A Novel Remote Sensing Spatiotemporal Data Fusion Framework Based on the Combination of Deep-Learning Downscaling and Traditional Fusion Algorithm

Dunyue Cui, Shidong Wang, Cunwei Zhao, and Hebing Zhang

***Abstract*—Traditional remote sensing spatiotemporal data fusion algorithms generally use upsampled low-resolution images (MODIS) to be fused with high-resolution images (Landsat); this makes both images less spatially consistent and many hybrid image elements in low-resolution images, so uncertainty errors propagate into the fusion results. To address this issue, we propose a framework for combining deep-learning-based super-resolution techniques with traditional spatiotemporal fusion methods. By reconstructing low-resolution images using super-resolution image reconstruction techniques, we obtain low-resolution images with more spatial details and better spatial consistency with high-resolution images. These reconstructed images are then fused using spatiotemporal fusion methods. In this study, we selected flexible spatiotemporal data fusion (FSDAF) and residual channel attention network (RCAN) to carry out a detailed study to prove the effectiveness of this kind of framework. That is, a new RCAN-FSDAF model is developed. After testing, RCAN-FSDAF has the following advantages: First, the band reflectance predicted by RCAN-FSDAF is closer to base reflectance than FSDAF, DMNet, and GAN-STFM, as shown by greater correlation and smaller error. Second, RCAN-FSDAF better decomposes image elements among heterogeneous features and more accurately identifies boundaries between different features and changes in land-cover type. Third, high spatial and temporal resolution NDVI data obtained by the inversion of the prediction results of RCAN-FSDAF are more accurate. The framework developed in this study can be extended to other spatial and temporal data fusion applications.**

***Index Terms*—Data fusion, reflectance, super-resolution techniques.**

I. Introduction

REMOTE sensing data observed by different satellite sensors have different temporal, spatial, and spectral resolutions [1], and high resolution in one mode may not imply high resolution in another. High temporal resolution remote sensing data, such as MODIS data, can have a temporal resolution of up to 1 d and a spatial resolution of up to 250 m. Such data are widely used by researchers due to their frequent revisiting and wide scanned area and have been used for a long time in applications, such as large area vegetation monitoring [2], crop yield estimation [3], and weather monitoring [4]. However, the low spatial resolution leads to a lack of definition in observing small surface features [5]. Single multispectral panchromatic satellite data products with long time series and high resolution (such as the Landsat satellite series, 16 d, 30 m) can be used in fine-scale studies of some regions, but the low temporal resolution and poor data quality disrupt temporal continuity and reduce overall coverage, which limit their application [6]. The support of high spatiotemporal resolution remote sensing data is needed for fine-scale studies of crop and forage yield estimation, disaster monitoring, and the surface dynamics of heterogeneous areas in specific regions and at specific times. Spatiotemporal data fusion techniques for multisource remote sensing data combine the detail of high spatial resolution remote sensing data with the temporal variation of high temporal resolution remote sensing data to generate high-resolution spatiotemporal data [7], [8]. Five types of fusion techniques are commonly used: weighted function-based fusion models, mixed image element decomposition-based fusion models, dictionary pair learning-based fusion models, combinatorial fusion methods, and neural-network-based fusion methods [9].

Gao et al. [10] developed a spatiotemporal adaptive reflection fusion model (STARFM) in 2006. It was the first spatiotemporal data fusion model to weigh fused data. The data fusion of remote sensing data was achieved by extracting spectrally similar neighboring information from the two types of data being fused and calculating weights for the surface reflectance of the central pixel. This approach was good for regions with homogeneous surface properties [9]. It performed badly in the fusion of data for heterogeneous areas and areas with land-cover changes because it did not accurately separate features from mixed

Manuscript received 19 August 2023; revised 9 November 2023, 28 December 2023, and 9 March 2024; accepted 22 March 2024. Date of publication 27 March 2024; date of current version 12 April 2024. This work was supported in part by the National Natural Science Foundation of China under Grant U22A20620 and Grant U21A20108, in part by the Natural Science Foundation of Henan Polytechnic University under Grant B2023-21, and in part by Henan Province Science and Technology Research Project under Grant 232102320276. *(Corresponding author: Shidong Wang.)*

Dunyue Cui, Shidong Wang, and Hebing Zhang are with the School of Surveying and Land Information Engineering, Henan Polytechnic University, Jiaozuo 454003, China (e-mail: c17832177427@163.com; wsd0908@163.com; jzitzhb@163.com).

Cunwei Zhao is with the Handan Guangkai Land Planning and Design Company Ltd., Handan 056004, China (e-mail: zhaocunwei2000@163.com).

Digital Object Identifier 10.1109/JSTARS.2024.3382136

image elements and, therefore, did not properly differentiate features and was prone to patchiness. Zhu et al. [11] developed an enhanced spatiotemporal adaptive reflection fusion model (ESTARFM) to improve the STARFM algorithm; it introduced conversion coefficients to mediate between the reflectances of coarse-resolution images and fine-resolution images. The model significantly improved the retention of spatial details and the accuracy of prediction for heterogeneous landscapes. Zhukov et al. [12] developed a multisensor multiresolution technique (MMT). It was the first method to fuse images for different times and spatial resolutions and also the first fusion model based on mixed image element decomposition. Many subsequent studies improved on MMT. For example, Gevaert and García-Haro [13] isolated changes in coarse pixels to predict changes in end elements and used Bayesian theory to constrain the predictions and produce more accurate results. Huang and Song [14] developed the sparse representation-based spatiotemporal reflection fusion model (SPSTFM). This was probably the first model to include dictionary learning techniques from image super-resolution in spatiotemporal data fusion. The model learned correspondences between high- and low-resolution images to predict high-resolution images for particular dates. The model accurately detected physical changes and land-cover type changes, but it was unstable and produced inaccurate predictions for highly heterogeneous regions [9]. Zhu et al. [15] developed a flexible spatiotemporal fusion model (FSDAF) that introduced the use of combinatorial fusion methods. This approach combined two or more different fusion methods and leveraged the advantages of each method to produce better fusion results. FSDAF used unsupervised classification to determine categories and then calculated low-resolution and high-resolution images of predicted land-cover changes and transformations. The model applied to heterogeneous landscapes and could process more spatial details than other methods. Many researchers, in China and elsewhere, have recently used deep-learning techniques in spatiotemporal data fusion. Song et al. [16] developed a method of spatiotemporal satellite image fusion based on a deep convolutional neural network. The key concept was similar to that underlying single-pair SPSTFM, but the researchers replaced dictionary pair learning with deep convolutional neural network learning and used a nonlinear model and a super-resolution model to learn the mapping relationships between the two kinds of data (MODIS data and Landsat data). The predicted images were treated as transition images, and the fusion model was used to improve the fusion results. Tan et al. [17], [18] proposed DSTFN and EDSTFN in 2018 and 2019, respectively, which utilize convolution to extract key features from the input data, change the size of the low spatial resolution image using the inverse convolution, and fuse the features extracted in the high spatial resolution image and the low spatial resolution image with the help of an equation that takes into account the change in ground cover in time. Li et al. [19] proposed DMNet in 2020, which combines dilation convolution and multiscale mechanism. Dilated convolution extends the receptive field of the convolution kernel, which is favorable for the extraction of small detail features, and the multiscale mechanism can extract the contextual information of the image at different scales, which makes the image details richer. Tan et al. [20] introduced conditional generative adversarial network and normalization techniques into the remote sensing spatiotemporal fusion problem to alleviate the strong temporal dependence between the reference and predicted images in the existing fusion model; GAN-STFM model is proposed and achieved better results. These show better performance compared with traditional methods, such as FSDAF.

Research into spatiotemporal data fusion of remote sensing images has made great progress until now, and many novel accurate spatiotemporal data fusion methods have been produced. We found in our literature survey that most traditional spatiotemporal data fusion methods relied on general resampling to upsample input data. However, there are more mixed pixels in heterogeneous areas and areas with changing land-cover types than in homogeneous areas; upsampling based on difference or reconstruction does not accurately decompose mixed pixels, and high-resolution images have poor spatial consistency with low-resolution images. In the conventional fusion methods, this leads to poor predictions of the boundaries between features. The accuracy of spatiotemporal data fusion depends largely on the accuracy of the input data, so an improved or innovative input data downscaling method can improve the performance of most spatiotemporal data fusion methods; for example, the authors in [21] and [22] produced good results, and Li et al. [23] proposed a MODIS strip noise cancelation strategy for spatiotemporal fusion methods, which better improves the accuracy of the fusion results. In addition, we found that compared with traditional spatiotemporal fusion methods, the deep-learning-based spatiotemporal fusion algorithms bring more noise to the image fusion process and ignore the advantages of traditional spatiotemporal fusion algorithms in terms of image element unmixing and land-use type change prediction. However, deep-learning-based methods have greater efficiency and generalization capabilities. So, considering the advantages of each, we found a new framework for combining the deep-learning-based super-resolution techniques used in deep-learning-based fusion methods with traditional spatiotemporal methods. For deep-learning-based super-resolution techniques, the more widely used ones mainly include channel-based attention mechanism, second-order attention mechanism, pixel-based attention mechanism, transformer self-attention mechanism, and multiscale large kernel-based attention mechanism. Typical examples of each of them are SRCNN proposed by Dong et al. [24] and residual channel attention network (RCAN) proposed by Zhang et al. [25], (pan sharpening in closed-loop regularization and modality-aware feature integration for pan sharpening) proposed by Zhou et al. [26], [27] and CUCaNet proposed by Zheng et al. [28], SwinIR proposed by Liang et al. [29], and MAN proposed by Wang et al. [30].

