


A Dual-Perspective Spatiotemporal Fusion Model for Remote Sensing Images by Discriminative Learning of the Spatial and Temporal Mapping

Zhonghua Qian, Linwei Yue , *Member, IEEE*, Xiao Xie , Qiangqiang Yuan , *Member, IEEE*, and Huanfeng Shen , *Senior Member, IEEE*

Abstract—Spatiotemporal fusion (STF) has received widespread attention as a cost-effective solution to the spatiotemporal conflicts in remote sensing images. Massive efforts have been made in the development of STF technology, and deep-learning methods have shown great potential in obtaining state-of-the-art results in recent years. However, it is still challenging to effectively fuse images with land-cover changes. The main problem is that the fine-resolution temporal changes are difficult to accurately model in the fusion stage due to the complex mapping relationships of the temporal features in different scale spaces. In this article, we propose a novel STF method with a dual-perspective framework, where the core idea is to predict the target information by discriminative learning of the spatial and temporal modeling for estimating heterogeneous temporal changes. Specifically, an encoder–decoder architecture based on a Swin transformer is designed to extract the global context information from the temporal change maps and predict the target image by learning the temporal mapping at a fine scale. A parallel subnetwork is further used to learn the spatial mapping across coarse-to-fine scales, considering the temporal changes. Channel-spatial attention is introduced to guide the model to focus on reconstructing the features delineating heterogeneous textures and temporal changes. The estimation from the dual perspectives is then fused to generate the final reconstructed image. Extensive experiments on three public datasets verified the superiority of the proposed method, compared with the mainstream STF algorithms, especially on image pairs with land-cover and phenological changes.

Index Terms—Deep learning, discriminative learning, remote sensing, spatiotemporal fusion (STF), Swin transformer.

Manuscript received 8 April 2024; revised 21 June 2024; accepted 7 July 2024. Date of publication 11 July 2024; date of current version 24 July 2024. This work was supported in part by the Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources under Grant KF-2023-08-23 and in part by Beijing Nova Program under Grant 20230484351. (*Corresponding authors: Linwei Yue; Xiao Xie.*)

Zhonghua Qian and Linwei Yue are with the Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources, Shenzhen 518034, China, and also with the School of Geography and Information Engineering, China University of Geosciences, Wuhan 430074, China (e-mail: qianzh@cug.edu.cn; yuelw@cug.edu.cn).

Xiao Xie is with the Key Laboratory for Environment Computation and Sustainability of Liaoning Province, Institute of Applied Ecology, Chinese Academy of Sciences, Shenyang 110016, China (e-mail: xiexiao@iae.ac.cn).

Qiangqiang Yuan is with the School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China (e-mail: qqyuan@sgg.whu.edu.cn).

Huanfeng Shen is with the School of Resource and Environmental Sciences, Wuhan University, Wuhan 430079, China (e-mail: shenhf@whu.edu.cn).

Source code is at <https://github.com/zhonghua/STM-STFNet>.

Digital Object Identifier 10.1109/JSTARS.2024.3426944

I. INTRODUCTION

REMOTE sensing images with high spatial and temporal resolutions are generally desirable in a broad range of applications, such as monitoring the fine-scale dynamics of urban resources [1], natural disasters [2], crop growth [3], etc. However, due to the technical limitations and cost constraints, there is always a tradeoff between the spatial and temporal resolutions of an individual satellite sensor. Therefore, the integration of observation information from multisource remote sensing images is necessary for capturing dynamic changes over heterogeneous land surfaces.

Spatiotemporal fusion (STF) of remote sensing images aims to generate a dense time series of high spatial resolution images by integrating the temporally sparse fine-resolution images and temporally dense coarse-resolution images. With the development of STF algorithms over the past decade, it has been shown that STF is an effective solution to the spatiotemporal conflicts in remote sensing images [4]. Massive efforts have been made in the development of STF technology, where the methods can be generally categorized as weight function-based methods, unmixing-based methods, Bayesian-based methods, learning-based methods, and hybrid methods [5].

In recent years, the booming development of deep-learning methods has resulted in impressive achievements in the field of STF [6], and these methods have shown great potential in obtaining a state-of-the-art performance. However, it is still a challenge to achieve the accurate estimation when the images acquired on the reference and target dates have heterogeneous land-cover changes. Until recently, the authors in [7], [8], [9], [10], [11], and [12] have been devoted to proposing various strategies to address the problem. For the deep-learning-based STF framework, the mainstream solutions have been aimed at improving the accuracy of the temporal feature mapping in the different-scale spaces, e.g., through spatiotemporal–spectral collaborative learning [7] or joint spatial and temporal modeling [9]. However, due to the complex relationships of the temporal changes across the different spatial scales, the fine-resolution temporal changes can be difficult to accurately predict without making full use of the inherent relevant information contained among the input data sources. Moreover, the mainstream convolutional neural network (CNN) based methods can only acquire the local dependencies of the image features, which

is not conducive to learning the real patterns of the temporal changes. Therefore, the appropriate modeling of the complementary information among the images is the key to a potential breakthrough.

In this article, we propose a dual-perspective STF method to tackle the challenge of fusing remote sensing images with abrupt land-cover changes. The core idea is to predict the target information by estimating the heterogeneous temporal changes from the dual perspectives of discriminative spatial and temporal modeling. In this way, the temporal difference mapping and the spatial mapping between coarse- and fine-scale space are integrated to infer the target fine-resolution textural information and improve the model performance to deal with images with temporal changes. The estimation results from the dual perspectives are then fused to generate the final reconstructed image. Overall, the main contributions of this article can be summarized as follows.

- 1) We propose a dual-perspective STF framework for remote sensing images with land-cover changes, which discriminatively learns the spatial mapping and temporal transfer features across coarse-to-fine spatial scales for prediction of the target image (referred to as STM-STFNet).
- 2) An encoder–decoder architecture is constructed based on a Swin transformer with a hierarchical structure to extract the global context information from the temporal change map so as to represent the multiscale temporal differences.
- 3) A multidimensional attention module is introduced to learn the spatial-channel weights, guiding the network to focus on the spatial features with heterogeneous textures and land-cover changes while establishing the spatial mapping for the target date considering the temporal changes of the input images.

The rest of this article is organized as follows. Section II summarizes the current approaches in the field of STF. Section III details the network structure of the proposed method. Section IV describes the experimental section, including the presentation and analysis of the results. Section V is the discussion section of the article. Finally, Section VI concludes this article.

II. RELATED WORKS

In this section, we introduce the main works related to STF techniques. In this article, we refer to the STF methods that do not use a learning-based mechanism as the traditional methods. In particular, the algorithms designed for modeling land-cover changes are analyzed.

A. Traditional Methods

Since the proposal of the spatial and temporal adaptive reflectance fusion model (STARFM) [13] in 2006, the weight function-based STF techniques have gained increased attention. STARFM assumes that the reflectance changes between two coarse images can represent the fine-resolution temporal changes and uses the weighted sum of the spatially and spectrally similar information to generate the fusion image. However, this assumption is only valid for the pure pixels and fails in heterogeneous landscapes. Inspired by the idea of STARFM,

advanced weight function-based STF algorithms were subsequently proposed, e.g., the enhanced spatial and temporal adaptive reflectance fusion model (ESTARFM) [14], the spatial and temporal nonlocal filter-based fusion model [15], Fit-FC [16], and other models, considering the physical characteristics in geoscience-related applications [17], [18]. These modified methods are designed to better process the heterogeneous areas with land-cover changes. However, the weight function-based methods construct a linear model for mapping the relationship between the coarse and fine images, and the prediction results are still unsatisfactory for scenes with complex landscapes and heterogeneous temporal changes.

This problem is also an issue for the unmixing-based STF methods. Although the coarse pixels are unmixed at the prediction date to obtain the reflectance change of each land-cover class, the linear mixing model commonly used for the unmixing-based methods [e.g., the multisensor multiresolution technique and the spatial–temporal data fusion approach] might not be well adapted to scenes with nonlinear spectral mixing effects caused by heterogeneous landscapes and distinct temporal changes [19].

The Bayesian-based methods [20], [21] use Bayesian theory to fuse the images in a probabilistic manner. In the Bayesian framework, STF can be considered as a maximum a posteriori problem, where the goal is to estimate the fine image at the prediction date by maximizing the conditional probability with the input fine and coarse image series [22]. However, the prediction accuracy of the Bayesian-based methods depends on the prior constraint on the model.

To combine the advantages of the above methods, hybrid-based methods have been introduced. For example, flexible spatiotemporal data adaptive fusion (FSDAF) [23] integrates the principles of the weighting and unmixing methods and has achieved good results for images with heterogeneous landscapes. With rigorous mathematical theory, the Bayesian methods are popularly combined with the weight function-based and unmixing-based methods [24]. The hybrid approaches enhance the generalization ability of the model by combining multiple methods to cope with the different land-cover changes, but this also increases the complexity of the model, which is not conducive to the large-scale application of the model.

B. Learning-Based Methods

Learning-based methods model the relationships among the temporal coarse images and the known fine image from the external training data samples, and thus predict the unknown fine-resolution information on the target date. Early attempts have been made to use dictionary pair learning [25], [26], regression algorithms [27], and traditional machine learning models (e.g., random forest [28] and extreme learning machine [29]) for STF. Following the shallow learning models starting with features manually extracted from the images, the deep-learning methods have a powerful ability for adaptive feature representation and nonlinear relation mapping, and have advanced the state-of-the-art performance of STF in recent years.

CNNs are the most popular framework used in STF, where the earlier works include STF using deep CNNs [30] and the deep

convolutional STF network [31]. Since then, researchers have attempted to use various strategies to improve the performance of STF methods based on CNNs, e.g., deeper networks, multiscale feature extraction [32], [33], [34], and attention-aware mechanisms [35], [36], [37]. However, the localized property of the convolutional kernels causes CNNs to lack the ability to capture global context information. As a result, the CNN-based models can fail to establish the proper relationship between pixels in local areas with spatially heterogeneous temporal changes. Generative adversarial networks (GANs) have also been introduced in STF [38]. Although the GAN-based methods have achieved good results in processing natural images, they are characterized as being difficult to train and converge to the optimal status, and result in visually pleasant but unrealistic texture details in the reconstructed images.

With increasing popularity in computer vision tasks, transformer networks have also achieved great attention in remote sensing image processing. In terms of STF modeling, Li et al. [39] introduced the vision transformer (ViT) [40] to model the temporal relationship between blocks within the coarse images. In particular, the Swin transformer constructs a hierarchical representation of the feature maps based on a self-attention mechanism computed with shifted windows and has obtained promising performances in various computer vision problems [41]. Chen et al. [42] proposed the Swin spatiotemporal fusion model (SwinSTFM) and conducted multilevel fusion of the extracted features by integrating unmixing theory into the Swin transformer model, which greatly improves the quality of the generated images. Recently, researchers attempted to combine a transformer and a CNN for remote sensing image STF. For example, MSNet [39] is designed with a multistream structure. It uses a transformer to learn the global temporal correlation of the images and employs a CNN to establish the mapping relationship between the input and the output. There are also other works that combine a transformer and a CNN for remote sensing image STF, e.g., STF-Trans [43], EMSNet [44], and MSFusion [45].

In addition to the architecture of the backbone networks, the other popular strategy is to manage the complex mapping relationships with multistream models. Distinguished with the above methods focusing on mapping the temporal [31], [33], [35], [36], [37] or spatial [30], [46], [47] dependencies across image pairs, Jia et al. [9] proposed a two-stream CNN-based STF method that performs the temporal change-based and spatial information-based mapping simultaneously. More recently, the MTDL-STF methods in [48] proposed a novel multitask DL framework that employs a super-resolution net and a fusion net for the spatial information-based mapping and temporal change-based mapping, respectively. The performance is overall promising; however, the methods are only tested on normalized difference vegetation index (NDVI) data. Compared with NDVI, the reflectance images are characterized by more detailed textures and heterogeneous temporal changes. In this method, both spatial and temporal mapping were learned using the CNN-based modules, where the redundant information in the extracted features might prevent efficient modeling on the mapping relationships for the task of image STF [41].

