Journal of Remote Sensing
A SCIENCE PARTNER JOURNAL

RESEARCH ARTICLE

RES-STF: Spatio temporal Fusion of Visible Infrared Imaging Radiometer Suite and Landsat Land Surface Temperature Based on Restormer

Qunming Wang* **and Ruijie Huang**

College of Surveying and Geo-Informatics, Tongji University, Shanghai 200092, China

*Address correspondence to: wqm11111@126.com

Citation: Wang Q, Huang R. RES-STF: Spatio temporal Fusion of Visible Infrared Imaging Radiometer Suite and Landsat Land Surface Temperature Based on Restormer. *J. Remote Sens.* 2024;4:Article 0208. https://doi.org/10.34133/remotesensing.0208

Submitted 10 April 2024
Accepted 11 July 2024
Published 21 August 2024

Fine spatial and temporal resolution land surface temperature (LST) data are of great importance for various researches and applications. Spatio-temporal fusion provides an important solution to obtain fine spatio-temporal resolution LST. For example, 100-m, daily LST data can be created by fusing 1-km, daily Moderate Resolution Imaging Spectroradiometer (MODIS) LST with 100-m, 16-day Landsat LST data. However, the quality of MODIS LST products has been decreasing noticeably in recent years, which has a great impact on fusion accuracy. To address this issue, this paper proposes to use Visible Infrared Imaging Radiometer Suite (VIIRS) LST to replace MODIS LST in spatio-temporal fusion. Meanwhile, to cope with the data discrepancy caused by the large difference in overpass time between VIIRS LST and Landsat LST, a spatio-temporal fusion method based on the Restormer (RES-STF) is proposed. Specifically, to effectively model the differences between the 2 types of data, RES-STF uses Transformer modules in Restormer, which combines the advantages of convolutional neural networks (CNN) and Transformer to effectively capture both local and global context in images. In addition, the calculation of self-attention is re-designed by concatenating CNN to increase the efficiency of feature extraction. Experimental results on 3 areas validated the effectiveness of RES-STF, which outperforms one non-deep learning- and 3 deep learning-based spatio-temporal fusion methods. Moreover, compared to MODIS LST, VIIRS LST data contain richer spatial texture information, leading to more accurate fusion results, with both RMSE and MAE reduced by about 0.5 K.

Introduction

Land surface temperature (LST) is an important factor in monitoring global change, serving as a vital part in researches such as global climate monitoring [1], crop evapotranspiration estimation [2], and urban heat island monitoring [3]. LST data can be acquired from thermal infrared (TIR) remote sensing images collected by various sensors, such as Landsat TM/ETM+/TIRS [4], Moderate Resolution Imaging Spectroradiometer (MODIS) [5], Advanced Very High-Resolution Radiometer (AVHRR) [6], Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) [7], and Visible Infrared Imaging Radiometer Suite (VIIRS) [8]. In practical applications, high spatio-temporal resolution LST data play a crucial role in guiding agricultural practices, irrigation, and drainage [9], as well as in studying urban heat environment changes [10,11]. However, due to limitations in hardware design of remote sensing sensors, trade-off always exists between temporal and spatial resolution in satellite-derived LST data.

Spatio-temporal fusion provides an important solution for obtaining high spatio-temporal resolution LST data. Currently, a number of spatio-temporal fusion methods have been developed, which can be classified into 4 main types [12]: spatial weighting methods, spatial unmixing methods, hybrid methods, and learning-based methods. The earliest spatial weighting method is the spatial and temporal adaptive reflectance fusion model (STARFM) [13], which integrates spectral, temporal, and spatial information of neighboring similar pixels to estimate high spatial resolution change information between different times. Zhu et al. [14] introduced an enhanced version, ESTARFM, which uses adaptive transformation coefficients for heterogeneous regions to modify the weights of neighboring similar pixels. Algorithms like Fit-FC [15] and Agri-Fuse [16] were also proposed to address seasonal changes and diverse phenological changes.

The multisensor multiresolution technique presented by Zhukov et al. [17] is one of the earliest spatial unmixing-based spatio-temporal fusion methods. The coarse proportions are produced from the fine spatial resolution image at the known time, assuming that there are no abrupt land cover changes between the known and prediction dates. Niu [18] introduced the spatial temporal data fusion approach (STDFA) method, which calculates the temporal changes of reflectance for each category by unmixing the temporal difference image at coarse spatial resolution. To reduce the feature differences between the known and predicted images, Wang et al. [19] introduced the concept of virtual image pairs and proposed VIPSTF-SU. Peng et al. [20] proposed a geographically weighted spatial unmixing (SU-GW) strategy to address spatial variation in land cover, effectively enhancing 12 existing methods. Xu et al. [21]

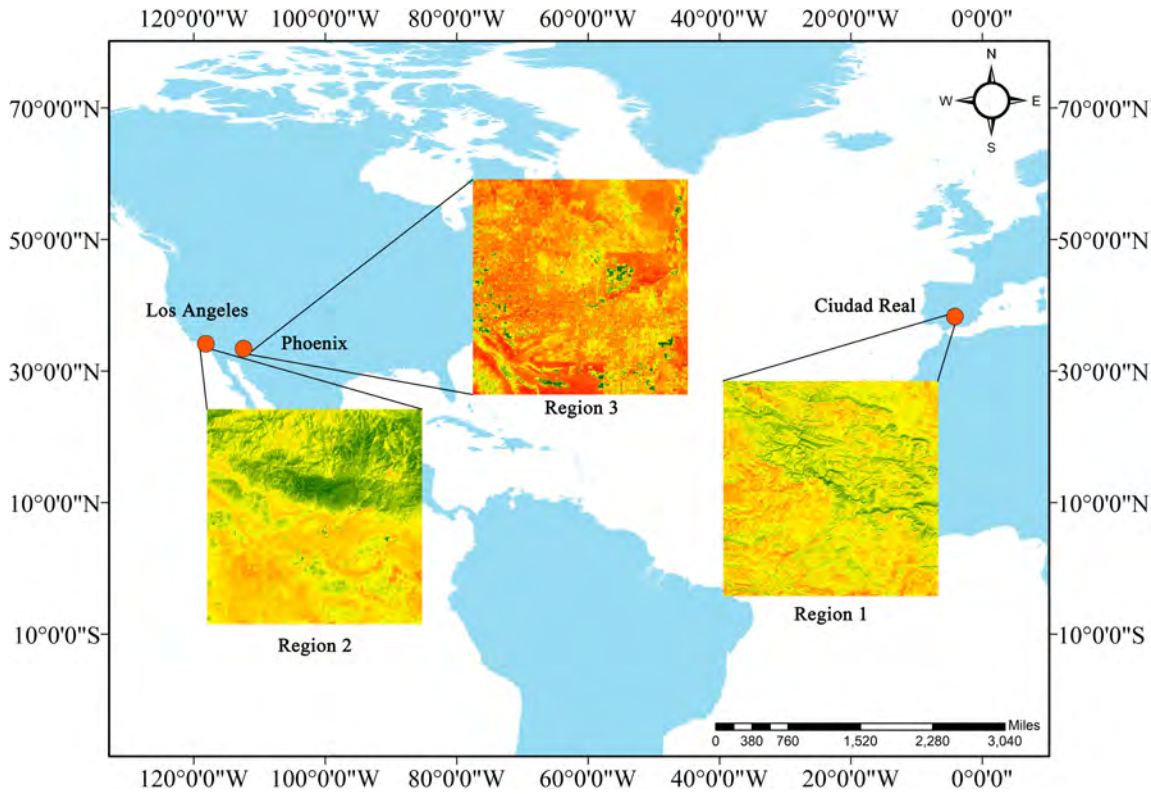

Fig. 1. The selected 3 study areas.

proposed a variation-based spatio-temporal data fusion algorithm, which quantifies the spectral difference between the fine and coarse spatial resolution datasets.

Hybrid methods combine the advantages of spatial weighting and spatial unmixing. FSDAF [22] integrated the basic ideas of spatial unmixing and STARFM into a single framework. Li et al. [23] introduced sub-pixel land cover change information into the FSDAF and proposed SFSDAF to address changes occurring in strong heterogeneous regions. FSDAF 2.0 [24] employs change detection algorithms to better handle pixels undergoing land cover changes.

Learning-based methods construct nonlinear relationships between images with different resolutions [25,26]. Song et al. [27] used deep convolutional neural networks (CNNs) to map the nonlinear relationship between Landsat and MODIS images. Liu et al. [28] considered temporal dependence among image sequences and proposed a 2-stream CNN (StfNet). Qin et al. [29] used multiscale features to overcome geometric registration errors between different resolution images. Zhang et al. [30] designed an enhanced cross-paired wavelet-based spatio-temporal fusion networks (ECPW-STFN) with fewer inputs. Generative Adversarial Network (GAN)-based methods [31], such as conditional-GAN [32], were proposed to overcome the challenge of selecting reference images in spatio-temporal fusion. Additionally, Transformer, which has the advantage of extracting long-distance global features, was also applied to spatio-temporal fusion [33].

Although spatio-temporal fusion methods were initially proposed for reflectance data, most of the principles are generally applicable to other environmental variables such as normalized difference vegetation index (NDVI), leaf area index (LAI), and LST [34]. To further enhance the performance of spatio-temporal fusion of LST data, it is important to model the dependency between the LST images with different spatial and temporal resolutions, and several attempts have been made based on the character of LST data. For example, Huang et al. [35] considered the pixel correlation of LST and used bilateral filtering to improve the weight function in STARFM. Weng et al. [36] proposed a spatio-temporal adaptive data fusion algorithm for temperature mapping by considering annual temperature cycle and the heterogeneity of urban LST landscapes. Quan et al. [37] combined the annual temperature cycle and diurnal temperature cycle models to characterize nonlinear temporal patterns of LST. To fuse the LST from different sensors, Wu et al. [38] harmonized the LST data with different spatial resolutions on the same date by a correction process.

Generally, to obtain high spatio-temporal resolution (e.g., 100 m, daily) LST data, the common solution is to fuse 1-km, daily MODIS data with 100-m, 16-day Landsat LST data. However, the quality of MODIS data has deteriorated after the long operation of over 20 years [39,40]. The presence of data noise and smooth effect in MODIS data cannot provide enough spatial information, thus severely limiting its applicability [41]. Therefore, the selection of suitable alternative data for spatio-temporal fusion of LST is a crucial issue.

The Suomi National Polar-orbiting Partnership (SNPP) satellite, launched in October 2011, routinely provides VIIRS data. As a next-generation satellite data product, its overall goal is to aid in the development of long-term climate data records [42]. The VIIRS data present similar spatio-temporal resolution to MODIS data, but generally with higher data quality [43]. Based on the advantage in data quality, this paper considered using VIIRS LST as alternative to MODIS LST in spatio-temporal

fusion with Landsat LST. However, the fusion of VIIRS LST with Landsat LST faces a new challenge due to their noticeable difference in overpass time. Specifically, the overpass time of VIIRS and Landsat is about 1:30 PM and 10:00 AM, respectively, leading to substantial difference between VIIRS and Landsat LST.

The uncertainty caused by the difference in acquisition times between VIIRS and Landsat LST images is one of the main challenges in spatio-temporal fusion. The methods to deal with this issue can be broadly classified into 2 categories: traditional methods and deep learning-based methods. The traditional methods adopt auxiliary data to harmonize multi-source data on the same day. For example, the intermediate spatial resolution data were introduced in [38]. Similarly, Ma et al. [44] employed temporally frequent assimilation data to perform temporal normalization to reduce the difference between multi-source LST on the same day. However, the need of the auxiliary data in the fusion process decreases the applicability in some cases, especially when such data are difficult to collect. For deep learning-based methods, they have great potential in characterizing the relation between multi-source data more explicitly through nonlinear models, such as the spatio-temporal temperature fusion network (STTFN) [45]. This is different from traditional methods that are typically performed based on simple linear transformation and exhibit notable limitations when great differences exist. However, the scheme of CNN suffers from the lack of global perception in images and it may fail to characterize the spatial heterogenous LST. Due to the parameter sharing mechanism, the performance of CNN can be compromised for construction of spatial details, especially when involving a large zoom factor for LST images with significant temporal and spatial variations.

Currently, visual Transformer models long-range dependencies in images by self-attention mechanism to solve the limitation in lack of global context awareness in CNN [46–48]. This makes it more suitable for downscaling involving large zoom factors. However, Transformer also suffers from the problem of large number of parameters, leading to heavy computational loads. Several studies have attempted to combine CNN with Transformer to design hybrid models [49–52]. Among them, Restormer is an effective network concatenating CNN and Transformer [53], which integrates the advantages of both models in its core module. Specifically, it can capture global context in the fine spatial resolution images based on the advantages of Transformer and can also capture local context based on the advantages of CNN. Moreover, by integration of both, Restormer is computationally more efficient. This makes Restormer suitable to extract high-level differential features, including temporal difference and sensor bias between multi-source data. Thus, Restormer has great potential for spatio-temporal fusion of LST images with great differences between multi-source data.

This paper proposes a Restormer-based spatio-temporal fusion network (RES-STF) for fusion of VIIRS LST and Landsat LST. The main contributions of this paper are as follows:

1. To create 100-m, daily LST, this paper proposes to use VIIRS LST instead of MODIS LST data as input in spatio-temporal fusion. In existing spatio-temporal fusion, MODIS LST is commonly used as coarse input data. Considering the decrease in data quality of MODIS LST, the strategy of using VIIRS LST is proposed as an alternative, which presents greater quality and is expected to lead to more accurate fusion results.