In this study, to solve the issue of spatial inconsistency between the upsampled low-resolution images and the high-resolution images, reduce the impact of hybrid image elements and uncertainty errors on fusion results. We propose a framework for combining deep-learning-based super-resolution techniques with traditional spatiotemporal fusion methods. FSDAF in the traditional spatiotemporal approach and RCAN in the

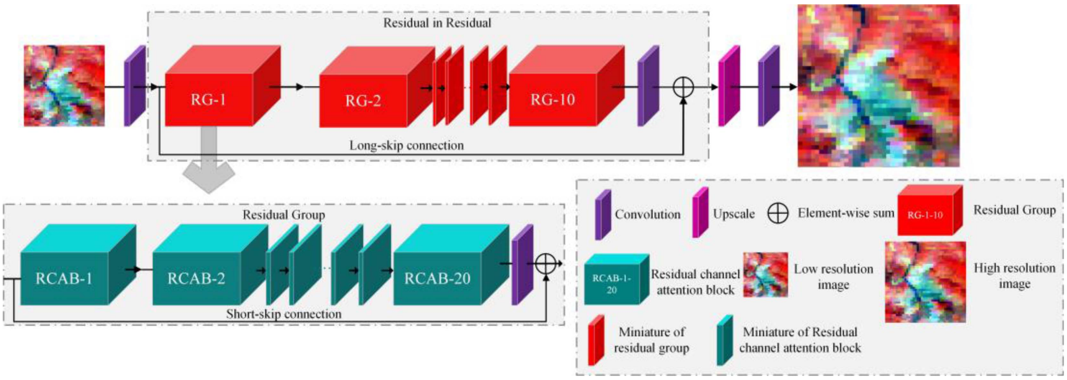

Fig. 1. Flowchart of RCAN (photograph referenced from Zhang et al. article [25]).

channel-based attention mechanism are selected to carry out a detailed study to prove the effectiveness of this kind of framework. That is, we developed an innovative RCAN-FSDAF model. RCAN-FSDAF has the following advantages. RCAN first learns the mapping relationship between high-resolution and low-resolution images and is trained to reach the optimal peak signal-to-noise ratio (PSNR). The trained RCAN can then downscale the low-resolution images, enable the downscaled low-resolution image to have higher spatial consistency with the high-resolution image, and FSDAF fuses spatiotemporal data. High spatiotemporal resolution data with higher accuracy and more spatial details are thereby produced. To critically evaluate and assess the reliability and applicability of the new framework, the deep-learning-based DMNet and GAN-STFM method is introduced as a comparative method and is tested in detail on three trials.

II. Materials and Methods

A. RCAN Downscaling Method

An RCAN is a deep-learning network used in super-resolution image reconstruction. It was developed by Zhang et al. [25] and it performs better and produces better super-resolution images than other networks. Low spatial resolution images contain both low-frequency information, which is flatter, and high-frequency information, which is relatively full of edges, textures, and other details. RCAN improved on super-resolution reconstruction techniques that use a residual network by introducing the residual channel attention (CA) mechanism. Interdependencies between channels adaptively adjust channel characteristics and ignore much of the low-frequency information, thus improving the performance of the network [31]. Fig. 1 shows the overall structure of RCAN. The four principal components of RCAN are a single-layer convolution for extracting shallow features, a residual-in-residual (RIR) module for extracting deep features, an upsampling module, and a reconstruction module. Convolution kernel size is 3 × 3, RIR is the most complex module and includes ten residual groups (RGs) and ten long jump connections. Each RG includes 20 residual channel attention modules (RCAB) and 20 short jump connections. RCAB is the CA integrated into the residual module. The RIR reaches a depth of over 400 layers, the other layers are single convolution layers or pixel alignment layers.

For the specific training settings, the learning rate set for RCAN in this study was 1e-8, the patch size was set to 192, the activation function used a rectified linear unit, and the L1 loss function was used as the optimizer by default.

B. FSDAF Spatiotemporal Fusion

Zhu et al. [15] developed an FSDAF approach in 2016. FSDAF requires only one high-resolution image and two low-resolution images to capture gradual and abrupt land-cover type changes to accurately predict high-resolution images in heterogeneous regions. The high-resolution image at time t_1 is classified to obtain the weight f_c of each feature class contained in each low spatial resolution image, which is calculated by

$$f_c\ (x_i, y_i) = N_c\ (x_i, y_i)/m \tag{1}$$

where c denotes the feature class, $N_c(x_i, y_i)$ is the number of high-resolution images (x_{ij}, y_{ij}) corresponding to feature category c in the low spatial resolution image element (x_i, y_i), and m is the total number of high-resolution images corresponding to (x_i, y_i) in the low spatial resolution image element.

The n ($n > l$) image elements for which the ground-cover type has not changed from time t_1 to t_2 are selected and the time change $\Delta F(c)$ of each band is obtained by linear spectral decomposition using

$$\Delta \mathrm{C}\ (x_i, y_i) = \sum_{c=1}^{l} f_c\,(x_i, y_i) \times \Delta F\ (c) \tag{2}$$

where l is the total number of categories obtained by unsupervised classification. The predicted value of change in F_2^{TP} over time for each band is given by

$$F_2^{TP}\ (x_{ij}, y_{ij}) = F_1\ (x_{ij}, y_{ij}) + \Delta F\ (c)\,. \tag{3}$$

Although high-resolution images at time t_2 (time-varying predicted values) are produced, F_2^{TP} does not accurately represent the predicted results when changes occur within a land-use class or when the type of ground-cover changes. The algorithm, therefore, introduces a residual $\mathrm{R}(x_i, y_i)$ to represent the difference between the base value and the predicted value F_2^{TP}. The parameter $\mathrm{R}(x_i, y_i)$ is given by

$$\mathrm{R}\ (x_i, y_i) = \ \Delta\mathrm{C}\,(x_i, y_i) - \frac{1}{m}\left[\sum_{j\,=\,1}^{m} F_2^{TP}\,(x_{ij}, y_{ij}) - \sum_{j\,=\,1}^{m} F_1\,(x_{ij}, y_{ij})\right]. \quad (4)$$

The algorithm introduces a thin slab spline function that is used to calculate the spatially varying prediction F_2^{TP} at time t_2 as follows. The basic thin plate spline function (TPS) is defined by

$$f_{TPS}\ (x, y) = a_0\ + a_1 x + a_2 y + \frac{1}{2}\sum_{i-1}^{N} b_i r_i^2 \mathrm{log} r_i^2 \quad (5)$$

$$\sum_{i\,=\,1}^{N} b_i = \sum_{i=1}^{N} b_i\, x_i = \sum_{i=1}^{N} b_i y_i\ = \ 0 \quad (6)$$

$$r_i^2 = (x - x_i)^2\ + (y - y_i)^2. \quad (7)$$

When $\sum_{i\,=\,1}^{N} [C_2(x_i, y_i, b) - f_{TPS-b}(x_i, y_i)]^2$ reaches a minimum, the coefficients in the equation are optimized, and after optimizing the parameters in the TPS function, the predicted values of the spatial variation of each high-resolution image element are predicted by

$$F_2^{SP}\ (x_{ij}, y_{ij}) = f_{TPS}\ (x_{ij}, y_{ij})\,. \quad (8)$$

Assigning the residuals to the corresponding high spatial resolution image elements within each low spatial resolution image element is key to increasing the accuracy of temporal prediction. The algorithm is designed with a new weighting function $\mathrm{CW}(x_{ij}, y_{ij})$ to better assign the residuals. The specific method is

$$\mathrm{CW}\ (x_{ij}, y_{ij}) = E_{ho}\ (x_{ij}, y_{ij}) \times HI\,(x_{ij}, y_{ij}) + \ R\,(x_i, y_i) \times [1 - HI\,(x_{ij}, y_{ij})] \quad (9)$$

where $HI\ (x_{ij}, y_{ij}) = (\sum_{k=1}^{m} I_k)/m$; $I_k = \ 1$ when high-resolution image element k in a moving window belongs to the same class as the central image element; otherwise, $I_k = 0$. HI ranges from 0 to 1, and larger values indicate a more homogeneous study area. Normalizing the weight $\mathrm{CW}(x_{ij}, y_{ij})$ gives $\mathrm{W}(x_{ij}, y_{ij})$, and the residual assigned to high-resolution image element *k* is $\mathrm{r}(x_{ij}, y_{ij})$, given by

$$\mathrm{W}\ (x_{ij}, y_{ij}) = \mathrm{CW}\,(x_{ij}, y_{ij})\,/\sum_{j=1}^{m} \mathrm{CW}\,(x_{ij}, y_{ij}) \quad (10)$$

$$\mathrm{r}\ (x_{ij}, y_{ij}) = \ m \times \mathrm{R}\,(x_i, y_i) \times \mathrm{W}\,(x_{ij}, y_{ij})\,. \quad (11)$$

