HCNNet: A Hybrid Convolutional Neural Network for Spatiotemporal Image Fusion

Zhuangshan Zhu, Yuxiang Tao, and Xiaobo Luo

***Abstract*— In recent years, leaps and bounds have developed spatiotemporal fusion (STF) methods for remote sensing (RS) images based on deep learning. However, most existing methods use 2-D convolution (Conv) to explore features. 3-D Conv can explore time-dimensional features, but it requires more memory footprint and is rarely used. In addition, the current STF methods based on convolutional neural networks (CNNs) are mainly the following two: 1) use 2-D Conv to extract features from multiple bands of the input image together and fuse the features to predict the multiband image directly and 2) use 2-D Conv to extract features from individual bands of the image, predict the reflectance data of individual bands, and finally stack the predicted individual bands directly to synthesize the multiband image. The former method does not sufficiently consider the spectral and reflectance differences between different bands, and the latter does not consider the similarity of spatial structures between adjacent bands and the spectral correlation. To solve these problems, we propose a 2-D/3-D hybrid CNN called HCNNet, in which the 2D-CNN branch extracts the spatial information features of single-band image, and the 3D-CNN branch extracts spatiotemporal features of single-band images. After fusing the features of the dual branches, we introduce neighboring band features to share spatial information so that the information is complementary to obtain single-band features and images, and finally stack each single-band image to generate multiband images. Visual assessment and metric evaluation of the three publicly available datasets showed that our method predicted better images compared with the five methods.**

***Index Terms*— Feature fusion, hybrid convolution (Conv), spatiotemporal fusion (STF), spectral correlation.**

I. Introduction

AS THE scientific research on the application of remote sensing (RS) images becomes more and more extensive and intensive [1], human beings can obtain data of electromagnetic radiation reflected or emitted from various landscapes. There is an increasing demand for high spatial resolution images to monitor rapid temporal changes on the Earth's surface [2]. Due to the limitations of satellite launch budget cost and key technologies, it is still not possible to obtain RS image data with high spatial and temporal resolution at the same time through a single satellite [3], [4]. In general, high spatial resolution images have finer spatial details and are widely used in urban spatial information extraction [5], forest change monitoring [6], [7], and human-made landscape monitoring [8], but such sensors have narrow widths and long revisit cycles on one hand, and the lack of surface data due to cloud cover on the other hand, making it difficult to achieve the purpose of continuous dynamic monitoring on a global scale with high spatial resolution image data in practical applications [9], [10]. In contrast, sensors that obtain high temporal resolution images usually have larger widths and shorter revisit periods, but their low spatial resolution is insufficient for quantitative monitoring of land cover change [11], [12]. If we can solve the problem of mutual constraints in the time and space of RS images and obtain features with both high temporal and high spatial resolution, it will help increase the value of RS data in practical applications [13], [14].

The current spatiotemporal fusion (STF) methods are mainly divided into weight function-based, unmixing-based, and learning-based algorithms. Among the weight function-based methods, the spatial and temporal adaptive reflectance fusion model (STARFM) proposed by Gao *et al.* [15] is the most influential. STARFM assumes that one coarse pixel only includes one land cover type. However, this ideal situation cannot be satisfied when coarse pixels are mixed, having a mixture of different land cover types [16]. The prediction performance of STARFM is affected by the characteristic patch size of the landscape. The STARFM was later modified and improved for more complex situations, resulting in the spatio-temporal adaptive algorithm for mapping reflection changes (STAARCH) [17], which improves the performance of the STARFM model in the presence of land-cover-type changes and disturbances. An enhanced spatial and temporal adaptive reflectance fusion model (ESTARFM) [18] based on the existing STARFM algorithm improves the accuracy of predicted fine-resolution reflectance, especially for heterogeneous landscapes.

The unmixing-based methods use spectral unmixing technology to estimate the selected endmember fractions of high temporal low spatial (HTLS) image pixels to reconstruct the corresponding low temporal high spatial (LTHS) image. Ming-Quan *et al.* [19] proposed the spatial and temporal data fusion model (STDFM) algorithm, which can obtain the reflectance of different endmembers. Zhang *et al.* [20] proposed an enhanced STDFM (ESTDFM) based on the STDFM

Manuscript received 5 February 2022; revised 7 April 2022; accepted 21 May 2022. Date of publication 25 May 2022; date of current version 28 July 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 41871226, in part by the Major Industrial Technology Research and Development Projects of High-Tech Industry in Chongqing under Grant D2018-82, and in part by the Intergovernmental International Scientific and Technological Innovation Cooperation Project of the National Key Research and Development Program under Grant 2021YFE0194700. *(Corresponding author: Yuxiang Tao.)*

The authors are with the School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China, and also with the Chongqing Engineering Research Center of Spatial Big Data Intelligent Technology, Chongqing 400065, China (e-mail: s200231013@stu.cqupt.edu.cn; taoyx@cqupt.edu.cn; luoxb@cqupt.edu.cn).

Digital Object Identifier 10.1109/TGRS.2022.3177749

algorithm by introducing a patch-based ISODATA classification method, the sliding window technology, and the temporal-weight concept. The spatial–temporal data fusion approach (STDFA) [21] is based on the assumption that the temporal variation properties of each land cover class are constant. To avoid the limitations including the constant window for disaggregation and sensor difference, Wu *et al.* [22] introduce an adaptive window size selection method and a modified spatial and temporal data fusion approach (MSTDFA) to generate daily synthetic Landsat imagery. The flexible spatiotemporal data fusion (FSDAF) approach proposed by Zhu *et al.* [16] combines two types of algorithms based on spatial pixels unmixing and spatiotemporal change filtering ideas. It introduces a thin plate spline (TPS) interpolation technique to identify feature type changes, significantly improving the fusion effect for heterogeneous ground cover and feature type changes. Furthermore, improved algorithms based on FSDAF have been proposed successively, which improve the accuracy of fused images to different degrees [23]–[25].

The learning-based approach builds association models by learning the mapping relationships between LTHS and HTLS images. SParse-representation-based SpatioTemporal reflectance Fusion Model (SPSTFM) [26] based on sparse representation is the first learning-based STF method. To explore spatio-spectral–temporal features, Zhao *et al.* [27] propose a novel sparse representation model to generate synthesized frequent high spectral and high spatial resolution data by blending multiple types. Techniques such as regression trees, random forests, and extreme learning machines (ELMs) have been widely used [28]. Boyte *et al.* [29] demonstrated that the regression tree model could be used to downscale 250 m enhanced moderate resolution imaging spectroradiometer normalized difference vegetation index (eMODIS NDVI) data using 30 m Landsat 8 operational land imager (OLI) data relatively easily and effectively. Ke *et al.* introduced machine learning approaches for MODIS evapotranspiration (ET) downscaling. In this study, random forests is used to implement the method [30]. Liu *et al.* [31] proposed a novel STF using a powerful learning technique, i.e., an ELM, which devotes itself to learning a mapping function on different images directly and obtains better fusion results while achieving much greater speed. Song *et al.* [32] proposed a convolutional neural network (CNN)-based fusion method, which first learns the nonlinear mapping model between MODIS and downsampled Landsat images and then learns the super-resolution (SR) model between the downsampled Landsat images and the original Landsat images. The deep convolution STF network (DCSTFN) proposed by Tan *et al.* [33] inputs a pair of LTHS and HTLS images for reference and a pair of HTLS images for prediction. The information is merged in the form of extracted feature images, and then the merged features are reconstructed into prediction images. Liu *et al.* [34] use temporal information in high-resolution image sequences and solve the STF problem with a two-stream CNN called StfNet. Tan *et al.* [35] propose an enhanced deep convolutional model (EDCSTFN), which focuses on reconstructing high-resolution images at reference time using two pairs of low-resolution images and high-resolution images at prediction time. Li *et al.* [36] proposed an STF method AMNet, in which an attention mechanism and a multiscale mechanism are incorporated. It differs from previous STF methods in which the residual images obtained from the MODIS images are twice subtracted and used directly for network training. Two special structures, the multiscale and attention mechanisms, are used to improve fusion accuracy.

Most of the previous STF methods use 2-D Conv for feature extraction, and there are few methods to design networks by combining both 2-D and 3-D convolutions (Convs). Li *et al.* [37] presented a new SR method for hyperspectral images, which alternately uses 2-D and 3-D Convs to solve the structural redundancy problem by sharing the spatial information in the reconstruction process of the existing models and improving the learning capability in the 2-D spatial domain. Wang *et al.* [38] proposed a new network structure for hyperspectral image SR and designed with depth splitting (DS) to jointly use the information of single and adjacent bands, which can effectively share spatial information compared with a single 3-D CNN, thus improving the learning ability in the 2-D spatial domain. The network structure uses the current band and its two adjacent bands to reconstruct the single-band SR and finally achieves the reconstruction of hyperspectral images by recursive means. Inspired by [37], [38], the HCNNet fusion model is designed to alleviate the mentioned STF problem. Considering that RS images are more challenging to acquire data in cloudy areas, HCNNet uses a pair of reference images near the prediction moment to make predictions on the target image. The novelty of the fusion model proposed in this article is mainly reflected in the following aspects.

1) A multiband–single-band–multiband structure is adopted in this article. The multiband–single-band process can fully consider the differences in reflectance of individual bands. The single-band–multiband process is also not a simple stacking of individual bands but fully considers each band's spatial structure similarity and spectral connection through feature iteration.
2) Building a dual-branch hybrid CNN architecture. The network uses the 2D-CNN branch to learn the spatial features of the LTHS image at the reference moment and the 3D-CNN branch to learn the spatiotemporal variation features of the LTHS and HTLS images. Finally, the dual-branch features are fused to improve single-band feature extraction and image reconstruction performance.
3) Linking 3D-CNN branching features with 2D-CNN branching features through the permute (Pme) modules and introducing convolutional block attention module (CBAM) to share spatial channel information and enhance spatial information feature extraction in the 2D-CNN branch.
4) The spatial and spectral information of adjacent bands of multiband images are associated, and the features of the former band are transferred to the feature reconstruction task of the next adjacent band after extraction. The spatial attention mechanism is introduced to enhance the spatial information extraction capability. With this inter-band feature iteration strategy, the features of every

single band are reconstructed, and the spatial–spectral information between bands is shared and complemented.

The remainder of this article is organized as follows. Section II presents the related research work. Section III elaborates on the proposed network structure. Section IV gives the experimental details, results, and performs ablation experiments for validation. Section V concludes this article.

II. Related Work

In this article, a novel hybrid convolutional module (HCNNet) is proposed to extract the potential features by 2-D/3-D Conv instead of one Conv, which enables the network to explore features of multiband images more. CBAM is an attention module for feed-forward CNN and can be seamlessly integrated into any CNN architecture by ignoring the overhead of the module and can be trained end-to-end with the base CNN. This section briefly introduces 2-D Conv, 3-D Conv, and attentional mechanisms.