2. To cope with the challenges brought by different overpass time when fusing VIIRS and Landsat LST, the RES-STF method is proposed. The core part of RES-STF is based on Restormer that can better explore high-level differential features, including temporal difference (between coarse VIIRS LST time-series) and sensor bias (between VIIRS and Landsat LST), to predict the temporal differences at fine spatial resolution. By combining the advantages of CNN in extracting multiscale information and Transformer in capturing long-range dependencies, the complex relationship between multi-scale and multi-temporal LST images can be characterized more reliably.

The rest of this paper is organized as follows. The "Study Area and Data" section lists the study areas and data used in this paper. The "Methods" section introduces the network structure of the proposed RES-STF method. In the "Experiments" section, experimental results are presented for validation of the RES-STF. The "Discussion" section further discusses the advantages of using VIIRS LST and the RES-STF method, and also opening issues for future research. Conclusions are drawn in the "Conclusion" section.

Study Area and Data

Study area

Three different types of regions were selected for study: Ciudad Real suburban areas in Spain (Region 1) and Phoenix (Region 2) and Los Angeles (Region 3) in the United States. The location of the study areas is exhibited in Fig. 1. All 3 areas cover the same spatial extent of 18 km × 18 km and exhibit strong spatial heterogeneity in temperature distribution. Additionally, land cover types vary across the 3 regions, leading to significant differences in texture information in the LST images. Specifically, Region 1 mainly covers scattered forests, farmland, and bare land. Region 2 is situated within a large urban center characterized by intricate spatial textures. Region 3 covers both urban and forested areas. These selected regions effectively represent most of the spatial heterogeneous scenarios in real world.

Data

Table 1 lists the information of the LST data used in this paper. The selected fine spatial resolution data are Landsat 8 LST, and the coarse data are VIIRS LST and MODIS LST data. It is seen that the overpass time of VIIRS and Landsat is quite different. This paper will further compare the performances of using VIIRS and MODIS LST as input in spatio-temporal fusion.

Landsat dataset

The Landsat 8 satellite was launched on 2013 February 12, and its TIR sensor provides LST data at 100 m. The satellite

Table1. Information on the 3 types of LST data used in this study

	Landsat 8	VIIRS	MODIS
Spatial resolution	100 m	750 m	1 km
Temporal resolution	16 days	1 day	1 day
Overpass time (equator)	10:00 AM	1:30 PM	10:30 AM (Terra)

passes the equator at approximately 10.00 am. The selected Landsat 8 LST data are level-2 temperature data product from the Collection 2 dataset, which used the single-channel algorithm to retrieve LST. It was downloaded from https://earthexplorer.usgs.gov/.

MODIS dataset

MODIS is operating on the Terra and Aqua satellite and has been on orbit over 2 decades. It provides LST data at 1-km spatial resolution. We selected the MOD11A1 product in this paper, which has the similar overpass time with Landsat. MOD11A1 is a daily surface temperature product (version 6) retrieved by the generalized split-window algorithm under clear sky.

VIIRS dataset

VIIRS onboard the SNPP satellite was launched on 2011 as a mission to promote the continuity of the Earth Observation System (EOS) mission. Compared with MODIS data, VIIRS has a larger bandwidth and, thus, there is almost no gap near the equator [54]. The overpass time of the satellite is 1:30 PM. The spatial resolution of VIIRS LST is 750 m, but it should be noted that the spatial resolution of the released VIIRS LST data (i.e., VNP21A1D product) is 1 km. This product uses the Temperature Emissivity Separation (TES) algorithm to retrieve LST. Research shows that the accuracy of VNP21A1D is around 1 Kelvin (K) on all surface types including vegetation, water, and desert [55]. The data source of both VIIRS LST and MODIS LST is https://ladsweb.modaps.eosdis.nasa.gov.

The detailed acquisition dates of all the LST images used in this paper are presented in Table 2. Note that the LST images of the 3 datasets were ensured to be acquired on the same dates.

Table 2. Acquisition dates of the used LST data in the 3 regions (presented by Julian date of the year)

Region 1 (Ciudad Real, Spain)	Region 2 (Phoenix, USA)	Region 3 (Los Angeles, USA)
2013166	2014152	2013153
2013342	2014296	2013169
2014121	2015043	2013329
2014297	2015107	2014076
2015028	2015299	2014156
2015316	2017288	2014172
2016159	2018259	2014284
2016319	2018291	2015111
2017097	2019182	2017228
2017193	2019278	
2017337		

Methods

Overview of RES-STF

Fig. 2 presents the architecture of RES-STF. For fusion of data from different sources, we focus on 2 types of differences: the sensor bias between coarse and fine spatial resolution images, and the temporal differences between coarse images. RES-STF aims to overcome the impact by these 2 types of differences in spatio-temporal fusion.

Specifically, in this paper, the input data for the model consist of coarse-fine image pairs at t_1 and t_3, along with the coarse data at t_2. The model predicts the fine spatial data at t_2. Here, coarse images are interpolated to match the spatial resolution of fine spatial resolution images in advance.

During the training process, 3 coarse-fine image pairs at another 3 times (denoted as T_1, T_2, and T_3), but covering the same area, are required. Accordingly, there are 2 coarse difference images ($\mathbf{C}_{12}$ and $\mathbf{C}_{23}$) and 2 fine difference images ($\mathbf{F}_{12}$ and $\mathbf{F}_{23}$)

$$\mathbf{C}_{12} = \mathbf{C}_2 - \mathbf{C}_1 \tag{1}$$

$$\mathbf{C}_{23} = \mathbf{C}_3 - \mathbf{C}_2 \tag{2}$$

$$\mathbf{F}_{12} = \mathbf{F}_2 - \mathbf{F}_1 \tag{3}$$

$$\mathbf{F}_{23} = \mathbf{F}_3 - \mathbf{F}_2 \tag{4}$$

We need to train the relation between coarse and fine difference images (i.e., $\mathbf{C}_{12}$ and $\mathbf{F}_{12}$, and $\mathbf{C}_{23}$ and $\mathbf{F}_{23}$). In this paper, we define the mapping processes between $\mathbf{C}_{12}$ and $\mathbf{F}_{12}$ as Forward training, and the mapping processes between $\mathbf{C}_{23}$ and $\mathbf{F}_{23}$ as Backward training.

The process of recovering fine spatial resolution images from coarse spatial resolution images can be viewed as an image super-resolution process. The fine spatial resolution images at T_1 and T_2 provide abundant texture information, serving as supplementary features for this process. Simultaneously, to deal with sensor differences in spatio-temporal fusion, this paper introduces the differences between fine and coarse spatial resolution images (i.e., sensor bias) to the network:

$$\Delta\mathbf{CF}_1 = \mathbf{C}_1 - \mathbf{F}_1 \tag{5}$$

$$\Delta\mathbf{CF}_3 = \mathbf{C}_3 - \mathbf{F}_3 \tag{6}$$

With the temporal difference features added into the network, the Forward and Backward training processes are optimized together. Finally, the fine spatial resolution image at T_2 is obtained upon a temporal weighting function:

$$\hat{\mathbf{F}}_2 = f\left(\mathbf{C}_{12}, \Delta\mathbf{CF}_1\right) + g\left(\mathbf{C}_{23}, \Delta\mathbf{CF}_3\right) \tag{7}$$

where f and g stand for Forward training and Backward training process, respectively.

Network architecture

As illustrated in Fig. 3, taking Forward training as an example, the proposed network can be divided into 3 major modules. The first stage is Globe Residual Learning, which uses information directly from the input images. The second stage is Multi-scale Bias Learning, which learns 2 types of deep bias features at different scales. Finally, Final Learning module aggregates information comprehensively from above.

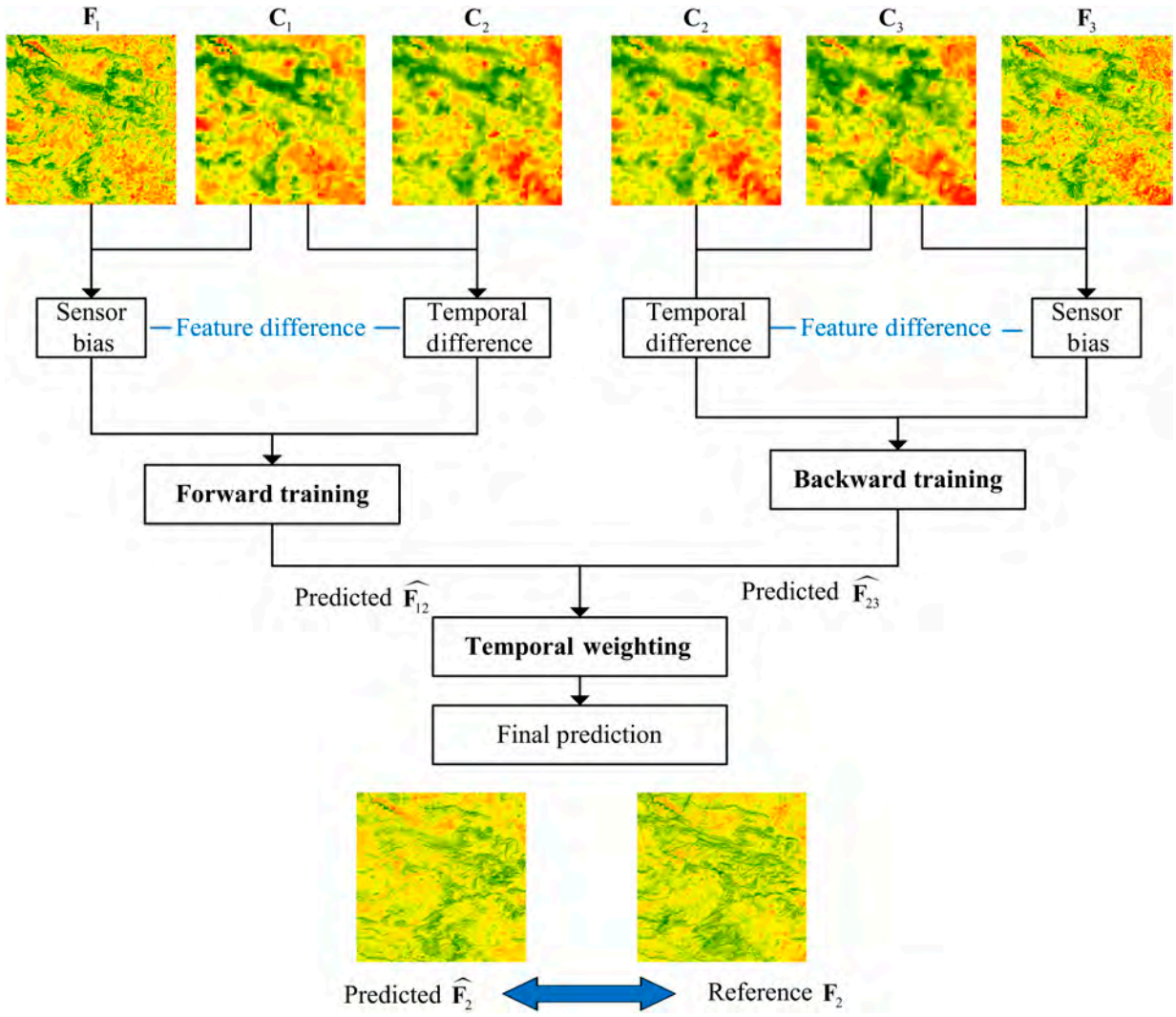

Fig. 2. Flowchart of RES-STF.

The detailed design of each module is as follows. Initially, 2 types of difference images (representing sensor bias and temporal difference) are produced, and high-dimensional features are extracted through patch embedding. Subsequently, 2 streams for the 2 difference images are concatenated and input into the Multi-scale Bias Learning module. This module is designed to extract features at different scales and employs skip connections strategy to enhance the interrelation of features. Simultaneously, in the Globe Residual Learning module, a 1×1 convolution is introduced to mitigate gradient explosion, serving as global learning to obtain the output of the module in high dimensions. Finally, the information extracted at multiple scales, coupled with the global residual, are fed into the Refinement Block (RB) for fine-tuning, yielding the ultimate temporal difference LST image at fine spatial resolution.

Multi-scale bias learning module

As indicated in Fig. 3, the Multi-scale Bias Learning module is the core part of RES-STF. The detailed design is introduced in this part. Recovering fine spatial resolution images from severely blurred counterparts introduces notable uncertainty, and conventional Transformers, with their extensive parameterization and time-intensive training, face limitations in this case. To address this issue, we incorporated the concept of a lightweight network Restormer into our design to extract multi-scale features. Simultaneously, to enhance the efficiency of extracting both global and local information, this network integrates CNN and Transformer modules. The Transformer block for multi-scale extraction adopts the Transformer encoder–decoder architecture, which transfers shallow features through a 4-level symmetric encoder-decoder to extract deep features. Each encoder and decoder contain multiple Transformer modules. In the down-sampling and up-sampling stages, we employed pixel shuffle and unshuffle operations, respectively. To mitigate gradient explosion and network degradation, the strategy of skip connection is implemented to concatenate the features from the encoder with those from the decoder.