Summing the distribution residuals and the changes over time, the total change in a high-resolution image element between times t_1 and t_2 is given by $\Delta\mathrm{F}(x_{ij}, y_{ij})$

$$\Delta\mathrm{F}\ (x_{ij}, y_{ij}) = \mathrm{r}\,(x_{ij}, y_{ij}) + \Delta F\,(c)\,. \quad (12)$$

Finally, an image element selection strategy similar to the STARFM method is used, and $\Delta F(c)$ is weighted by the spatial distance $\mathrm{D_k} = \ 1 + \sqrt{(\mathrm{x_k} - \mathrm{x_{ij}})^2 + (\mathrm{y_k} - \mathrm{y_{ij}})^2}/(\mathrm{w}/2)$, where *w* is the window size; in this study, *w* = 25 and k symbolizes the *k*th high-resolution image pixels. The final result $\widehat{F_2}(x_{ij}, y_{ij})$ is

$$\widehat{F_2}\ (x_{ij}, y_{ij}) = F_1\ (x_{ij}, y_{ij}) + \sum_{k-1}^{n} w_k \times \Delta F\,(x_{ij}, y_{ij}) \quad (13)$$

$$w_k = (1/D_k)\,/\sum_{k-1}^{n} (1/D_k)\,. \quad (14)$$

C. RCAN-FSDAF Implementation

The standard FSDAF method uses resampling methods based on difference or reconstruction for the upsampling method for the raw data. This upsampling method is likely to lead to problems with neighboring image elements having the same image elements as well as spatial inconsistencies between high-resolution images and low-resolution images, and the accuracy of spatiotemporal data fusion depends on the accuracy of the input data [17]. We, therefore, combined the standard FSDAF model with the RCAN downscaling method. The trained RCAN can downscale the low-resolution images to obtain input data with increased spatial information, and FSDAF fuses spatiotemporal data. High spatiotemporal resolution data with higher accuracy and more spatial details are, thereby, produced. Fig. 2 is a flowchart of RCAN-FSDAF.

Iterative training of RCAN is based on Landsat data using RCAN to learn the mapping relationship between low-resolution images and high-resolution images. MODIS data at times t_1 and t_2 are initially downscaled using well-trained RCAN to complement the feature information at coarse resolution. This produces MODIS data with increased spatial heterogeneity details. The MODIS data at times t_1 and t_2 after downscaling are fused with the Landsat data at time t_1 using FSDAF to produce high spatial and temporal resolution fused data at time t_2. RCAN-FSDAF consists primarily of three steps.

1) The 100 Landsat4-8 TM/OLI image blocks of any time and regions were selected, and they were all downsampled to 240 m (×8) and 120 m (×4). The mapping relationships between 240 m images and 120 m images and between 120 m images and the 30 m original images were trained using RCAN. Optimal PSNR was reached after 1200 iterations for each training run.
2) The preprocessed MODIS data at times t_1 and t_2 were upsampled to 240 m and reconstructed twice by the RCAN model trained in the preceding step. The first reconstruction was to convert the 240 m MODIS data into 120 m data, and the second reconstruction was to convert this 120 m MODIS data to 30 m.
3) The Landsat data at time t_1 and the MODIS data at times t_1 and t_2 after reconstruction by RCAN were fused using the FSDAF model to produce the 30 m high-resolution spatiotemporal data at time t_2.

D. Datasets and Preprocessing

The high-resolution images used in this study were all from Landsat 8 OLI datasets Level 1 digital product [blue, red, and near-infrared (NIR) band], provided by the U.S. Geological

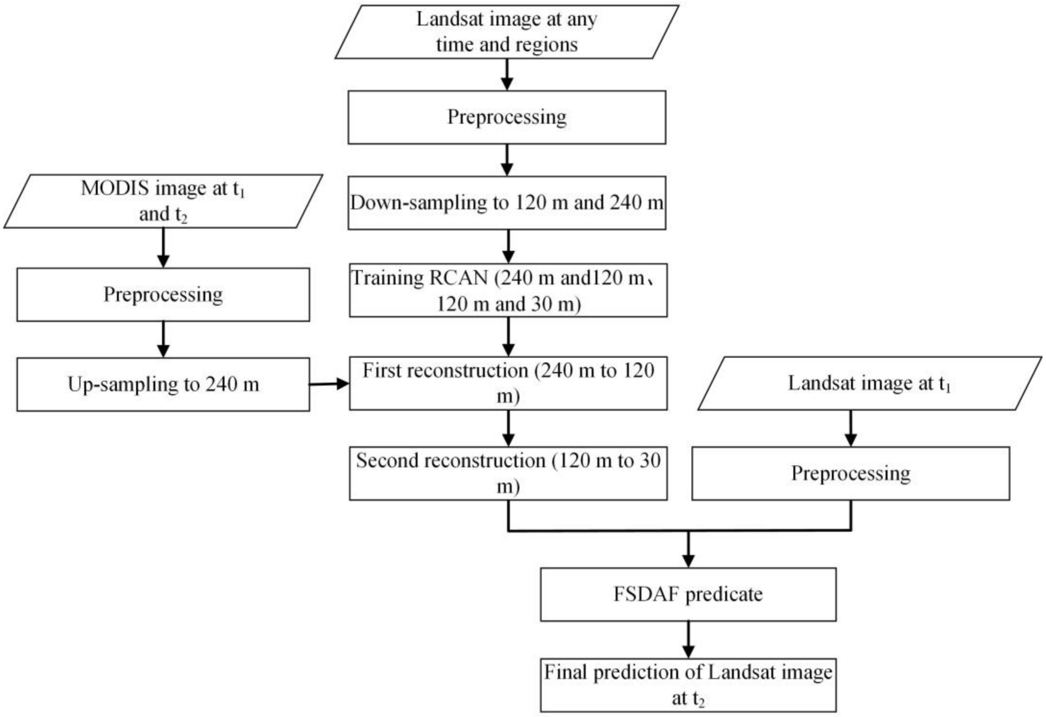
Landsat image at any time and regions
Preprocessing
Down-sampling to 120 m and 240 m
Training RCAN (240 m and120 m、120 m and 30 m)
MODIS image at t_1 and t_2
Preprocessing
Up-sampling to 240 m
First reconstruction (240 m to 120 m)
Second reconstruction (120 m to 30 m)
Landsat image at t_1
Preprocessing
FSDAF predicate
Final prediction of Landsat image at t_2

Fig. 2. Flowchart of RCAN-FSDAF.

Survey[1] open-source website. The Landsat ecosystem disturbance adaptive processing system (LEDAPS) was used for calibration and atmospheric correction of Landsat images. Because LEDAPS uses an atmospheric correction method (the 6S method) similar to the MODIS surface reflectivity product, and the reflectance of the two sensors is consistent and comparable [32].

For the low-resolution images, two MODIS datasets were used in this study. For the first dataset, the MODIS13Q1 dataset, provided by NASA,[2] was reprojected and resampled using the projection tool (MRT) to match the resolution and range of the Landsat image. For the second dataset, MODIS-like images were collected from the original Landsat images, thus eliminating the effects of different sensors. The purpose of using two datasets is to ensure the accuracy realism of RCAN-FSDAF, on the one hand, to eliminate the influence of sensor differences on the accuracy of the RCAN-FSDAF method, and on the other hand, to reflect the real accuracy of the RCAN-FSDAF method on satellite data. Finally, MATLAB software was used to smooth and denoise the three datasets using S-G filtering to meet the demand for high reflectance in elements of the Landsat and MODIS images.

III. Experiments

A. Experimental Implementation

Three different tests were set up in this study, namely the highly heterogeneous regions test, the land-use change regions test, and the NDVI inversion test. Random points were selected to quantitatively evaluate the test results using root-mean-square error (RMSE), correlation coefficient (R), and average deviation (AD) as quantitative evaluation metrics. The test regions were selected as follows.

For the highly heterogeneous regions test, the region (33°06′–33°10′N, 111°25′–111°29′E) is located in the southwestern part of Nanyang City, Henan Province, with complex features, which was selected as the study area. The two Landsat 8 OLI images were collected on 18 June 2020 and 30 September 2020, both in the vegetation growing season. The starting observation dates of the two MODIS13Q1 images were, respectively, 9 June 2020 and 29 September 2020, as shown in Fig. 3, and the Landsat 8 OLI image collected on 30 September 2020 was used as the base images for comparison with the predicted FSDAF, DMNet, GAN-STFM, and RCAN-FSDAF images to assess the performance of RCAN-FSDAF.