A. CNNs Based on 2-D Conv

CNNs can automatically learn features from data (usually large-scale) and generalize the results to unknown data of the same type. In 2D-CNN, the Conv operation is implemented by computing the sum of dot products between the input data and the corresponding convolution kernel (CK). Each channel of the 2-D convolutional layer can be represented as

$$D_i = \Omega\left(\sum_j (w_i \cdot x_j) + b_i\right) \tag{1}$$

where D_i is the ith channel of the convolutional feature map, w_i represents the ith CK, and b_i is the bias term of the ith feature map. x_j is the jth channel of the previous layer, $\cdot$ operator represents the Conv operation, and $\Omega(\cdot)$ represents the nonlinear Relu activation function. This function can be used to improve CNN's nonlinearity and speed up the training process of the network model and can be expressed by the following equation:

$$\Omega(x) = \begin{cases} x, & \text{if } x > 0 \\ 0, & \text{if } x \le 0. \end{cases} \tag{2}$$

B. CNNs Based on 3-D Conv

To fully use the temporal variation features of HTLS images at different moments, we use 3-D Conv to extract spatiotemporal features from input images. 3-D Conv has an additional depth channel compared with 2-D Conv, and this depth channel is the temporal channel in the STF. A distinctive feature of hyperspectral images is the strong correlation of neighboring bands, and therefore there is much recent literature on SR of hyperspectral images using 3-D Conv [39]–[41]. There is also a strong correlation between different resolutions of images in the STF task in terms of bands, so in this article, we use 3-D Conv to extract spatiotemporal features. In the 3D-CNN branch, the input data are convolved with 3-D CK. The activation value at spatial position (x, y, z) can be formulated as

$$v_{ij}^{xyz} = \Phi\left(\sum_t \sum_{f=0}^{F_i-1} \sum_{g=0}^{G_i-1} \sum_{h=0}^{H_i-1} w_{ijt}^{fgh} \cdot v_{(i-1)t}^{(x+f)(y+g)(z+h)} + b_{ij}\right) \tag{3}$$

where $\Phi(x)$ is the activation function, and G_i and F_i are the height and width of the spatial dimension of CK, respectively. H_i is the spectral dimension of CK and the connection index of the current (jth) feature map to the feature map of layer $i-1$. w_{ijt}^{fgh} represents the connection value with the tth feature map at position (f, g, h). $v_{(i-1)t}^{(x+f)(y+g)(z+h)}$ is the value of the tth feature map of layer $i-1$ at position $(x+f, y+g, z+h)$, and b_{ij} is the bias value.

C. Attentional Mechanisms

CBAM [42] introduces both spatial and channel attention to enhance the representation of feature regions and determine what the network model should focus on. The attention process is divided into two parts: the channel attention module and the spatial attention module, which saves parameters and computational power and ensures that it can be integrated into the existing network architectures as a plug-and-play module. The two modules can be used separately or in combination. The CBAM network structure is shown in Fig. 1. In the CBAM module, the input feature maps are first subjected to global max-pooling (denoted by MaxPool2d) and global average pooling (denoted by AvgPool2d). Then the resulting feature maps are sent to a linear layer (denoted by Linear) to reduce the number of feature maps, which are activated by the Relu function (denoted by Relu) and then restored to the original number of features by a linear layer. After that, the output features are summed element by element (represented by $\oplus$), and the final channel attention feature map is obtained by the sigmoid activation function (denoted by Sigmoid). The spatial attention feature map is operated to find the maximum value (represented by Max) and the mean value (represented by Mean), and the two obtained feature maps are concatenated (represented by ©) by concatenation operation, and then 2-D Conv (represented by Conv2d) is performed to reduce the number of feature maps. The spatial attention features are generated after the sigmoid activation function and multiplied (represented by $\otimes$) with the input to get the final generated features.

III. Proposed Methodology

In this section, we first introduce the principle of STF and then introduce the architecture of the proposed HCNNet, including the overall network structure, the dual-branch network structure, the dual-branch feature fusion, and the band feature iteration. Finally, the loss function of the model is introduced.

A. Principle of STF

The STF methods in this article use satellite sensor data from two different sources. M and L represent the HTLS

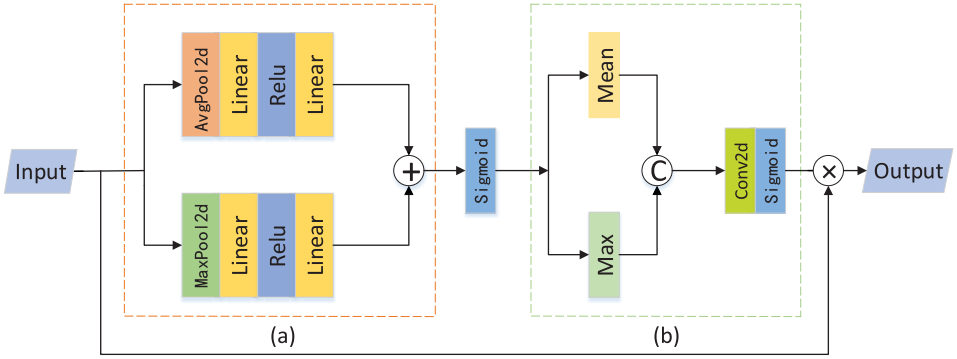

Fig. 1. Convolutional block attention module. (a) Channel attention. (b) Spatial attention.

and LTHS images, respectively. We have now acquired HTLS image M_1 at the moment of forecast date t_1, HTLS image M_k, and Landsat image L_k for the same geographical area at the moment of reference date t_k. The task of STF is to use the already acquired images to predict the LTHS image L_1 at the moment t_1. The reference moment t_k is selected near the predicted moment, and the timespan should not be too large so that the predicted image L_1 contains both the time variation in the HTLS image and LTHS image detail texture information. The above process can be abstracted to establish a mapping relationship between the target image and the acquired image, and this mapping relationship can be expressed by the following equation:

$$L_1 = \varphi(M_1, L_k, M_k|\theta), \quad k \neq 1 \tag{4}$$

where the parameters θ represent a set of learnable parameters that can be learned by training the STF model to build a nonlinear mapping to approximate the actual function. In this article, $k = 0$.

B. Overall Network Structure

The general architecture of HCNNet is shown in Fig. 2. Unlike previous approaches, HCNNet deals with each band separately but considers the associations between each band. The overall structure includes 2D-CNN, 3D-CNN, Pme modules, spatiotemporal feature fusion (TSFF), and spatial–spectral feature fusion (SSFF) modules. The images we input to the network are LTHS image L_0 at the moment t_0, HTLS image M_0 at the moment t_0, and HTLS image M_1 at the moment t_1, and the image to be predicted is L_1 at the moment t_1. Each image consists of i bands. All the fine spatial texture features of L_1 come from the L_0 image, so we input each band of L_0 into the 2D-CNN branch for spatial detail feature extraction when rebuilding each band feature. Time change information needs to be extracted from M_0 and M_1. L_0 can also reflect time change compared with M_0. To make full use of the observed data, we input M_0, M_1, and L_0 together into the 3D-CNN branch for spatiotemporal feature extraction when reconstructing the time-varying features in each band. The weights w_1 and w_2 are set for the dual branches, respectively, to control the temporal and spatial feature learning flexibly, and the outputs of the dual branches are connected by the TSFF and SSFF modules.

The results of reconstructing the feature of each band are shown below

$$F_1^i = \begin{cases} C\left[w_1 * \left(f_{2D}\left(L_0^i\right) + P\left(f_{3D}\left(M_0^i, M_1^i, L_0^i\right)\right)\right), \right. \\ \quad \left. w_2 * P\left(f_{3D}\left(M_0^i, M_1^i, L_0^i\right)\right)\right], \qquad i = 1 \\ C\left[w_1 * \left(f_{2D}\left(L_0^i\right) + P\left(f_{3D}\left(M_0^i, M_1^i, L_0^i\right)\right)\right), \right. \\ \quad \left. w_2 * P\left(f_{3D}\left(M_0^i, M_1^i, L_0^i\right)\right), F_1^{i-1}\right], \quad 1 < i \leq B \end{cases} \tag{5}$$

where F_1^i represents the ith band feature reconstructed at the moment t_1, C represents the concatenation operation, $f_{2D}(\cdot)$ and $f_{3D}(\cdot)$ denote the operation of two channels, respectively, and w_1 and w_2 represent the weights of the dual branches, respectively. B represents the total number of bands of the image. P represents the Pme module. After the 3D-CNN branch goes through the Split-3d module, the output graph size is $B \times C \times 3 \times W \times H$, where B is the batch size, C denotes the number of channels, and W and H represent the width and height, respectively. The Pme performs the permutation operation, and the size of the output graph will become $3 \times B \times C \times W \times H$ by the Pme module. The original output of the 3D-CNN branch is converted into three feature maps of size $B \times C \times W \times H$ and the feature maps match the feature map size of the 2D-CNN branch. The reconstructed image bands are spatially and spectrally connected. In feature reconstruction of F_1^i, we need to obtain the band feature F_1^{i-1} before the band, and we implement each band feature reconstruction by multiple such steps. After that, the single-band image reconstruction is completed by a 2-D Conv operation, and the single-band image rebuild will be introduced in detail in Section III. Finally, each single-band image is used to generate multi-band RS images by C operation, as shown below

$$L_1 = C\left(L_1^1, L_1^2, \ldots L_1^i\right), \quad 1 \leq i \leq B \tag{6}$$

where C represents the concatenation operation, and B denotes the total number of bands of RS images. HCNNet adopts a multiband–single-band–multiband structure to accomplish the STF.

C. Dual-Branch Network Structure

Most previous CNNs use 2-D Conv to extract features, and few have extracted spatiotemporal features by combining both

