and introducing some physical models.

REFERENCES

[1] D. Eleftheriou *et al*., "Determination of annual and seasonal daytime and nighttime trends of MODIS LST over Greece- climate change implications," *Science of The Total Environment*, vol. 616, pp. 937-947, 2018.

[2] Q. Weng, "Thermal infrared remote sensing for urban climate and environmental studies: Methods, applications, and trends," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 64, no. 4, pp. 335-344, 2009.

[3] M. C. Anderson, J. M. Norman, W. P. Kustas, R. Houborg, P.J.Starks and N. Agam., "A thermal-based remote sensing technique for routine mapping of land-surface carbon, water and energy fluxes from field to regional scales," *Remote Sensing of Environment*, vol. 112, no. 12, pp. 4227-4241, 2008.

[4] J. D. Kalma, T. R. McVicar, and M. F. McCabe, "Estimating Land Surface Evaporation: A Review of Methods Using Remotely Sensed Surface Temperature Data," *Surveys in Geophysics*, vol. 29, no. 4, pp. 421-469, 2008.

[5] S. Elsayed, M. Elhoweity, H. Ibrahim, Y. Dewir, H. Migdadi and U. Schmidhalter, "Thermal imaging and passive reflectance sensing to estimate the water status and grain yield of wheat under different irrigation regimes," *Agricultural Water Management*, vol. 189, pp. 98-110, 2017.

[6] J. Guo *et al*., "Evaluation of the grain yield and nitrogen nutrient status of wheat (*Triticum aestivum L.*) using thermal imaging," *Field Crops Research*, vol. 196, pp. 463-472, 2016.

[7] M. Herrero-Huerta, S. Lagüela, S. Alfieri and M.Menenti, "Generating high-temporal and spatial resolution TIR image data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 79, pp. 149-162, 2019.

[8] Y. Zhou, Q. Weng, K. R. Gurney, Y. Shuai and X. Hu, "Estimation of the relationship between remotely sensed anthropogenic heat discharge and building energy use," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 67, pp. 65-72, 2012.

[9] W. Zhan *et al.*, "Disaggregation of remotely sensed land surface temperature: Literature survey, taxonomy, issues, and caveats," *Remote Sensing of Environment*, vol. 131, pp. 119–139, 2013.

[10] A. Dominguez, J. Kleissl, J. Luvall and D. Rickman, "High-resolution urban thermal sharpener (HUTS)," *Remote Sensing of Environment*, vol. 115, pp. 1772-1780, 2011.

[11] P. Wu, H. Shen, L. Zhang and F.M. Göttsche, "Integrated fusion of multi-scale polar-orbiting and geostationary satellite observations for the mapping of high spatial and temporal resolution land surface temperature," *Remote Sensing of Environment*, vol. 156, pp. 169-181, 2015.

[12] J. Kim and T. Hogue, "Evaluation and sensitivity testing of a coupled Landsat-MODIS downscaling method for land surface temperature and vegetation indices in semi-arid regions," *Journal of Applied Remote Sensing*, vol. 6, pp. 1-17, 2012.

[13] D. Liu, and R. Pu, "Downscaling Thermal Infrared Radiance for Subpixel Land Surface Temperature Retrieval," *Sensors*, vol. 8, no. 4, pp. 2695-2706, 2008.

[14] R. Mechri, C. Ottlé, O. Pannekoucke and A. Kallel, "Genetic particle filter application to land surface temperature downscaling," *Journal of Geophysical Research: Atmospheres*, vol. 119, no. 5, pp. 2131-2146, 2014.

[15] A. Kallel, C. Ottlé, S. L. Hegarat-Mascle, F. Maignan and D. Courault, "Surface Temperature Downscaling From Multiresolution Instruments Based on Markov Models," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 3, pp. 1588-1612, 2013.

[16] J. Quan, W. Zhan, T. Ma, Y. Du, Z. Guo and B. Qin, "An integrated model for generating hourly Landsat-like land surface temperatures over heterogeneous landscapes," *Remote Sensing of Environment*, vol. 206, pp. 403-423, 2018.

[17] H. Xia, Y. Chen, Y. Li and J. Quan, "Combining kernel-driven and fusion-based methods to generate daily high-spatial-resolution land surface temperatures," *Remote Sensing of Environment*, vol. 224, pp. 259-274, 2019.

[18] X. Zhu, F. Cai, J. Tian and T. Williams, "Spatiotemporal Fusion of Multisource Remote Sensing Data: Literature Survey, Taxonomy, Principles, Applications, and Future Directions," *Remote Sensing*, vol. 10, no. 4, pp. 527, 2018.

[19] J. Li, Y. Li, L. He, J. Chen and A.Plaza, "Spatio-temporal fusion for remote sensing data: an overview and new benchmhiark," *Science China Information Sciences*, vol. 63, no. 4, pp. 140301, 2020.

[20] X. Li, G. M. Foody, D. S. Boyd, Y. Ge, Y. Zhang, Y. Du and F.Ling, "SFSDAF: An enhanced FSDAF that incorporates sub-pixel class fraction change information for spatio-temporal image fusion," Remote Sensing of Environment, vol. 237, pp. 111537, 2020.

[21] 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.

[22] 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.

[23] X. Zhu, J. Chen, F. Gao, X. Chen and J. 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.

[24] X. Zhu, E. H. Helmer, F. Gao, D. Liu, J. Chen and M. Lefsky, "A flexible spatiotemporal method for fusing satellite images with different resolutions," *Remote Sensing of Environment*, vol. 172, pp. 165-177, 2016.

[25] H. Liu, and Q. Weng, "Enhancing temporal resolution of satellite imagery for public health studies: A case study of West Nile Virus outbreak in Los Angeles in 2007," *Remote Sensing of Environment*, vol. 117, pp. 57-71, 2012.