The regional test for land-use change in this study consists of two tests, the regional test for the flood hazard and the regional test for phenological change. The flood hazard region is a relatively homogeneous area of 18 × 18 km near Dongting Lake (29°09′–29°18′N, 1132°55′–113°04′E) in the western part of Yueyang City, Hunan Province. The area is in the Yangtze River basin, which experienced the largest flood disaster in the 21st century during the 2016 flood season (May–October), especially after June. The flooding exceeded the predicted warning level in several areas after the second half of June. The two Landsat 8 OLI images used in this test were acquired on 5 June 2016 and 23 July 2016, and two corresponding 240 m MODIS-like images were produced, as shown in Fig. 4. The Landsat 8 OLI image collected on 5 June 2016 was used as the base images

[1] [Online]. Available: http://glovis.usgs.gov/
[2] [Online]. Available: http://www.nasa.gov/

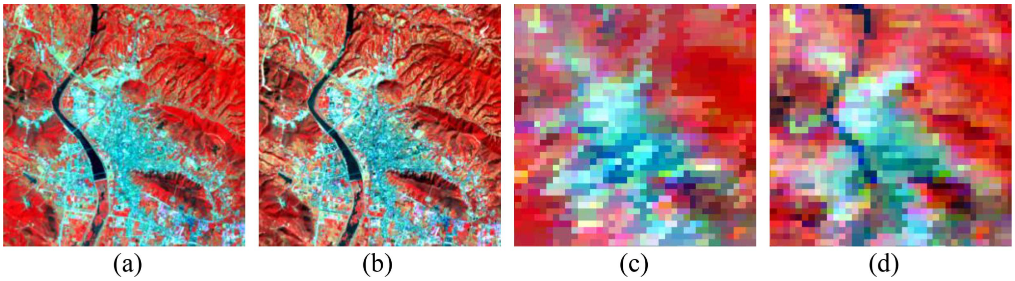

Fig. 3. Heterogeneous regions. (a) and (b) Landsat8 OLI images (300 × 300 pixels) acquired on 18 June 2020 (t_1) and 30 September 2020 (t_2). (c) and (d) 240 m MODIS13Q1 images with 9 June 2020 and 29 September 2020 as the start observation dates, respectively, all images using NIR–red–blue as RGB.

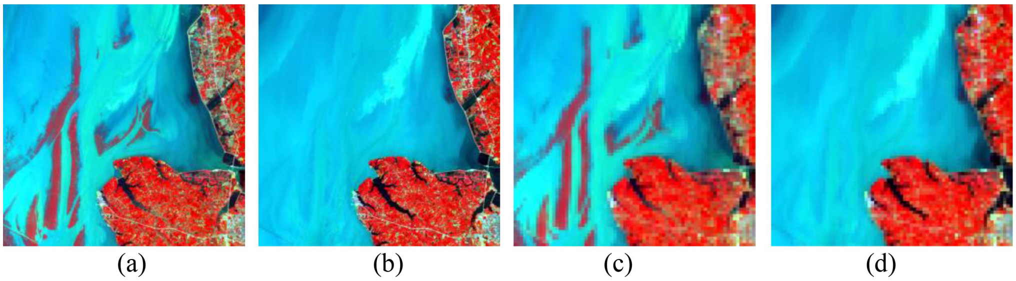

Fig. 4. Flood hazard regions. (a) and (b) Landsat 8 OLI images (600 × 600 pixels) acquired on 5 June 2020 (t_2) and 23 July 2020 (t_1). (c) and (d) 240 m MODIS1-like images, all images using NIR–red–blue as RGB.

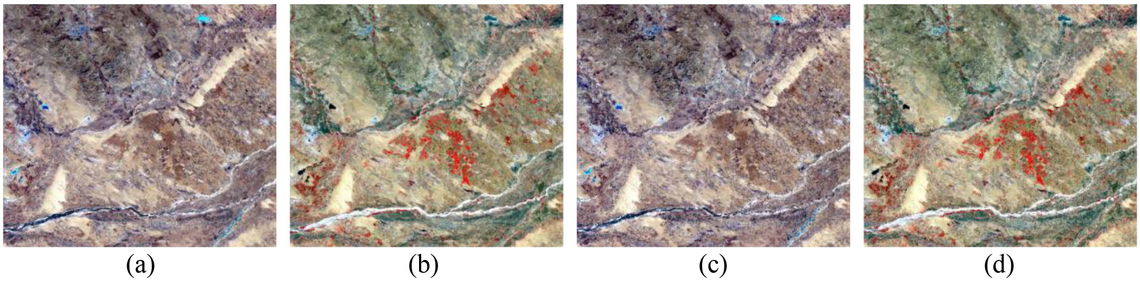

Fig. 5. Phenological change regions. (a) and (b) Landsat 8 OLI images (2480 × 2800 pixels) acquired on 17 March 2015 (t_1) and 4 May 2015 (t_2). (c) and (d) 240 m MODIS-like images.

for comparison. The selection of phenological change regions was based on the AHB dataset released by Prof. Jun Li's team in the spatial and temporal fusion dataset [9], [33]. The region (43°14′–44°02′N, 119°07′–120°07′E) is located in the northeast of China, for which there are a lot of circular pastures and farmlands, the region size is 74.4 km×84 km. The two Landsat 8 OLI images used in this test were acquired on 17 March 2015 and 4 May 2015, and two corresponding 240 m MODIS-like images were produced, as shown in Fig. 5. The Landsat 8 OLI image collected on 4 May 2015 was used as the base images for comparison.

In this study, the upsampling method used by the conventional FSDAF method in all three tests is the bicubic method. Figs. 6 and 7 show the comparison of the results of MODIS-like datasets tested in land-use change areas after downscaling by RCAN and upsampling by bicubic method, and it can be clearly seen that the boundaries of MODIS-like images after downscaling by the RCAN method are clearer and have better spatial consistency with the Landsat images.

B. Heterogeneous Regions Test Results

1) Visual comparison: Fig. 8 shows the (base image) Landsat image [see Fig. 8(a)] and the Landsat 8 OLI images [see Fig. 8(b), (c), (d), and (e)], respectively, produced from the FSDAF, DMNet, GAN-STFM, and RCAN-FSDAF predictions for 30 September 2020. It is clear from the enlarged scenes that the images predicted by RCAN-FSDAF had more spatial detail and that they more clearly identified the boundaries between features than those predicted by FSDAF, DMNet, and GAN-STFM. RCAN-FSDAF predicted parts of the shadows and topography of the mountains, which are more similar to the base satellite images, while in the images predicted by FSDAF, DMNet, and GAN-STFM, many spatial details are lost and appear somewhat

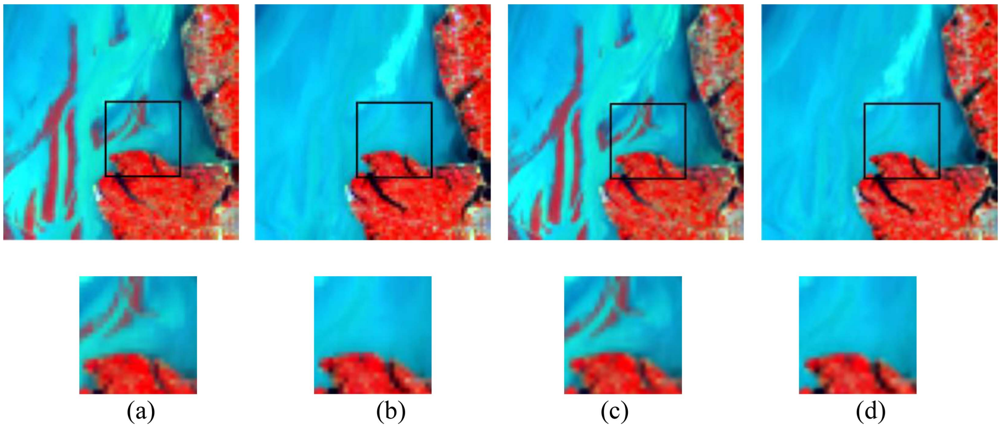

Fig. 6. Flood hazard regions. (a) and (b) 30 m MODIS-like images of 5 June 2020 (t_2) and 23 July 2020 (t_1) using the bicubic method. (c) and (d) 30 m MODIS-like images of 5 June 2020 (t_2) and 23 July 2020 (t_1) using RCAN method.