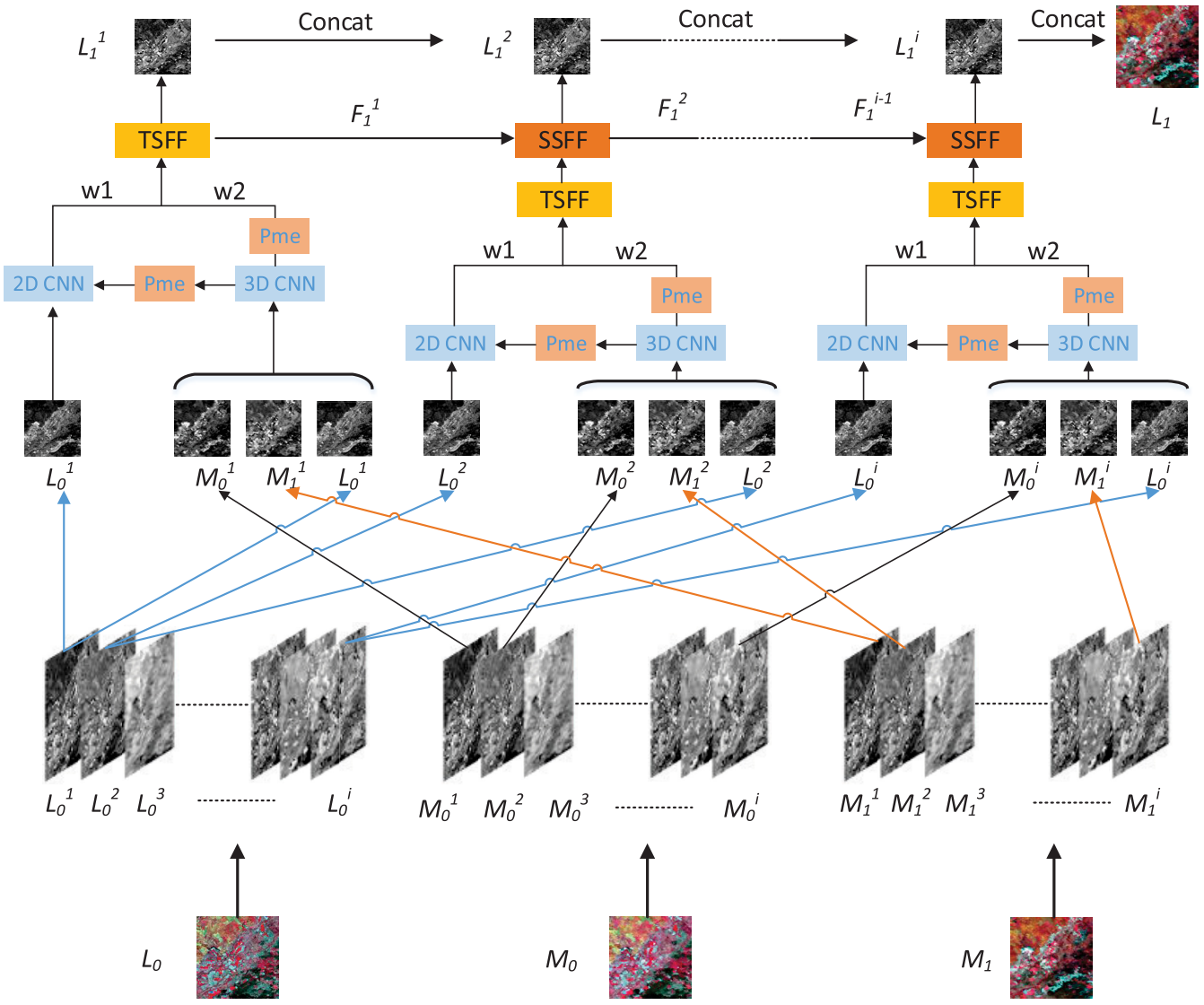

Fig. 2. General structure of our proposed HCNNet. First, we feed the first band L_0^1 of the L_0 image into the 2D-CNN to learn spatial detail features and feed the first band M_0^1, M_1^1, and L_0^1 of each M_0, M_1, and L_0 image together into the 3D-CNN branch to extract both spatial and temporal features. The Pme module enhances the feature learning capability of the 2D-CNN branch. Then, the reconstruction of the first band features F_1^1 and the first band image L_1^1 is completed by the TSFF module. The second-band feature extraction process of the L_1 image is the same as that of the first band, and both are completed by the 2-D, 3-D branch, and the TSFF module. The difference is that the spatial and spectral features F_1^1 learned in the first band are incorporated into the second band feature F_1^2 of L_1, and the second band features F_1^2 of L_1 and the second band image L_1^2 are finally completed by the spatial–spectral feature fusion (SSFF) module. The remaining features and images of each band of L_1 images are reconstructed similar to the second band. Finally, we merge the single-band images by concatenation (Concat) operation to obtain L_1.

2-D and 3-D Convs. The architecture includes 3D-CNN and 2D-CNN branches which are shown in Fig. 3. Next, we present the details of the network structure.

1) 2D-CNN: For the STF, the aim is to reflect the spatiotemporal variation information finely. In the first band, for example, we perform 2-D Conv to extract spatial features, and then these features are followed by several CBAM modules to extract deeper features further. Afterward, these features are enhanced by branching features of 3D-CNN, which is done by the Pme module. The output L_D after the dth CBAM module is as follows:

$$L_D = C(X_D(\ldots C(X_{D-1}(C(X_1(L_0), P(M_1)) + L_0), P(M_D)) + L_0 \ldots)), P(M_{D+1})) + L_0 \quad (7)$$

where X_D denotes the operation of the dth CBAM module, M_D is the result after processing by the dth Split-3d module, C represents the concatenation operation, and $P(\cdot)$ denotes the operation of the Pme module. Because shallow features retain more edge and texture features [38], we use a jump connection to feed L_0 into each CBAM module. To enhance the spatial channel feature extraction capability of the CBAM module, we connect the features extracted from the 3D-CNN branch through the Pme module followed by the C operation to the CBAM module output results. The outputs from different CBAM modules are connected by the C operation and then passed through a convolutional layer with a CK size of 1×1 to reduce the number of feature maps and improve the model's computational efficiency. The network details of generation L_D are shown in Fig. 4(a).

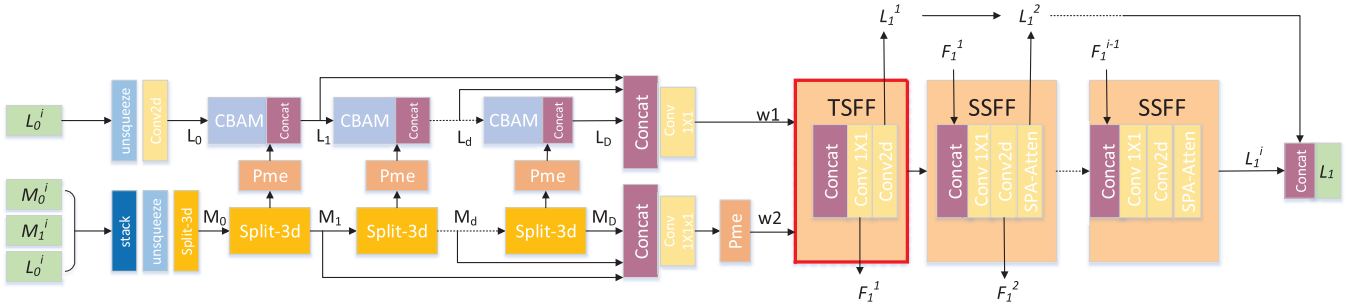

Fig. 3. Dual-branch network structure.

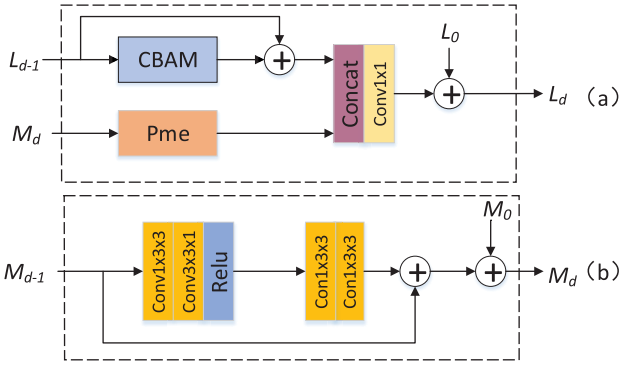

Fig. 4. Network details. (a) Network details for generating L_D. (b) Split-3d module.

2) 3D-CNN: For the STF, we use the 3D-CNN branch to perform feature extraction for each band of M_0, M_1, and L_0 images. If conventional 3-D Conv is used, then the network parameters will increase significantly and consume much memory. Therefore, instead of regular 3-D Conv, we used a separable 3-D Conv (represented by Split-3d) by splitting the $3 \times 3 \times 3$ CK into two sets of $3 \times 1 \times 1$ and $1 \times 3 \times 3$ CK, and the Split-3d has essentially the same effect as directly using $3 \times 3 \times 3$ CK [41]. Take the first band of M_0, M_1, and L_0 as an example. At the beginning of the network, we stack them (denoted by stack) in dimension ($3 \times H \times W$), where H and W denote the image height and width, respectively. The dimension is expanded by twice decompression (represented by unsqueeze) to $1 \times 1 \times 3 \times H \times W$, the first dimension represents the batch size, and the second dimension represents the number of feature maps, after which the shallow features are obtained by Split-3d module. Afterward, the in-depth features are further extracted after several Split-3d modules, and the in-depth features are connected to the shallow features by jump connection and the output formula as

$$M_D = Y_D(Y_{D-1}(\ldots Y_1(M_0) + M_0 \ldots) + M_0) + M_0 \quad (8)$$

where Y_D denotes the operation of the dth Split-3d module, M_D is the result after processing by the dth Split-3d module, and the outputs from different Split-3d modules are connected by Concat and then passed through a convolutional layer with a CK size of $1 \times 1 \times 1$ to reduce the number of feature maps and improve the efficiency of model computation.

In each Split-3d module [shown in Fig. 4(b)], spatiotemporal features are extracted using $1 \times 3 \times 3$ and $3 \times 1 \times 1$ CKs, and local residual connections are added to the module. Similarly, the initial features M_0 are connected to the end of each Split-3d module. Compared with using 2-D Conv, 3-D Conv can improve the reconstruction performance of individual bands using the spatiotemporal variation information of both M_0^1 and M_1^1 bands and L_0^1 delicate spatial information, which will be demonstrated in the ablation experiments in Section IV.

D. Dual-Branch Feature Fusion and Band Feature Iteration

The reconstruction of the first-band feature F_1^1 and the first-band image L_1^1 is completed by the TSFF module. After the dual-branch feature extraction is completed and the number of feature maps is reduced by 1×1 and $1 \times 1 \times 1$ CK, respectively, we use the Pme module in the 3D-CNN branch to achieve dimensionality reduction and match the size of the 2D-CNN branch feature map. Then we use the Concat module to realize the dual-branch feature fusion and use 2-D Conv to complete feature reconstruction and reduce the number of feature maps to 1 to complete the first band image reconstruction. There are certain spatial structural similarities and spectral correlations between adjacent bands, and we use these properties to introduce adjacent band features to improve the reconstruction performance of individual band features. The subsequent image reconstruction of each band feature set will be done by the SSFF module. Take the rebuild of the second-band feature F_1^2 and the second-band image L_1^2 as an example. After the dual-branch feature extraction is completed, the first-band feature F_1^1 will be fused with it by the Concat module. Then, the rebuild of the second-band feature F_1^2 is completed by the two-layer 2-D Conv, and the reconstruction of the second-band L_1^2 is completed by the SPA-Atten module. The image reconstruction for each band is shown as follows:

$$L_1^i = \begin{cases} \mathrm{Conv}_3(\mathrm{Conv}_1(F_1^i)), & i = 1 \\ S(\mathrm{Conv}_3(\mathrm{Conv}_1(C(F_1^{i-1}, F_1^i)))), & 1 < i \le B \end{cases} \quad (9)$$

where $\mathrm{Conv}_1(\cdot)$ and $\mathrm{Conv}_3(\cdot)$, respectively, represent the dot product operation using 1×1 and 3×3 CKs, $S(\cdot)$ represents the SPA-Atten module, and the SPA-Atten module is the green rectangular frame part of the CBAM module in Fig. 1.

E. Loss Function

Zhao *et al.* [43] showed the importance of structural similarity (SSIM) loss and proposed a loss function consisting of mean absolute error (MAE) loss and multi-scale SSIM

(MS-SSIM) after comparing several loss functions, which has a significant improvement in image restoration results without changing the network structure [44]. Inspired by this, we use a compound loss function for the STF, and this compound loss function consists of content loss and vision loss, and the formula is shown below

$$L_{\text{HCNNet}} = L_{\text{content}} + \alpha L_{\text{vision}}. \tag{10}$$

L_{content} is calculated using MAE. α is a balance factor controlling visual loss, which is empirically set to 0.8. SSIM belongs to the perceptual correlation index. In the image reconstruction task, SSIM takes a value between 0 and 1, and the closer to 1 means the more similar the two images are. The SSIM index algorithm is a single-scale approach, which is a drawback of the method because the correct scale depends on viewing conditions (e.g., display resolution and viewing distance). MS-SSIM is an approach for image quality assessment, which provides more flexibility than the single-scale approach in incorporating the variations in image resolution and viewing conditions. The essence of multiscale is to continuously downsample the generated image and the actual image by a factor of 2 to obtain images with multiple resolutions. Furthermore, these images with different resolutions are evaluated for SSIM in turn, and finally, these SSIMs are fused into one value somehow. MS-SSIM is an improvement of SSIM. It has been shown that MS-SSIM can effectively preserve the high-frequency details of images [35]. Therefore, we use MS-SSIM for visual loss assessment and define the L_{vision} loss as the following, where P is the patch block of the predicted image

$$L_{\text{vision}} = 1 - \text{MS-SSIM}(P). \tag{11}$$

IV. Experimental Results and Analysis

In this section, to validate the effectiveness of the proposed method, we first present the dataset used by the model. After that, we give the specific parameter settings of the HCNNet network. Then we briefly introduce the image evaluation metrics. The experimental results are given and the performance of HCNNet is evaluated by comparing it with five STF models. Finally, we demonstrate the effectiveness of the proposed network structure and the related modules in it by ablation experiments.