III. METHODOLOGY

A. Overall Framework

The proposed STM-STFNet method is an end-to-end network for fusing remote sensing images with heterogeneous land-cover changes. The input of the proposed method is made up of five images, i.e., the coarse image (C_2) at the target moment (t_2) and two pairs of reference images [i.e., the coarse and fine images at the neighboring moments (C_1, F_1, C_3 , and F_3)]. The output is the fine-resolution image (F_2) at the target moment t_2 . The overall structure of the network is shown in Fig. 1.

The basic assumption is that the fine-resolution details can be predicted from two perspectives, i.e., spatial reconstruction of the coarse image at the target date or temporal estimation using the change maps from multiple coarse images. Accordingly, STM-STFNet consists of two dual streams, which focus on temporal modeling with a Swin transformer and spatial modeling with a multidimensional attention mechanism. For the temporal mapping, the Swin transformer-based encoder–decoder module is utilized to learn the fine-resolution change map. The modeling process can be expressed as follows:

$$\widehat{F}_2^{t,1} = M_{t,\Phi_t} (F_1, C_1, C_2) + F_1 \quad (1)$$

where $\widehat{F}_2^{t,1}$ is the result estimated from the perspective of temporal modeling through forward prediction with the aid of F_1 , C_1 , and C_2 . M_t denotes the branch for mapping the temporal relationships between imaging dates t_1 and t_2 , and Φ_t is the corresponding parameter set learned by the subnetwork.

The dual-branch structure is constructed from the perspective of spatial modeling, which serves to reconstruct the spatial details of the target image by learning the relationships between the low-frequency and high-frequency feature space. The basic assumption is that the spatial difference information at t_2 , which is expressed as $F_2 - C_2$, can be inferred from the appropriate modeling of the known prior from $F_1 - C_1$, as well as the temporal difference $C_2 - C_1$. The process can be expressed as follows:

$$\widehat{F}_2^{s,1} = M_{s,\Phi_s} (C_2 - C_1, F_1 - C_1, C_2) + C_2 \quad (2)$$

where $\widehat{F}_2^{s,1}$ is the spatial modeling result. M_s and Φ_s denote the mapping function and parameter set for the spatial modeling branch, respectively. The ultimate goal is to reconstruct the fine-resolution image details with full integration of the reference information, considering the complex landscapes and heterogeneous temporal differences among the input data. The forward prediction with the reference images acquired on t_1 is obtained by combining the dual-branch estimations

$$\widehat{F}_2^1 = \text{Average} \left(\widehat{F}_2^{t,1}, \widehat{F}_2^{s,1} \right) \quad (3)$$

In the symmetric part of the network, similar strategies are used to obtain the backward prediction with the reference images acquired on t_3 , i.e., $\widehat{F}_2^{s,3}$, $\widehat{F}_2^{t,3}$, and \widehat{F}_2^3 . The final reconstructed

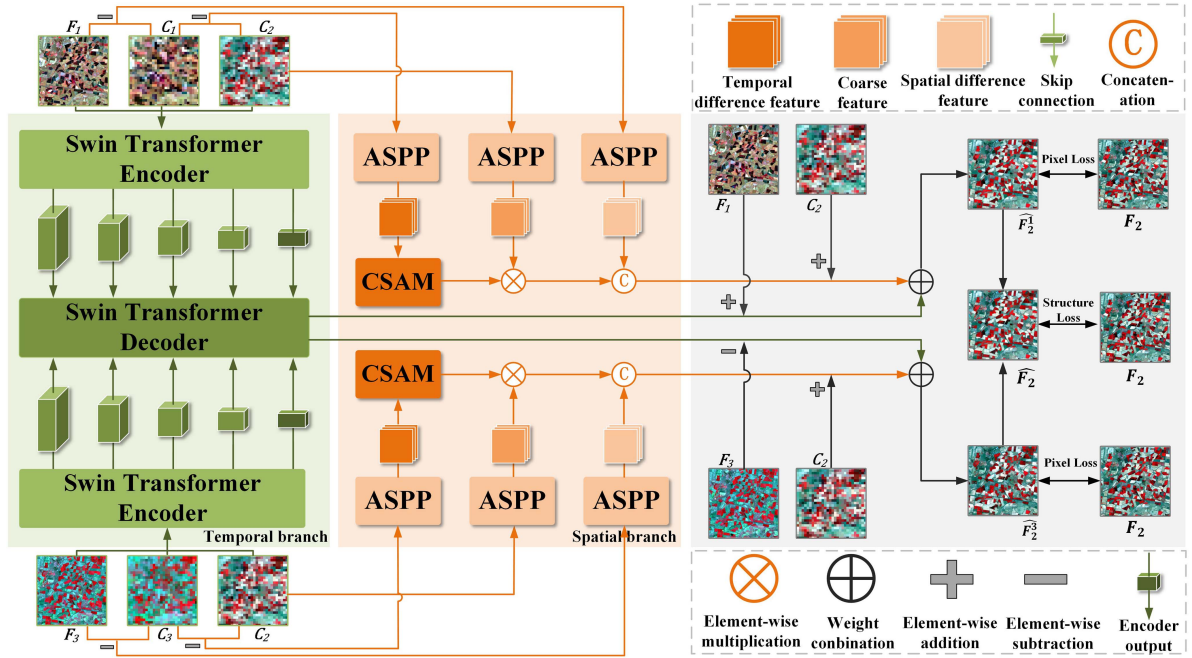


Fig. 1. Architecture of the proposed STM-STFNet method. For the temporal branch, the input is the stacked tensor of the reference images (i.e., F_1 , C_1 , and C_2 for the forward modeling and F_3 , C_3 , and C_2 for the backward modeling). For the spatial branch, the inputs are the temporal difference between the target moment and the reference moment (i.e., $C_2 - C_1$ and $C_2 - C_3$), the spatial difference between the coarse and fine images at the reference moment (i.e., $F_1 - C_1$ and $F_3 - C_3$), and the coarse image at the target moment (C_2).

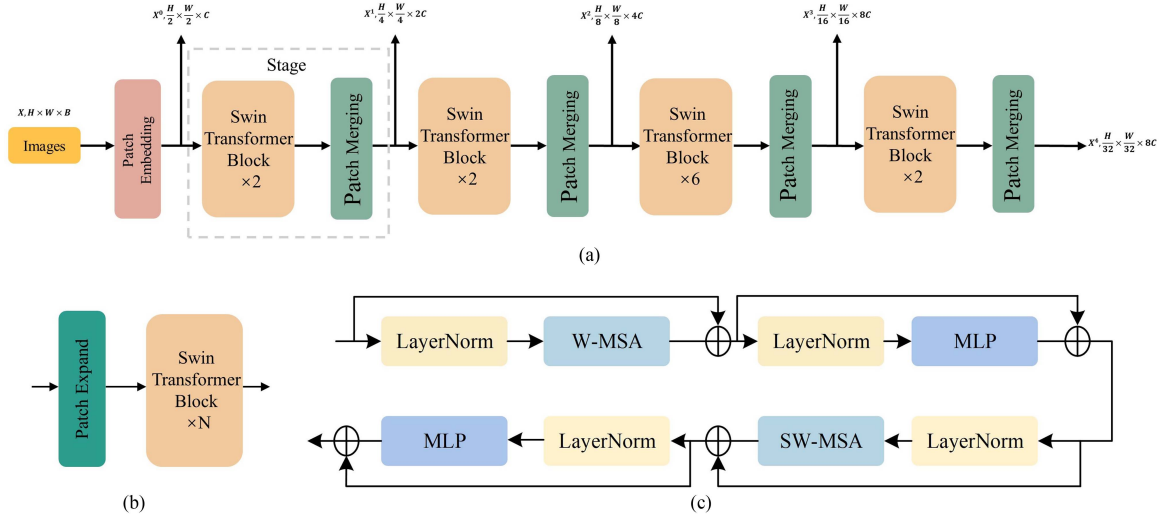


Fig. 2. Flowchart of the Swin transformer-based encoder. (a) Architecture of the encoder is made up of four stages, where each stage in (b) contains N Swin transformer blocks ($N_i \in [2, 6, 2, 2]$, $i = 1, 2, 3, 4$). (c) Structure of two successive Swin transformer blocks, where W-MSA and SW-MSA are alternately used.

image \widehat{F}_2 is then calculated by

$$\widehat{F}_2 = \text{Average}(\widehat{F}_2^1, \widehat{F}_2^3). \quad (4)$$

B. Swin Transformer-Based Encoder-Decoder for Temporal Modeling

To deal with temporal changes, we designed a Swin transformer-based encoder-decoder network (as shown in Fig. 2) for modeling the temporal relationships between neighboring dates across different spatial scales. As shown in (1),

the input for the forward and backward prediction module is the feature set constructed from the three reference images (i.e., F_1 , C_1 , and C_2 for the forward prediction and F_3 , C_3 , and C_2 for the backward prediction) concatenated along the channel dimension. With the powerful capability of the Swin transformer for local-to-global modeling, the encoder-decoder network learns the temporal transfer information from the input data, where no additional modeling assumptions are required [49], [50].

Following the instructions in [41], the encoder consists of four stages, where each stage is composed of an even number of consecutive Swin transformer blocks, followed by a patch

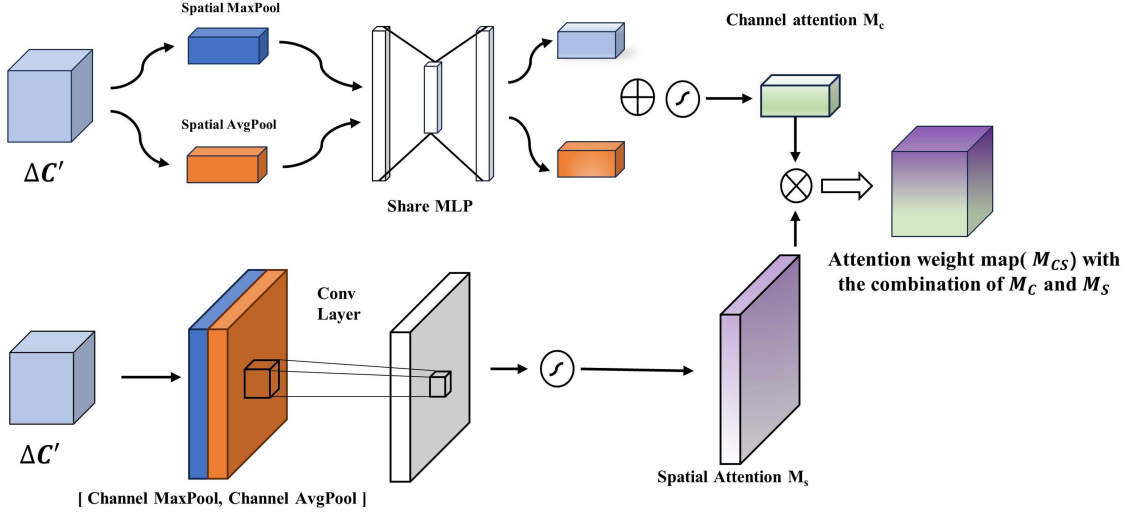


Fig. 3. Architecture of the CSAM module.

merging process. The Swin transformer block is mainly used to calculate the self-attention, and the patch merging layer reduces the resolution of the input feature map for downsampling. The four stages constitute a hierarchical structure, which reduces the feature resolution and increases the receptive field layer-by-layer to obtain the global information.

The core component of the Swin transformer block is the multihead self-attention (MSA) module, which adaptively learns the relationships between different regions of the input. Specifically, the whole image is divided into multiple nonoverlapping windows in the self-attention calculation, where each window contains $M \times M$ patches ($M = 8$ here) with the size of 2×2 . The image patches are then used as the input of several stacked transformer blocks. The self-attention within the local window is calculated by introducing the relative position information

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V \quad (5)$$

$$Q = XW_Q, K = XW_K, V = XW_V \quad (6)$$

where $X \in \mathbb{R}^{n \times d}$ is the vector matrix of the input patches, while n is the number of patches, and d is the feature dimension. The query Q , key K , and value V matrices are obtained by multiplying X and the parameter matrices W_Q , W_K , and $W_V \in \mathbb{R}^{d \times d}$, respectively. Q and K are then used to obtain the normalized weights, and then multiplied with V to obtain the feature of interest, as shown in (5). B is the relative position bias, which denotes the relative position between patches within a single window. While the window partitioning scheme of the Swin transformer greatly reduces the computational complexity, compared with the traditional ViT, the model adopts a shifted window scheme to realize the cross-window fusion of patch features. In the successive Swin transformer blocks, local window MSA (W-MSA) and shifted window MSA (SW-MSA) are alternately used in the successive Swin transformer blocks. In this way, the self-attention computation in the new windows partitioned in

layer $l + 1$ crosses the boundaries of the previous windows in layer l , and thus provides connections across the local windows. Moreover, the patch merging layer is also applied to merge adjacent patches and further increase the receptive field.