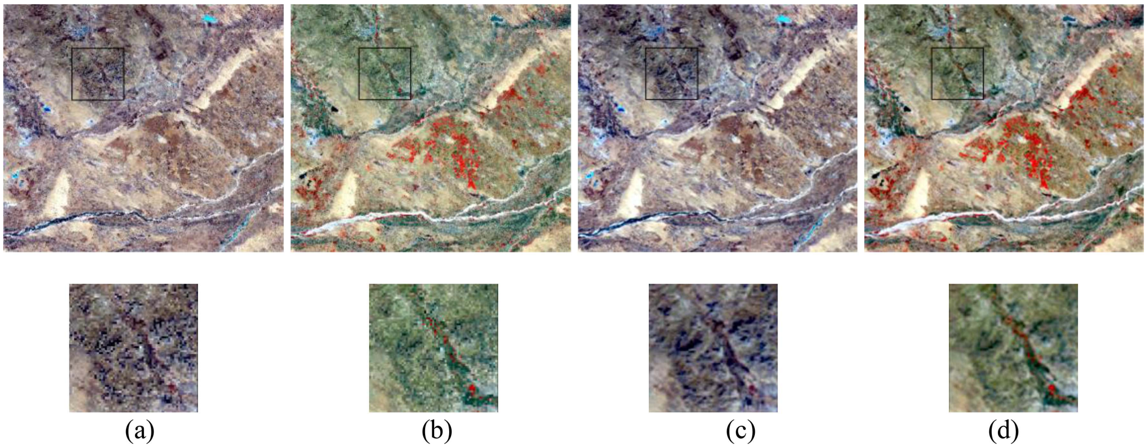

Fig. 7. Phenological change regions. (a) and (b) 30 m MODIS-like images of 17 March 2015 (t_1) and 4 May 2015 (t_2) using bicubic method. (c) and (d) 30 m MODIS-like images of 17 March 2015 (t_1) and 4 May 2015 (t_2) using RCAN method.

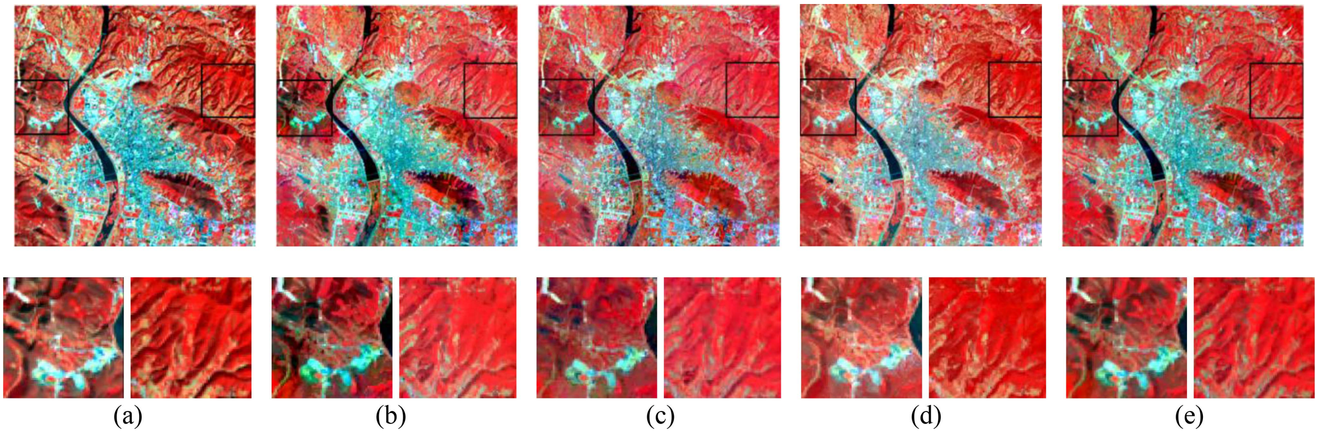

Fig. 8. Heterogeneous regions test. (a) Is the base image acquired on 30 September 2020 (t_2). (b), (c), (d), and (e) are, respectively, the images predicted by FSDAF, DMNet, GAN-STFM, and RCAN-FSDAF.

blurred. In general, RCAN-FSDAF better predicts boundaries between features in complex heterogeneous regions than FSDAF and DMNet.

2) Quantitative comparison: 4000 sample points were randomly selected within the study area to determine the reliability and accuracy of the RCAN-FSDAF method in predicting heterogeneous landforms. The predictions were quantified using the RMSE, correlation coefficient (R), and mean deviation (AD). The scatterplots (see Fig. 9) show the correlation between predicted and base reflectance. It can be seen that the correlation

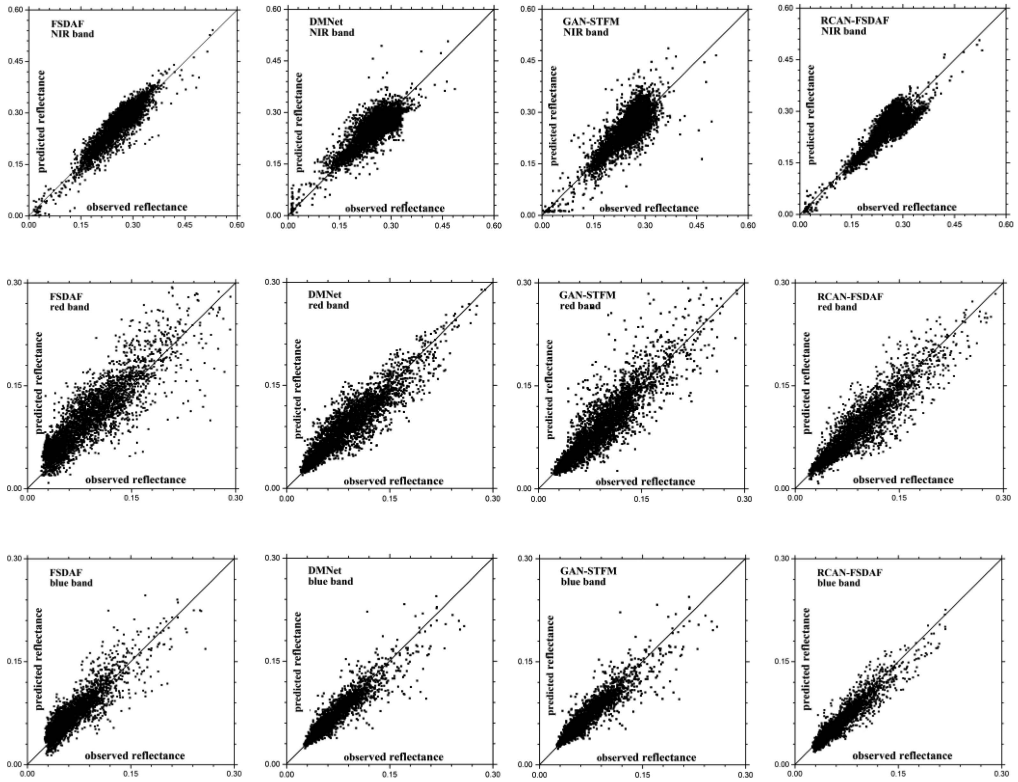

Fig. 9. Scatter plots of the base reflectance and the predicted ones product by the FSDAF, DMNet, GAN-STFM, and RCAN-FSDAF for NIR, red, and blue bands. (Darker colors indicate a higher density of points, and the diagonal lines are 1:1 lines).

TABLE I
INDICATORS OF THE DATA FUSION METHODS USED TO QUANTIFY PREDICTIONS OF HIGHLY HETEROGENEOUS REGIONS

OLI	RMSE					Related coefficient (*R*)					Average difference (*AD*)				
Band	t_1	Prediction				t_1	Prediction				t_1	Prediction			
		FSDAF	DMNet	GAN-STFM	RCAN-FSDAF		FSDAF	DMNet	GAN-STFM	RCAN-FSDAF		FSDAF	DMNet	GAN-STFM	RCAN-FSDAF
blue	0.0981	0.0648	0.0502	0.0454	0.0463	0.4745	0.8851	0.9102	0.9089	0.9015	0.2042	0.0055	0.0051	0.0048	0.0042
red	0.1045	0.0689	0.0549	0.0486	0.0415	0.3272	0.8781	0.8854	0.8875	0.8862	0.1432	0.0046	0.0032	0.0036	0.0035
NIR	0.0754	0.0561	0.0251	0.0210	0.0206	0.5048	0.8968	0.9103	0.9201	0.9241	0.0641	0.0048	0.0024	0.0027	0.0021

Note: The values in the t_1 column of the table represent the quantitative evaluation metrics of the predicted band reflectance compared with the known moment t_1.

between predicted reflectance and base reflectance in NIR, the red, and blue bands of RCAN-FSDAF were closer to the 1:1 line, with the NIR band being the most accurate. This result is explained by the images used all being observed during the vegetation growth cycle. The large variation in reflectance indicates that RCAN-FSDAF better identifies changes in the features. Table I presents the results of the quantitative comparison, and the overall reflectance predicted by FSDAF, DMNet, GAN-STFM, and RCAN-FSDAF for all three bands is less than the base reflectance (AD $<$ 0). For the NIR, red, and blue bands, RCAN-FSDAF has lower values of RMSE and AD and a greater value of *R* than FSDAF and GAN-STFM, except for DMNet. For DMNet, the RCAN-FSDAF prediction for the red band is not superior, with a lower *R*-value than the DMNet method, and all other indices are optimal. The difference in prediction accuracy of NIR band reflectance is greater (RMSE 0.0206 versus 0.0210, 0.0251, and 0.0561, AD 0.0021 versus 0.0024, 0.0027, and 0.0048, and *R* 0.9241 versus 0.9103, 0.9201, and 0.8968), so NIR band reflectance predicted by RCAN-FSDAF is closer to the base reflectance, as indicated the scatterplot for each band. In summary, RCAN-FSDAF predictions are more accurate than those of FSDAF, DMNet, and GAN-STFM, with excellent performance in both visual and quantitative comparisons.