A. Dataset Introduction

We conducted experiments on three publicly available datasets. The first dataset is located in the Coleambally irrigation area (CIA) in southern New South Wales, Australia. The CIA dataset consists of 17 cloud-free Landsat-MODIS image pairs taken from October 2001 to May 2002, with an image size of 3200 × 2720 and six bands per image. The size of the fields within the CIA region is relatively small and phenological variability dominates [45]. The second dataset, the lower gwydir catchment (LGC), is located in northern NSW, Australia. The LGC dataset consists of 14 cloud-free Landsat-MODIS image pairs captured from April 2004 to April 2005, with an image size of 1720 × 2040 and six bands per image. The region experienced a significant flood in mid-December 2004, which subsided in late December and was dominated by the changes in surface cover type. The third dataset, the Daxing dataset, includes 29 cloud-free Landsat-MODIS image pairs from September 2013 to November 2019, collected from the Daxing district located in the south of Beijing city, with an image size of 1640 × 1640 and six bands per image. The primary purpose of this dataset is to provide a benchmark for evaluating the performance of STF in detecting land cover changes [46]. The two short-wave infrared bands in the MODIS images in the Daxing dataset have significant band noise, and we did not consider these two bands in our experiments. Both the public dataset images have been atmospherically corrected. During the cropping process, the images can be cropped randomly while ensuring that the study area contains the main study target (e.g., fields, floods, buildings), but the cropping area of the images should be kept consistent across different datasets. We cropped three datasets separately to a size of 1600 × 1600. Totally, 17 Landsat and MODIS image pairs are available in the CIA dataset, and each reference image pair (moment t_0) is used to predict the image at the closest future moment (moment t_1) to it, which can be divided into 16 datasets, each consisting of two Landsat-MODIS image pairs, where M_0, M_1, and L_0 are used as training and L_1 is used as the target for validation. We randomly select 12 sets of data from the 16 sets as the training set and four sets of data as the test set. Similar to the CIA dataset, 14 Landsat and MODIS image pairs are available in the LGC dataset, and we randomly select ten sets of data from the 13 groups as the training set and three sets of data as the test set. In all, 29 Landsat and MODIS image pairs are available in the Daxing dataset, and we randomly select 23 sets of data from the 28 sets as the training set and five sets of data as the test set.

B. Parameter Settings

HCNNet is implemented in the PyTorch architecture. The network mainly uses 1 × 3 × 3, 3 × 1 × 1, 1 × 1 × 1, 1 × 1, and 3 × 3 CKs, with a small number of 7 × 7 CKs in the CBAM module and SPA-Atten module. The whole fusion model has 907 205 trainable weight parameters. We chose the Adam-optimized stochastic gradient descent method to optimize the network training parameters. The initial learning rate is set to 1e-4, and the learning rate will decrease to 0.1 of the initial learning rate if the loss does not improve for five consecutive epochs during the training process. Forty epochs are trained for the LGC dataset, and 50 epochs are trained for the CIA and Daxing datasets, which are larger than the LGC dataset. w_1 is taken as 0.4, and w_2 is taken as 0.6 for the dual-branch weights. The number of CBAM modules and split-3d modules is set to 5. The size of RS images is significant, and to reduce the memory occupancy, we cut the original 1600 × 1600 image into smaller 160 × 160 images, and the sliding step size is set to 160 × 160. These hyperparameter settings can be adjusted according to the experimenter's hardware device conditions and dataset. HCNNet is run in a Windows 10 Professional environment

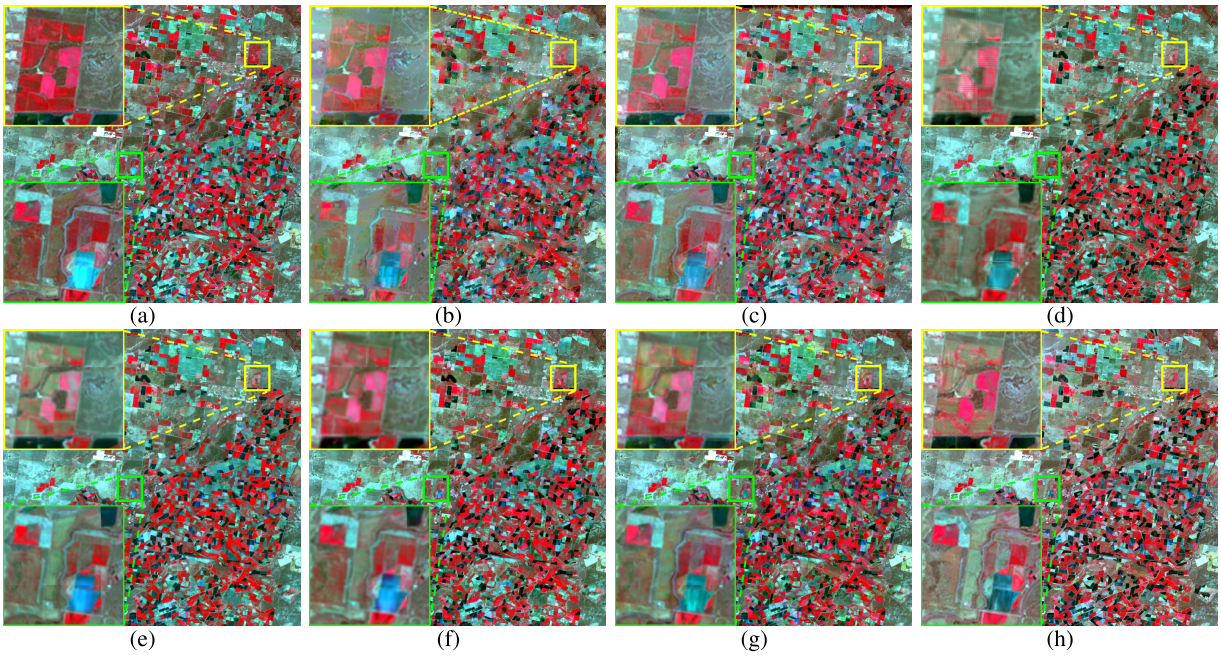

Fig. 5. Predicted results for the target Landsat image in CIA dataset. (a) Observed Landsat image (t_0). (b) Predicted by STARFM. (c) Predicted by FSDAF. (d) Predicted by DCSTFN. (e) Predicted by AMNet. (f) Predicted by EDCSTFN. (g) Predicted by the proposed method. (h) Observed Landsat image (t_1).

with a hardware environment including 32 GB RAM, Intel(R) Core(TM) i7-8700K CPU at 3.70 GHz, and an NVIDIA GeForce RTX 3090 with 24 GB RAM.

C. Comparison and Evaluation

To verify the effectiveness of the proposed models in this article, we compare the HCNNet model with the weighting-based fusion model STARFM, the unmixing-based model FSDAF, and the CNN-based DCSTFN, EDCSTFN, and AMNet models. For the objectivity and fairness of the experiments, we implemented the above fusion models in the same environment. We used root mean square error (RMSE) [33], SSIM, correlation coefficient (CC) [34], spectral angle mapper (SAM) [47], and erreur relative global adimensionnelle de synthese (ERGAS) [48] as objective evaluation metrics.

The CIA dataset has four sets of test data, and we selected a set of 20 011 017-20 011 102 test data to visualize the experimental results (20 011 017 represents the image observed at the reference moment t_0, i.e., October 17, 2001, and 20 011 102 represents the image observed at the prediction moment t1, i.e., November 2, 2001), as shown in Fig. 5. The LGC dataset has three sets of test data, and we selected a set of 20 041 212–20 041 228 test data to visualize the experimental results. The number sequence represents the same meaning as above, as shown in Fig. 6. The Daxing dataset has five sets of test data, and we selected a set of 20 181 001–20 181 204 test data to visualize the experimental results, as shown in Fig. 7.

In Fig. 5, the yellow and green rectangular boxes (both 650 × 650 in size) in the upper left and lower right corners of each image are obtained by enlarging the small yellow and central green rectangular boxes (both 130 × 130 in size) in the upper right corner of the figure by a factor of 5, respectively. Most of the area in the yellow rectangular box is planted with rice. With the change in Landsat images from t_0 to t_1 moments, we find that part of the crop in this region has been harvested. Both the models, STARFM and FSDAF, are less effective than DCSTFN, AMNet, and HCNNet in reflecting this changing trend in this region, but both are better than EDCSTF, with FSDAF slightly better than STARFM. The DCSTFN and AMNet models lose some spatial details and edge information in some local areas, although they better rebuild this overall change trend in the region. In addition to reflecting the overall trend of change, HCNNet is more effective in rebuilding the spatial details of some local areas.

The CIA dataset is mainly characterized by the change in phenology, with the color of the lower right area of the green rectangular box zoomed in on the Landsat image changing from blue to dark green from moment t_0 to t_1. The difference in image color predicted by all the models compared with the Landsat image at the prediction moment reflects the presence of some degree of spectral distortion in all the models. Among them, AMNet and EDCSTFN have the most severe distortion, and the color of this area is blue, which differs significantly from the actual surface, dark green. The STARFM and FSDAF models rebuild the spectra closely, but both are inferior to the DCSTFN and HCNNet models. HCNNet is close to the spectral results compared with DCSTFN but outperforms DCSTFN in local spatial detail reconstruction. From the above visual comparison, it can be seen that the prediction results of our proposed model are closest to the authentic Landsat images compared with other models.

In Fig. 6, the green rectangular box in the upper left corner of each image (both 1275 × 675 in size) is obtained by enlarging the small green rectangular box in the lower right corner (both 850 × 450 in size) by 1.5 times, respectively. The LGC dataset is located in a study area dominated by the changes in the land cover type, which experienced a

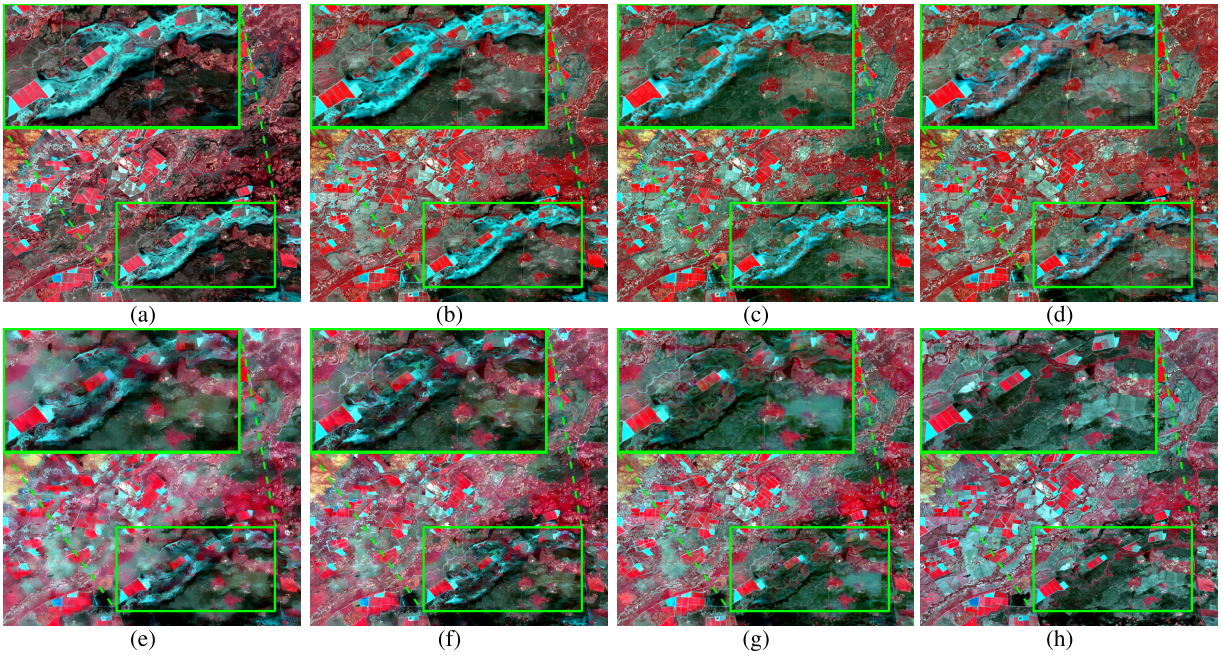

Fig. 6. Predicted results for the target Landsat image in the LGC dataset. (a) Observed Landsat image (t_0). (b) Predicted by EDCSTFN. (c) Predicted by DCSTFN. (d) Predicted by AMNet. (e) Predicted by FSDAF. (f) Predicted by STARFM. (g) Predicted by the proposed method. (h) Observed Landsat image (t_1).