Each MSA module is followed by a two-layer multilayer perceptron (MLP). Moreover, a LayerNorm (LN) layer is employed before each MSA module and each MLP. The computation process can be summarized as follows:

$$\hat{z}^l = W - \text{MSA}(\text{LN}(z^{l-1})) + z^{l-1} \quad (7)$$

$$z^l = \text{MLP}(\text{LN}(\hat{z}^l)) + \hat{z}^l \quad (8)$$

$$\hat{z}^{l+1} = \text{SW} - \text{MSA}(\text{LN}(z^l)) + z^l \quad (9)$$

$$z^{l+1} = \text{MLP}(\text{LN}(\hat{z}^{l+1})) + \hat{z}^{l+1} \quad (10)$$

where \hat{z}^l and z^l denote the output features of the W-MSA module and the MLP module for block l , respectively.

The Swin transformer decoder is the inverse process of the Swin transformer encoder, which serves to recover the global temporal difference features to the input resolution and fuse the features at different scales from the encoder through a skip connection. Differing from the encoder structure, patch merging is replaced by a patch expanding layer. The Swin transformer block in the decoder also helps to establish long-range dependencies and global context connections during upsampling.

C. Multidimensional Attention Module for Spatial Modeling

The goal of spatial modeling for STF is to reconstruct the spatial details of the target moment, where an attention-based module (as shown in Fig. 3) is designed to tackle the challenge of predicting the unknown high-frequency information in the areas with complex image textures and abrupt changes between the reference and target dates. For the spatial modeling, atrous spatial pyramid pooling (ASPP) [51] is first used to extract multiscale features from the input feature maps. Taking the forward prediction as an example, the inputs for the spatial

branch are three image maps, i.e., $C_2 - C_1$, $F_1 - C_1$, and C_2 (2). To predict the unknown $F_2 - C_2$ using the prior information from $F_1 - C_1$, the basic assumptions are as follows.

- 1) The changed areas between the reference and target dates with significant textural differences should be paid more attention.
- 2) In the regions with more detailed textures in the images, the spatial differences can be more informative.

Accordingly, the features extracted from the temporal difference map $C_2 - C_1$ (i.e., $\Delta C'$) are input into the channel-spatial attention module (CSAM) to obtain the attention weight, with full consideration of the spatial and temporal importance. The process can be expressed as follows:

$$M_{cs}(\Delta C') = M_c(\Delta C') \otimes M_s(\Delta C') \quad (11)$$

where \otimes denotes the elementwise multiplication, while $M_c \in \mathbb{R}^{C \times 1 \times 1}$ and $M_s \in \mathbb{R}^{1 \times H \times W}$ denote the channel and spatial attention maps, respectively. In this module, the input feature maps are processed in parallel with the channel and spatial attention calculation. Following the attention module introduced in [52], the spatial attention map is calculated by

$$M_s(\Delta C') = \sigma(f^{7 \times 7}([\Delta C'_{Avg}; \Delta C'_{Max}])) \quad (12)$$

where σ denotes the sigmoid function, and $f^{7 \times 7}$ represents a convolution operation with the filter size of 7×7 . $[\]$ represents the concatenation operation. $\Delta C'_{Avg}$ and $\Delta C'_{Max}$ denote the average-pooled features and max-pooled features across the channels, respectively. They are used to aggregate the channel information of $\Delta C'$ and highlight the spatially heterogeneous information. Furthermore, the incorporation of channel attention is based on the fact that the surface changes can be distinct for different spectral bands (e.g., phenological changes). For the calculation of the channel attention map, the channelwise spatial information is first aggregated using the average-pooling and max-pooling operations, and the two spatial feature descriptors are generated, i.e., $\Delta C'_{Avg}$ and $\Delta C'_{Max}$. The two descriptors are then fed into a shared MLP with one hidden layer, and the channel attention map can be obtained by merging the output feature vector using elementwise summation. The calculation can be expressed as follows:

$$M_c(\Delta C') = \sigma(\text{MLP}(\Delta C'_{Avg}) + \text{MLP}(\Delta C'_{Max})) \quad (13)$$

After the channel and spatial attention maps are obtained, the multidimensional map $M_{CS} \in \mathbb{R}^{C \times H \times W}$ can be obtained with elementwise multiplication, and the refined features can be obtained by multiplying the multiscale features extracted from $F_1 - C_1$ (referring to $\Delta F_1 C'_1$) with M_{CS} . The mapping relationships are then established, as follows:

$$F_2 = f\left(\left[M_{cs}(\Delta C') \otimes (\Delta F_1 C'_1); C'_2\right]\right) + C_2 \quad (14)$$

where f represents the convolution operation. C'_2 indicates the feature set extracted from image C_2 using ASPP, which it is concatenated with $\Delta F_1 C'_1$ and refined with channel-spatial attention to form the input of the convolutional mapping function.

D. Loss Function

The loss function consists of two loss terms for optimizing the forward and backward modeling: a structural loss to preserve the image textures and an edge loss focusing on enhancing the high-frequency details in the reconstructed result. The Charbonnier loss function [53] is used for pixel-level image supervision. Moreover, the structural similarity (SSIM) method [54] is introduced for SSIM supervision. The specific formulae for the loss function formula used in the proposed method are as follows:

$$L_{\text{structure}} = 1 - \text{SSIM}(\widehat{F}_2, F_2) \quad (15)$$

$$L_{\text{pixel1}} = \frac{1}{N} \sqrt{(\widehat{F}_2^1 - F_2)^2 + \varepsilon^2} \quad (16)$$

$$L_{\text{pixel2}} = \frac{1}{N} \sqrt{(\widehat{F}_2^3 - F_2)^2 + \varepsilon^2} \quad (17)$$

$$L_{\text{edge}} = \frac{1}{N} \sqrt{(\text{Sobel}(\widehat{F}_2) - \text{Sobel}(F_2))^2 + \varepsilon^2} \quad (18)$$

$$L_{\text{total}} = L_{\text{pixel1}} + L_{\text{pixel2}} + L_{\text{structure}} + L_{\text{edge}} \quad (19)$$

where \widehat{F}_2^1 is the estimated fine-resolution image with forward prediction using F_1 , C_1 , and C_2 , while \widehat{F}_2^3 denotes the corresponding backward prediction result. The parameter $\varepsilon = 0.001$ is used to stabilize the process of error backpropagation. N denotes the total number of pixels in the image. The edge loss is implemented on the edge information extracted by the Sobel operator.

IV. EXPERIMENTS

A. Experimental Datasets

Three public datasets were used to validate the proposed method: the lower Gwydir catchment (LGC) dataset [55], the Coleambally irrigation area (CIA) dataset [55], and the Daxing (DX) [4] dataset. The three datasets were collected from different spatial regions and feature phenological and land-cover changes. As such, they can be considered as representative to compare the STF methods' performance in different cases.

The LGC dataset was collected in southern New South Wales (NSW) in Australia and consists of 14 pairs of cloud-free Landsat-moderate resolution imaging spectroradiometer (MODIS) images from 16 April 2004 to 3 April 2005, where the size of each image is $3200 \times 2720 \times 6$. The Landsat images were acquired by the thematic mapper (TM) sensor of Landsat 5, and the MODIS images were acquired by the Terra satellite. This dataset is typically characterized by abrupt land-cover changes due to flooding in the area in mid-December 2004.

The CIA dataset was collected in the northern part of NSW in rice-growing areas with modern irrigation systems. It consists of 17 cloud-free Landsat-MODIS data pairs acquired between 8 October 2001 and 4 May 2002. The Landsat images were acquired by the enhanced thematic mapper plus (ETM+) sensor onboard Landsat 7, and the MODIS images were acquired by the Terra satellite. The image size is $1720 \times 2040 \times 6$. There was no significant land-cover change in this area during the data

TABLE I
DETAILS OF THE TEST IMAGES FOR EACH DATASET

| Dataset | Reference date | | Prediction date |
|---------|----------------|--------------------------|-----------------|
| | One pair | Two pairs | |
| LGC | 2004/11/26 | 2004/11/26 2004/12/28 | 2004/12/12 |
| CIA | 2001/11/25 | 2001/11/25 2002/02/22 | 2002/01/12 |
| DX | 2017/05/07 | 2017/05/07 2017/11/15 | 2017/09/12 |

collection period; however, phenological change can be clearly observed in the farmland regions.

The DX dataset collected in the DX area of Beijing in China features both phenological and land-cover changes. This dataset consists of 27 cloud-free Landsat–MODIS data pairs collected during the period of September 01, 2013 to November 05, 2019. Landsat 8 operational land imager and Terra MODIS images were used, and the size of each image was $1640 \times 1640 \times 6$. Due to the obvious band noise in the fifth and sixth bands of this dataset, only the first four bands of the DX dataset were selected.

B. Experimental Design and Evaluation

In the experiments, each dataset was divided into a training set and a test set. The details of the test images for each dataset are provided in Table I, while the remaining image pairs were used for the model training. The proposed STM-STFNet method uses two reference image pairs as the input.

We chose five representative STF methods for comparison: ESTARFM [14], FSDAF [23], the GAN-based spatiotemporal fusion model (GAN-STFM) [46], spatiotemporal fusion method using a recurrent neural network (STFRNN) [7], and SwinSTFM [42]. Among these methods, ESTARFM and FSDAF belong to the traditional methods, while the other three methods are based on deep-learning frameworks. Except for ESTARFM and STFRNN, which require two reference image pairs as the input, the other comparative methods use only one image pair as the reference and obtain the fusion image in the manner of forward prediction. We slightly modified the structure of the SwinSTFM algorithm to support the input of two reference image pairs, which is referred to here as SwinSTFM-B.

For a fair comparison, the hyperparameters of the methods were empirically adjusted to achieve the optimal results. The size of the sliding window for searching for similar pixels was uniformly set to 41×41 for ESTARFM and FSDAF, while the number of similar pixels was set to 20. For the deep-learning methods, the patch size of GAN-STFM was set to 128×128 with a stride length of 100. The other parameters were determined following the original design. In terms of the proposed STM-STFNet, the input patch size was 128×128 , and the epochs were set to 80. The initial learning rate was 1×10^{-4} , and the learning rate was dropped by half when the accuracy of the test set no longer increased for three epochs.

In addition to the visual examination, the quantitative performance of the methods is evaluated here through six evaluation metrics: root-mean-square error (RMSE), SSIM [54], universal image quality index (UIQI) [56], correlation coefficient (CC), relative global synthesis error (ERGAS) [57], and spectral angle mapper (SAM) [58]. The formulae for these metrics are as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N}} \quad (20)$$

$$\text{SSIM} = \frac{2\mu_{\hat{y}}\mu_y + a_1}{\mu_{\hat{y}}^2 + \mu_y^2 + a_1} \cdot \frac{2\sigma_{\hat{y}y} + a_2}{\sigma_{\hat{y}}^2 + \sigma_y^2 + a_2} \quad (21)$$

$$\text{UIQI} = \frac{\sigma_{\hat{y}y} 2\mu_{\hat{y}}\mu_y 2\sigma_{\hat{y}}\sigma_y}{\sigma_{\hat{y}}\sigma_y\mu_{\hat{y}}^2 + \mu_y^2\sigma_{\hat{y}}^2 + \sigma_y^2} \quad (22)$$

$$\text{CC} = \frac{\sum_{i=1}^N (y_i - \mu_y)(\hat{y}_i - \mu_{\hat{y}})}{\sqrt{\sum_{i=1}^N (y_i - \mu_y)^2} \sqrt{\sum_{i=1}^N (\hat{y}_i - \mu_{\hat{y}})^2}} \quad (23)$$

$$\text{ERGAS} = 100 \frac{h}{l} \sqrt{\frac{1}{M} \sum_{i=1}^{N_B} \left(\frac{\text{RMSE}_i}{\mu_i} \right)^2} \quad (24)$$

$$\text{SAM} = \frac{1}{N} \sum_{i=1}^N \cos^{-1} \frac{\sum_{j=1}^{N_B} y_i^j \hat{y}_i^j}{\sqrt{\sum_{j=1}^{N_B} (y_i^j)^2} \sqrt{\sum_{j=1}^{N_B} (\hat{y}_i^j)^2}} \quad (25)$$

where N denotes the total number of pixels in the image. y_i is the i th pixel in the reference image, and \hat{y}_i denotes the value of the predicted image. Moreover, $\mu_{\hat{y}}$ and μ_y denote the mean value of the predicted image and real image, respectively. $\sigma_{\hat{y}y}$ denotes the covariance of the predicted and real image, and $\sigma_{\hat{y}}$ and σ_y denote the variances of the predicted image and real image, respectively. a_1 and a_2 are the constants used to maintain stability. h and l denote the spatial resolution of the fine and coarse image, respectively. N_B denotes the number of bands. μ_i denotes the mean of the i th band of the reference image. Larger values of SSIM, UIQI, and CC, and smaller values of RMSE, ERGAS, and SAM represent the better results.