Specifically, each Transformer module can be divided into 2 parts: a Multi-Dconv Head Transposed Nonlinear (MDTA) module and a Gated-Dconv Feed-Forward Network module (GDFN). MDTA module introduces convolution operations before the calculation of multi-head self-attention by 1×1 and 3×3 convolutions to aggregate high-dimensional spatial features. GDFN is a convolutional modification of feedback network, and it contains 2 convolution-based branches. We

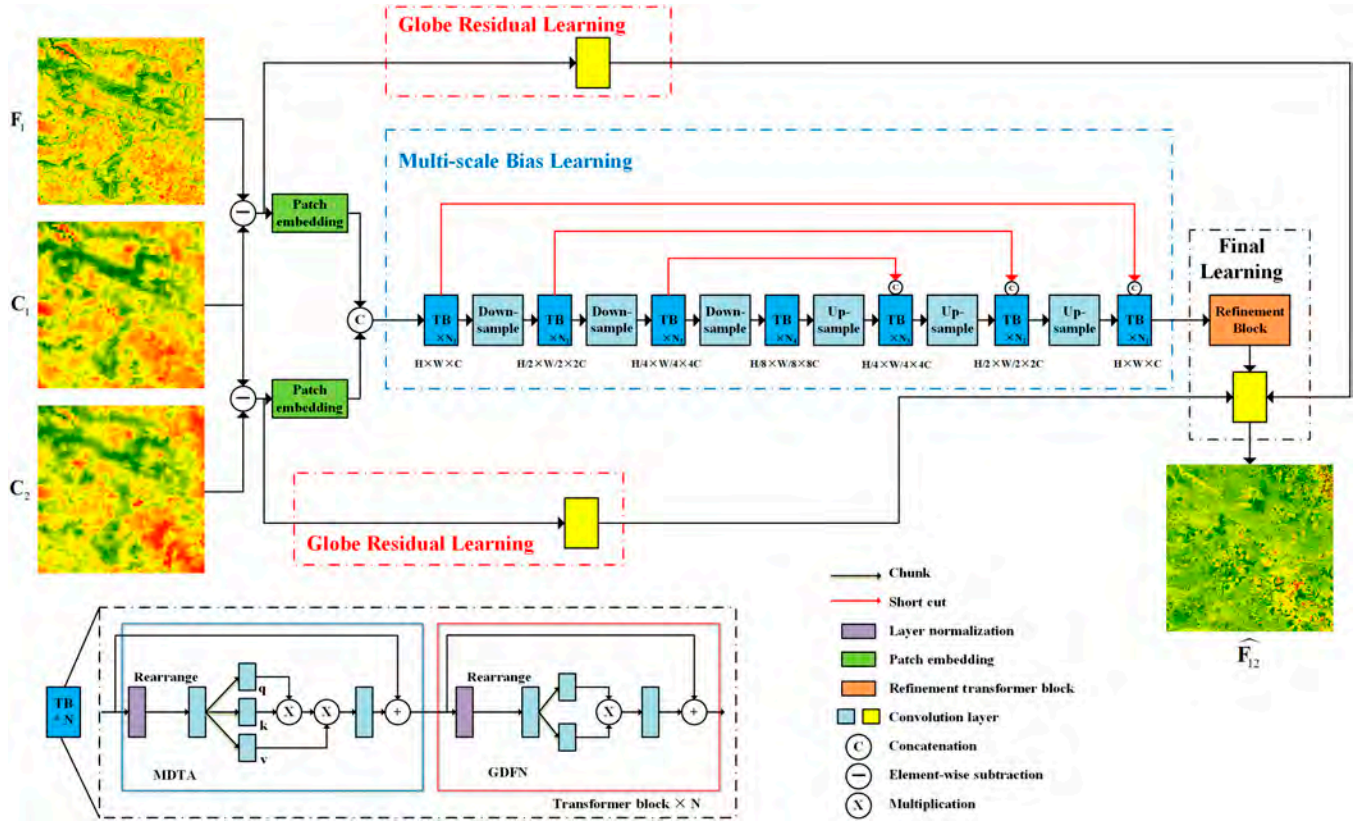

Fig. 3. Detailed structure of RES-STF (illustrating forward training as an example; backward training is the same except the input and out images).

performed it in 2 parallel paths and then used an element-wise product, of which the GELU nonlinear activation function is used. Finally, a depth-wise convolutional layer is used to restore the local structural information of the image. These 2 modules reduce the computational complexity and can efficiently expand high-dimensional feature information.

The network concludes with a comprehensive restoration process involving multi-scale sampled information and global residual information. To preserve detailed texture information, Transformer modules are re-used for feature integration. In the end, a convolutional layer is applied to obtain the predicted fine spatial resolution difference image.

Loss function

The proposed RES-STF chooses L2 loss as loss function. To reduce the consumption of memory, we took the strategy of patching images to train our network.

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\mathbf{F}}_i - \mathbf{F}_i\right)^2 \tag{8}$$

To increase training efficiency and achieve global optimal results, Forward training and Backward training are conducted simultaneously. Therefore, the final loss function can be divided into 3 parts, namely, the difference predicted between the first 2 dates, the difference predicted between the last 2 dates, and the fine spatial resolution image on the prediction date through temporal weighting.

$$\begin{aligned}\mathcal{L} = {} & \mathcal{L}\left(f\left(\mathbf{C}_{12}, \Delta\mathbf{CF}_1\right), \mathbf{F}_{12}\right) \\ & + \mathcal{L}\left(g\left(\mathbf{C}_{23}, \Delta\mathbf{CF}_3\right), \mathbf{F}_{23}\right) + \mathcal{L}\left(f+g, \mathbf{F}_2\right)\end{aligned} \tag{9}$$

Temporal weighting

As we predict the 2 difference images based on the Forward and Backward parts in the network, we need to make full use of both predictions to reconstruct the final $\mathbf{F}_2$. Accordingly, a temporal weighting method is employed in RES-STF:

$$\hat{\mathbf{F}}_2 = \alpha \cdot \left(\mathbf{F}_1 + \hat{\mathbf{F}}_{12}\right) + (1-\alpha) \cdot \left(\mathbf{F}_3 - \hat{\mathbf{F}}_{23}\right) \tag{10}$$

where α is the weighting factor. If the change between 2 coarse spatial resolution images $\mathbf{C}_2$ and $\mathbf{C}_k$ ($k = 1$ or 3) is smaller, then $\mathbf{F}_2$ may be more similar to the adjacent fine spatial resolution difference image $\mathbf{F}_k$. That is, we should give more emphasis to $\mathbf{F}_k$. Thus, the coefficient α can be defined by:

$$\alpha = \frac{1/\mathbf{C}_{12}}{1/\mathbf{C}_{12} + 1/\mathbf{C}_{23}} \tag{11}$$

Considering the spatial heterogeneity in the LST image, the weight is calculated based on the data in the window size of 3×3 fine spatial resolution pixels.

Experiments

Evaluation metrics

For all the study areas, we compared the predicted 100-m Landsat-Like LST with the reference Landsat LST. To show the advantage of RES-STF, we compared it with 4 spatio-temporal fusion algorithms, including 3 deep learning-based methods (StfNet, STTFN, and ECPW-STFN) and one traditional method (ESTARFM). Three metrics, including correlation coefficient (CC), root mean square error (RMSE), and mean absolute error (MAE) were chosen for accuracy evaluation [56].

CC is an objective evaluation indicator that reflects the correlation between the predicted and reference images. The ideal value is 1. The closer CC is to 1, the more similar the 2 images are. CC can be defined by the following equation:

$$\mathrm{CC} = \frac{\sum_{i=1}^{n}\left(\mathbf{F}_i - \mu_{\mathbf{F}}\right)\left(\hat{\mathbf{F}}_i - \mu_{\hat{\mathbf{F}}}\right)}{\sqrt{\sum_{i=1}^{n}\left(\mathbf{F}_i - \mu_{\mathbf{F}}\right)^2}\sqrt{\sum_{i=1}^{n}\left(\hat{\mathbf{F}}_i - \mu_{\hat{\mathbf{F}}}\right)^2}} \quad (12)$$

where $\hat{\mathbf{F}}$ is the predicted image, $\mathbf{F}$ is the reference image, n is the number of pixels in the image, and $\mu_{\hat{\mathbf{F}}}$ and μF are the mean value of $\hat{\mathbf{F}}$ and $\mathbf{F}$, respectively.

RMSE is a typical indicator for evaluating errors. The larger the error, the greater the RMSE:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{F}_i - \hat{\mathbf{F}}_i\right)^2} \quad (13)$$

MAE is used to measure the average absolute error between the predicted and reference images:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\mathbf{F}_i - \hat{\mathbf{F}}_i\right| \quad (14)$$

Experimental setup

The experimental data in this study are divided into training and prediction sets. Each group of data consists of 2 pairs of coarse-fine spatial resolution images on 2 known dates, and a coarse spatial resolution image and a reference fine spatial resolution image on the prediction date. Images in each pair are ensured to be acquired on the same day. It is worth noting that for each study area, training sets and predicting sets both cover the same region, but the dates are different. Before inputting to the model, all the coarse images were resampled to the spatial size of Landsat LST data (i.e., 600 × 600 pixels) with the same geographical reference.

For prediction, we selected 4 dates for each region, as illustrated in Figs. 4 to 6. In each data group, t_2 represents the prediction date. The accuracy of the predicted Landsat-like LST was evaluated by comparing it to the reference Landsat LST image at t_2.

All the deep learning-based methods were implemented under the PyTorch framework, and the NVIDIA GeForce RTX 2080 Ti GPU was used to accelerate the training process. To optimize

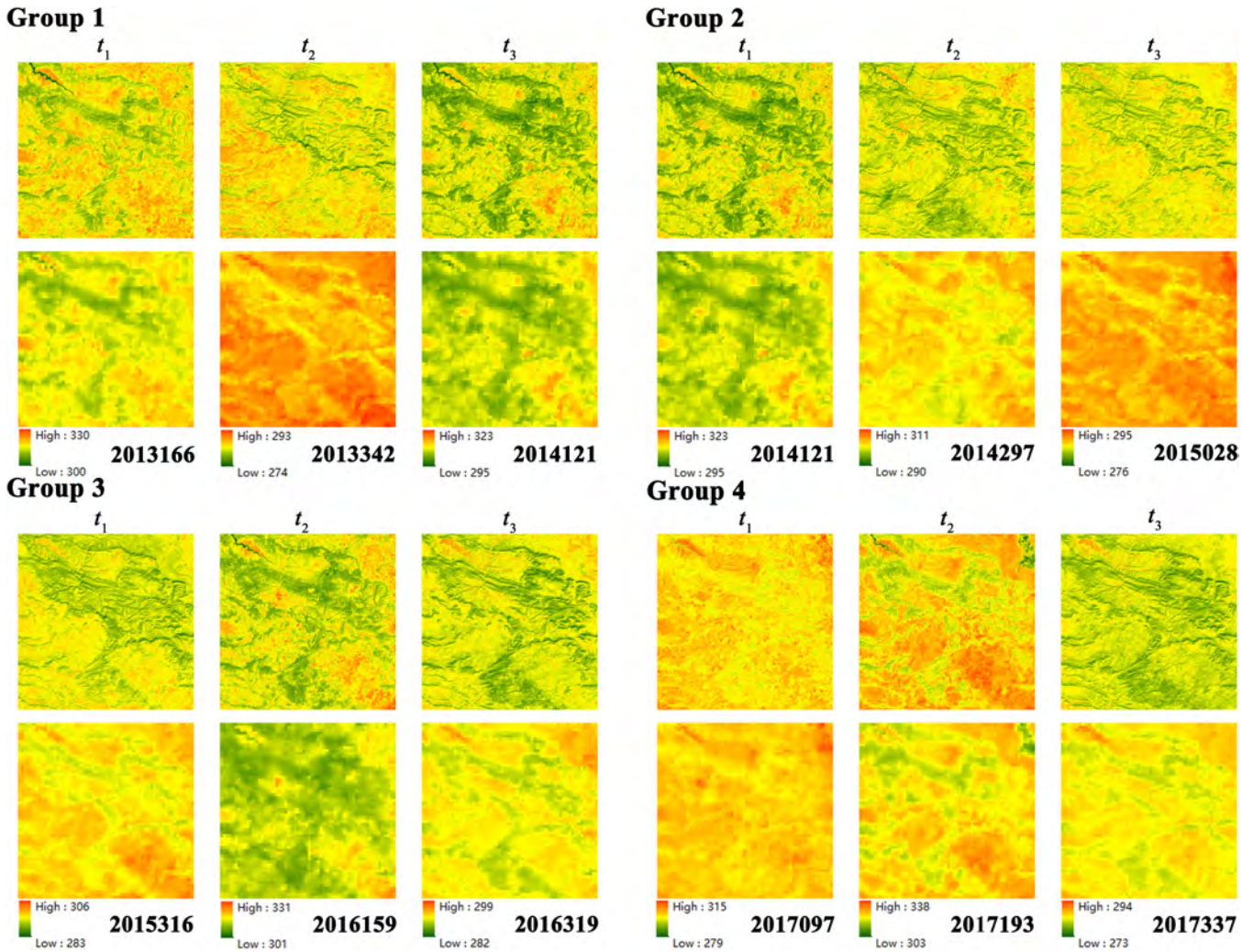

Fig. 4. Data used for prediction in Region 1 (the first and second lines in each group represent Landsat and VIIRS LST images, respectively; the Landsat LST image at t_2 in each group was predicted).