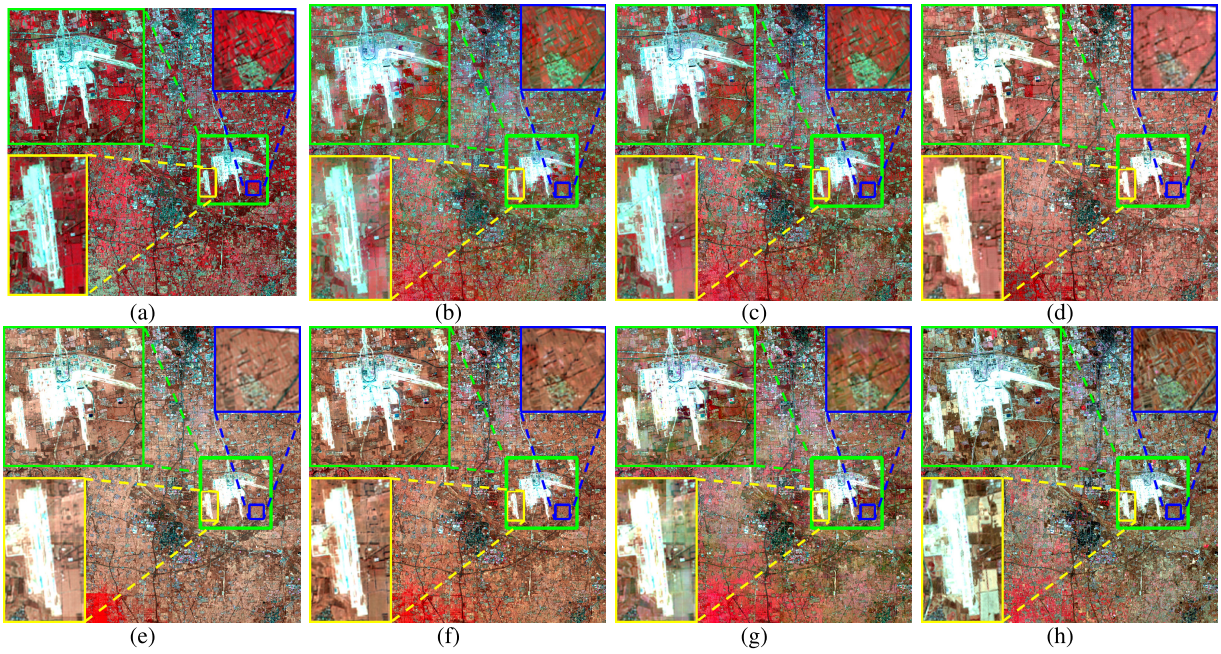

Fig. 7. Predicted results for the target Landsat image in the Daxing dataset. (a) Observed Landsat image (t_0). (b) Predicted by STARFM. (c) Predicted by FSDAF. (d) Predicted by DCSTFN. (e) Predicted by AMNet. (f) Predicted by EDCSTFN. (g) Predicted by the proposed method. (h) Observed Landsat image (t_1).

significant flood in mid-December 2004, followed by receding floods in late December. We selected a set of 20 041 212–20 041 228 test data to test the reconstruction effect of each model for the area of land-cover-type change. The image rebuilt by EDCSTFN is the worst, indicating that it has not learned the complex mapping relationship between MODIS and Landsat data, and the reconstruction result is influenced by the reference image, and the fusion result is very similar to the reference image. The DCSTFN model has a better result than the EDCSTFN model in that it predicts the process of flood receding in the local area. The AMNte model reconstruction results are better than the DCSTFN model, which predicts a larger area of flood recession. Both the STARFM and FSDAF models outperformed EDCSTFN, DCSTFN, and AMNet in

TABLE I
QUANTITATIVE ASSESSMENT OF DIFFERENT SPATIOTEMPORAL FUSION METHODS FOR THE CIA DATASET

Evaluation	Band	STARFM	FSDAF	DCSTFN	AMNet	EDCSTFN	Proposed
RMSE(↓)	Band1	0.0109	0.0156	0.0119	0.0111	0.0106	**0.0100**
	Band2	0.0138	0.0164	0.0135	0.0141	0.0137	**0.0125**
	Band3	0.0206	0.0224	**0.0193**	0.0221	0.0220	0.0201
	Band4	0.0404	0.0390	0.0325	0.0385	0.0336	**0.0322**
	Band5	0.0551	0.0552	0.0470	0.0544	0.0487	**0.0435**
	Band7	0.0465	0.0477	0.0400	0.0465	0.0424	**0.0368**
	Average	0.0312	0.0327	0.0274	0.0311	0.0285	**0.0258**
SSIM(↑)	Band1	0.9499	0.9407	0.9454	0.9488	0.9545	**0.9566**
	Band2	0.9368	0.9313	0.9355	0.9394	0.9432	**0.9465**
	Band3	0.8846	0.8854	0.8908	0.8882	0.8977	**0.9007**
	Band4	0.8296	0.8421	0.8627	0.8713	0.8771	**0.8831**
	Band5	0.7477	0.7563	0.7894	0.8020	0.8068	**0.8229**
	Band7	0.7479	0.7541	0.7923	0.8005	0.8099	**0.8249**
	Average	0.8494	0.8516	0.8693	0.8750	0.8815	**0.8891**
SAM(↓)		0.1917	0.2031	0.1764	0.1903	0.1711	**0.1651**
ERGAS(↓)		2.8537	3.0158	2.7873	2.9160	2.8882	**2.6945**
CC(↑)		0.9366	0.9356	0.9531	0.9410	0.9515	**0.9589**

TABLE II
QUANTITATIVE ASSESSMENT OF DIFFERENT SPATIOTEMPORAL FUSION METHODS FOR THE LGC DATASET

Evaluation	Band	STARFM	FSDAF	DCSTFN	AMNet	EDCSTFN	Proposed
RMSE(↓)	Band1	0.0224	0.0135	0.0156	0.0169	0.0157	**0.0109**
	Band2	0.0289	0.0177	0.0194	0.0230	0.0204	**0.0153**
	Band3	0.0390	0.0225	0.0251	0.0286	0.0258	**0.0209**
	Band4	0.0688	0.0332	0.0421	0.0507	0.0437	**0.0322**
	Band5	0.0824	0.0523	0.0442	0.0499	0.0471	**0.0358**
	Band7	0.0770	0.0486	0.0423	0.0438	0.0425	**0.0357**
	Average	0.0531	0.0313	0.0314	0.0355	0.0325	**0.0251**
SSIM(↑)	Band1	0.9286	0.9281	0.9196	0.9286	0.9300	**0.9475**
	Band2	0.9101	0.9017	0.8986	0.9068	0.9093	**0.9219**
	Band3	0.8729	0.8628	0.8467	0.8682	0.8719	**0.8828**
	Band4	0.8161	0.8162	0.7679	0.8113	0.8169	**0.8279**
	Band5	0.7369	0.7098	0.7288	0.7568	0.7656	**0.7977**
	Band7	0.7405	0.7117	0.7277	0.7624	0.7713	**0.7894**
	Average	0.8342	0.8217	0.8149	0.8390	0.8442	**0.8612**
SAM(↓)		0.1878	0.1907	0.1970	0.2060	0.1949	**0.1696**
ERGAS(↓)		2.8702	2.9098	2.9516	3.1546	3.0393	**2.6947**
CC(↑)		0.9415	0.9398	0.9227	0.9016	0.9202	**0.9508**

reflecting this changing trend of flood recession in this area, indicating that these two models have some advantages in rebuilding areas with abrupt changes in land cover types, with the STARFM model predicting better results than FSDAF. The HCNNet model, which is closest to the actual Landsat observation data, predicts the flood receding process best compared with other models and outperforms other models in reconstructing spatial detail information in local areas.

The building in the center of the green rectangle (380 × 380 in size) at the bottom right of Fig. 7 is Beijing Daxing Airport, and the enlarged rectangle (760 × 760 in size) at the top left corner shows that the land cover changes around the airport during the construction process from t_0 to t_1. Both STARFM and FSDAF perform less well than the deep-learning-based models in predicting the change in the extensive range of land cover changes around the airport, but DCSTFN, AMNet, and EDCSTFN reconstruct the colors of the area around the airport with significant differences compared with the t_1 image moment, and there is a problem of spectral distortion. We selected two areas near the airport for zooming in to compare the reconstruction effect of each model, represented by a yellow rectangular box (88 × 156 in size) and a blue rectangular box (77 × 77 in size) zoomed in five times and six times, respectively. Compared with the above methods, the proposed model performs better predicts both land-type changes and mitigating spectral distortions.

The magnified yellow rectangular box (440 × 780 in size) has buildings in the left half of the area, and its color changes from light blue to white from t_0 to t_1 moments, while most of the crops in the right half of the area are harvested and the color of the change area changes from red to green. In this region reconstruction process, the STARFM and FSDAF models reconstructed images are more influenced by the reference moment image and closer to the reference moment image. The proposed model has less spectral distortion than DCSTFN, AMNet, and EDCSTF and retains more spatial details in the local area. The magnified blue rectangular box region (462 × 462 in size) shows significant phenological changes. The reconstructed images from the STARFM and FSDAF models are closer to the reference moment images and fail

TABLE III

QUANTITATIVE ASSESSMENT OF DIFFERENT SPATIOTEMPORAL FUSION METHODS FOR THE DAXING DATASET

Evaluation	Band	STARFM	FSDAF	DCSTFN	AMNet	EDCSTFN	Proposed
RMSE(↓)	Band1	0.0199	0.0190	0.0189	0.0193	**0.0181**	0.0184
	Band2	0.0232	0.0229	0.0229	0.0230	0.0233	**0.0220**
	Band3	0.0320	0.0318	0.0286	**0.0277**	0.0281	0.0279
	Band4	0.0559	0.0562	0.0585	0.0556	0.0497	**0.0464**
	Average	0.0327	0.0325	0.0322	0.0314	0.0298	**0.0287**
SSIM(↑)	Band1	0.8768	0.8821	0.8697	0.8798	0.8856	**0.8861**
	Band2	0.8351	0.8364	0.8348	0.8459	0.8436	**0.8462**
	Band3	0.7316	0.7402	0.7669	**0.7806**	0.7766	0.7726
	Band4	0.5858	0.6009	0.6285	0.6480	0.6605	**0.6748**
	Average	0.7573	0.7649	0.7750	0.7886	0.7916	**0.7949**
SAM(↓)		0.2173	0.2141	0.2060	0.1973	0.1960	**0.1949**
ERGAS(↓)		3.0351	3.0101	2.9878	3.0058	2.9799	**2.8744**
CC(↑)		0.9451	0.9458	0.9488	0.9568	0.9582	**0.9597**