C. Land-Use Change Test Results

1) Visual comparison: For the flood hazard regions, Fig. 10(a) shows the base image on 5 June 2016, and Fig. 10(b), (c), (d), and (e) show, respectively, the image obtained using FSDAF, DMNet, GAN-STFM, and RCAN-FSDAF predictions. Closer inspection of the zoomed images reveals that RCAN-FSDAF successfully delineates the boundaries between the unflooded and flooded regions. Furthermore, the shapes of the unflooded regions predicted by RCAN-FSDAF are more similar

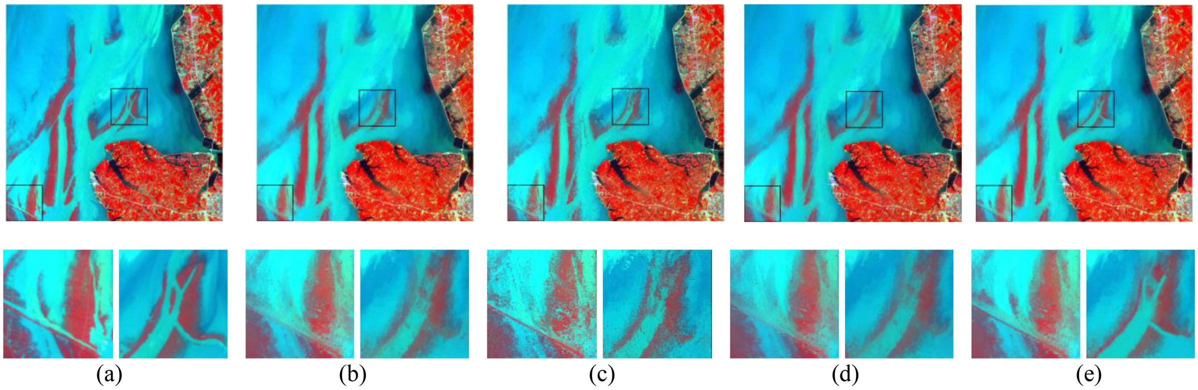

Fig. 10. Flood hazard regions test. (a) Is the base image acquired on 5 June 2016 (t_2). (b), (c), (d), and (e) are, respectively, the images predicted by FSDAF, DMNet, GAN-STFM, and RCAN-FSDAF.

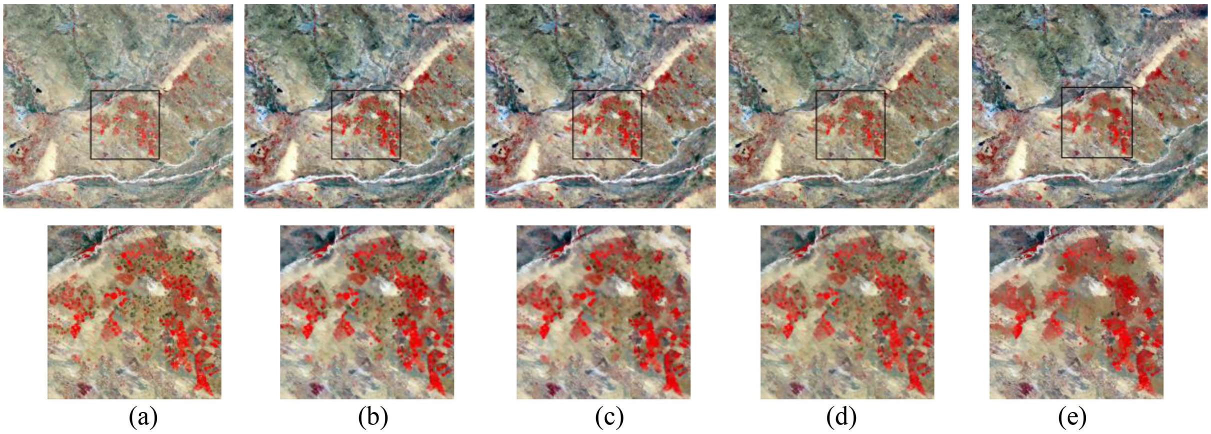

Fig. 11 Phenological change regions. (a) Is the base image acquired on 4 May 2015 (t_2). (b), (c), (d), and (e) are, respectively, the images predicted by FSDAF, DMNet, GAN-STFM, and RCAN-FSDAF.

TABLE II
Indicators of Three Data Fusion Methods Used to Predict the Flood Hazard Regions

OLI	RMSE					Related coefficient (R)					Average difference (AD)				
Band	t_1	Prediction				t_1	Prediction				t_1	Prediction			
		FSDAF	DMNet	GAN-STFM	RCAN-FSDAF		FSDAF	DMNet	GAN-STFM	RCAN-FSDAF		FSDAF	DMNet	GAN-STFM	RCAN-FSDAF
blue	0.0868	0.0485	0.0435	0.0430	0.0421	0.4781	0.8972	0.9047	0.8901	0.9115	0.0948	-0.0024	-0.0016	-0.0014	-0.0018
red	0.0664	0.0421	0.0384	0.0351	0.0321	0.7595	0.8825	0.9104	0.9124	0.9243	0.0655	0.0038	0.0025	0.0018	0.0010
NIR	0.1042	0.0462	0.0346	0.0305	0.0241	0.6757	0.8948	0.9046	0.8942	0.8991	0.0881	0.0027	0.0021	0.0024	0.0019

Note: The values in the t_1 column of the table represent the quantitative evaluation metrics of the predicted band reflectance compared with the known moment t_1.

to the base image than those predicted by FSDAF, DMNet, and GAN-STFM. Although FSDAF, DMNet, and GAN-STFM manage to capture the general shape of the unflooded regions, they fail to clearly define the boundaries or to identify the smaller unflooded regions. For the phenological change regions, Fig. 11 shows the (base image) Landsat image [see Fig. 11(a)] and the Landsat 8 OLI images [see Fig. 11(b), (c), (d), and (e)], respectively, produced from the FSDAF, DMNet, GAN-STFM, and RCAN-FSDAF predictions for 4 May 2015. Many circular pastures exist in the regions and crops change with the seasons, i.e., phenological changes. RCAN-FSDAF, FSDAF, DMNet, and GAN-STFM demonstrate the ability to predict phenological changes in circular pastures. However, RCAN-FSDAF stands out by offering a more distinct prediction of the boundaries delineating these circular pastures.

2) Quantitative comparison (8000 points were randomly sampled): The scatterplots [see Figs. 12 and 13] show the correlation between predicted and base reflectance. For the flooded hazard regions, all three bands of RCAN-FSDAF are closer to the 1:1 line, with the blue band being the most pronounced. For the phenological change regions, there is no major difference between the four methods. However, Tables II and III display the indicator values; evidently, for the two test regions, RCAN-FSDAF produces more accurate predictions compared