C. Fusion Results for the LGC Dataset

Table II provides the quantitative comparison of the fusion results obtained by the different STF methods for the LGC dataset. The target is to reconstruct the fine-resolution image from 12 December 2004, where the real Landsat image was used as the ground truth. The main challenge was to model the dramatic land-cover changes caused by the sudden flooding that occurred during the period between the reference and target dates in this area. Overall, the results show that the deep-learning methods outperform the traditional methods. The proposed STM-STFNet method achieves the best results in terms of all the indices in this case. Among the comparative methods, SwinSTFM shows a good performance, and SwinSTFM-B with two pairs of reference images as the input shows an improvement over the original version. Although ESTARFM and STFRNN also include two image pairs as the reference, the unsatisfactory results indicate

TABLE II
QUANTITATIVE EVALUATION RESULTS OF THE DIFFERENT METHODS FOR THE LGC DATASET

| Conference | Method | RMSE | SSIM | UIQI | CC | ERGAS | SAM |
|------------|------------|---------------|---------------|---------------|---------------|---------------|----------------|
| One pair | FSDAF | 0.0354 | 0.7349 | 0.7086 | 0.7364 | 2.2122 | 17.0154 |
| | GAN-STFM | 0.0321 | 0.7621 | 0.7366 | 0.7536 | 1.9686 | 15.9986 |
| | SwinSTFM | 0.0289 | 0.8058 | 0.7989 | 0.8006 | 1.7781 | 14.7437 |
| Two pairs | ESTARFM | 0.0363 | 0.7403 | 0.7080 | 0.7137 | 2.2870 | 17.7563 |
| | STFRNN | 0.0298 | 0.8067 | 0.7665 | 0.7729 | 1.8361 | 15.2593 |
| | SwinSTFM-B | 0.0270 | 0.8254 | 0.8186 | 0.8223 | 1.6646 | 13.6926 |
| | STM-STFNet | 0.0264 | 0.8298 | 0.8229 | 0.8308 | 1.6184 | 13.3486 |

The best results are highlighted in bold.

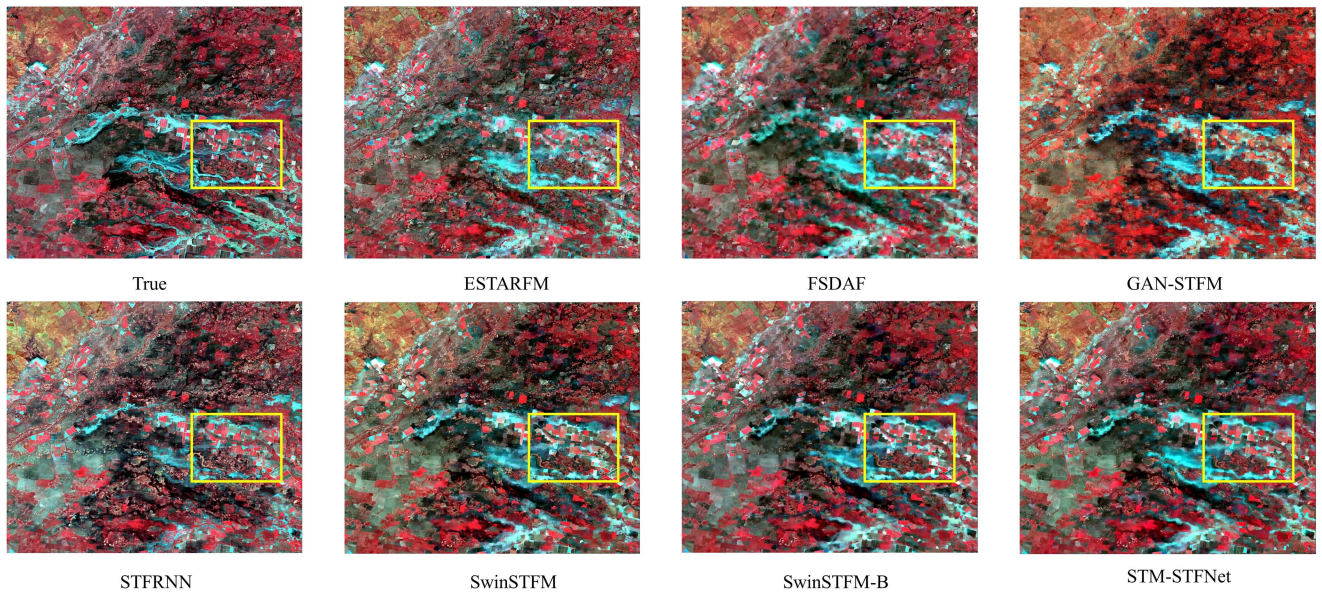


Fig. 4. Fusion images for 12 December 2004 obtained by the different methods with the LGC dataset.

that the accurate modeling is the primary element required to reconstruct a high-accuracy result.

The fusion results for the LGC dataset on 12 December 2004 were visually compared with the real reference image. It can be clearly observed in Fig. 4 that GAN-STFM produces serious spectral distortion in the results. Moreover, all the STF methods are inevitably influenced by the textures on the reference dates and generate fake textures in the reconstructed images, especially in flooded areas with dramatic land-cover changes. However, the zoomed-in views of a typical flooded region, as shown in Fig. 5, indicate that the comparative methods show different performances in reconstructing the image details. There are obvious artifacts and grid effects in the results of ESTARFM, STFRNN, and GAN-STFM, which are probably caused by the processing within local windows (or receptive fields) in the modeling process. Compared with SwinSTFM and SwinSTFM-B, the proposed STM-STFNet method is less affected by the reference images and is capable of reconstructing natural features in the fusion image. In Fig. 6, we present the absolute differences between the fusion images and the

reference, where the error distribution maps show that large errors generally exist in the results of ESTARFM, FSDAF, GAN-STFM, and STFRNN. This is basically consistent with the visual examination, where the fusion images are contaminated with artifacts and blurring effects. The proposed STM-STFNet obtains the best results among all the methods, especially over the boundaries along the flooded fields.

D. Fusion Results for the CIA Dataset

Table III provides the quantitative evaluation results of the different fusion methods for the CIA dataset. The target was to reconstruct the fine image from 12 January 2002. Since no significant land-cover changes occurred in the CIA dataset, the quality of the fusion images is much better than that for the LGC dataset. However, the spectral distortion brought by the phenological changes brings challenges to the STF methods. Similar to the results for LGC dataset, the SwinSTFM-B achieves an overall good performance. However, the proposed STM-STFNet again obtains the best accuracy scores, which verifies the effectiveness

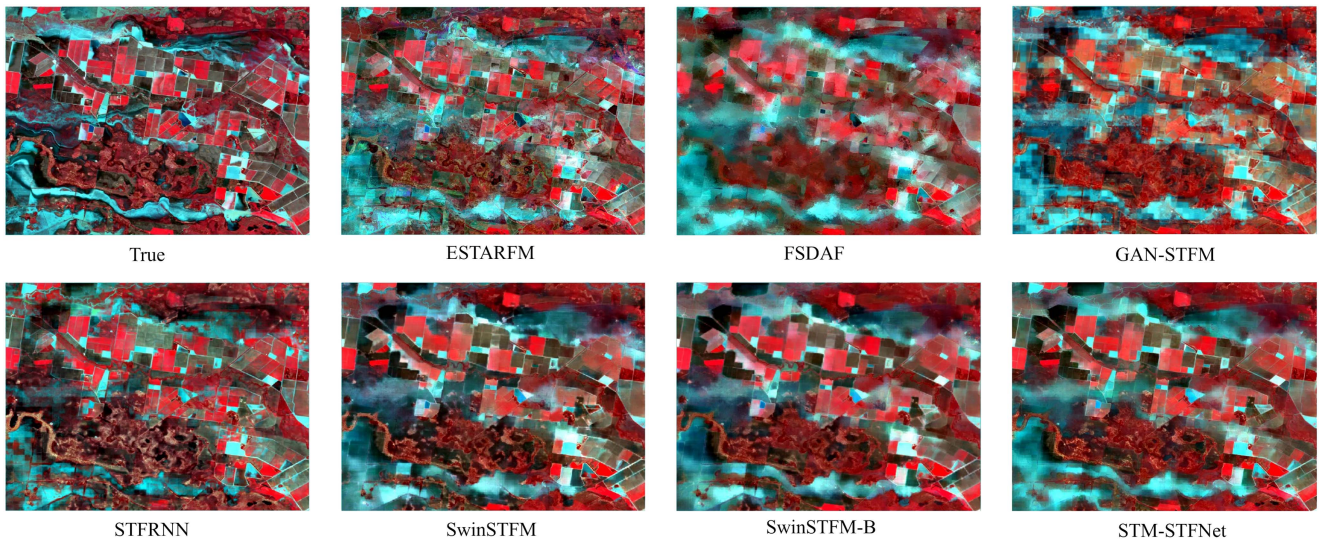


Fig. 5. Subregions of the fusion images for 12 December 2004 obtained by the different methods with the LGC dataset.

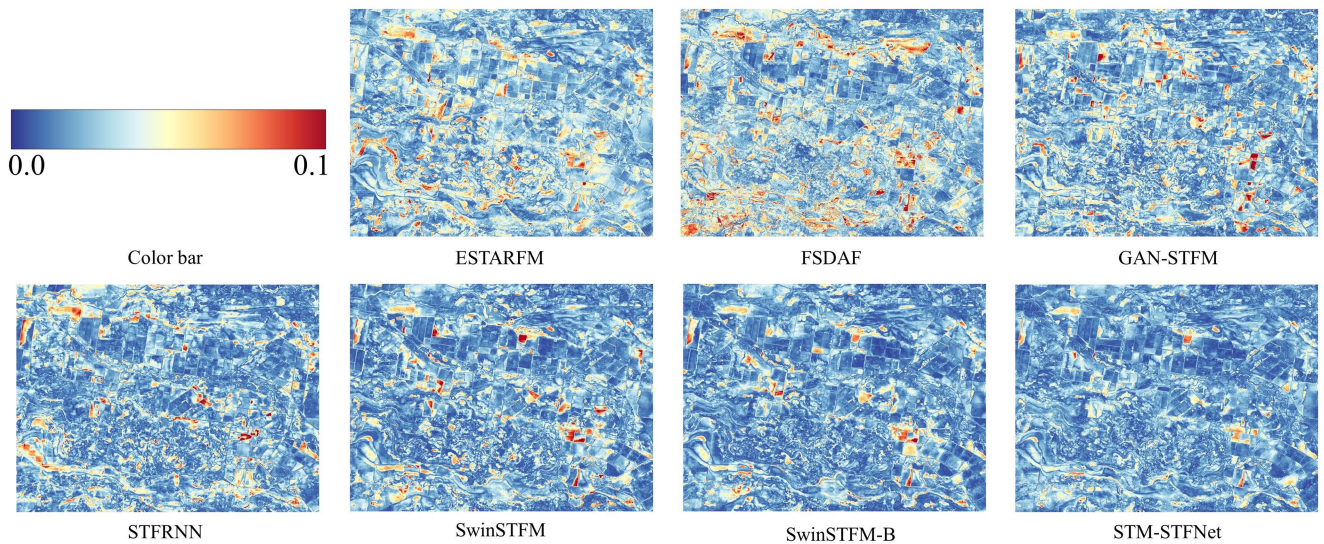


Fig. 6. Average absolute difference maps of the subregions within the fusion images for 12 December 2004 with the LGC dataset.