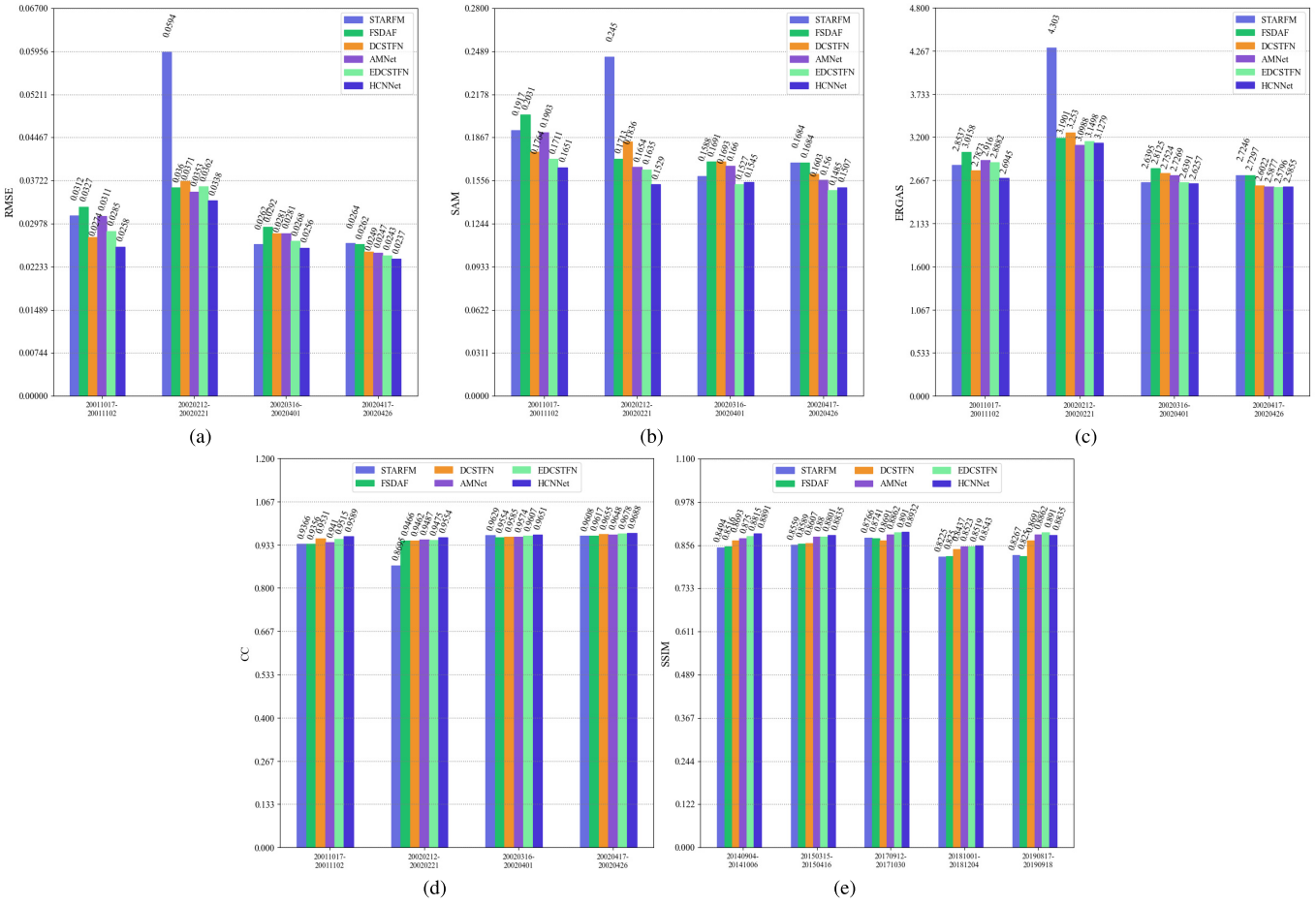

Fig. 8. Quantitative evaluation results of CIA four groups of test data (for root mean square error (RMSE), spectral angle mapper (SAM), relative dimensionless global error (ERGAS), correlation coefficient (CC), and structural similarity (SSIM), the values are averaged among all the six bands). (a) RMSE. (b) SAM. (c) ERGAS. (d) CC. (e) SSIM.

to reflect the significant phenological changes. The DCSTFN, AMNet, and EDCSTF models reflect the phenological changes to some degree but have severe spectral distortion compared with the observed images at t_1. Compared with the above models, the proposed model reflects the phenological changes in the region and alleviates the spectral distortion problem and has a better visual effect than other models.

Tables I–III give the evaluation results of each algorithm for the test data 20011017–20011102 in the CIA dataset, 20041212–20041228 in the LGC dataset, and 20181001–20181204 in the Daxing dataset, respectively. Based on the three sets of test data from the three datasets, we conclude that HCNNet can predict phenological changes and predict more accurately the areas where abrupt changes in

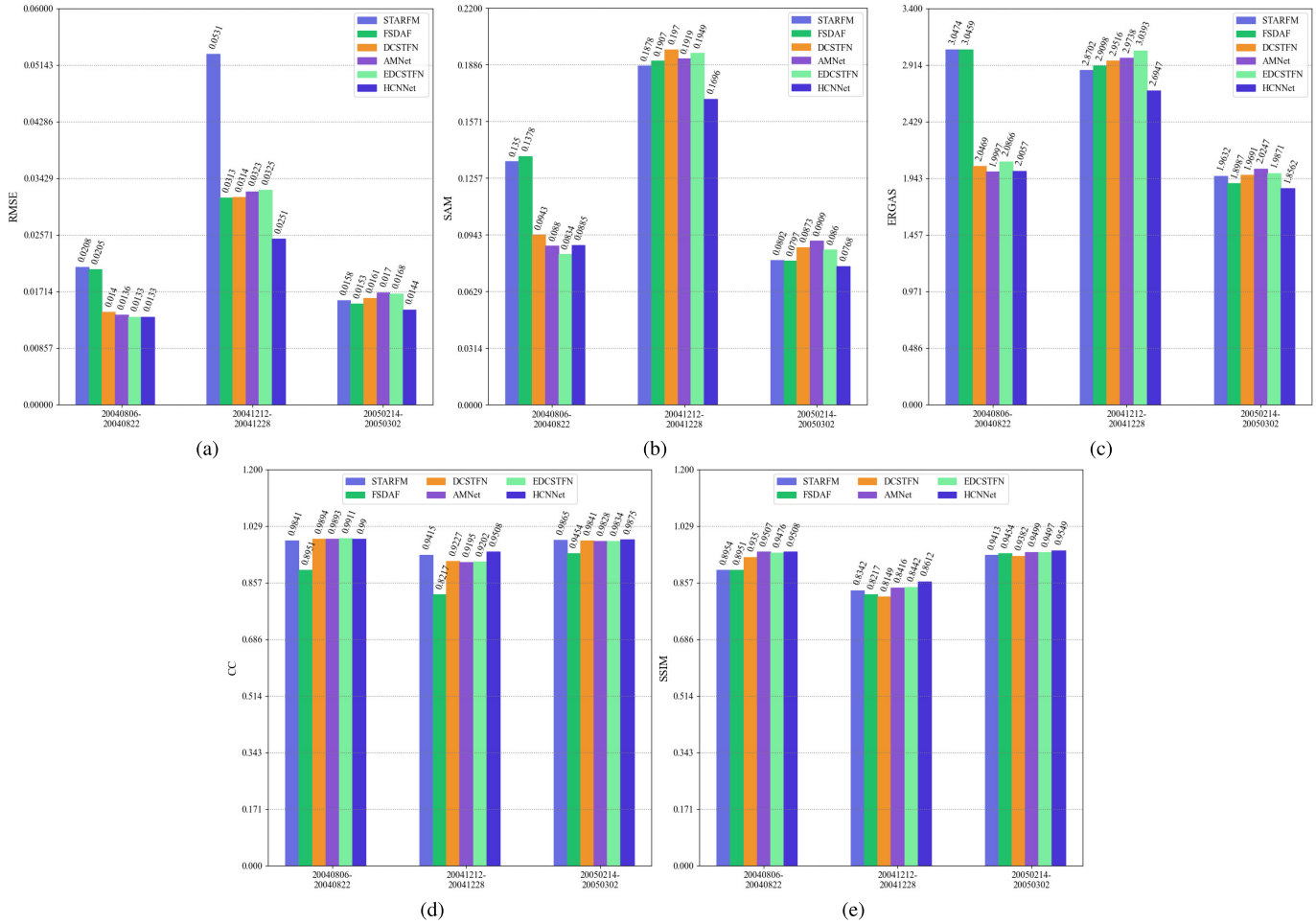

Fig. 9. Quantitative evaluation results of LGC three groups of test data (for root mean square error (RMSE), spectral angle mapper (SAM), relative dimensionless global error (ERGAS), correlation coefficient (CC), and structural similarity (SSIM), the values are averaged among all the six bands). (a) RMSE. (b) SAM. (c) ERGAS. (d) CC. (e) SSIM.

TABLE IV
AVERAGED QUANTITATIVE METRICS FOR CIA AREA ON THE TEST DATASET

Method	RMSE	SAM	ERGAS	CC	SSIM
STARFM	0.0358	0.1910	3.1302	0.9325	0.8511
FSDAF	0.0310	0.1780	2.9370	0.9498	0.8524
DCSTFN	0.0293	0.1724	2.8487	0.9558	0.8607
AMNet	0.0298	0.1694	2.8323	0.9530	0.8734
EDCSTFN	0.0289	0.1589	2.8142	0.9569	0.8761
Proposed	**0.0272**	**0.1558**	**2.7584**	**0.9621**	**0.8800**
Reference	0	0	0	1	1

TABLE V
AVERAGED QUANTITATIVE METRICS FOR LGC AREA ON THE TEST DATASET

Method	RMSE	SAM	ERGAS	CC	SSIM
STARFM	0.0299	0.1343	2.6269	0.9707	0.8903
FSDAF	0.0224	0.1361	2.6181	0.9704	0.8874
DCSTFN	0.0205	0.1262	2.3225	0.9654	0.8960
AMNet	0.0209	0.1236	2.3327	0.9639	0.9141
EDCSTFN	0.0209	0.1215	2.3710	0.9649	0.9138
Proposed	**0.0176**	**0.1116**	**2.1855**	**0.9761**	**0.9223**
Reference	0	0	0	1	1

land cover types occur while retaining more spatial detail information.

To verify the overall performance of each model on the three publicly available datasets, we further compare the performance of each algorithm. Fig. 8 shows the quantitative metrics for the four sets of CIA test data. These metrics are calculated for the entire image of each group. From the perspective of statistical error, the HCNNet prediction results have higher accuracy than the other models. Second, the prediction results of HCNNet remain relatively stable and show strong robustness compared with other methods. Table IV presents the average quantitative metrics for the CIA region on the entire test dataset. Each quantitative metric shows that the HCNNet model outperforms other methods, which indicates that the proposed model does improve the accuracy of the fusion.