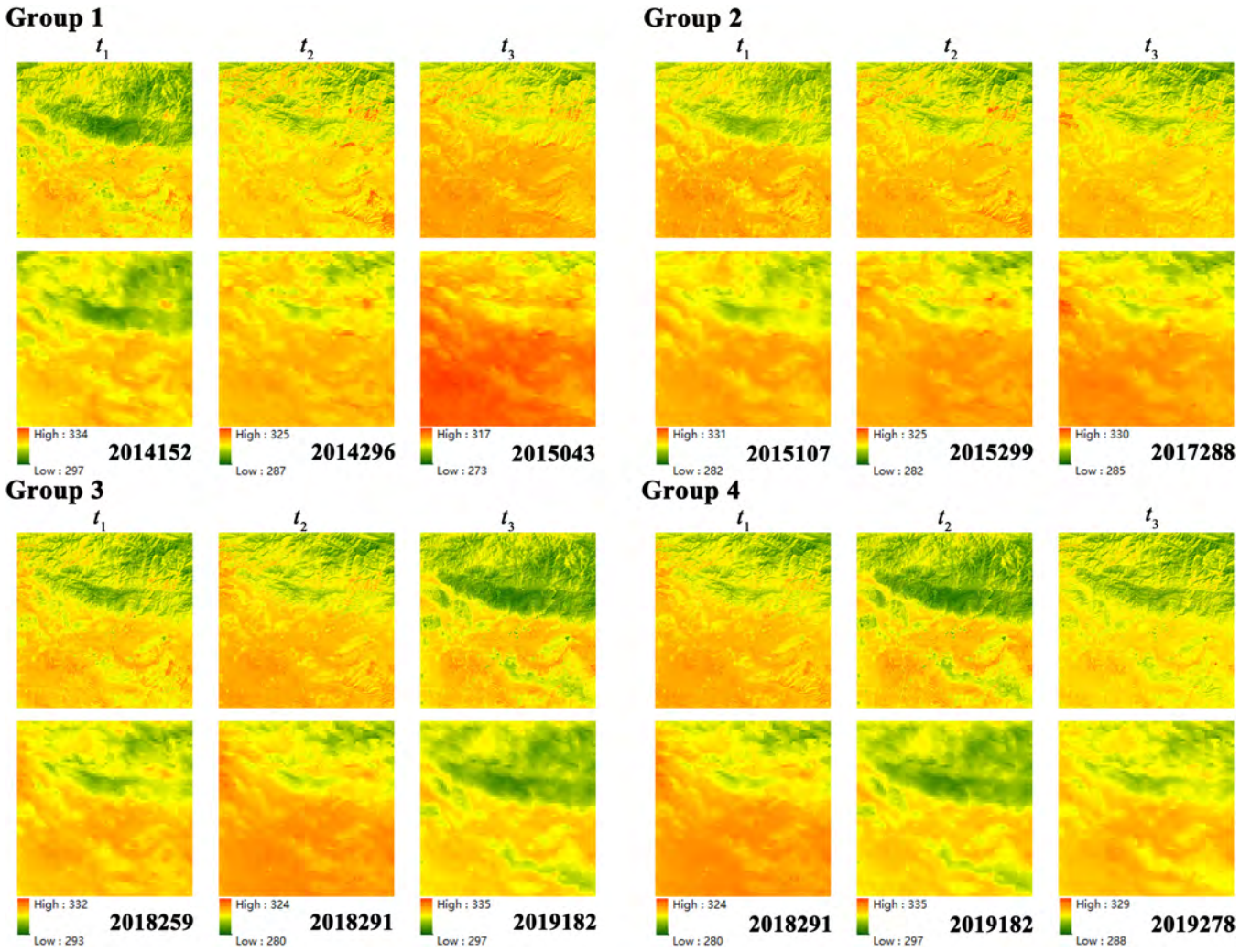

Fig. 5. Data used for prediction in Region 2 (the first and second lines in each group represent Landsat and VIIRS LST images, respectively; the Landsat LST image at t_2 in each group was predicted).

memory usage and improve training efficiency, the images were clipped into blocks with a spatial size of 48 × 48 pixels.

Validation of the RES-STF method

Qualitative evaluation

Figs. 7 to 9 show the results of fusion of VIIRS and Landsat based on the proposed RES-STF method and the 4 benchmark methods. Overall, benefiting from the construction of high-dimensional nonlinear relationship among different resolution LST data in RES-STF, the images predicted by RES-STF in the 3 regions are most satisfying, which exhibit the greatest similarity to the reference images in terms of color and texture information across. The other 4 algorithms show varying degrees of deviation. For example, ESTARFM and STTFN exhibit over-estimated temperature on the date 2013342 in Region 1, while ECPW-STFN and StfNet show under-estimated temperature on the date 2019182 in Region 2 and on the date 2013169 in Region 3, respectively. Additionally, it is evident that RES-STF maintains stable fusion performance across different scenarios, demonstrating its robustness in spatio-temporal fusion of LST data.

Quantitative evaluation

The scatter density plots for the 3 regions are shown in Figs. 10 to 12. Overall, RES-STF demonstrates superior accuracy across all 3 regions compared to the other 4 algorithms. Examining the scatterplot in Fig. 10 for Region 1, it is evident that both the ESTARFM and STTFN exhibit significant deviations in their results, particularly noticeable on date 2013342, with numerous scattered data points departing from the expected trend. In contrast, the scatter distribution of RES-STF predictions presents a narrow line, and the density center of the scatter points is closer to the central line compared to the other 4 algorithms, indicating that the results of the proposed algorithm have smaller disparities from the reference images.

The qualitative evaluation results for the 3 metrics (CC, RMSE, and MAE) are presented in Tables 3 to 5. Overall, it can be seen that the STTFN presents slightly weaker performance in fusion of VIIRS LST and Landsat LST, with larger deviations in terms of the metrics (e.g., RMSE and MAE range from 3 to 4 K). While the fusion accuracy of ESTARFM and StfNet is relatively greater than STTFN, their performances are not robust enough. For example, in Region 2, their performances are comparable to RES-STF, but in Region 1, the accuracy of ESTARFM and StfNet is lower, with

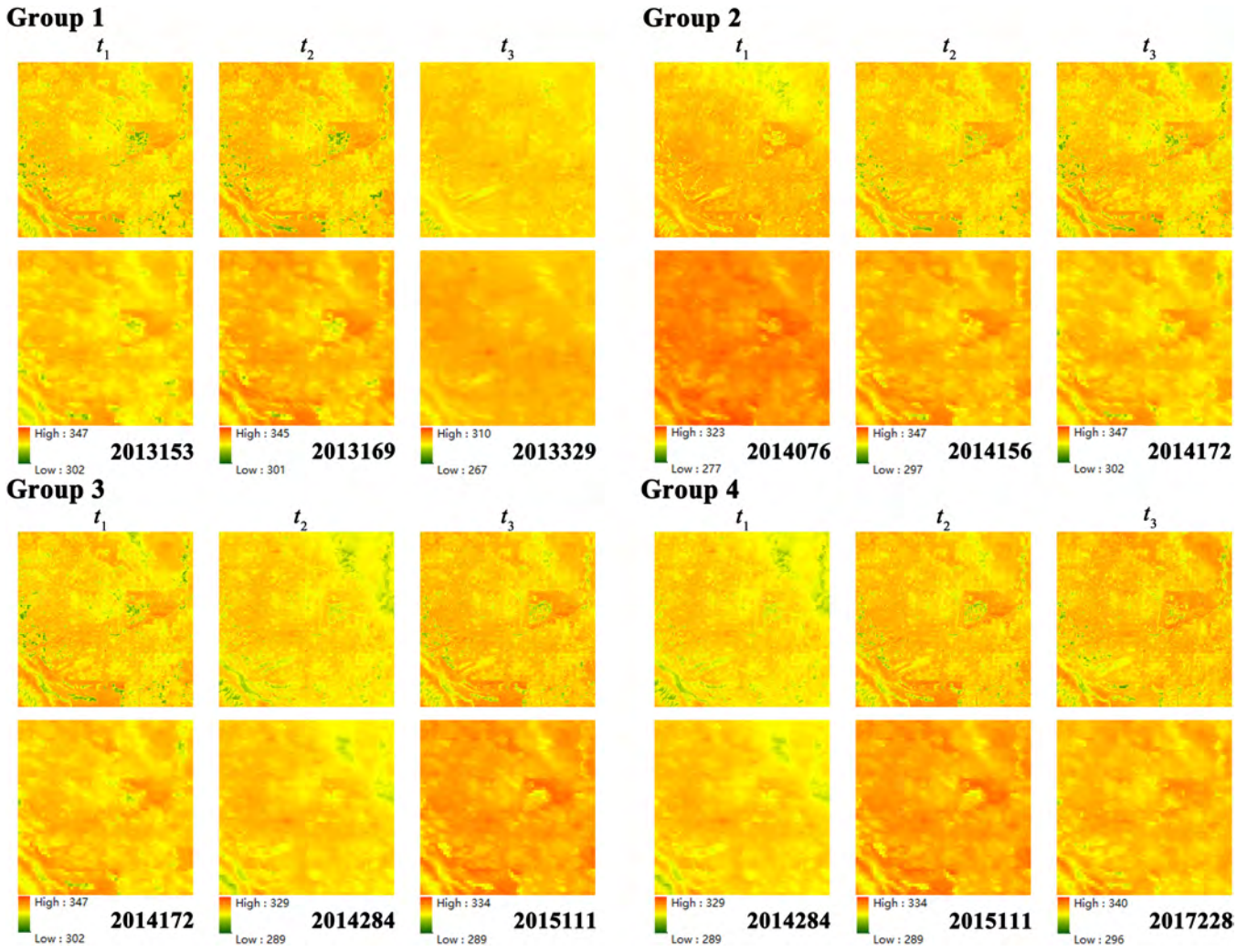

Fig. 6. Data used for prediction in Region 3 (the first and second lines in each group represent Landsat and VIIRS LST images, respectively; the Landsat LST image at t_2 in each group was predicted).

CC values almost below 0.800. ECPW-STFN produces relatively lower accuracy than the other methods in Regions 2 and 3. In general, RES-STF exhibits the greatest accuracy across all 3 regions, with mean RMSE and MAE values around 2 K in all cases.

Computational efficiency

In Table 6, we evaluated the computational efficiency of ESTARFM and 4 deep learning-based methods. As can be seen from the results, for the training stage, RES-STF is the most time-consuming among these methods. However, considering the noticeable gain in accuracy by RES-STF (as shown in Tables 3 to 6), the longer training time is generally acceptable.

Validation of the use of VIIRS LST data

In this paper, the VIIRS LST data were proposed as alternative to MODIS LST for fusion. To validate the advantage of using VIIRS data as coarse data, we made a brief comparison between the use of VIIRS LST and MODIS LST in spatio-temporal fusion. Region 2 was selected for study here.

First, a comparison is conducted for direct quality assessment of the 2 datasets, with the Landsat LST at the same time as reference. To harmonize the spatial resolution, both Landsat and VIIRS LST were degraded to 1 km. As illustrated in Fig. 13, MODIS LST fails to effectively present the spatial temperature distribution, while VIIRS LST exhibits richer textural details, demonstrating greater similarity with Landsat data in spatial pattern. The quantitative assessment results are shown in Table 7. It is seen that compared to MODIS LST data, VIIRS LST exhibits larger RMSE and MAE values due to greater discrepancies in satellite overpass times with Landsat. However, benefiting from better quality, VIIRS LST presents much larger CC than MODIS LST, with gains of about 0.060 in terms of mean CC. Previous study reveals that in spatio-temporal fusion, one of the most important factors affecting the accuracy is the spatial correlation between coarse and fine spatial resolution images [57]. The VIIRS LST contains richer spatial details and is closer to fine spatial resolution data in terms of spatial texture and, thus, can be served as effective alternative data source in spatio-temporal fusion tasks.

To compare the fusion accuracy of using VIIRS LST and MODIS LST, RES-STF was applied for fusion of each dataset separately with Landsat LST, and the results are shown in Table 8. Across all 3 metrics, the accuracy produced by fusing with MODIS LST is consistently lower than that achieved by fusing with VIIRS LST.

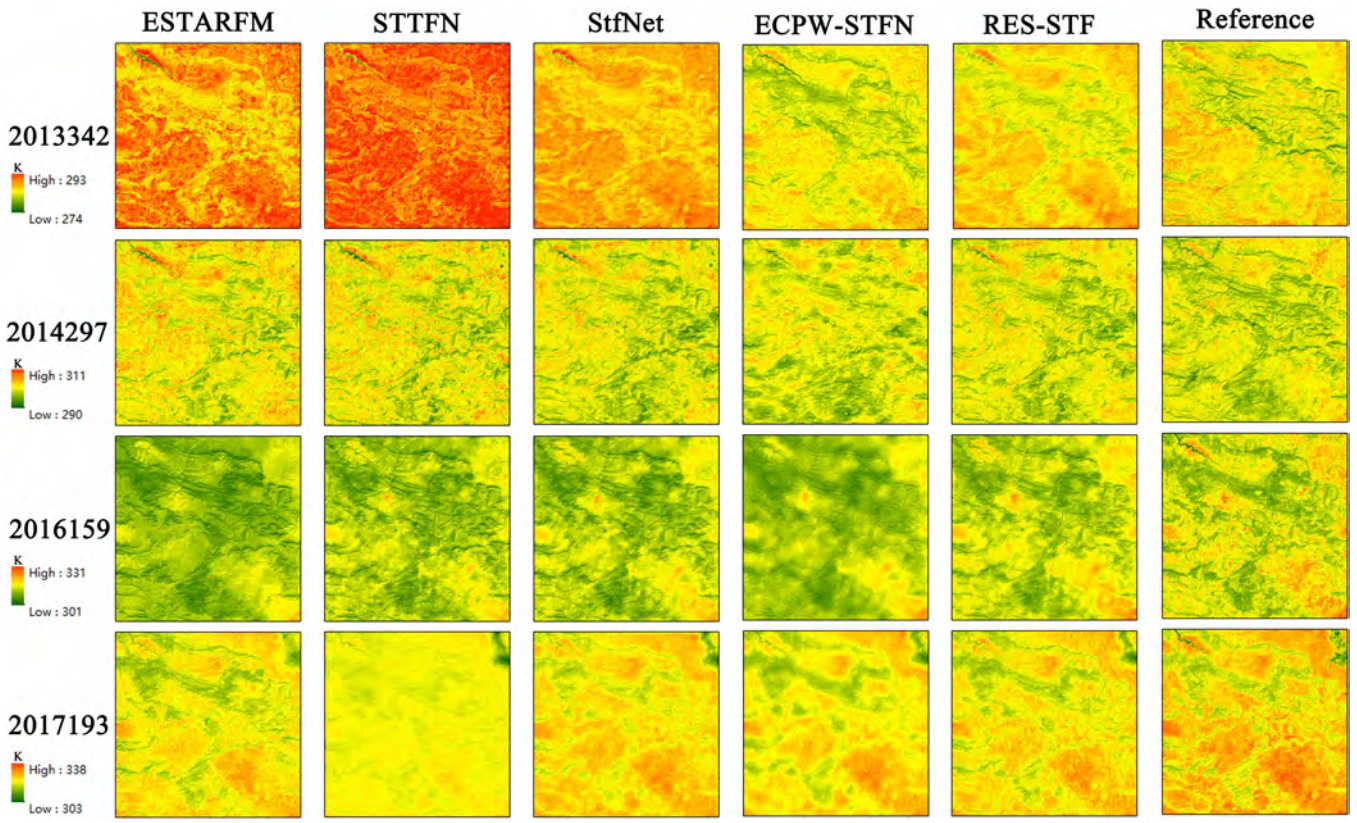

Fig. 7. Landsat-like prediction of the 4 algorithms for Region 1.