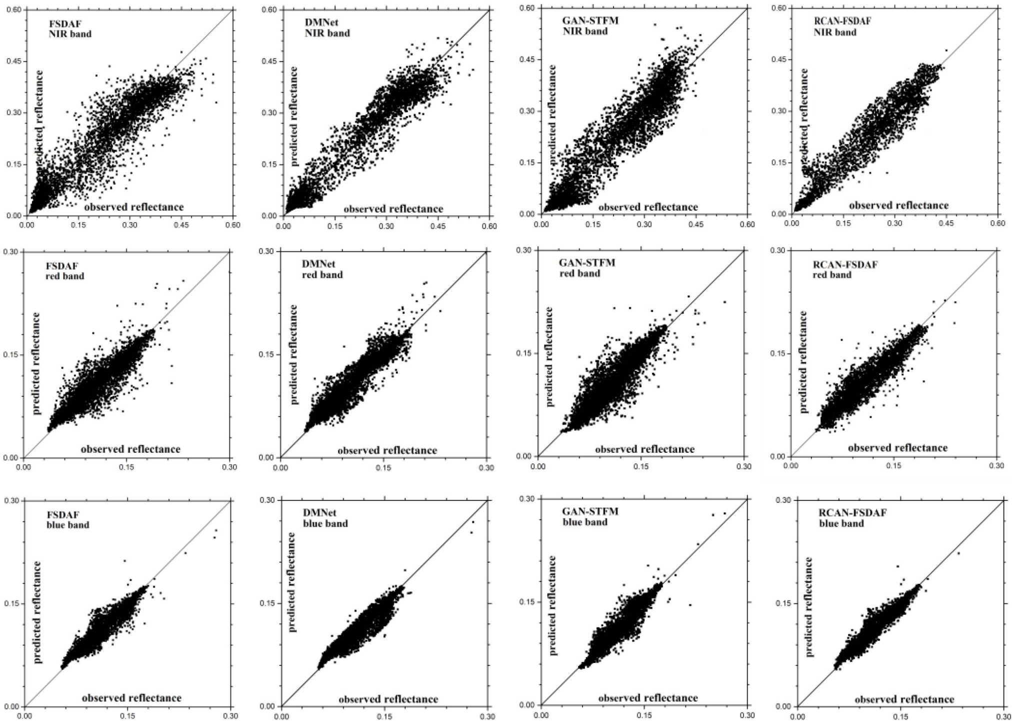

Fig. 12. Scatter plots of the base reflectance and the predicted ones product by the FSDAF, DMNet, GAN-STFM, and RCAN-FSDAF for NIR, red, and blue bands.

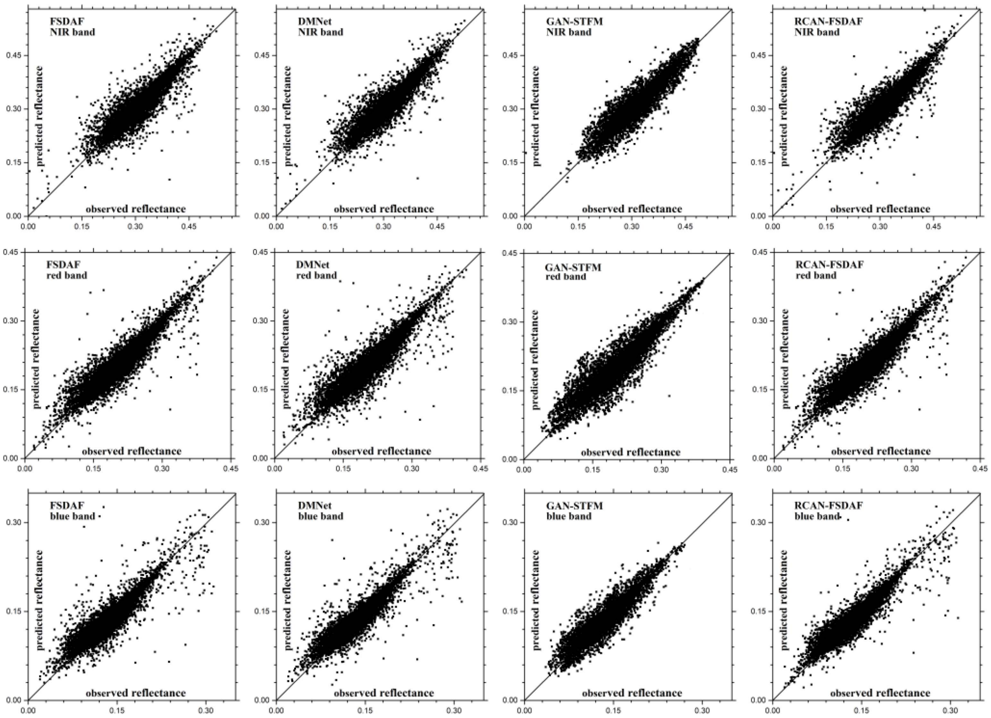

Fig. 13. Scatter plots of the base reflectance and the predicted ones product by the FSDAF, DMNet, GAN-STFM, and RCAN-FSDAF for NIR, red, and blue bands.

with FSDAF, DMNet, and GAN-STFM. For the three bands, RCAN-FSDAF showed significantly lower values of RMSE and AD and a notably higher value of *R* than FSDAF, DMNet, and GAN-STFM. Separately, for the flood hazard regions, the most prominent differences were found in the red band (*R* 0.9243 versus 0.9124, 0.9104, and 0.8825, and AD 0.0010 versus 0.0018, 0.0025, and 0.0038). Regarding the overall variance, both models predict the reflectance for the blue band to be greater than the base reflectance (AD $<$ 0). Additionally, the reflectance for the NIR and red bands predicted by both models

TABLE III
INDICATORS OF THREE DATA FUSION METHODS USED TO PREDICT THE PHENOLOGICAL CHANGE REGIONS

OLI	RMSE					Related coefficient (*R*)					Average difference (*AD*)				
		Prediction					Prediction					Prediction			
Band	t_1	FSDAF	DMNet	GAN-STFM	RCAN-FSDAF	t_1	FSDAF	DMNet	GAN-STFM	RCAN-FSDAF	t_1	FSDAF	DMNet	GAN-STFM	RCAN-FSDAF
blue	0.0651	0.0411	0.0394	0.0383	0.0387	0.641	0.8744	0.8801	0.8854	0.8941	0.0087	0.0037	0.0031	0.0026	0.0025
red	0.0624	0.0324	0.0304	0.0256	0.0241	0.754	0.9129	0.9147	0.9102	0.9248	0.0091	0.0051	0.0050	0.0053	0.0048
NIR	0.0548	0.0241	0.0347	0.0210	0.0182	0.628	0.9247	0.9251	0.9256	0.9289	0.0076	0.0039	0.0037	0.0038	0.0034

Note: The values in the t_1 column of the table represent the quantitative evaluation metrics of the predicted band reflectance compared with the known moment t_1.

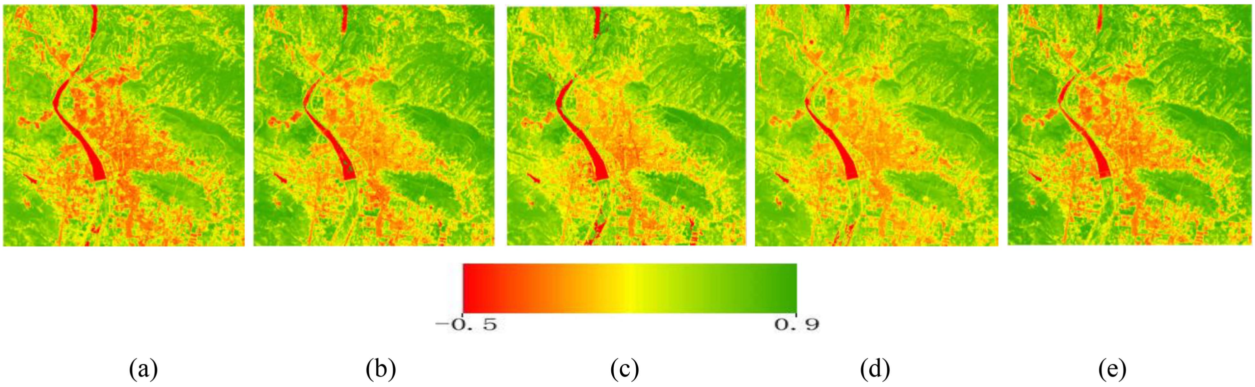

Fig. 14. (a) Shows the NDVI inversion results of the base image acquired in the first test. (b), (c), (d), and (e) show the respective NDVI inversion results of FSDAF, DMNet, GAN-STFM, and RCAN-FSDAF predicted images.

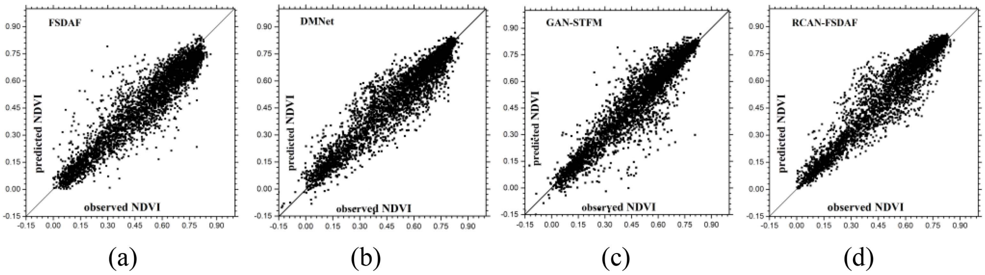

Fig. 15. Scatter plots of the base NDVI and the predicted ones product by the FSDAF, DMNet, GAN-STFM, and RCAN-FSDAF.

is less than the base reflectance (AD > 0). For the phenological changes' regions, the most notable differences were found in the red band (RMSE 0.0241 versus 0.0256, 0.0304, and 0.0324, *R* 0.9248 versus 0.9147, 0.9129, and 0.9102, and AD 0.0048 versus 0.0050, 0.0051, and 0.0053). Overall, RCAN-FSDAF demonstrates superior accuracy in predicting regions with changes in land-cover types.