Fig. 9 shows the quantitative metrics for the three sets of LGC test data. The HCNNet model shows significantly high scores for the metrics, and predicted results remain stable remain stable for all test data. The classical STARFM and

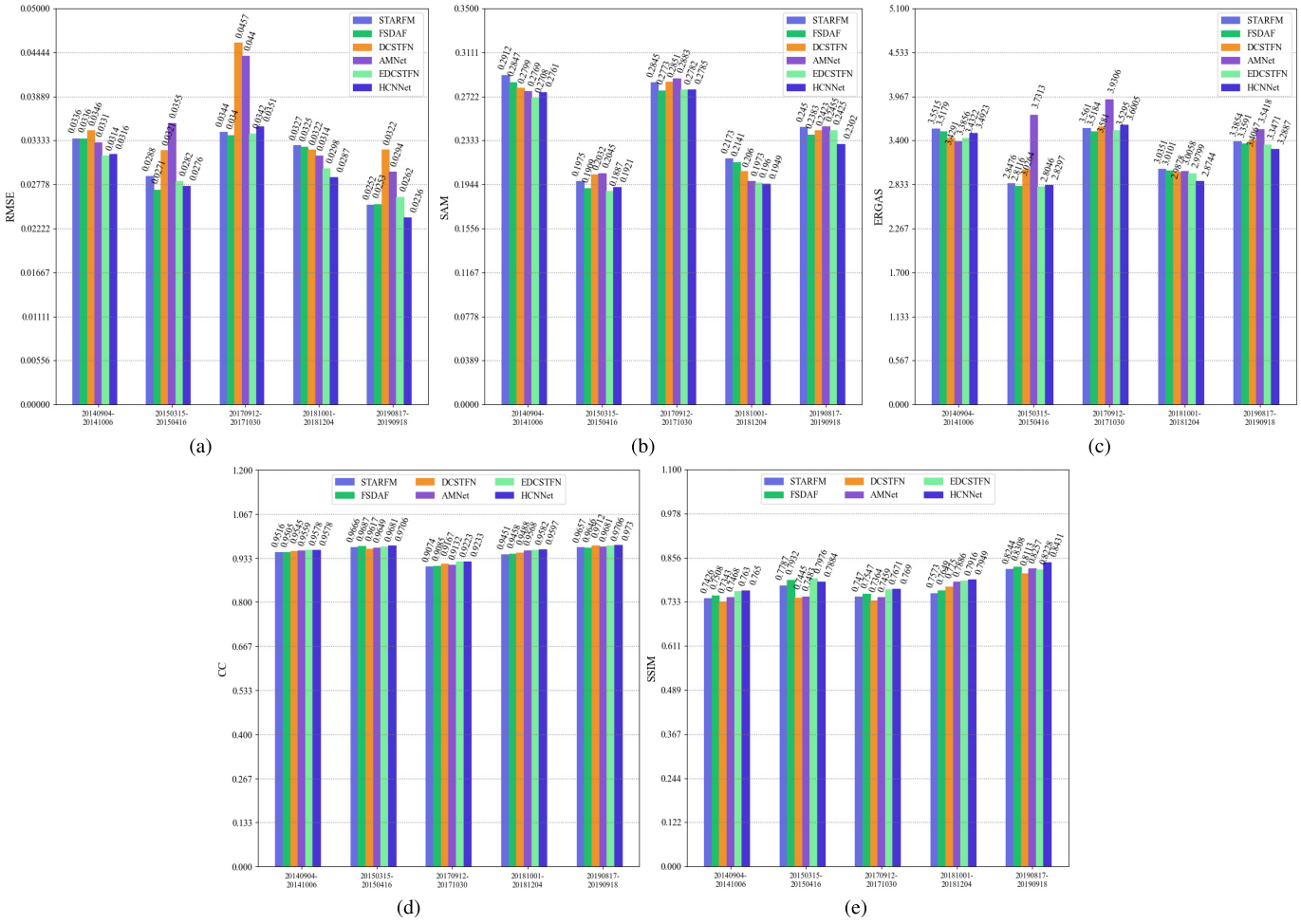

Fig. 10. Quantitative evaluation results of Daxing five groups of test data (for root mean square error (RMSE), spectral angle mapper (SAM), relative dimensionless global error (ERGAS), correlation coefficient (CC), and structural similarity (SSIM), the values are averaged among all the four bands). (a) RMSE. (b) SAM. (c) ERGAS. (d) CC. (e) SSIM.

TABLE VI

AVERAGED QUANTITATIVE METRICS FOR DAXING AREA ON THE TEST DATASET

Method	RMSE	SAM	ERGAS	CC	SSIM
STARFM	0.0309	0.2471	3.2761	0.9473	0.7700
FSDAF	0.0305	0.2411	3.2434	0.9476	0.7789
DCSTFN	0.0354	0.2433	3.2862	0.9506	0.7603
AMNet	0.0347	0.2425	3.5190	0.9518	0.7711
EDCSTFN	0.0299	0.2352	3.2187	0.9554	0.7884
Proposed	**0.0293**	**0.2343**	**3.2171**	**0.9569**	**0.7921**
Reference	0	0	0	1	1

FSDAF models have large fluctuations on different test data. The large fluctuations in the datasets indicate that they are not as robust as HCNNet. Table V presents the average quantitative metrics for LGC regions over the entire test dataset. The rows of the HCNNet model are shown in bold. Each quantitative metric shows that the HCNNet model outperforms the other methods, which is further evidence that our proposed approach can truly improve fusion accuracy.

Fig. 10 shows the quantitative metrics for the five Daxing test datasets. The predictions of HCNNet remain stable on all test data and outperform most of the models in all test sets. The classical STARFM and FSDAF models perform more consistently on this test data. Table VI lists the average quantitative metrics for the Daxing region on the entire test dataset. The best metrics among each of the metrics measured by each model are shown in bold, and the quantitative metrics all show that the HCNNet model outperforms the other methods.

D. Ablation Experiments

We conducted ablation experiments on the three datasets to analyze their impact on network performance by replacing or removing different modules. To demonstrate the effect of 3D-CNN on the structure of the HCNNet network, we replaced Split-3d with 2D-CNN, where the Pme module has no meaningful existence, and removed it, leaving the rest of the structure unchanged, defining this experimental method as "No-3D-CNN." To verify the effect of CBAM on the network structure, we replace the CBAM module with 2D-CNN and keep the rest of the structure unchanged, defining the sub-method as "No-CBAM." To verify the effect of Pme on the network structure, we remove the Pme module and leave the rest of the structure unchanged, and define this

TABLE VII
PERFORMANCE OF DIFFERENT ABLATION METHODS ON THE THREE DATASETS

Evaluation		No-3D-CNN(1)	No-CBAM(2)	No-Pme(3)	No-SPA(4)	No-SSFF(5)	Proposed(6)
RMSE(↓)	CIA	0.0289	0.0302	0.0284	0.0274	0.0282	**0.0272**
	LGC	0.0177	0.0177	0.0179	0.0178	0.0186	**0.0176**
	Daxing	0.0293	**0.0289**	0.0293	0.0292	0.0293	0.0293
SAM(↓)	CIA	0.1584	0.1695	0.1614	0.1565	0.1613	**0.1558**
	LGC	0.1146	0.1121	0.1128	0.1132	0.1155	**0.1116**
	Daxing	0.2387	0.2353	0.2381	0.2359	0.2369	**0.2343**
ERGAS(↓)	CIA	2.8621	2.8952	2.7904	2.7632	2.7914	**2.7584**
	LGC	2.2098	2.1886	2.2046	2.1964	2.2660	**2.1855**
	Daxing	3.2688	3.2358	3.2664	3.2362	**3.1996**	3.2171
CC(↑)	CIA	0.9610	0.9521	0.9574	0.9616	0.9582	**0.9621**
	LGC	0.9736	0.9755	0.9748	0.9749	0.9733	**0.9761**
	Daxing	0.9561	**0.9572**	0.9565	0.9566	0.9535	0.9569
SSIM(↑)	CIA	0.8788	0.8726	0.8764	0.8799	0.8754	**0.8800**
	LGC	0.9195	0.9214	0.9207	0.9209	0.9167	**0.9223**
	Daxing	0.7887	0.7913	0.7891	0.7908	0.7791	**0.7921**

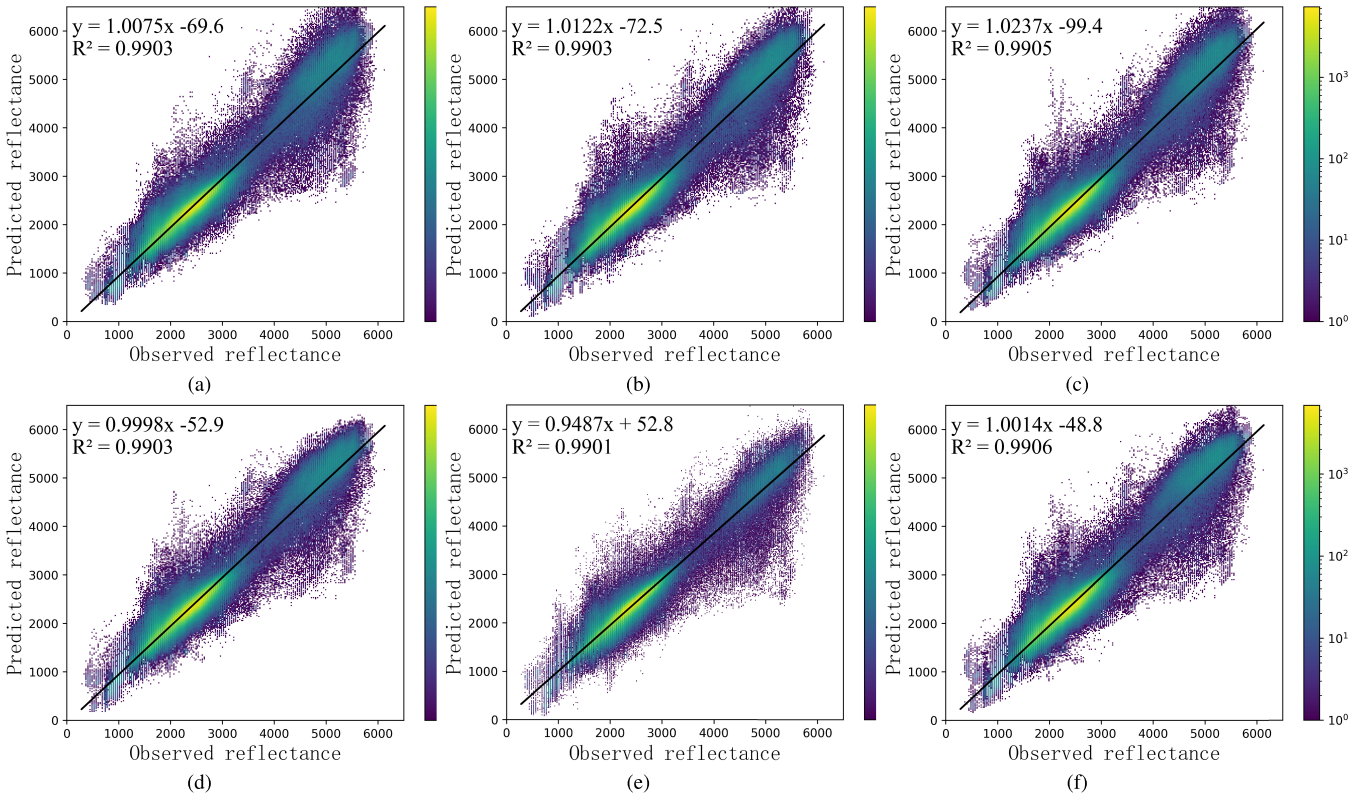

Fig. 11. Correlation between the NIR-band-predicted image and observed image on March 2, 2005, of LGC. (a) Result of No-3D-CNN. (b) Result of No-CBAM. (c) Result of No-Pme. (d) Result of No-SPA. (e) Result of No-SSFF. (f) Result of the proposed.

method as "No-Pme." To verify the effect of the SPA-Atten module on the network structure, we remove the SPA-Atten module and leave the rest of the structure unchanged, defining this method as "No-SPA." To verify the information sharing and complementarity of adjacent bands, we omit the step of feature transfer between adjacent bands, keep the rest of the structure unchanged, and define this method as "No-SSFF." The experimental results are shown in Table VII.