TABLE III
QUANTITATIVE EVALUATION RESULTS OF THE DIFFERENT METHODS FOR THE CIA DATASET

| Conference | Method | RMSE | SSIM | UIQI | CC | ERGAS | SAM |
|------------|------------|---------------|---------------|---------------|---------------|---------------|---------------|
| One pair | FSDAF | 0.0382 | 0.7406 | 0.7372 | 0.7677 | 1.3119 | 11.5507 |
| | GAN-STFM | 0.0357 | 0.7818 | 0.8094 | 0.8169 | 1.1986 | 10.4628 |
| | SwinSTFM | 0.0242 | 0.8644 | 0.9178 | 0.9190 | 0.8029 | 7.0830 |
| Two pairs | ESTARFM | 0.0344 | 0.7959 | 0.8230 | 0.8330 | 1.1351 | 10.0101 |
| | STFRNN | 0.0269 | 0.8627 | 0.8995 | 0.8935 | 0.9064 | 7.7886 |
| | SwinSTFM-B | 0.0222 | 0.8815 | 0.9312 | 0.9333 | 0.7502 | 6.3826 |
| | STM-STFNet | 0.0211 | 0.8946 | 0.9377 | 0.9394 | 0.7155 | 6.1024 |

The best results are highlighted in bold.

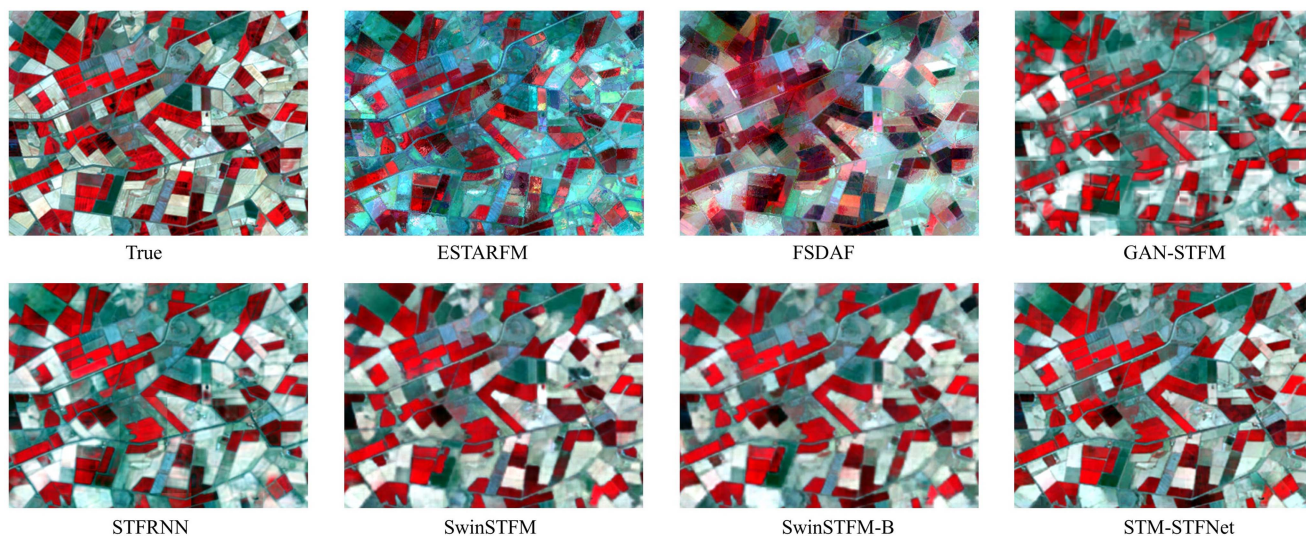


Fig. 7. Subregions of the fusion images of the different methods for 12 January 2002 with the CIA dataset.

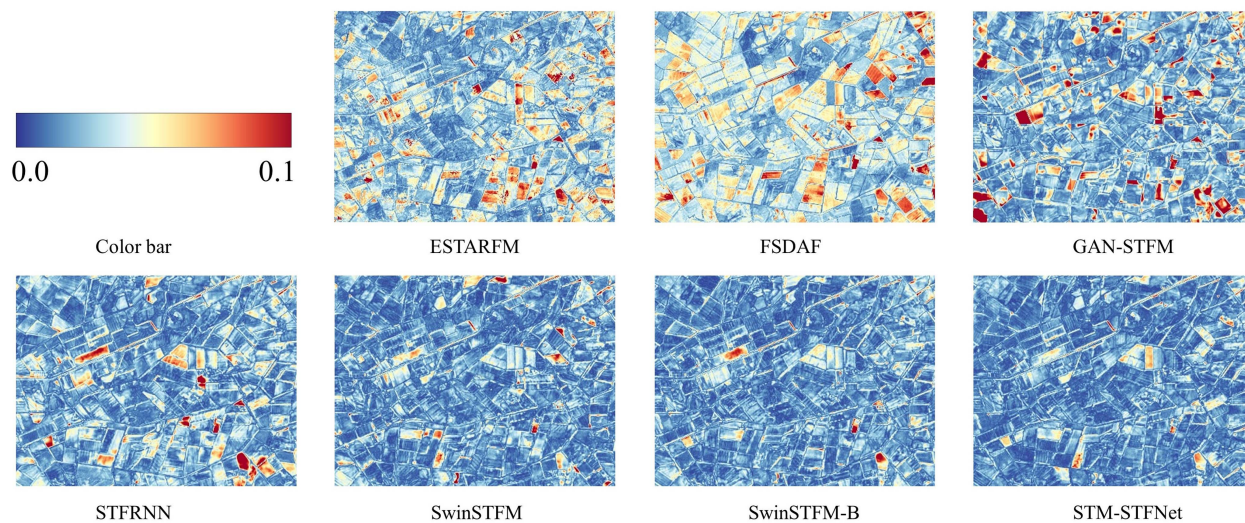


Fig. 8. Average absolute difference maps of the subregions within the fusion images for 12 January 2002 with the CIA dataset.

of the proposed framework with a dual-perspective learning mechanism.

We further present the visual images obtained by the STF methods for a typical farmland region in Fig. 7, and the absolute difference maps are shown in Fig. 8. The results show that FSDAF shows the most significant spectral distortion, which is consistent with the SAM values in Table III and the error maps in Fig. 8. Moreover, ESTARFM and GAN-STFM also introduce obvious spectral bias into the fusion results, which can further influence the interpretation accuracy. The other four deep-learning-based methods show a good performance in this case. STM-STFNet achieves better results in terms of spectral fidelity, compared with STFRNN, and outperforms SwinSTFM and SwinSTFM-B when considering the reconstruction of fine-scale spatial details. Overall, the fusion image obtained by the proposed STM-STFNet method has the best quality from the perspective of both the spatial reconstruction and spectral fidelity.

The results also indicate that the proposed method performs well in processing image scenes with phenological changes.

E. Fusion Results for the DX Dataset

The DX dataset is a relatively new dataset composed of images collected during 2013–2017, where the time span is much longer than that of the LGC and CIA datasets. As noted in Section IV, the test images feature both phenological and land-cover changes, which can effectively evaluate the generalization ability of the proposed method. Table IV provides a comparison of the quantitative results of the different fusion methods for the DX dataset. The target was to reconstruct the fine image acquired on 12 September 2017. The visual performance of the comparative methods over a subregion containing DX airport is shown in Figs. 9 and 10. The results show that STFRNN generates a relatively dark image, where the spectral bias is also reflected in

TABLE IV
QUANTITATIVE EVALUATION RESULTS OF THE DIFFERENT METHODS FOR THE DX DATASET

| Conference | Method | RMSE | SSIM | UIQI | CC | ERGAS | SAM |
|------------|------------|---------------|---------------|---------------|---------------|---------------|----------------|
| One pair | FSDAF | 0.0335 | 0.7274 | 0.6931 | 0.7030 | 2.2733 | 18.3086 |
| | GAN-STFM | 0.0306 | 0.7760 | 0.7131 | 0.7517 | 2.1396 | 16.6724 |
| | SwinSTFM | 0.0280 | 0.7960 | 0.7868 | 0.8065 | 1.8946 | 14.7442 |
| Two pairs | ESTARFM | 0.0304 | 0.7753 | 0.7813 | 0.7826 | 1.9016 | 15.3557 |
| | STFRNN | 0.0293 | 0.7739 | 0.7576 | 0.7834 | 1.9273 | 15.2644 |
| | SwinSTFM-B | 0.0264 | 0.8105 | 0.8062 | 0.8330 | 1.7309 | 13.3932 |
| | STM-STFNet | 0.0244 | 0.8362 | 0.8379 | 0.8583 | 1.5266 | 12.2901 |

The best results are highlighted in bold.

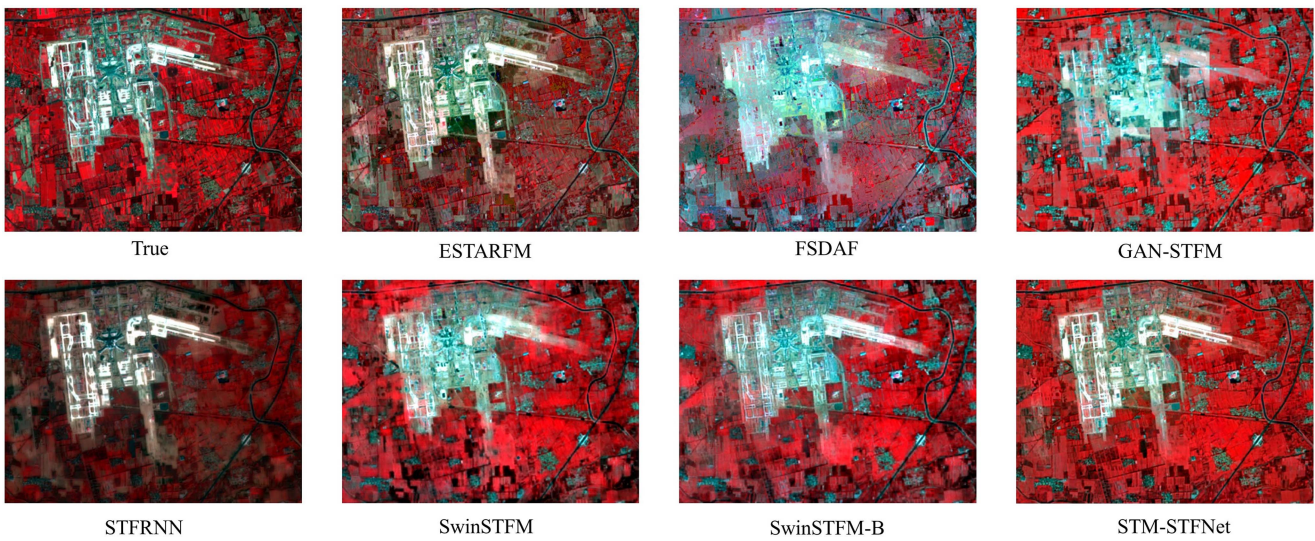


Fig. 9. Subregions of the fusion images of the different methods for 12 September 2017 with the DX dataset.