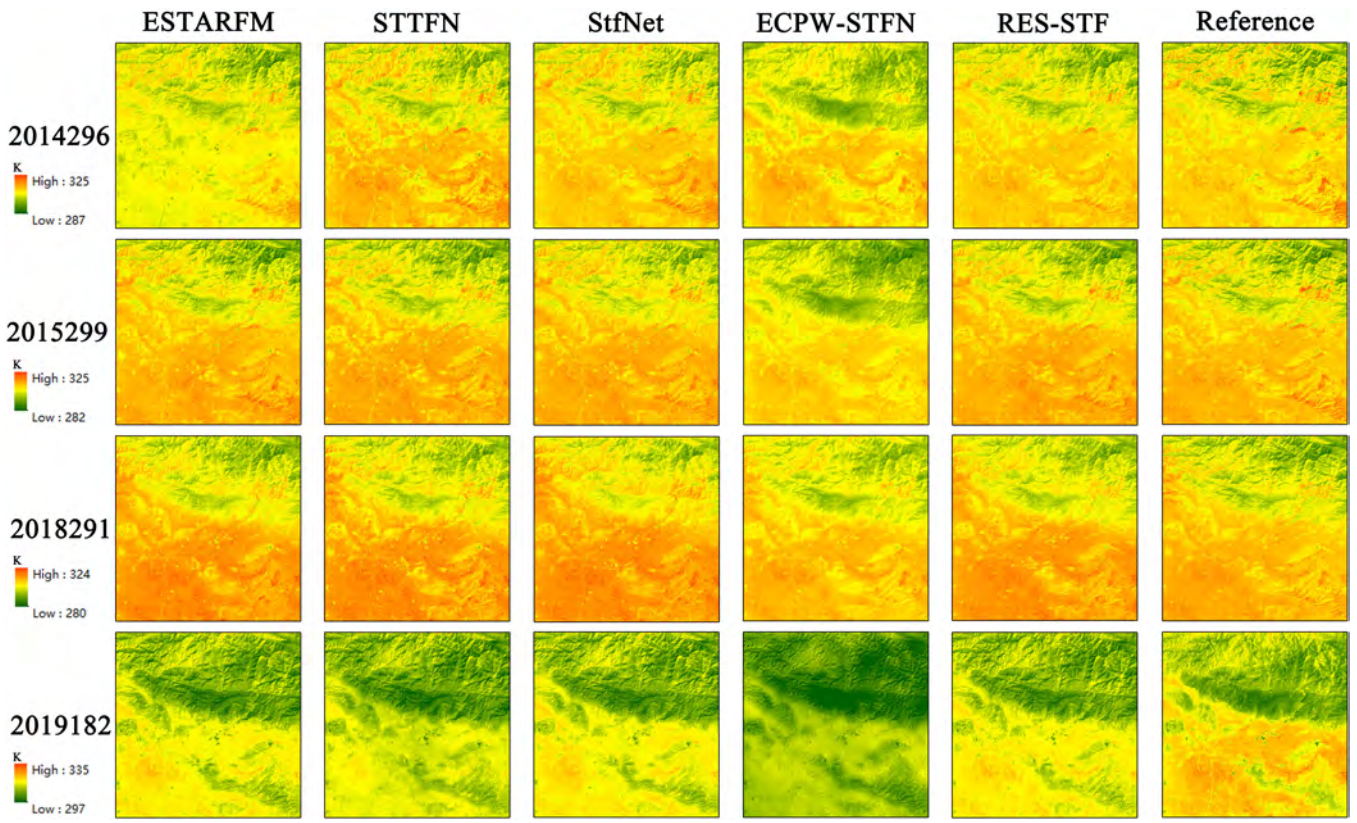

Fig. 8. Landsat-like prediction of the 4 algorithms for Region 2.

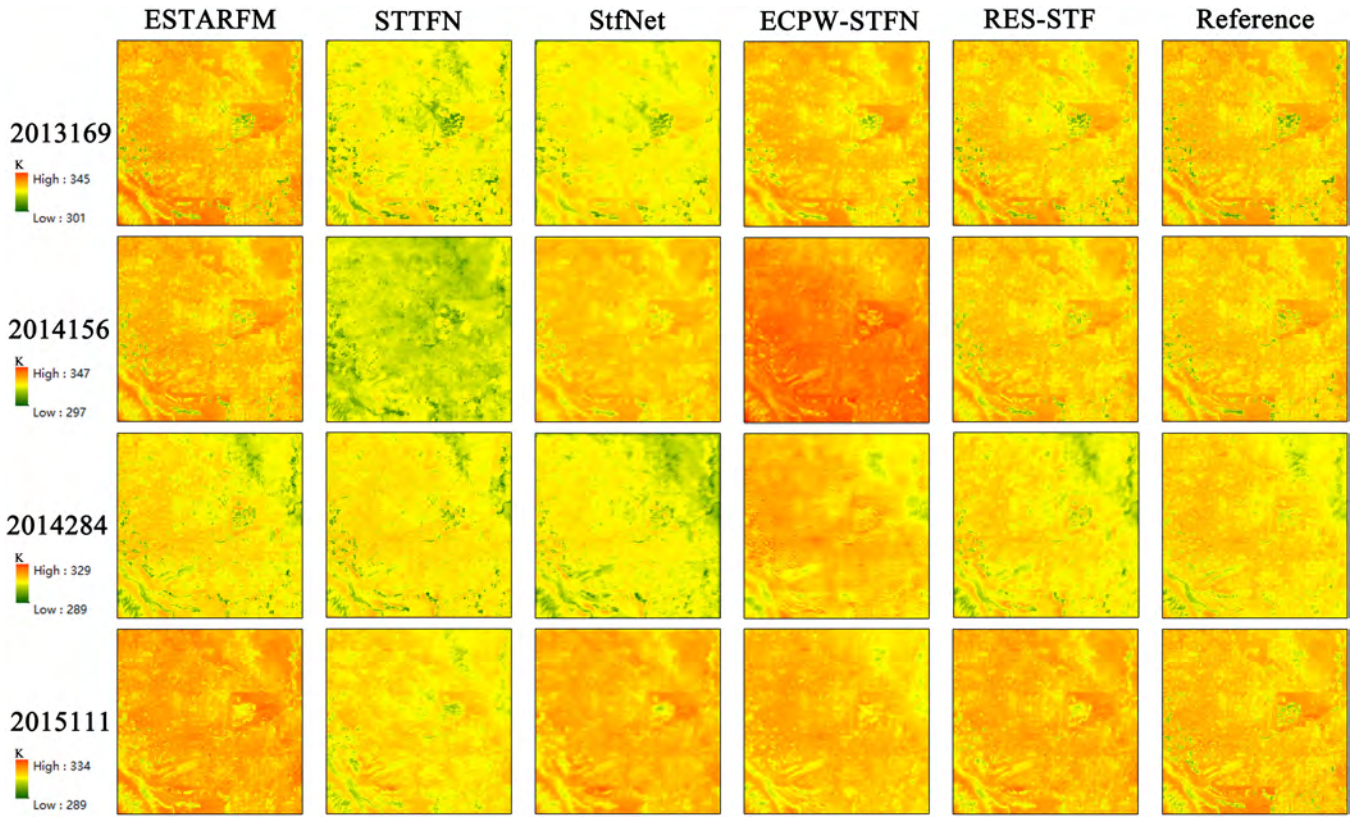

Fig. 9. Landsat-like prediction of the 4 algorithms for Region 3.

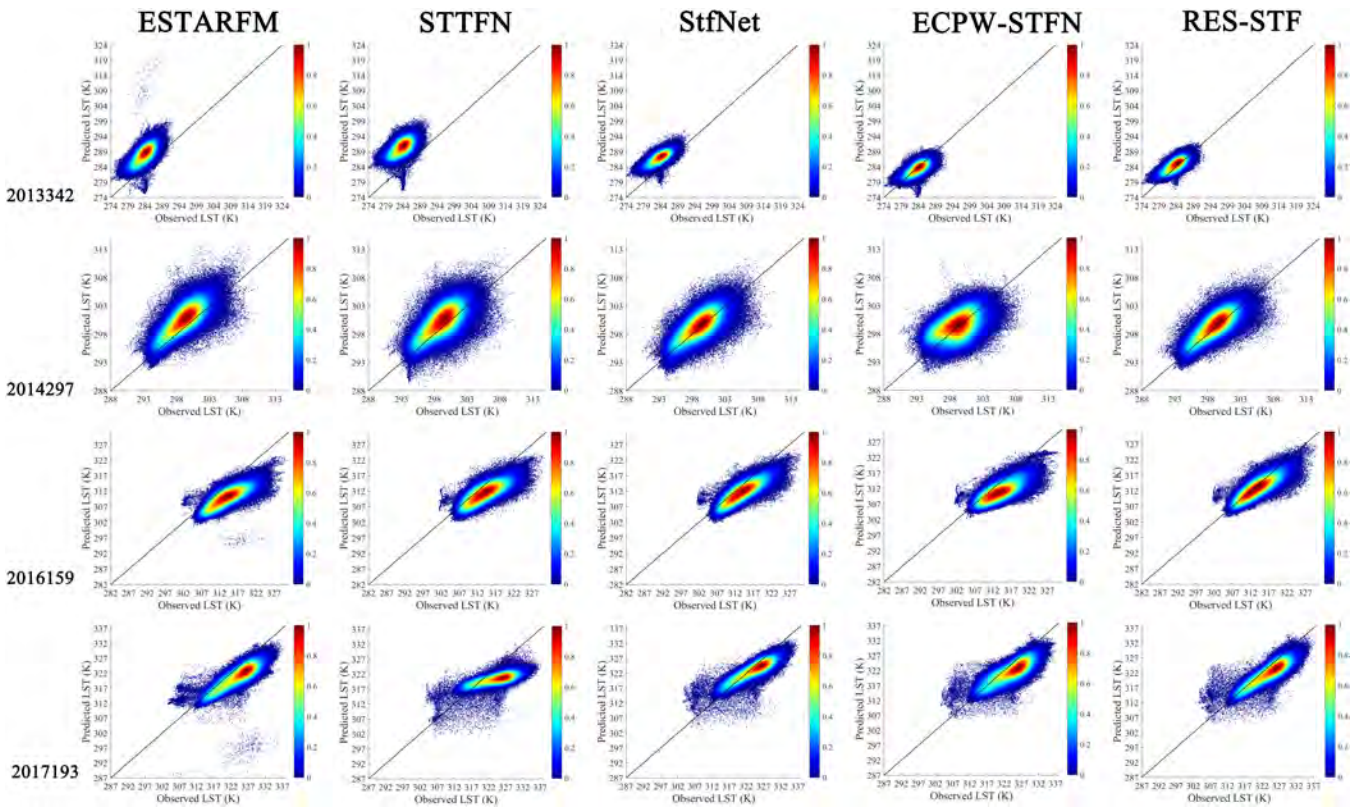

Fig. 10. Scatterplots of the 4 algorithms for Region 1.