[26] Y. Ma *et al.*, "Estimation of daily evapotranspiration and irrigation water efficiency at a Landsat-like scale for an arid irrigation area using multi-source remote sensing data," *Remote Sensing of Environment*, vol. 216, pp. 715-734, 2018.

[27] B. Mohamadi, S. Chen, T. Balz, K. Gulshad and S.C. McClure, "Normalized Method for Land Surface Temperature Monitoring on Coastal Reclaimed Areas," *Sensors*, vol. 19, no. 22, pp. 4836, 2019.

[28] B. Huang, J. Wang, H. Song, D. Fu and K. Wong, "Generating High Spatiotemporal Resolution Land Surface Temperature for Urban Heat Island Monitoring," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 5, pp. 1011-1015, 2013.

[29] P. Wu, H. Shen, T. Ai and Y. Liu, "Land-surface temperature retrieval at high spatial and temporal resolutions based on multi-sensor fusion," *International Journal of Digital Earth*, vol. 6, no. sup1, pp. 113-133, 2013.

[30] Q. Weng, P. Fu and F. Gao, "Generating daily land surface temperature at Landsat resolution by fusing Landsat and MODIS data," *Remote Sensing of Environment*, vol. 145, pp. 55-67, 2014.

[31] J. Quan, W. Zhan, Y. Chen, M. Wang and J. Wang, "Time series decomposition of remotely sensed land surface temperature and investigation of trends and seasonal variations in surface urban heat islands," *Journal of Geophysical Research: Atmospheres*, vol. 121, no. 6, pp. 2638-2657, 2016.

[32] N. Upadhyaya and M. Dixit, "A Review: Relating Low Level Features to High Level Semantics in CBIR," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 9, no. 3, pp. 433-444, 2016.

[33] Q. Yuan *et al.*, "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sensing of Environment*, vol. 241, pp. 111716, 2020.

[34] H. Song, Q. Liu, G. Wang R. Hand 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

[35] Z. Tan, P. Yue, L. Di and J. Tang, "Deriving High Spatiotemporal Remote Sensing Images Using Deep Convolutional Network," *Remote Sensing*, vol. 10, no. 7, pp. 1066, 2018.

[36] 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.

[37] J. Li, G. Yuan and H. Fan, "Multifocus Image Fusion Using Wavelet-Domain-Based Deep CNN " *Computational Intelligence and Neuroscience*, vol. 2019, pp. 23, 2019.

[38] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *in Proceedings of the 32nd International Conference on Machine Learning*, pp. 448-456, 2015.

[39] F. Isikdogan, A. C. Bovik and P. Passalacqua, "Surface Water Mapping by Deep Learning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 11, pp. 4909-4918, 2017.

[40] N. Yokoya, "Texture-Guided Multisensor Superresolution for Remotely Sensed Images," *Remote Sensing*, vol. 9, no. 4, pp. 316, 2017.

[41] J. Yu *et al.*, "Wide Activation for Efficient and Accurate Image Super-Resolution," *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 2621-2624, 2018.

[42] Y. Zhang, Y. Tian, Y. Kong, B. Zhong and Y. Fu, "Residual Dense Network for Image Super-Resolution," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2472-2481, 2018.

[43] B. Lim, S. Son, H. Kim, S. Nah and K. Mu-Lee *et al.*, "Enhanced Deep Residual Networks for Single Image Super-Resolution," *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 1132-1140, 2017.

[44] P. J. Huber, "Robust Estimation of a Location Parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73-101, 1964.

[45] H. Leung and S. Haykin, "The complex backpropagation algorithm," *IEEE Transactions on Signal Processing*, vol. 39, no. 9, pp. 2101-2104, 1991.

[46] Y. Yang, A. Zhan, L. Cao, F. Meng and W. Xu, "Selection of a marker gene to construct a reference library for wetland plants, and the application of metabarcoding to analyze the diet of wintering herbivorous waterbirds," *PeerJ*, vol. 4, pp. e2345, 2016.

[47] S.-B. Duan *et al.*, "Validation of Collection 6 MODIS land surface temperature product using in situ measurements," *Remote Sensing of Environment*, vol. 225, pp. 16-29, 2019.

[48] S.-B. Duan, Z.-L. Li, H. Wu, P. Leng, M. Gao and C. Wang, "Radiance-based validation of land surface temperature products derived from Collection 6 MODIS thermal infrared data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 70, pp. 84-92, 2018.

[49] J. C. Jiménez-Muñoz, J. A. Sobrino, D. Skoković, C. Mattar and J. Cristóbal, "Land Surface Temperature Retrieval Methods From Landsat-8 Thermal Infrared Sensor Data," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 10, pp. 1840-1843, 2014.

[50] M. A. Friedl *et al.*, "Global land cover mapping from MODIS: algorithms and early results," *Remote Sensing of Environment*, vol. 83, no. 1, pp. 287-302, 2002.

[51] F. Gao *et al.*, "Toward mapping crop progress at field scales through fusion of Landsat and MODIS imagery," *Remote Sensing of Environment*, vol. 188, pp. 9-25, 2017.

[52] W. Zhao, H. Wu, G. Yin and S.-B. Duan, "Normalization of the temporal effect on the MODIS land surface temperature product using random forest regression," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 152, pp. 109-118, 2019.

[53] P. Wu, Z. Yin, H. Yang, Y. Wu and X. Ma, "Reconstructing Geostationary Satellite Land Surface Temperature Imagery Based on a Multiscale Feature Connected Convolutional Neural Network," *Remote Sensing*, vol. 11, no. 3, pp. 300, 2019.