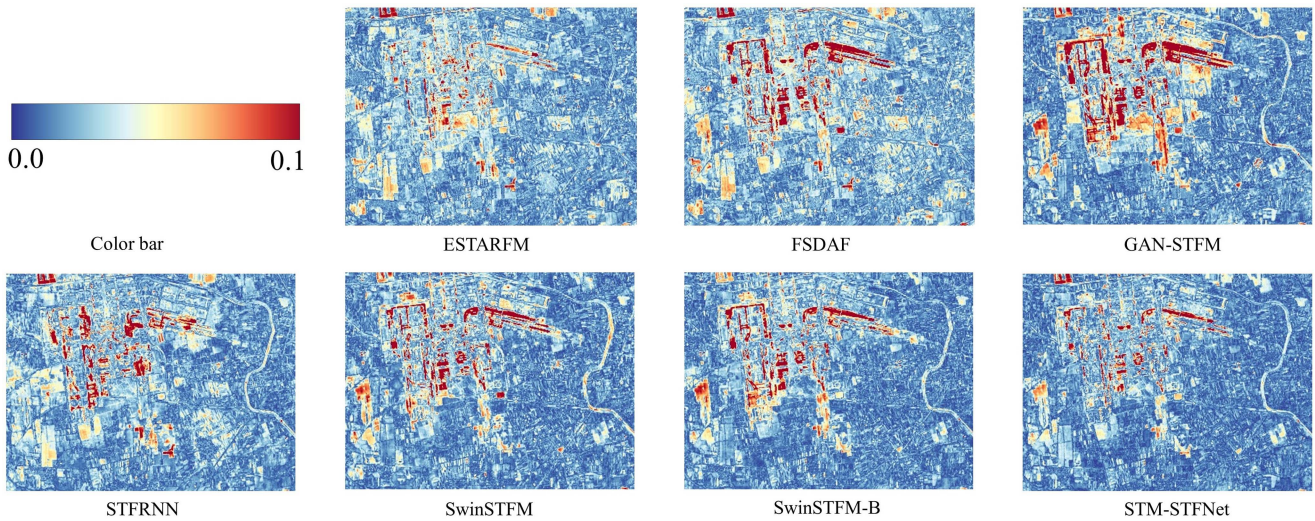


Fig. 10. Average absolute difference maps of the subregions within the fusion images for 12 September 2017 with the DX dataset.

TABLE V
QUANTITATIVE EVALUATION RESULTS OF THE DIFFERENT METHODS FOR THE CIA AND DX DATASETS

| Dataset | Method | RMSE | SSIM | UIQI | CC | ERGAS | SAM |
|---------|---------------------|---------------|---------------|---------------|---------------|---------------|----------------|
| CIA | w/o spatial branch | 0.0217 | 0.8918 | 0.9345 | 0.9366 | 0.7391 | 6.2338 |
| | w/o temporal branch | 0.0318 | 0.8099 | 0.8196 | 0.8426 | 1.1031 | 9.6539 |
| | STM-STFNet | 0.0211 | 0.8946 | 0.9377 | 0.9394 | 0.7155 | 6.1024 |
| DX | w/o spatial branch | 0.0251 | 0.8301 | 0.8284 | 0.8519 | 1.6270 | 12.5680 |
| | w/o temporal branch | 0.0263 | 0.8174 | 0.8017 | 0.8348 | 1.6409 | 13.0306 |
| | STM-STFNet | 0.0244 | 0.8362 | 0.8379 | 0.8583 | 1.5266 | 12.2901 |

The best results are highlighted in bold.

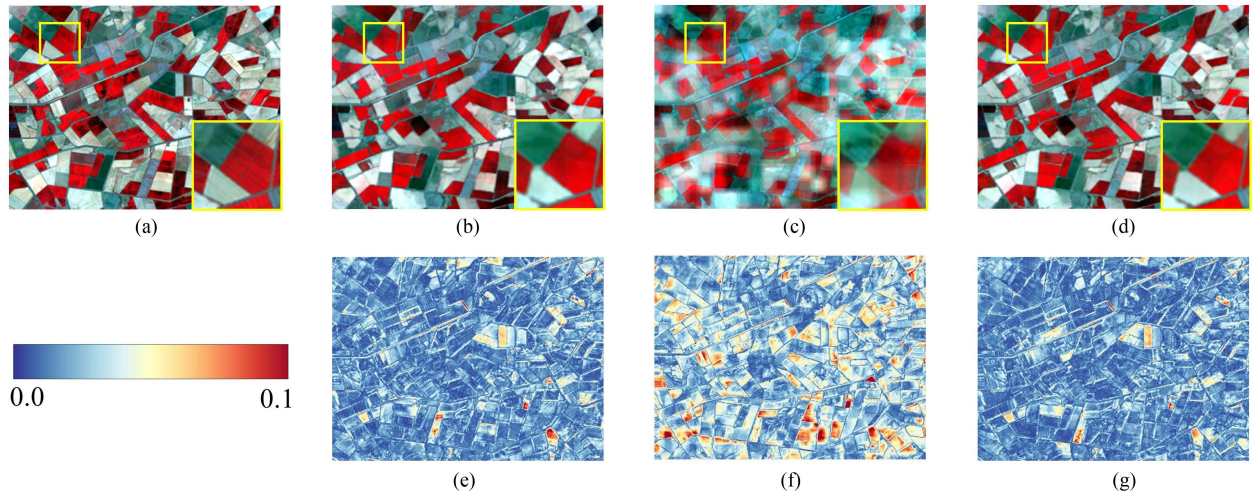


Fig. 11. Fusion results of the different methods on the CIA dataset. (a) Subregion on 12 January 2004. (b) Subregion results of STM-STFNet w/o the spatial branch. (c) Subregion results of STM-STFNet w/o the temporal branch. (d) Subregion results of STM-STFNet. (e) Average absolute difference map of STM-STFNet w/o the spatial branch in the subregion. (f) Average absolute difference map of STM-STFNet w/o the temporal branch in the subregion. (g) Average absolute difference map of STM-STFNet in the subregion.

the large SAM values in Table IV. Comparatively, ESTARFM shows a good performance in this case and outperforms FSDAF and GAN-STFM in terms of both the visual and quantitative evaluation, especially over the bright artificial surface. These results indicate that the STF methods perform distinctively for different image datasets with various characteristics. However, the proposed method has a robust performance over all the datasets and achieves the best quantitative scores. The visual maps in Figs. 9 and 10 also show that STM-STFNet achieves a balance in processing image patches with diverse textures (e.g., buildings, roads, and farmland), which results in the lowest errors over the whole image.

V. DISCUSSION

A. Ablation Study

To verify the effectiveness of the proposed STM-STFNet method, we conducted ablation experiments on the CIA and DX datasets, which represent phenological and land-cover changes, respectively. First, the core modules (i.e., the spatial modeling branch and temporal modeling branch) were subsequently removed from the overall framework, and the corresponding

evaluation results are provided in Table V. The results show that the absence of the modules has a negative impact on the evaluation scores, but to different degrees. Both modules are important, while the temporal modeling branch plays a critical role in estimating the temporal changes and preserving the spectral fidelity. Figs. 11 and 12 are the visual results of the ablation experiments with the CIA and DX datasets, respectively, where the overall quality of the reconstruction results is consistent with the quantitative results.

In addition, we also conducted ablation experiments for the loss terms, and the results obtained with the DX dataset are presented in Table VI. Both the structure loss and edge loss result in a certain improvement in the results, which proves the effectiveness of the designed hybrid loss function. Between the loss functions, the edge loss only pays attention to the high-frequency details, and thus has relatively little impact on the overall results.

B. Impact of the Spatial and Temporal Modeling

In this part, we describe the analysis conducted on the impact of the spatial and temporal modeling on the STF performance.

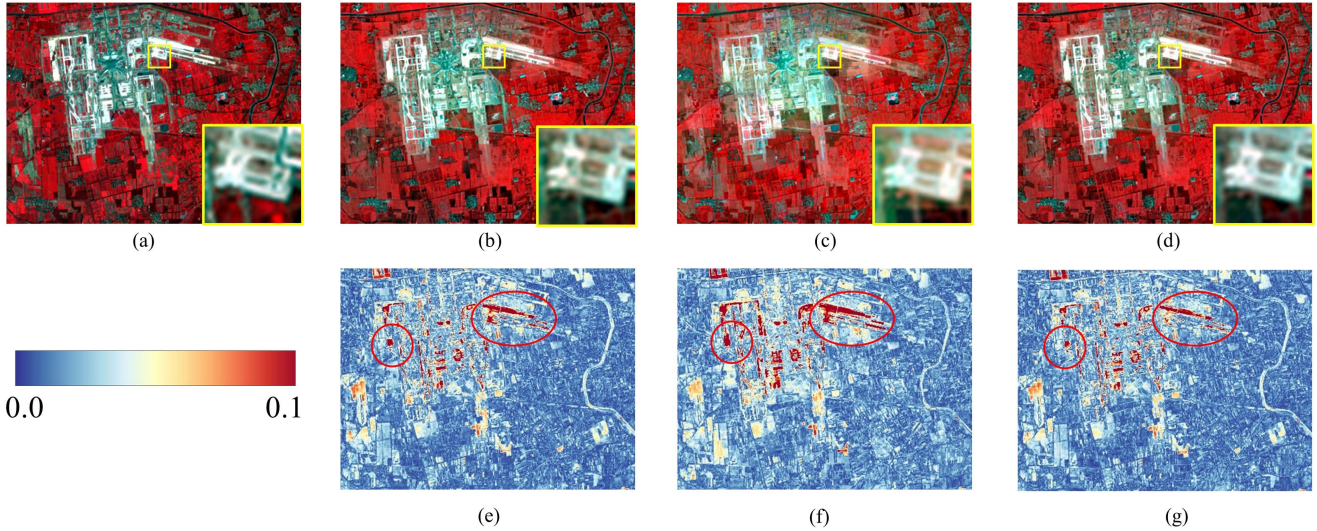


Fig. 12. Fusion results of the different methods on the DX dataset. (a) Subregion on 12 September 2017. (b) Subregion results of STM-STFNet w/o the spatial branch. (c) Subregion results of STM-STFNet w/o the temporal branch. (d) Subregion results of STM-STFNet. (e) Average absolute difference map of STM-STFNet w/o the spatial branch in the subregion. (f) Average absolute difference map of STM-STFNet w/o the temporal branch in the subregion. (g) Average absolute difference map of STM-STFNet in the subregion.

TABLE VI
QUANTITATIVE EVALUATION RESULTS FOR THE DX DATASET WITH DIFFERENT LOSS FUNCTIONS

| | L_{Pixel} | $L_{\text{Structure}}$ | L_{Edge} | RMSE | SSIM | UIQI | CC | ERGAS | SAM |
|--------|--------------------|------------------------|-------------------|---------------|---------------|---------------|---------------|---------------|----------------|
| Method | ✓ | ✓ | | 0.0248 | 0.8321 | 0.8307 | 0.8535 | 1.5900 | 12.4198 |
| | ✓ | | ✓ | 0.0251 | 0.8217 | 0.8263 | 0.8526 | 1.5998 | 12.3814 |
| | ✓ | ✓ | ✓ | 0.0244 | 0.8362 | 0.8379 | 0.8583 | 1.5266 | 12.2900 |

The best results are highlighted in bold.

The results in Table V and Figs. 11 and 12 show that the temporal modeling achieves results that are closer to those of the proposed STM-STFNet method in the quantitative and visual results. This indicates that temporal prediction plays a more critical role in STF. In the highlighted regions in Fig. 11(b)–(d), it can be seen that the spatial modeling has advantages in reconstructing the spatial details. The visualization of the attention weight maps of the spatial branch [referring to $M_{cs}(\Delta C')$ in (11)] in Fig. 13 also shows that the spatial modeling pays special attention to the heterogeneous information related to the spatial details and land-cover changes, which is consistent with the role of the spatial modeling branch, as described in Section III. The spatial modeling helps the model in dealing with land-cover changes within the multitemporal images and serves to reconstruct the fine-scale textures in the results.

However, significant prediction errors and spectral distortion are the main problems in the case of abrupt land-cover changes, as shown in Fig. 12. In this case, the spatial modeling has difficulty in achieving satisfactory results without the involvement of the temporal modeling branch for change prediction.

C. Complexity Analysis

Table VII presents the frames per second (FPS) and parameter size of different deep-learning methods on LGC dataset. All the

TABLE VII
COMPUTATIONAL EFFICIENCY AND PARAMETER SIZE OF DIFFERENT DEEP-LEARNING METHODS ON LGC DATASET

| Methods | Dataset | FPS | Parameters(MB) |
|------------|---------|--------|----------------|
| GAN-STFM | | 211.63 | 16.19 |
| STFRNN | | 348.19 | 4.06 |
| SwinSTFM | LGC | 20.59 | 143.19 |
| SwinSTFM-B | | 12.49 | 200.77 |
| STM-STFNet | | 14.11 | 113.32 |

codes were run on an NVIDIA GeForce RTX 4090 GPU and 13th Gen Intel Core i7-13700F CPU. The patch size for the computational complexity experiments was set to 128×128 for a fair comparison of FPS values.