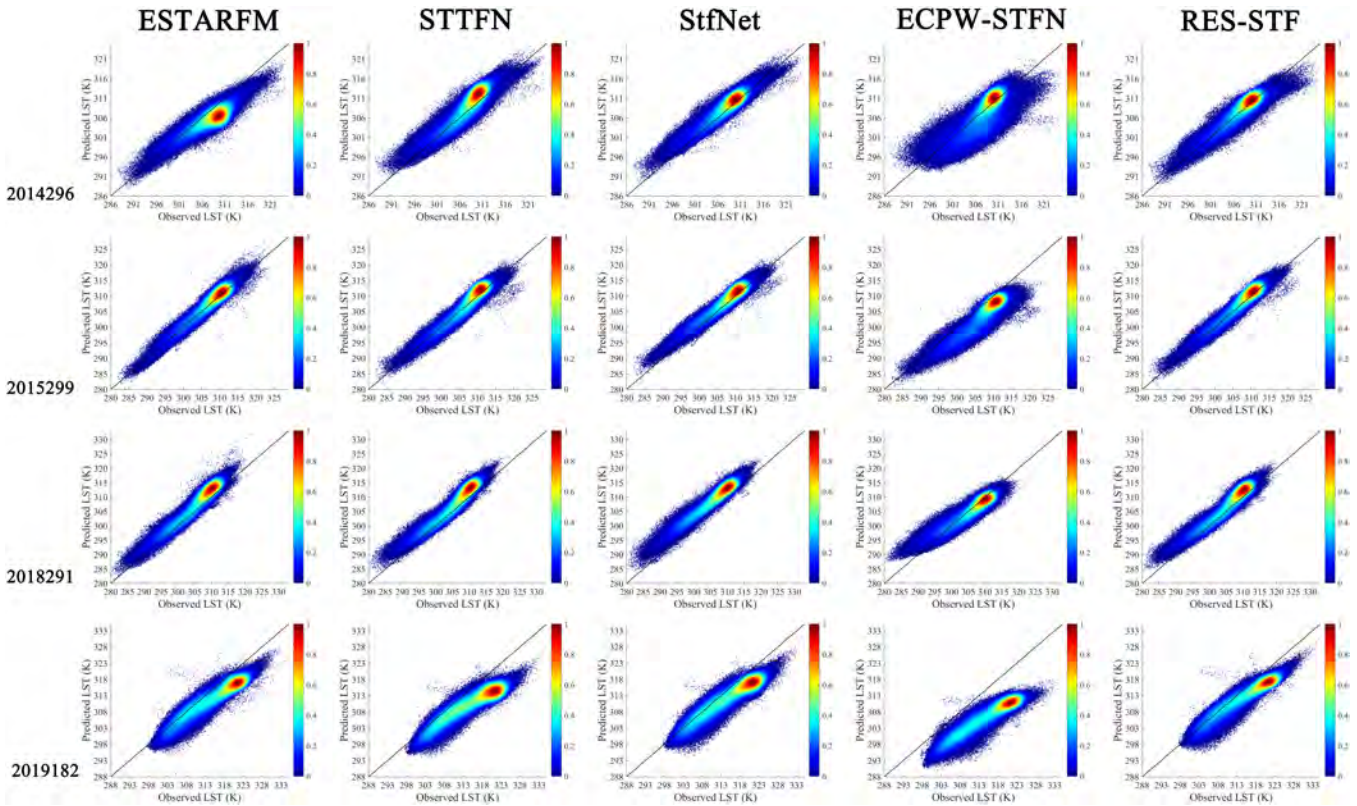

Fig. 11. Scatterplots of the 4 algorithms for Region 2.

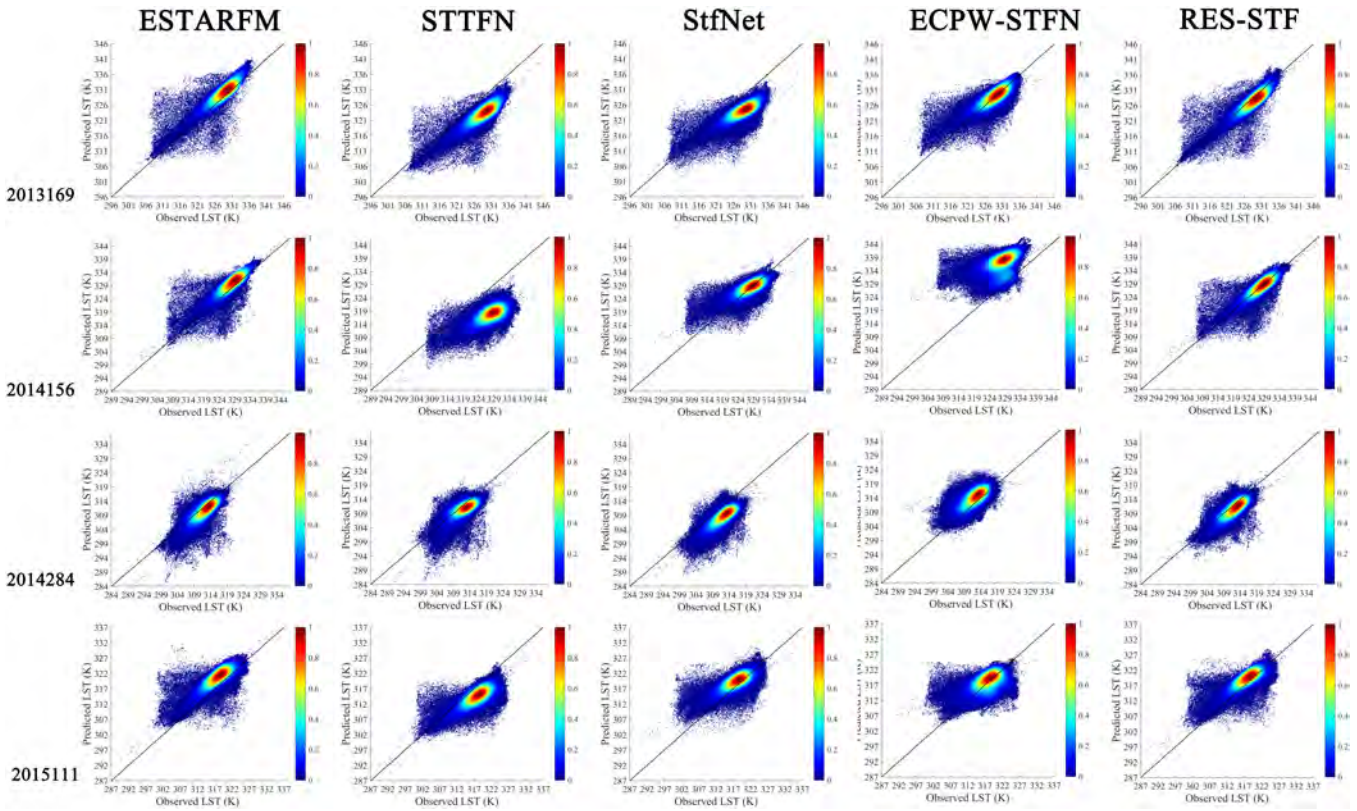

Fig. 12. Scatterplots of the 4 algorithms for Region 3.

Table 3. Accuracy assessment for Region 1 (with most accurate results highlighted in bold)

Date	ESTARFM	STTFN	Stf Net	ECPW-STFN	RES-STF
			CC		
2013342	0.649	0.557	0.704	0.668	**0.704**
2014297	0.737	0.634	0.697	0.476	**0.823**
2016159	0.775	0.763	0.812	0.732	**0.816**
2017193	0.845	0.814	**0.879**	0.805	0.876
Mean	0.751	0.692	0.773	0.670	**0.804**
			RMSE (K)		
2013342	4.760	6.995	3.587	**1.827**	1.891
2014297	2.368	2.380	1.923	2.375	**1.545**
2016159	4.883	3.822	3.375	4.091	**2.616**
2017193	3.927	3.146	**2.648**	3.653	3.065
Mean	3.984	4.085	2.883	2.986	**2.279**
			MAE (K)		
2013342	4.272	6.648	3.227	**1.432**	1.527
2014297	1.892	1.871	**1.490**	1.913	1.722
2016159	4.230	3.123	2.738	3.337	**2.012**
2017193	3.217	2.383	**2.053**	2.954	2.423
Mean	3.402	3.506	2.377	2.409	**1.921**

Table 4. Accuracy assessment for Region 2 (with most accurate results highlighted in bold)

Date	ESTARFM	STTFN	Stf Net	ECPW-STFN	RES-STF
			CC		
2014296	0.898	0.921	**0.945**	0.787	0.933
2015299	**0.974**	0.965	0.971	0.918	0.970
2018291	**0.969**	0.967	0.969	0.937	0.964
2019182	0.937	0.920	0.931	0.934	**0.944**
Mean	0.944	0.943	**0.954**	0.894	0.952
			RMSE (K)		
2014296	2.906	2.204	1.750	3.222	**1.647**
2015299	**1.418**	1.670	1.770	4.131	1.540
2018291	2.652	2.681	3.604	**2.345**	2.395
2019182	3.678	5.313	3.633	8.993	**3.341**
Mean	2.663	2.967	2.689	4.672	**2.230**
			MAE (K)		
2014296	2.422	1.845	1.356	2.456	**1.274**
2015299	**1.103**	1.358	1.432	3.485	1.233
2018291	2.246	2.254	3.273	**1.794**	1.986
2019182	3.117	4.607	3.015	8.614	**2.819**
Mean	2.222	2.516	2.269	4.087	**1.828**

Comparison with the linear correction scheme

In this section, RES-STF was compared with the linear correction scheme for dealing with the difference between multiresolution data on the same day. Specifically, a linear model was constructed between VIIRS and Landsat LST. For the known time (i.e., t_1 or t_3), the VIIRS LST image was linear corrected using the coefficients estimated from the linear model constructed at the corresponding time. For the prediction time, as the Landsat LST is unknown, thus, VIIRS and Landsat LST at both known times were input to the linear model, and the VIIRS LST image at the prediction time was corrected using the coefficients estimated from the linear model. After linear correction, the ESTARFM method was employed for fusion, and the scheme is denoted as ESTARFM-lc.

The results are shown in Table 9. It is seen from the results that RES-STF is more accurate than the linear correction scheme. The linear correction is an effective solution but is not very stable due to the great spatial and temporal variations. That is, the relation between VIIRS and Landsat LST cannot be simply characterized by a linear model. Moreover, the linear relation obtained from the known time may not be directly applicable to the prediction time.

Ablation experiments

RES-STF mainly focuses on modeling differential information, including sensor bias and temporal difference. When constructing the network, the integration of data at different scales unavoidably requires super-resolution, which introduces uncertainty in the fusion process. To address this challenge, we designed a series of structures in RES-STF for optimizing the deep feature extraction process during the modeling of differential features. It is necessary to conduct ablation experiments to validate the effectiveness of the network structure.

Training strategy

In the "Network architecture" section, we mentioned that the network of RES-STF is optimized for both Forward and Backward training through simultaneous training instead of separate training. Table 10 compares the accuracy of these 2 training methods. Overall, the simultaneous training strategy is preferable. More precisely, compared with the separate training strategy, the CC of the proposed simultaneous training strategy is 0.200 larger, and the RMSE and MAE are about 0.300 K and 0.100 K lower, respectively. In addition, it is obvious that the separate training strategy is relatively more time-consuming. Therefore, considering both efficiency and accuracy, the simultaneous training strategy is suggested.

Advantages of the structure

To demonstrate the stability of the proposed network, we removed some structures and evaluated the corresponding performances. Tables 11 and 12 list the accuracy after removing Globe Learning (GL) and RB from the network structure,

Table 5. Accuracy assessment for Region 3 (with most accurate results highlighted in bold)

Date	ESTARFM	STTFN	StfNet	ECPW-STFN	RES-STF
			CC		
2013169	**0.915**	0.878	0.820	0.843	0.908
2014156	**0.891**	0.683	0.787	0.398	0.884
2014284	0.797	0.721	0.777	0.675	**0.801**
2015111	**0.826**	0.749	0.754	0.516	0.808
Mean	**0.857**	0.757	0.784	0.608	0.850
			RMSE (K)		
2013169	2.262	5.595	5.043	2.195	**1.975**
2014156	2.539	5.481	2.459	9.490	**1.873**
2014284	2.616	2.775	4.190	3.180	**2.449**
2015111	3.422	4.151	2.845	3.280	**2.537**
Mean	2.709	4.500	3.634	4.536	**2.208**
			MAE (K)		
2013169	1.714	5.334	4.64	**1.412**	1.424
2014156	2.036	4.957	1.660	8.565	**1.011**
2014284	1.981	2.163	3.758	2.500	**1.939**
2015111	2.994	3.587	2.187	2.425	**1.973**
Mean	2.181	4.010	3.061	3.725	**1.586**

respectively. The results indicate that both GL and RB structures are beneficial.

Discussion

Advantage of using VIIRS

In the "Validation of the use of VIIRS LST data" section, we validated the advantage of VIIRS LST based on RES-STF. To further validate the value of VIIRS LST over MODIS LST data in different spatio-temporal fusion models, this section compares the fusion results of using the 2 types of data based on the ESTARFM method, as shown in Table 13. It can be observed that the fusion results produced by VIIRS LST are still superior to those by MODIS LST across all 3 metrics. This indicates that under the ESTARFM method, VIIRS LST also offers more advantages than MODIS LST. Therefore, the VIIRS LST data hold significant potential in spatio-temporal fusion applications.

Adaptation of current models

The RES-STF method proposed in this paper aims to deal with the data differences caused by sensors. For such problems, several non-deep learning methods have also been developed in existing studies. For example, LiSTF was developed to enhance current spatio-temporal fusion models by simulating a MODIS-like image closer to the Landsat image at the prediction time [58]. It considers amplitude difference between fine and coarse spatial resolution reflectance data pairs at the same time.

Table 6. The efficiency of different models

	ESTARFM	STTFN	StfNet	ECPW-STFN	RES-STF
Training time (s)	-	2,012	1,792	2,004	12,007
Prediction time (s)	49	0.28	0.22	0.99	1.27

Specifically, for each land cover type, a linear mapping model is established between the fine and coarse spatial resolution images at the known time. The relation characterized by the model is transferred to the images at the prediction time. This method assumes that land cover does not change between the known and prediction times.