The coefficient of determination R^2 can effectively evaluate the correlation between the observed and predicted images. Its purpose is to fit the pixel values of the predicted and observed images to obtain a regression function and to determine a statistical indicator of their proximity [40], defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(x_i - y_i)^2}{\sum_{i=1}^{N}(x_i - \bar{x_i})^2} \tag{12}$$

where $\bar{x_i}$ represents the average pixel value of the observed image, x_i represents a certain pixel value of the observed image, y_i represents the corresponding pixel of the predicted image, and N represents the total number of pixels in the image. The performance of the predicted image improves when the value is closer to 1.

The metrics in Table VII show that the proposed complete HCNNet has the best prediction results. By comparing the results of objective evaluation metrics of the ablation methods

corresponding to serial numbers (2), (3), (4), and (5) with those of serial number (6), respectively, we can directly conclude that the inclusion of CBAM, Pme, SPA, and SSFF modules can effectively improve prediction results. By comparing the serial numbers (1) and (3), we can indirectly demonstrate that the inclusion of 3D-CNN can effectively improve prediction accuracy. Fig. 11 depicts the observed surface reflectance and prediction results for each ablation method. This figure visually illustrates how well the predicted results match the observed results under different experiments. We have chosen the near-infrared (NIR) band as a representative for a detailed presentation of the experimental samples. Although the "point clouds" of each comparison experiment are more similar, the point cloud of the proposed full HCNNet is more concentrated on the fit straight line, with fewer points scattered outside and smoother contour edges. The values of R^2 and the slope of the fit straight line with intercept in the figure also indicate that the results predicted by the complete HCNNet have a higher correlation with the observed images. The subjective inspection of the statistical metrics and correlation graphs indicates that the predicted images of the proposed complete HCNNet are closer to the observed images.

V. Conclusion

The short timespan of the CIA and LGC datasets and the relatively simple land cover scenarios make the prediction less complicated, and the indicators obtained by each model in the experiments of the CIA and LGC datasets were better than those of the Daxing dataset. The land cover of the LGC dataset changed abruptly, making prediction difficult, and the prediction results of each model in this dataset were more different, while the prediction results of HCNNet were optimal in terms of indicator evaluation and visualization. The experimental data in Tables I–III show that the strategy of band feature iteration gradually improves the reconstruction of subsequent bands. Compared with the six-band data from CIA and LGC, only four bands were used for experiments in the Daxing dataset. The prediction results of the proposed model in this dataset are affected to some extent, and if the two short-wave infrared bands of the Daxing dataset can be subsequently denoised to introduce more band feature information, the image spatial detail reconstruction effect may be improved.

The proposed method differs from previous STF methods in model data input and output, a multiband–single-band–multiband network structure was used. In the process of "multiband–single-band," we make full use of the observed data, consider the spectral and reflectance differences of RS images in each adjacent band, use the 2D-CNN branch to extract the spatial detail features of single-band images, and use 3D-CNN branch to extract the temporal and spatial variation features of images at the same time. The CBAM mechanism is introduced, and the Pme module is added to link the dual branches to further improve the feature extraction capability of the dual-branch network. In the process of "single-band–multiband," we consider the similarity of spatial structure and spectral correlation of each neighboring band, transfer each single-band feature between neighboring bands iteratively, and introduce the spatial attention mechanism to achieve spatial information sharing and complementarity between bands finally. The experimental results on three publicly available datasets show that HCNNet can effectively extract time change and spectral information while maintaining as many spatial details as possible. Compared with other fusion models, HCNNet is also more robust and has great potential to improve the prediction accuracy in landscapes with heterogeneous and large-scale abrupt land cover changes.

References

[1] C. Toth and G. Jóźków, "Remote sensing platforms and sensors: A survey," *ISPRS J. Photogramm. Remote Sens.*, vol. 115, pp. 22–36, May 2016.

[2] Z. Yang, C. Diao, and B. Li, "A robust hybrid deep learning model for spatiotemporal image fusion," *Remote Sens.*, vol. 13, no. 24, p. 5005, Dec. 2021.

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

[4] G. Patanè and M. Spagnuolo, "Heterogenous spatial data: Fusion, modeling, and analysis for GIS applications," *Synth. Lectures Vis. Comput.*, vol. 8, no. 2, pp. 1–155, Apr. 2016.

[5] A. Puissant, J. Hirsch, and C. Weber, "The utility of texture analysis to improve per-pixel classification for high to very high spatial resolution imagery," *Int. J. Remote Sens.*, vol. 26, no. 4, pp. 733–745, 2005.

[6] J. Pisek, M. Lang, and J. Kuusk, "A note on suitable viewing configuration for retrieval of forest understory reflectance from multi-angle remote sensing data," *Remote Sens. Environ.*, vol. 156, pp. 242–246, Jan. 2015.

[7] Y. Li, C. Huang, and X. Ye, "Rural ecosystem planning in high-resolution remote sensing and GIS environment: Case study of beautiful countryside planning of Wuyi County, China," in *Proc. 19th Int. Conf. Geoinform.*, Jun. 2011, pp. 1–5.

[8] X. Li, Y. Zhou, G. R. Asrar, J. Mao, X. Li, and W. Li, "Response of vegetation phenology to urbanization in the conterminous United States," *Global Change Biol.*, vol. 23, no. 7, pp. 2818–2830, Jul. 2017.

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

[10] 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, Aug. 2014.

[11] Y. Ma, F. Chen, J. Liu, Y. He, J. Duan, and X. Li, "An automatic procedure for early disaster change mapping based on optical remote sensing," *Remote Sens.*, vol. 8, no. 4, p. 272, Mar. 2016.

[12] W. Shi and M. Hao, "A method to detect earthquake-collapsed buildings from high-resolution satellite images," *Remote Sens. Lett.*, vol. 4, no. 12, pp. 1166–1175, Dec. 2013.

[13] Z. Fu, Y. Sun, L. Fan, and Y. Han, "Multiscale and multifeature segmentation of high-spatial resolution remote sensing images using superpixels with mutual optimal strategy," *Remote Sens.*, vol. 10, no. 8, p. 1289, 2018.

[14] V. R. Pandit and R. J. Bhiwani, "Image fusion in remote sensing applications: A review," *Int. J. Comput. Appl.*, vol. 120, no. 10, pp. 22–32, Jun. 2015.

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

[16] 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.

[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.

[18] 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.

[19] W. Ming-Quan, W. Jie, N. Zheng, Z. Yong-Qing, and W. Chang-Yao, "A model for spatial and temporal data fusion," *J. Infr. Millim. Waves*, vol. 31, no. 1, p. 80, 2012.

[20] W. Zhang *et al.*, "An enhanced spatial and temporal data fusion model for fusing Landsat and MODIS surface reflectance to generate high temporal Landsat-like data," *Remote Sens.*, vol. 5, no. 10, pp. 5346–5368, 2013.

[21] Z. Niu, "Use of MODIS and landsat time series data to generate high-resolution temporal synthetic landsat data using a spatial and temporal reflectance fusion model," *J. Appl. Remote Sens.*, vol. 6, no. 1, Mar. 2012, Art. no. 063507.

[22] M. Wu, W. Huang, Z. Niu, and C. Wang, "Generating daily synthetic landsat imagery by combining landsat and modis data," *Sensors*, vol. 15, no. 9, pp. 24002–24025, 2015.

[23] M. Liu *et al.*, "An improved flexible spatiotemporal DAta fusion (IFSDAF) method for producing high spatiotemporal resolution normalized difference vegetation index time series," *Remote Sens. Environ.*, vol. 227, pp. 74–89, Jun. 2019.

[24] Shi *et al.*, "A comprehensive and automated fusion method: The enhanced flexible spatiotemporal DAta fusion model for monitoring dynamic changes of land surface," *Appl. Sci.*, vol. 9, no. 18, p. 3693, Sep. 2019.

[25] D. Guo, W. Shi, M. Hao, and X. Zhu, "FSDAF 2.0: improving the performance of retrieving land cover changes and preserving spatial details," *Remote Sens. Environ.*, vol. 248, Oct. 2020, Art. no. 111973.

[26] 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.

[27] C. Zhao, X. Gao, W. J. Emery, Y. Wang, and J. Li, "An integrated spatio-spectral–temporal sparse representation method for fusing remote-sensing images with different resolutions," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3358–3370, Jun. 2018.

[28] L. H. J. C. A. P. Yunfei Li and J. Li, "A sensor-bias driven spatiotemporal fusion model based on convolutional neural networks," *Sci. China Inf. Sci.*, vol. 63, no. 4, pp. 1–16, Apr. 2019.

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

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

[31] 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.

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

[33] 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, p. 1066, Jul. 2018.

[34] X. Liu, C. Deng, J. Chanussot, D. Hong, and B. Zhao, "StfNet: A two-stream convolutional neural network for spatiotemporal image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6552–6564, Sep. 2019.

[35] 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, p. 2898, Dec. 2019.

[36] 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.

[37] Q. Li, Q. Wang, and X. Li, "Exploring the relationship between 2D/3D convolution for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8693–8703, Oct. 2021.

[38] Q. Wang, Q. Li, and X. Li, "Hyperspectral image superresolution using spectrum and feature context," *IEEE Trans. Ind. Electron.*, vol. 68, no. 11, pp. 11276–11285, Nov. 2021.

[39] Q. Li, Q. Wang, and X. Li, "Mixed 2D/3D convolutional network for hyperspectral image super-resolution," *Remote Sens.*, vol. 12, no. 10, p. 1660, May 2020.

[40] J. Li, R. Cui, Y. Li, B. Li, Q. Du, and C. Ge, "Multitemporal hyperspectral image super-resolution through 3D generative adversarial network," in *Proc. 10th Int. Workshop Anal. Multitemporal Remote Sens. Images (MultiTemp)*, Aug. 2019, pp. 1–4.

[41] Q. Wang, Q. Li, and X. Li, "Spatial-spectral residual network for hyperspectral image super-resolution," 2020, *arXiv:2001.04609*.

[42] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[43] 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.

[44] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 47–57, Mar. 2017.

[45] 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.

[46] 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, pp. 1–17, Apr. 2020.

[47] R. Yuhas, A. F. H. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm," in *Proc. Summaries 3rd Annu. JPL Airborne Geosci. Workshop*, vol. 1, Jun. 1992, pp. 147–149.

[48] M. M. Khan, L. Alparone, and J. Chanussot, "Pansharpening quality assessment using the modulation transfer functions of instruments," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 11, pp. 3880–3891, Nov. 2009.

Zhuangshan Zhu received the B.S. degree from the China University of Mining and Technology, Beijing, China, in 2018. He is pursuing the master's degree with the Chongqing Engineering Research Center of Spatial Big Data Intelligence Technology, Chongqing, China.

His research focuses on spatiotemporal fusion of remote sensing images.

Yuxiang Tao received the Ph.D. degree in geochemistry from the Chinese Academy of Sciences, Beijing, China, in April 1994, focusing on remote sensing intelligent computing, resource, and environmental economics.

Xiaobo Luo received the B.S. degree in GIS from the Institute of Geophysics and Geomatics, China University of Geosciences, Wuhan, China, in 1999, the M.S. degree in GIS from the School of Information Engineering, Chinese Academy of Sciences, Wuhan, in 2004, and the Ph.D. degree in cartography and GIS from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China, in 2010.

He is a Professor with the Chongqing University of Posts and Telecommunications, Chongqing, China. His research interests include urban thermal infrared remote sensing, remote sensing image processing, and ecological environment monitoring and evaluating.