The results show that GAN-STFM and STFRNN are relatively light, in terms of both parameter size and FPS values. However, the STM-STFNet has a considerable computational complexity with SwinSTFM while achieving a better result. Moreover, the trained STM-STFNet takes 134.28 s to generate one target image with the size of $3200 \times 2720 \times 6$ with two reference image pairs. Compared with 13745.45 s required for ESTARFM and 3416.25 s for FSDAF, the proposed method

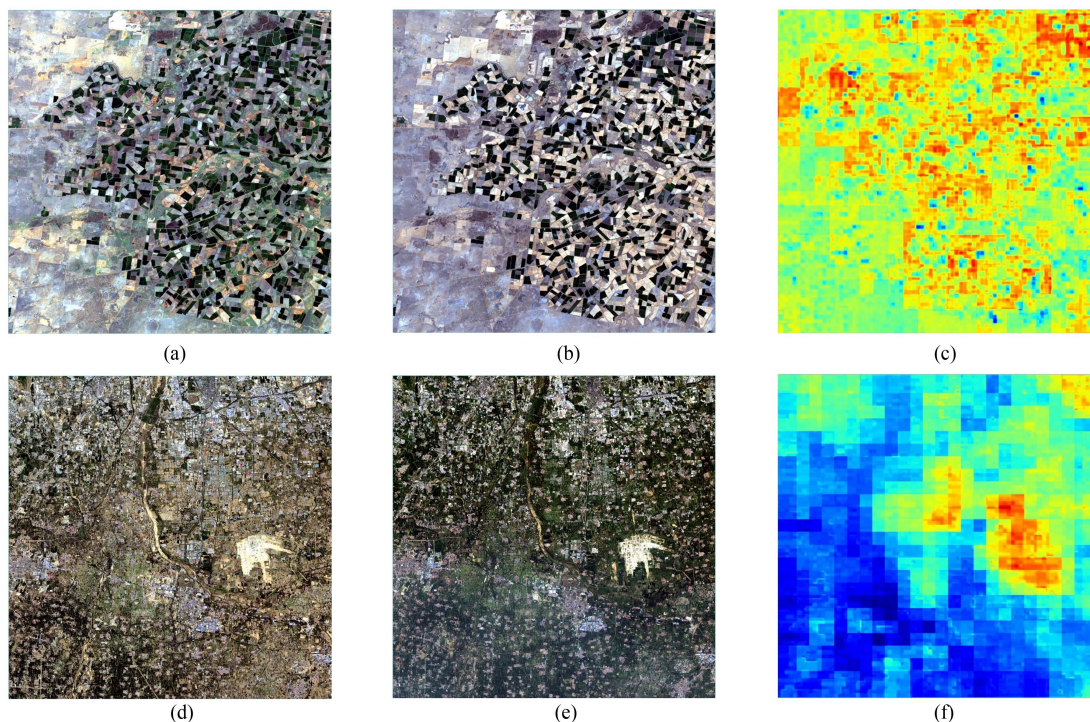


Fig. 13. (a) Image from 22 February 2002 in the CIA dataset. (b) Image from 12 January 2002 in the CIA dataset. (c) Attention weight map of the spatial branch for the CIA dataset. (d) Image from 7 May 2017 in the DX dataset. (e) Image from 12 September 2017 in the DX dataset. (f) Attention weight map of the spatial branch for the DX dataset.

shows a better efficiency. Overall, given the superior performance and computational cost of the proposed method, the proposed STM-STFNet can be a good candidate method for STF task.

VI. CONCLUSION

In this article, we have proposed a dual-perspective STF method by discriminative learning of spatial and temporal mapping to tackle the challenge of fusing remote sensing images with abrupt land-cover changes. The extensive experiments showed that the proposed method can achieve state-of-the-art results on three datasets (i.e., the LGC, CIA, and DX datasets) for image STF. These results show that the proposed STM-STFNet method can adapt well to datasets characterized by land-cover changes, phenological changes, and even both. Moreover, the ablation experiments validated the effectiveness of the dual-perspective modeling. The temporal modeling plays a vital role in the accurate change prediction and spectral preservation, while the spatial modeling helps to reconstruct the high-frequency components and capture the change information between the reference and target dates.

This article has provided a new solution for STF modeling with complex temporal changes. In terms of the limitations, the spatial modeling with significant temporal changes from STF is still challenging. To further promote the reconstruction accuracy, our future studies will focus on incorporating physical knowledge related to the spatial, temporal, and spectral characteristics into the learning-based STF framework, and we will attempt to

develop STF methods adapted to monitoring different terrestrial processes (e.g., crop growth, urban construction, and the water environment).

REFERENCES

- [1] A. Schneider, "Monitoring land cover change in urban and peri-urban areas using dense time stacks of Landsat satellite data and a data mining approach," *Remote Sens. Environ.*, vol. 124, pp. 689–704, 2012, doi: [10.1016/j.rse.2012.06.006](https://doi.org/10.1016/j.rse.2012.06.006).
- [2] V. Heimhuber, M. G. Tulbure, and M. Broich, "Addressing spatio-temporal resolution constraints in Landsat and MODIS-based mapping of large-scale floodplain inundation dynamics," *Remote Sens. Environ.*, vol. 211, pp. 307–320, 2018, doi: [10.1016/j.rse.2018.04.016](https://doi.org/10.1016/j.rse.2018.04.016).
- [3] T. Hwang, C. Song, P. V. Bolstad, and L. E. Band, "Downscaling real-time vegetation dynamics by fusing multi-temporal MODIS and Landsat NDVI in topographically complex terrain," *Remote Sens. Environ.*, vol. 115, no. 10, pp. 2499–2512, 2011, doi: [10.1016/j.rse.2011.05.010](https://doi.org/10.1016/j.rse.2011.05.010).
- [4] J. Li, Y. Li, L. He, J. Chen, and A. Plaza, "Spatio-temporal fusion for remote sensing data: An overview and new benchmark," *Sci. China Inf. Sci.*, vol. 63, no. 4, Mar. 2020, Art. no. 140301, doi: [10.1007/s11432-019-2785-y](https://doi.org/10.1007/s11432-019-2785-y).
- [5] X. Zhu, F. Cai, J. Tian, and T. K.-A. Williams, "Spatiotemporal fusion of multisource remote sensing data: Literature survey, taxonomy, principles, applications, and future directions," *Remote Sens.*, vol. 10, no. 4, Apr. 2018, Art. no. 527, doi: [10.3390/rs10040527](https://doi.org/10.3390/rs10040527).
- [6] Z. Wang, Y. Ma, and Y. Zhang, "Review of pixel-level remote sensing image fusion based on deep learning," *Inf. Fusion*, vol. 90, pp. 36–58, Feb. 2023, doi: [10.1016/j.inffus.2022.09.008](https://doi.org/10.1016/j.inffus.2022.09.008).
- [7] X. Meng, Q. Liu, F. Shao, and S. Li, "Spatio-temporal-spectral collaborative learning for spatio-temporal fusion with land cover changes," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jun. 2022, Art. no. 5704116, doi: [10.1109/tgrs.2022.3185459](https://doi.org/10.1109/tgrs.2022.3185459).
- [8] B. Huang and H. Zhang, "Spatio-temporal reflectance fusion via unmixing: Accounting for both phenological and land-cover changes," *Int. J. Remote Sens.*, vol. 35, no. 16, pp. 6213–6233, 2014, doi: [10.1080/01431161.2014.951097](https://doi.org/10.1080/01431161.2014.951097).