However, in spatio-temporal fusion of LST, for scenes with complex spatial distribution of land cover and strong temporal changes, challenges are encountered when characterizing the differences between fine and coarse spatial resolution data through traditional linear models. Taking VIIRS data as an example, to address the data differences due to inconsistent overpass times with Landsat, it is necessary to account for the spatio-temporal heterogeneity of LST data and describe the nonlinear mapping relationship between different sensors by combining physical mechanism.

Future research

For the deep learning-based algorithms mentioned in this paper, training data are a crucial factor that affects model performance. That is, the selection of appropriate training data is an important issue. For the study areas in this paper, choosing different types of training data, such as areas with different land cover types or spatial texture, can lead to different results (e.g., STTFN relies heavily on the selected training data and areas since it is a pure CNN framework). Therefore, RES-STF still has room for enhancement in the future, especially when training data are difficult to collect. For example, it would be worthwhile to develop some pre-training strategies and self-supervised methods like contrastive learning to improve the performance of RES-STF.

The recovery of texture details (e.g., in urban areas) is a key issue in spatio-temporal fusion, especially for LST with great spatio-temporal heterogeneity. By further integrating other relevant environmental variables, such as digital elevation model (DEM), NDVI, and other multimodal data [59], it may be possible to reduce uncertainties in the fusion process. It is also important to consider more challenges in practical scenarios, such as data gaps caused by cloud contamination. In such cases, it would be interesting to consider the use of passive microwave brightness temperature data that are not affected by cloud cover or even applying crowdsourced geospatial data with fine spatio-temporal resolution [60].

Conclusion

Spatio-temporal fusion is an important solution to create fine spatio-temporal resolution LST data by fusion of LST data with

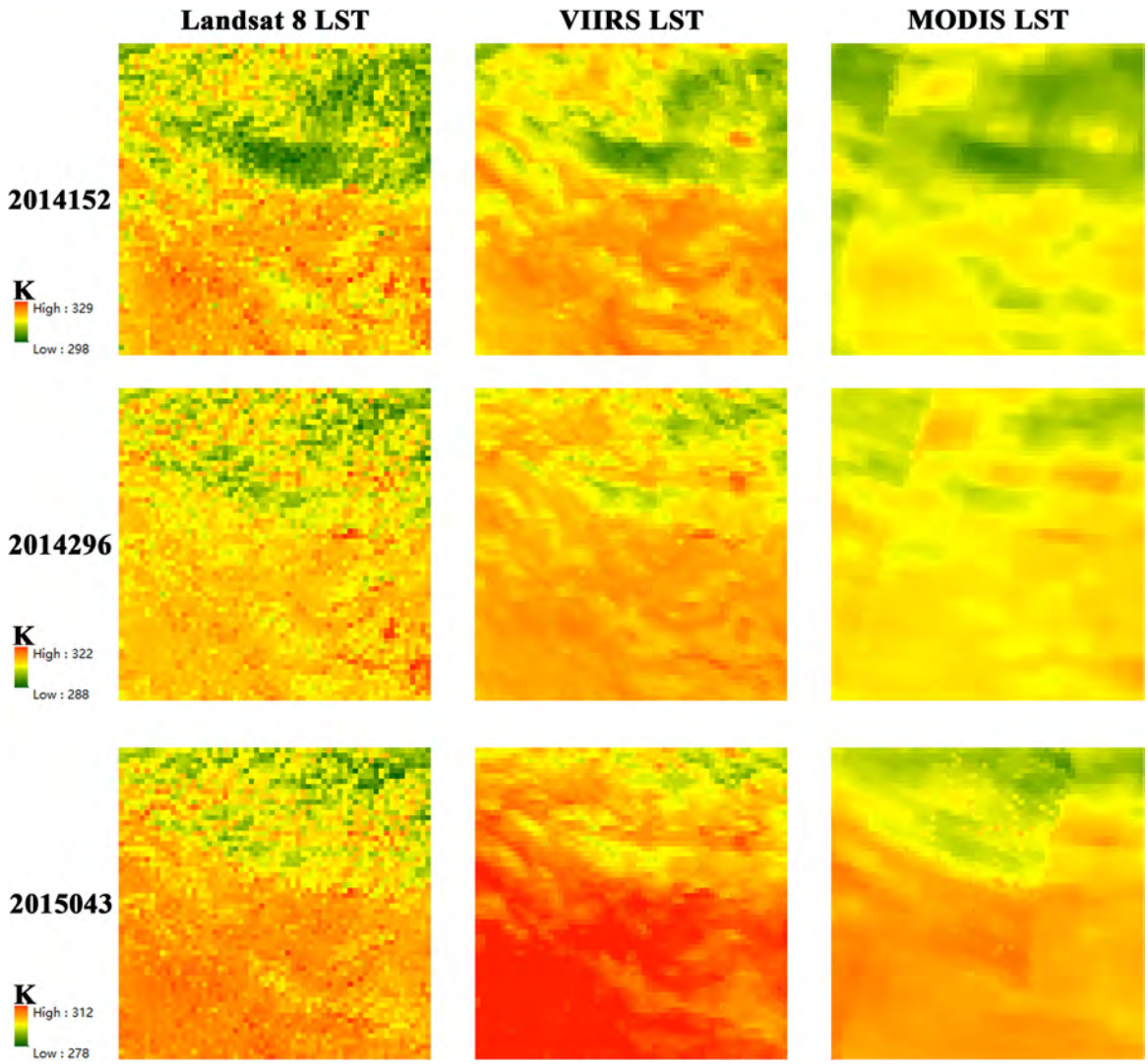

Fig. 13. Comparison among Landsat LST, VIIRS LST, and MODIS LST in Region 2.

Table 7. Evaluation of VIIRS and MODIS LST by referring to Landsat LST (taking Region 2 as an example; with most accurate results highlighted in bold)

	CC		RMSE (K)		MAE (K)	
	VIIRS	MODIS	VIIRS	MODIS	VIIRS	MODIS
2014152	**0.876**	0.829	**3.189**	4.834	**2.498**	4.132
2014296	**0.747**	0.586	3.981	**3.902**	3.287	**3.155**
2015043	**0.874**	0.825	6.712	**3.278**	6.139	**2.463**
2015107	**0.904**	0.889	**3.624**	5.409	**2.778**	4.695
2015299	**0.845**	0.804	4.301	**3.827**	3.526	**3.138**
Mean	**0.849**	0.787	4.361	**4.250**	3.646	**3.517**

either fine spatial or fine temporal resolution (but not both). Considering the issue of data quality in the coarse spatial resolution input data (i.e., MODIS LST) in spatio-temporal fusion, this paper proposed to utilize VIIRS LST as replacement of MODIS LST data. Furthermore, to deal with the data discrepancy caused by the different overpass time between VIIRS LST

Table 8. Comparison between the use of VIIRS LST and MODIS LST for prediction in region 2 (with most accurate results highlighted in bold)

	CC		RMSE (K)		MAE (K)	
	VIIRS	MODIS	VIIRS	MODIS	VIIRS	MODIS
2014296	**0.933**	0.918	**1.647**	2.210	**1.274**	1.824
2015299	**0.970**	0.968	**1.540**	2.431	**1.233**	2.001
2018291	**0.964**	0.959	**2.395**	2.693	**1.986**	2.176
2019182	**0.944**	0.933	**3.341**	3.770	**2.819**	3.243
Mean	**0.952**	0.944	**2.230**	2.776	**1.828**	2.311

Table 9. Accuracy assessment for 3 regions (with most accurate results highlighted in bold; lc denotes the linear correction scheme; * denotes linear correction makes positive effect)

		CC			RMSE (K)			MAE (K)		
		ESTARFM	ESTARFM-lc	RES-STF	ESTARFM	ESTARFM-lc	RES-STF	ESTARFM	ESTARFM-lc	RES-STF
Region 1	2013342	0.649	0.656*	**0.704**	4.760	3.669*	**1.891**	4.272	3.140*	**1.527**
	2014297	0.737	0.744*	**0.823**	2.368	2.310*	**1.545**	1.892	1.860*	**1.722**
	2016159	0.775	0.767	**0.816**	4.883	9.260	**2.616**	4.230	8.875	**2.012**
	2017193	0.845	0.849*	**0.876**	3.927	**2.704***	3.065	3.217	**1.969***	2.423
	Mean	0.751	0.754*	**0.804**	3.984	4.485	**2.279**	3.402	3.961	**1.921**
Region 2	2014296	0.898	0.895	**0.933**	2.906	2.942	**1.647**	2.422	2.449	**1.274**
	2015299	**0.974**	0.974	0.970	**1.418**	1.430	1.540	**1.103**	1.117	1.233
	2018291	**0.969**	0.920	0.964	2.652	2.883	**2.395**	2.246	2.351	**1.986**
	2019182	0.937	0.939*	**0.944**	3.678	3.748	3.341	3.117	3.181	2.819
	Mean	0.944	0.932	**0.952**	2.663	2.750	**2.230**	2.222	2.274	**1.828**
Region 3	2013169	0.915	0.913	0.908	2.262	2.369	**1.975**	1.714	1.833	**1.424**
	2014156	**0.891**	0.888	0.884	2.539	2.751	**1.873**	2.036	2.262	**1.011**
	2014284	0.797	**0.816***	0.801	2.616	3.872	**2.449**	1.981	3.427	**1.939**
	2015111	0.826	**0.829***	0.808	3.422	3.432	**2.537**	2.994	3.005	**1.973**
	Mean	0.857	**0.861***	0.850	2.709	3.106	**2.208**	2.181	2.631	**1.586**

Table 10. Accuracy assessment of different training strategies (taking Region 1 as an example; Sep-training means that Forward and Backward training are separated; the most accurate results are highlighted in bold)

	CC		RMSE (K)		MAE (K)	
	Training	Sep-training	Training	Sep-training	Training	Sep-training
2013342	0.704	**0.738**	1.891	**1.842**	1.527	**1.480**
2014297	**0.823**	0.702	**1.545**	1.816	1.722	**1.405**
2016159	0.816	**0.819**	**2.616**	3.246	**2.012**	2.586
2017193	**0.876**	0.876	**3.065**	3.302	**2.423**	2.670
Mean	**0.804**	0.783	**2.264**	2.551	**1.920**	2.035

Table 11. Accuracy assessment between with and without the GL block (taking region 1 as an example; with most accurate results highlighted in bold)

	CC		RMSE (K)		MAE (K)	
	With GL	Without GL	With GL	Without GL	With GL	Without GL
2013342	0.704	**0.734**	**1.891**	2.119	**1.527**	1.740
2014297	**0.823**	0.649	**1.545**	2.084	1.722	**1.629**
2016159	**0.816**	0.811	**2.616**	3.311	**2.012**	2.647
2017193	**0.876**	0.871	3.065	**2.569**	2.423	**1.925**
Mean	**0.804**	0.766	**2.279**	2.520	**1.921**	1.985

Table 12. Accuracy assessment between with and without the RB (taking region 1 as an example; with most accurate results highlighted in bold)

	CC		RMSE (K)		MAE (K)	
	With RB	Without RB	With RB	Without RB	With RB	Without RB
2013342	0.704	**0.747**	1.891	**1.579**	1.527	**1.246**
2014297	**0.823**	0.663	**1.545**	1.888	1.722	**1.468**
2016159	**0.816**	0.796	**2.616**	2.958	**2.012**	2.282
2017193	**0.876**	0.864	**3.065**	3.307	**2.423**	2.644
Mean	**0.804**	0.767	**2.279**	2.433	1.921	**1.910**

Table 13. Comparison between the use of VIIRS LST and MODIS LST based on ESTARFM (taking region 2 as an example; with most accurate results highlighted in bold)

	CC		RMSE (K)		MAE (K)	
	VIIRS	MODIS	VIIRS	MODIS	VIIRS	MODIS
2014296	**0.898**	0.887	2.906	**2.108**	2.422	**1.566**
2015299	**0.974**	0.951	**1.418**	2.478	**1.103**	1.912
2018291	**0.969**	0.953	**2.652**	3.641	**2.246**	3.200
2019182	0.937	**0.940**	**3.678**	4.116	**3.117**	3.632
Mean	**0.944**	0.932	**2.663**	3.085	**2.222**	2.577

and Landsat LST, the RES-STF method was proposed. RES-STF combines the advantages of CNN in multi-scale feature extraction and Transformer in global dependency to effectively capture both local and global information in the images. Through experiments in 3 different regions, the effectiveness of RES-STF was validated. RES-STF shows strong robustness in various areas, and it is more accurate than one typical non-deep learning-based method (i.e., ESTARFM) and 3 typical deep learning-based methods (i.e., STTFN, StfNet and ECPW-STFN). Additionally, the VIIRS LST data present richer spatial details and higher data quality compared to the MODIS LST data, and the use of VIIRS LST leads to greater fusion accuracy. In summary, VIIRS LST data are an effective alternative to MODIS LST in the spatio-temporal fusion task.

Acknowledgments

Funding: This research was supported by the National Natural Science Foundation of China under grants 42171345 and 42222108.
Author contributions: Q.W. and R.H. designed the research, analyzed the dataset, wrote the original manuscript, and revised the whole manuscript. R.H. produced the dataset. Q.W. provided the funding to support the research. All authors approved the final manuscript.
Competing interests: The authors declare that they have no competing interests.

Data Availability

The MODIS and VIIRS data are available through https://ladsweb.modaps.eosdis.nasa.gov. The Landsat data are available through https://earthexplorer.usgs.gov/.