- [9] D. Jia, C. Song, C. Cheng, S. Shen, L. Ning, and C. Hui, "A novel deep learning-based spatiotemporal fusion method for combining satellite images with different resolutions using a two-stream convolutional neural network," *Remote Sens.*, vol. 12, no. 4, Feb. 2020, Art. no. 698, doi: [10.3390/rs12040698](https://doi.org/10.3390/rs12040698).
- [10] Q. Wang, Y. Tang, X. Tong, and P. M. Atkinson, "Virtual image pair-based spatio-temporal fusion," *Remote Sens. Environ.*, vol. 249, Nov. 2020, Art. no. 112009, doi: [10.1016/j.rse.2020.112009](https://doi.org/10.1016/j.rse.2020.112009).
- [11] X. Jiang and B. Huang, "Unmixing-based spatiotemporal image fusion accounting for complex land cover changes," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2022, Art. no. 5623010, doi: [10.1109/tgrs.2022.3173172](https://doi.org/10.1109/tgrs.2022.3173172).
- [12] Q. Liu, X. Meng, X. Li, and F. Shao, "Detail injection-based spatiotemporal fusion for remote sensing images with land cover changes," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Mar. 2023, Art. no. 5401514, doi: [10.1109/tgrs.2023.3252054](https://doi.org/10.1109/tgrs.2023.3252054).
- [13] F. Gao, J. Masek, M. Schwaller, and F. G. Hall, "On the blending of the Landsat and MODIS surface reflectance: Predicting daily Landsat surface reflectance," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 8, pp. 2207–2218, Aug. 2006, doi: [10.1109/tgrs.2006.872081](https://doi.org/10.1109/tgrs.2006.872081).
- [14] X. Zhu, J. Chen, F. Gao, X. Chen, and J. G. Masek, "An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions," *Remote Sens. Environ.*, vol. 114, no. 11, pp. 2610–2623, Nov. 2010, doi: [10.1016/j.rse.2010.05.032](https://doi.org/10.1016/j.rse.2010.05.032).
- [15] Q. Cheng, H. Liu, H. Shen, P. Wu, and L. Zhang, "A spatial and temporal nonlocal filter-based data fusion method," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4476–4488, Aug. 2017, doi: [10.1109/tgrs.2017.2692802](https://doi.org/10.1109/tgrs.2017.2692802).
- [16] Q. Wang and P. M. Atkinson, "Spatio-temporal fusion for daily Sentinel-2 images," *Remote Sens. Environ.*, vol. 204, pp. 31–42, 2018, doi: [10.1016/j.rse.2017.10.046](https://doi.org/10.1016/j.rse.2017.10.046).
- [17] T. Hilker et al., "A new data fusion model for high spatial- and temporal-resolution mapping of forest disturbance based on Landsat and MODIS," *Remote Sens. Environ.*, vol. 113, no. 8, pp. 1613–1627, Aug. 2009, doi: [10.1016/j.rse.2009.03.007](https://doi.org/10.1016/j.rse.2009.03.007).
- [18] 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 Sens. Environ.*, vol. 206, pp. 403–423, Mar. 2018, doi: [10.1016/j.rse.2017.12.003](https://doi.org/10.1016/j.rse.2017.12.003).
- [19] R. Heylen, M. Parente, and P. Gader, "A review of nonlinear hyperspectral unmixing methods," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 1844–1868, Jun. 2014, doi: [10.1109/JSTARS.2014.2320576](https://doi.org/10.1109/JSTARS.2014.2320576).
- [20] J. Wei, L. Wang, P. Liu, X. Chen, W. Li, and A. Y. Zomaya, "Spatiotemporal fusion of MODIS and Landsat-7 reflectance images via compressed sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 7126–7139, Dec. 2017, doi: [10.1109/TGRS.2017.2742529](https://doi.org/10.1109/TGRS.2017.2742529).
- [21] J. Xue, Y. Leung, and T. Fung, "A Bayesian data fusion approach to spatiotemporal fusion of remotely sensed images," *Remote Sens.*, vol. 9, no. 12, 2017, Art. no. 1310.
- [22] H. Shen, X. Meng, and L. Zhang, "An integrated framework for the spatio-temporal-spectral fusion of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7135–7148, Dec. 2016, doi: [10.1109/tgrs.2016.2596290](https://doi.org/10.1109/tgrs.2016.2596290).
- [23] X. Zhu, E. H. Helmer, F. Gao, D. Liu, J. Chen, and M. A. Lefsky, "A flexible spatiotemporal method for fusing satellite images with different resolutions," *Remote Sens. Environ.*, vol. 172, pp. 165–177, Jan. 2016, doi: [10.1016/j.rse.2015.11.016](https://doi.org/10.1016/j.rse.2015.11.016).
- [24] C. M. Gevaert and F. J. García-Haro, "A comparison of STARFM and an unmixing-based algorithm for Landsat and MODIS data fusion," *Remote Sens. Environ.*, vol. 156, pp. 34–44, Jan. 2015, doi: [10.1016/j.rse.2014.09.012](https://doi.org/10.1016/j.rse.2014.09.012).
- [25] B. Huang and H. Song, "Spatiotemporal reflectance fusion via sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 10, pp. 3707–3716, Oct. 2012, doi: [10.1109/tgrs.2012.2186638](https://doi.org/10.1109/tgrs.2012.2186638).
- [26] B. Wu, B. Huang, and L. Zhang, "An error-bound-regularized sparse coding for spatiotemporal reflectance fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6791–6803, Dec. 2015, doi: [10.1109/TGRS.2015.2448100](https://doi.org/10.1109/TGRS.2015.2448100).
- [27] S. P. Boyte, B. K. Wylie, M. B. Rigge, and D. Dahal, "Fusing MODIS with Landsat 8 data to downscale weekly normalized difference vegetation index estimates for central Great Basin rangelands, USA," *GISci. Remote Sens.*, vol. 55, no. 3, pp. 376–399, May 2018, doi: [10.1080/15481603.2017.1382065](https://doi.org/10.1080/15481603.2017.1382065).
- [28] Y. Ke, J. Im, S. Park, and H. Gong, "Downscaling of MODIS one kilometer evapotranspiration using Landsat-8 data and machine learning approaches," *Remote Sens.*, vol. 8, no. 3, Mar. 2016, Art. no. 215, doi: [10.3390/rs8030215](https://doi.org/10.3390/rs8030215).
- [29] X. Liu, C. Deng, S. Wang, G.-B. Huang, B. Zhao, and P. Lauren, "Fast and accurate spatiotemporal fusion based upon extreme learning machine," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 2039–2043, Dec. 2016, doi: [10.1109/lgrs.2016.2622726](https://doi.org/10.1109/lgrs.2016.2622726).
- [30] H. Song, Q. Liu, G. Wang, R. Hang, and B. Huang, "Spatiotemporal satellite image fusion using deep convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 3, pp. 821–829, Mar. 2018, doi: [10.1109/jstars.2018.2797894](https://doi.org/10.1109/jstars.2018.2797894).
- [31] Z. Tan, P. Yue, L. Di, and J. Tang, "Deriving high spatiotemporal remote sensing images using deep convolutional network," *Remote Sens.*, vol. 10, no. 7, 2018, Art. no. 1066.
- [32] P. Qin, H. Huang, H. Tang, J. Wang, and C. Liu, "MUSTFN: A spatiotemporal fusion method for multi-scale and multi-sensor remote sensing images based on a convolutional neural network," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 115, Dec. 2022, Art. no. 103113, doi: [10.1016/j.jag.2022.103113](https://doi.org/10.1016/j.jag.2022.103113).
- [33] W. Li, C. Yang, Y. Peng, and X. Zhang, "A multi-cooperative deep convolutional neural network for spatiotemporal satellite image fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10174–10188, Sep. 2021, doi: [10.1109/JSTARS.2021.3113163](https://doi.org/10.1109/JSTARS.2021.3113163).
- [34] Z. Yin et al., "Spatiotemporal fusion of land surface temperature based on a convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1808–1822, Feb. 2021, doi: [10.1109/TGRS.2020.2999943](https://doi.org/10.1109/TGRS.2020.2999943).
- [35] W. Li, X. Zhang, Y. Peng, and M. Dong, "Spatiotemporal fusion of remote sensing images using a convolutional neural network with attention and multiscale mechanisms," *Int. J. Remote Sens.*, vol. 42, no. 6, pp. 1973–1993, Mar. 2021, doi: [10.1080/01431161.2020.1809742](https://doi.org/10.1080/01431161.2020.1809742).
- [36] D. Lei, G. Ran, L. Zhang, and W. Li, "A spatiotemporal fusion method based on multiscale feature extraction and spatial channel attention mechanism," *Remote Sens.*, vol. 14, no. 3, 2022, Art. no. 461.
- [37] H. Jiang, Y. Qian, G. Yang, and H. Liu, "MLKNet: Multi-stage for remote sensing image spatiotemporal fusion network based on a large kernel attention," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 1257–1268, Dec. 2024, doi: [10.1109/JSTARS.2023.3338978](https://doi.org/10.1109/JSTARS.2023.3338978).
- [38] P. Liu, J. Li, L. Wang, and G. He, "Remote sensing data fusion with generative adversarial networks: State-of-the-art methods and future research directions," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 295–328, Jun. 2022, doi: [10.1109/MGRS.2022.3165967](https://doi.org/10.1109/MGRS.2022.3165967).
- [39] W. S. Li, D. W. Cao, Y. D. Peng, and C. Yang, "MSNet: A multi-stream fusion network for remote sensing spatiotemporal fusion based on transformer and convolution," *Remote Sens.*, vol. 13, no. 18, Sep. 2021, Art. no. 3724, doi: [10.3390/rs13183724](https://doi.org/10.3390/rs13183724).
- [40] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
- [41] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.
- [42] G. Chen, P. Jiao, Q. Hu, L. Xiao, and Z. Ye, "SwinSTFM: Remote sensing spatiotemporal fusion using Swin transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jun. 2022, Art. no. 5410618, doi: [10.1109/tgrs.2022.3182809](https://doi.org/10.1109/tgrs.2022.3182809).
- [43] T. Benzenati, A. Kallel, and Y. Kessentini, "STF-trans: A two-stream spatiotemporal fusion transformer for very high resolution satellites images," *Neurocomputing*, vol. 563, Jan. 2024, Art. no. 126868, doi: [10.1016/j.neucom.2023.126868](https://doi.org/10.1016/j.neucom.2023.126868).
- [44] W. S. Li, D. W. Cao, and M. H. Xiang, "Enhanced multi-stream remote sensing spatiotemporal fusion network based on transformer and dilated convolution," *Remote Sens.*, vol. 14, no. 18, Sep. 2022, Art. no. 4544, doi: [10.3390/rs14184544](https://doi.org/10.3390/rs14184544).
- [45] G. Yang et al., "MSFusion: Multistage for remote sensing image spatiotemporal fusion based on texture transformer and convolutional neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4653–4666, Jun. 2022, doi: [10.1109/jstars.2022.3179415](https://doi.org/10.1109/jstars.2022.3179415).
- [46] Z. Tan, M. Gao, X. Li, and L. Jiang, "A flexible reference-insensitive spatiotemporal fusion model for remote sensing images using conditional generative adversarial network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5601413, doi: [10.1109/tgrs.2021.3050551](https://doi.org/10.1109/tgrs.2021.3050551).
- [47] Z. Tan, M. Gao, J. Yuan, L. Jiang, and H. Duan, "A robust model for MODIS and Landsat image fusion considering input noise," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5407217, doi: [10.1109/TGRS.2022.3145086](https://doi.org/10.1109/TGRS.2022.3145086).

- [48] D. Jia, C. Cheng, S. Shen, and L. Ning, "Multitask deep learning framework for spatiotemporal fusion of NDVI," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5616313, doi: [10.1109/TGRS.2021.3140144](https://doi.org/10.1109/TGRS.2021.3140144).
- [49] Z. Tan, L. Di, M. Zhang, L. Guo, and M. Gao, "An enhanced deep convolutional model for spatiotemporal image fusion," *Remote Sens.*, vol. 11, no. 24, Dec. 2019, Art. no. 2898, doi: [10.3390/rs11242898](https://doi.org/10.3390/rs11242898).
- [50] Z. Ao, Y. Sun, X. Pan, and Q. Xin, "Deep learning-based spatiotemporal data fusion using a patch-to-pixel mapping strategy and model comparisons," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 2022, Art. no. 5407718, doi: [10.1109/tgrs.2022.3154406](https://doi.org/10.1109/tgrs.2022.3154406).
- [51] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
- [52] S. H. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. 15th Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
- [53] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep Laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5835–5843.
- [54] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: [10.1109/tip.2003.819861](https://doi.org/10.1109/tip.2003.819861).
- [55] I. V. Emelyanova, T. R. McVicar, T. G. Van Niel, L. T. Li, and A. I. J. M. van Dijk, "Assessing the accuracy of blending Landsat-MODIS surface reflectances in two landscapes with contrasting spatial and temporal dynamics: A framework for algorithm selection," *Remote Sens. Environ.*, vol. 133, pp. 193–209, Jun. 2013, doi: [10.1016/j.rse.2013.02.007](https://doi.org/10.1016/j.rse.2013.02.007).
- [56] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002, doi: [10.1109/97.995823](https://doi.org/10.1109/97.995823).
- [57] L. Wald, *Data Fusion: Definitions and Architectures: Fusion of Images of Different Spatial Resolutions*. Paris, France: Presses de l'École des Mines, 2002.
- [58] R. H. Yuhas, A. F. H. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm," in *Proc. JPL, Summaries 3rd Annu. JPL Airborne Geosci. Workshop*, 1992, pp. 147–149.



Xiao Xie received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2016.

She was a Research Fellow with the Department of Cartography, Technical University of Munich, Munich, Germany, from 2014 to 2016. She is currently a Professorate-Senior-Engineer with the Key Lab of Environmental Computing and Sustainability, as well as an Assistant Professor of urban and environmental computation with the Institute of Applied Ecology, Chinese Academy of Sciences, Beijing, China. Her

research interests include 3D GIS and smart cities.



Qiangqiang Yuan (Member, IEEE) received the B.S. degree in surveying and mapping engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2006 and 2012, respectively.

In 2012, he joined the School of Geodesy and Geomatics, Wuhan University, where he is currently a Professor. He has authored or coauthored more than 90 research articles, including more than 70 peer-reviewed articles in international journals, such as *IEEE TRANSACTIONS ON IMAGE PROCESSING* and *IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING*. His research interests include image reconstruction, remote sensing image processing and application, and data fusion.

Dr. Yuan was a recipient of the Top-Ten Academic Star of Wuhan University in 2011 and the Youth Talent Support Program of China in 2019, and the Hong Kong Scholar Award from the Society of Hong Kong Scholars and China National Postdoctoral Council in 2014. He is on the editorial board of nine international journals and has frequently served as a referee for more than 50 international journals for remote sensing and image processing.



Zhonghua Qian received the B.E. degree in surveying and mapping engineering from the School of Surveying and Mapping, Henan Polytechnic University, Jiaozuo, China, in 2022. He is currently working toward the M.E. degree in resources and environment engineering with the School of Geography and Information Engineering, China University of Geosciences, Wuhan, China.

His research interests include multisource remote sensing data fusion and deep learning.



Linwei Yue (Member, IEEE) received the B.S. degree in geographic information system and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2012 and 2017, respectively.

She is currently an Associate Professor with the School of Geography and Information Engineering, China University of Geosciences, Wuhan, China. Her research interests include multisource remote sensing data fusion and water environment applications.



Huanfeng Shen (Senior Member, IEEE) received the B.S. degree in surveying and mapping engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2002 and 2007, respectively.

He is currently a Hongyi Distinguished Professor with Wuhan University, where he serves as a Dean of the School of Resource and Environmental Sciences. His research interests include remote sensing image processing, multisource data fusion, and intelligent environmental sensing. He was or is the PI of two

projects supported by the National Key Research and Development Program of China, and eight projects supported by the National Natural Science Foundation of China. He has authored or coauthored more than 200 journal citation report papers and published five books as a Chief Editor. His papers received 16 269 citations in Web of Science (as of May 2024).

Dr. Shen is a Fellow of IET, the Deputy Director of the Remote Sensing Geography Committee of the Geographical Society of China, and the Deputy Director of the Big Data and Artificial Intelligence Committee of the Chinese Society for Geodesy, Photogrammetry, and Cartography. He is currently a Senior Regional Editor for the *Journal of Applied Remote Sensing*, and an Editorial Board Member of the *ISPRS Journal of Photogrammetry and Remote Sensing* and *Geo-Spatial Information Science*.