References

1. Hansen J, Ruedy R, Sato M, Lo K. Global surface temperature change. *Rev Geophys*. 2010;48:RG4004.
2. Song L, Liu S, Kustas WP, Nieto H, Sun L, Xu Z, Skaggs TH, Yang Y, Ma M, Xu T, et al. Monitoring and validating spatially and temporally continuous daily evaporation and transpiration at river basin scale. *Remote Sens Environ*. 2018;219:72–88.
3. Shen H, Huang L, Zhang L, Wu P, Zeng C. Long-term and fine-scale satellite monitoring of the urban heat island effect by the fusion of multi-temporal and multi-sensor remote sensed data: A 26-year case study of the city of Wuhan in China. *Remote Sens Environ*. 2016;172:109–125.
4. Roy DP, Wulder MA, Loveland TR, C.E. W, Allen RG, Anderson MC, Helder D, Irons JR, Johnson DM, Kennedy R, et al. Landsat-8: Science and product vision for terrestrial global change research. *Remote Sens Environ*. 2014;145:154–172.
5. Justice CO, Vermote E, Townshend JRG, Defries R, Roy DP, Hall DK, Salomonson VV, Privette JL, Riggs G, Strahler A, et al. The Moderate Resolution Imaging Spectroradiometer (MODIS): Land remote sensing for global change research. *IEEE Trans Geosci Remote Sens*. 1998;36(4):1228–1249.
6. Reiners P, Asam S, Frey C, Holzwarth S, Bachmann M, Sobrino J, Göttsche FM, Bendix J, Kuenzer C. Validation of AVHRR land surface temperature with MODIS and in situ LST—A timeline thematic processor. *Remote Sens*. 2021;13(17):3473.
7. Hulley GC, Hook SJ, Abbott E, Malakar N, Islam T, Abrams M. The ASTER global emissivity dataset (ASTER GED): Mapping Earth's emissivity at 100 meter spatial scale. *Geophys Res Lett*. 2015;42(19):7966–7976.

8. Rogers MA, Miller SD, Seaman CJ, Torres J, Hillger D, Szoke E, Line WE. VIIRS after 10 years—A perspective on benefits to forecasters and end-users. *Remote Sens*. 2023;15(4):976.
9. Hu T, Renzullo LJ, van Dijk AIJM, He J, Tian S, Xu Z, Zhou J, Liu T, Liu Q. Monitoring agricultural drought in Australia using MTSAT-2 land surface temperature retrievals. *Remote Sens Environ*. 2020;236:Article 111419.
10. Li L, Zhan W, Hu L, Chakraborty TC, Wang Z, Fu P, Wang D, Liao W, Huang F, Fu H, et al. Divergent urbanization-induced impacts on global surface urban heat island trends since 1980s. *Remote Sens Environ*. 2023;295:Article 113650.
11. Shi W, Goodchild M, Batty M, Li Q, Liu X, Zhang A. Prospective for urban informatics. *Urban Inform*. 2022;1:2.
12. Wang Q, Tang Y, Ge Y, Xie H, Tong X, Atkinson PM. A comprehensive review of spatial-temporal-spectral information reconstruction techniques. *Sci Remote Sens*. 2023;8:Article 100102.
13. Gao F, Masek J, Schwaller M, Hall F. On the blending of the Landsat and MODIS surface reflectance: Predicting daily Landsat surface reflectance. *IEEE Trans Geosci Remote Sens*. 2006;44(8):2207–2218.
14. Zhu X, Chen J, Gao F, Chen X, Masek JG. An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions. *Remote Sens Environ*. 2010;114(11):2610–2623.
15. Wang Q, Atkinson PM. Spatio-temporal fusion for daily Sentinel-2 images. *Remote Sens Environ*. 2018;204:31–42.
16. Gu Z, Chen J, Chen Y, Qiu Y, Zhu X, Chen X. Agri-fuse: A novel spatiotemporal fusion method designed for agricultural scenarios with diverse phenological changes. *Remote Sens Environ*. 2023;299:Article 113874.
17. Zhukov B, Oertel D, Lanzl F, Reinhackel G. Unmixing-based multisensor multiresolution image fusion. *IEEE Trans Geosci Remote Sens*. 1999;37(3):1212–1226.
18. Niu Z. 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*. 2012;6(1):Article 063507.
19. Wang Q, Tang Y, Tong X, Atkinson PM. Virtual image pair-based spatio-temporal fusion. *Remote Sens Environ*. 2020;249:Article 112009.
20. Peng K, Wang Q, Tang Y, Tong X, Atkinson PM. Geographically weighted spatial unmixing for spatiotemporal fusion. *IEEE Trans Geosci Remote Sens*. 2022;60:5404217.
21. Xu C, Du X, Yan Z, Zhu J, Xu S, Fan X. VSDF: A variation-based spatiotemporal data fusion method. *Remote Sens Environ*. 2022;283:Article 113309.
22. Zhu X, Helmer EH, Gao F, Liu D, Chen J, Lefsky MA. A flexible spatiotemporal method for fusing satellite images with different resolutions. *Remote Sens Environ*. 2016;172:165–177.
23. Li X, Foody GM, Boyd DS, Ge Y, Zhang Y, du Y, Ling F. SFSDAF: An enhanced FSDAF that incorporates sub-pixel class fraction change information for spatio-temporal image fusion. *Remote Sens Environ*. 2020;237:Article 111537.
24. Guo D, Shi W, Hao M, Zhu X. FSDAF 2.0: Improving the performance of retrieving land cover changes and preserving spatial details. *Remote Sens Environ*. 2020;248:Article 111973.
25. Huang B, Song H. Spatiotemporal reflectance fusion via sparse representation. *IEEE Trans Geosci Remote Sens*. 2012;50(10):3707–3716.
26. Chen G, Lu H, Di D, Li L, Emam M, Jing W. StfMLP: Spatiotemporal fusion multilayer perceptron for remote-sensing images. *IEEE Geosci Remote Sens Lett*. 2023;20:5000105.
27. Song H, Liu Q, Wang G, Hang R, Huang B. Spatiotemporal satellite image fusion using deep convolutional neural networks. *IEEE J Sel Top Appl Earth Obs Remote Sens*. 2018;11(3):821–829.
28. Liu X, Deng C, Chanussot J, Hong D, Zhao B. *StfNet*: A two-stream convolutional neural network for spatiotemporal image fusion. *IEEE Trans Geosci Remote Sens*. 2019;57(9):6552–6564.
29. Qin P, Huang H, Tang H, Wang J, Liu C. MUSTFN: A spatiotemporal fusion method for multi-scale and multi-sensor remote sensing images based on a convolutional neural network. *Int J Appl Earth Obs Geoinf*. 2022;115:Article 103113.
30. Zhang X, Li S, Tan Z, Li X. Enhanced wavelet based spatiotemporal fusion networks using cross-paired remote sensing images. *ISPRS J Photogramm Remote Sens*. 2024;211:281–297.
31. Chen J, Wang L, Feng R, Liu P, Han W, Chen X. CycleGAN-STF: Spatiotemporal fusion via CycleGAN-based image generation. *IEEE Trans Geosci Remote Sens*. 2021;59(7):5851–5865.
32. Tan Z, Gao M, Li X, Jiang L. A flexible reference-insensitive spatiotemporal fusion model for remote sensing images using conditional generative adversarial network. *IEEE Trans Geosci Remote Sens*. 2022;60:5601413.
33. Chen G, Jiao P, Hu Q, Xiao L, Ye Z. SwinSTFM: Remote sensing spatiotemporal fusion using Swin transformer. *IEEE Trans Geosci Remote Sens*. 2022;60:5410618.
34. Zhan W, Chen Y, Zhou J, Wang J, Liu W, Voogt J, Zhu X, Quan J, Li J. Disaggregation of remotely sensed land surface temperature: Literature survey, taxonomy, issues, and caveats. *Remote Sens Environ*. 2013;131:119–139.
35. Huang B, Wang J, Song H, Fu D, Wong K. Generating high spatiotemporal resolution land surface temperature for urban heat island monitoring. *IEEE Geosci Remote Sens Lett*. 2013;10(5):1011–1015.
36. Weng Q, Fu P, Gao F. Generating daily land surface temperature at Landsat resolution by fusing Landsat and MODIS data. *Remote Sens Environ*. 2014;145:55–67.
37. Quan J, Zhan W, Ma T, Du Y, Guo Z, Qin B. An integrated model for generating hourly Landsat-like land surface temperatures over heterogeneous landscapes. *Remote Sens Environ*. 2018;206:403–423.
38. Wu P, Shen H, Zhang L, Göttsche F-M. Integrated fusion of multi-scale polar-orbiting and geostationary satellite observations for the mapping of high spatial and temporal resolution land surface temperature. *Remote Sens Environ*. 2015;156:169–181.
39. Zhang T, Zhou Y, Zhu Z, Li X, Asrar GR. A global seamless 1 km resolution daily land surface temperature dataset (2003–2020). *Earth Syst Sci Data*. 2022;14(2):651–664.
40. Shen H, Li X, Cheng Q, Zeng C, Yang G, Li H, Zhang L. Missing information reconstruction of remote sensing data: A technical review. *IEEE Geosci Remote Sens Mag*. 2015;3(3):61–85.
41. Zhao W, Wu H, Yin G, Duan S-B. Normalization of the temporal effect on the MODIS land surface temperature product using random forest regression. *ISPRS J Photogramm Remote Sens*. 2019;152:109–118.
42. Li H, Sun D, Yu Y, Wang H, Liu Y, Liu Q, du Y, Wang H, Cao B. Evaluation of the VIIRS and MODIS LST products

in an arid area of Northwest China. *Remote Sens Environ*. 2014;142:111–121.
43. Román MO, Justice C, Paynter I, Boucher PB, Devadiga S, Endsley A, Erb A, Friedl M, Gao H, Giglio L, et al. Continuity between NASA MODIS collection 6.1 and VIIRS collection 2 land products. *Remote Sens Environ*. 2024;302:Article 113963.
44. Ma J, Shen H, Wu P, Wu J, Gao M, Meng C. Generating gapless land surface temperature with a high spatio-temporal resolution by fusing multi-source satellite-observed and model-simulated data. *Remote Sens Environ*. 2022;278:Article 113083.
45. Yin Z, Wu P, Foody GM, Wu Y, Liu Z, du Y, Ling F. Spatiotemporal fusion of land surface temperature based on a convolutional neural network. *IEEE Trans Geosci Remote Sens*. 2021;59(2):1808–1822.
46. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai, Unterthiner XT, Dehghani M, Minderer M, Heigold G, Gelly S, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv. 2020. https://doi.org/10.48550/arXiv.2010.11929.
47. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. arXiv. 2017. https://doi.org/10.48550/arXiv.1706.03762.
48. Chen Z, Zhang Y, Gu J, Kong L, Yang X. Recursive generalization Transformer for image super-resolution. arXiv. 2023. https://doi.org/10.48550/arXiv.2303.06373.
49. Wang W, Xie E, Li X, Fan D-P, Song K, Liang D, Lu T, Luo P, Shao L. Pyramid Vision Transformer: A versatile backbone for dense prediction without convolutions. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada: IEEE; 2021. p. 548–558.
50. Wang Z, Cun X, Bao J, Zhou W, Liu J, Li H. Uformer: A general U-shaped Transformer for image restoration. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans (LA): IEEE; 2022. p. 17662–17672.
51. Mehta S, Rastegari M. MobileViT: Light-weight, general-purpose, and mobile-friendly Vision Transformer. arXiv. 2021. https://doi.org/10.48550/arXiv.2110.02178.
52. Hao M, Chen S, Lin H, Zhang H, Zheng N. A prior knowledge guided deep learning method for building extraction from high-resolution remote sensing images. *Urban Inform*. 2024;3(1):6.
53. Zamir SW, Arora A, Khan S, Hayat M, Khan FS, Yang M-H. Restormer: Efficient Transformer for high-resolution image restoration. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans (LA): IEEE; 2022. p. 5718–5729.
54. Hillger D, Seaman C, Liang C, Miller S, Lindsey D, Kopp T. Suomi NPP VIIRS Imagery evaluation. *J Geophys Res*. 2014;119(11):6440–6455.
55. Guillevic PC, Biard JC, Hulley GC, Privette JL, Hook SJ, Olioso A, Göttsche FM, Radocinski R, Román MO, Yu Y, et al. Validation of land surface temperature products derived from the Visible Infrared Imaging Radiometer suite (VIIRS) using ground-based and heritage satellite measurements. *Remote Sens Environ*. 2014;154:19–37.
56. Zhu X, Zhan W, Zhou J, Chen X, Liang Z, Xu S, Chen J. A novel framework to assess all-round performances of spatiotemporal fusion models. *Remote Sens Environ*. 2022;274:Article 113002.
57. Tang Y, Wang Q, Zhang K, Atkinson PM. Quantifying the effect of registration error on spatio-temporal fusion. *IEEE J Sel Top Appl Earth Obs Remote Sens*. 2020;13:487–503.
58. Li J, Li Y, Cai R, He L, Chen J, Plaza A. Enhanced spatiotemporal fusion via MODIS-like images. *IEEE Trans Geosci Remote Sens*. 2022;60:5610517.
59. Li X, Peng Q, Zheng Y, Lin S, He B, Qiu Y, Chen J, Chen Y, Yuan W. Incorporating environmental variables into spatiotemporal fusion model to reconstruct high-quality vegetation index data. *IEEE Trans Geosci Remote Sens*. 2024;62:4401812.
60. Huang X, Wang S, Yang D, Hu T, Chen M, Zhang M, Zhang G, Biljecki F, Lu T, Zou L, et al. Crowdsourcing geospatial data for earth and human observations: A review. *J Remote Sens*. 2024;4:0105.