D. NDVI Inversion Test Results

Until now, spatiotemporal data fusion has been widely used and has produced good results in the inversion of various quantitative remote sensing parameters. In order to establish the applicability of RCAN-FSDAF, NDVI inversion of the fused data produced in the first test was performed, and five inversion results were obtained, as shown in Fig. 14. A Total of 4000 random points were randomly selected to determine the accuracy of the inversion results using RMSE, *R,* and AD. From the scatterplot in Fig. 15, we find that the correlation between NDVI obtained from the inversion of the predicted data using FSDAF and NDVI obtained from the inversion of the base data was biased below the 1:1 line; a slight improvement in NDVI from the inversion of DMNet and GAN-STFM forecasts, while the correlation between NDVI obtained from the inversion of the predicted data using RCAN-FSDAF and the NDVI obtained from the inversion of the base data was more uniformly distributed at the center of the 1:1 line. Table IV shows the values of the indicators of the three NDVI inversions, and it can be seen that the results produced by RCAN-FSDAF were closer to the base inversion results than the results of FSDAF, DMNet, and GAN-STFM, with smaller values of RMSE and AD and a greater value of *R*. These values indicate that the results obtained by the inversion of the prediction data using RCAN-FSDAF were more accurate than FSDAF. This result will be

TABLE IV
INDICATORS OF THE FUSION RESULTS OF THE THREE METHODS USED FOR NDVI INVERSION

Method	RMSE	Related coefficient (R)	Average difference (AD)
FSDAF	0.1847	0.8549	0.0059
DMNet	0.1126	0.8741	0.0053
GAN-STFM	0.1085	0.8645	0.0034
RCAN-FSDAF	0.1024	0.8956	0.0024

of value in subsequent quantitative remote sensing analysis and application.

IV. Discussion

In this study, a new spatiotemporal data fusion method RCAN-FSDAF is proposed by combining the downscaling method based on deep learning with the traditional spatiotemporal data fusion method. Based on Landsat and MODIS data, three tests of the RCAN-FSDAF method were conducted: the highly heterogeneous area test, the land-use change area test, and the NDVI inversion result test, and the RCAN-FSDAF method showed good fusion effect and accuracy in all three tests in this study. Among them, the highly heterogeneous region test and the land-use change regions test are representative tests to test the spatiotemporal data fusion algorithm. Similarly, the authors in [6] and [15] set up highly heterogeneous regions and land-use change regions to test FSDAF and PSTAF-GAN. However, there are still some uncertainties, which can be described as follows.

The proposed method provides a new framework to solve the current problem of limited fusion accuracy due to the large uncertainty error of input data by traditional spatiotemporal data fusion methods. The resampling methods based on difference or reconstruction are poor in decomposing the mixed image elements of the input data and have poor spatial consistency, which leads to the traditional spatiotemporal data fusion methods to distinguish the boundaries between different features difficulty. Deep-learning methods have advantages in extracting deep features and texture features, etc., and are capable of extracting more spatial information. However, compared with the traditional spatiotemporal data fusion methods, the spatiotemporal data fusion algorithm based solely on deep learning brings more noise to the image fusion process and ignores the advantages of the traditional spatiotemporal data fusion algorithm in terms of image element unmixing and land-use type change prediction, etc. The basic principle of the RCAN-FSDAF method is to combine the deep-learning-based super-resolution technology with the traditional spatiotemporal data fusion method. By complementing the feature information at coarse resolution with RCAN, the input data with more spatial heterogeneity details are obtained, and thus, the accuracy of FSDAF is improved. However, the improvement of the prediction performance of the RCAN-FSDAF method depends on the downscaling effect of the RCAN to a certain extent. Methods with better downscaling effects will obtain better fusion results. Similarly, Zhai et al. [21] reduced the scale of input MODIS data by introducing a linear spectral mixing model, which replaced the resampled low-resolution data in FSDAF and generated high-precision predicted LAI data with high spatial and temporal resolution. Xie et al. [22] proposed a CDSTARFM algorithm based on the combination of the image element decomposition downscaling method and STARFM model; first, the MODIS data were decomposed and downscaled using the image element decomposition downscaling method, and then the downscaled MODIS data were used to replace the directly resampled MODIS data in the STARFM model for data fusion. The results show that the CDSTARFM algorithm has better accuracy; therefore, the RCAN-FSDAF method has some alternatives. In future work, we will choose to use super-resolution techniques with higher efficiency and accuracy, such as (pan sharpening in closed-loop regularization and modality-aware feature integration for pan sharpening) proposed by Zhou et al. [26], [27] and CUCaNet proposed by Zheng et al. [28]. However, this study takes the combination of RCAN and FSDAF as an example to carry out a detailed study, which verifies the reliability of the framework of combining the deep-learning-based downscaling method with the traditional spatiotemporal data fusion method, which can also be extended to the performance improvement of other traditional spatiotemporal data fusion methods, so the RCAN-FSDAF method has also certain scalability.

Different MODIS data are used for the first test and the second test in this study. The use of real MODIS data on the one hand is to ensure the practicality of the RCAN-FSDAF method proposed in this study for the fusion effect of MODIS and Landsat data. On the other hand, the use of similar MODIS data is to eliminate the influence of errors caused by sensor differences on the test accuracy and to ensure the authenticity of the fusion accuracy because sensor differences are a bigger problem for spatiotemporal data fusion algorithms [34], [35]. The MODIS data used in this study is the 16-day synthetic global vegetation index product (MODIS13Q1), and the daily surface reflectance product is not used. Therefore, the temporal phase difference can affect the fusion accuracy of the RCAN-FSDAF method. In addition to this, the data used in this study have some alignment errors. Therefore, they need to be further investigated to achieve better fusion results.

Overall, the RCAN-FSDAF method proposed in this study can provide a new framework and reference for the current traditional spatiotemporal data fusion methods for accuracy improvement. The validation results and analysis of the three tests conducted on the RCAN-FSDAF method in this study can provide data support for the framework of this study. However, the errors caused by temporal phase discrepancies and alignment errors need further specialized research.

V. Conclusion

Aiming at the problem that the traditional spatiotemporal fusion method reduces the fusion accuracy due to the large uncertainty error of input data, this study proposes a framework of combining deep-learning-based super-resolution technology with traditional spatiotemporal fusion method and explores in detail its applicability in areas of high land-use heterogeneity and areas of land-cover change, and its effectiveness in the subsequent inversion of quantitative remotely sensed data by taking RCAN and FSDAF as examples. The results show that our proposed new framework has a large improvement in fusion accuracy. Our main conclusions are summarized as follows.

1) The predicted reflectance of various bands by RCAN-FSDAF was closer to the base reflectance than FSDAF, GAN-STFM, and DMNet, as shown by the higher correlation and smaller error with the base reflectance. We found that RCAN-FSDAF more accurately predicted complex features or complex mountainous environments, and more accurately identified boundaries between different feature changes in land cover, because it better decomposes elements consisting of mixed image features.
2) Indicators of the NDVI inversion of the prediction results of the two methods showed that the high spatiotemporal resolution NDVI data produced by the inversion of the RCAN-FSDAF prediction results were more accurate than other data. This result indicates the possibility of obtaining more accurate results by fusing long time-series data and, subsequently, inverting remote sensing results. This method can be extended to other remote sensing inversions of biophysical parameters.
3) This study is the first published record of RCAN being combined with FSDAF, the combination increased prediction accuracy. Training RCAN is not complicated and is relatively efficient. If this virtue can be extended to other spatiotemporal data fusion methods, it will improve the prediction accuracy of the other spatiotemporal data fusion methods.

References

[1] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[2] D. Bian, B. Pu, Z. Ni, and K. Liu, "Spatial and temporal change of grassland coverage based on time-series of MODIS-NDVI data in ALI region," *Chin. J. Grassland*, vol. 36, no. 3, pp. 73–78, May 2014.

[3] L. Zhou, H. Mu, H. Ma, and L. Chen, "Remote sensing estimation on yield of winter wheat in North China based on convolutional neural network," *Trans. Chin. Soc. Agricultural Eng.*, vol. 35, no. 15, pp. 119–128, Aug. 2019.

[4] S. Liu et al., "Vegetation phenology in the Tibetan plateau using MODIS data from 2000 to 2010," *Remote Sens. Inf.*, vol. 29, no. 6, pp. 25–30, Dec. 2014.

[5] 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.

[6] Q. Liu, X. Meng, F. Shao, and S. Li, "PSTAF-GAN: Progressive spatio-temporal attention fusion method based on generative adversarial network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2022, Art. no. 5408513.

[7] 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.

[8] Y. Zheng, H. Song, L. Sun, Z. Wu, and B. Jeon, "Spatio-temporal fusion of satellite images via very deep convolutional networks," *Remote Sens.*, vol. 11, no. 2, Dec. 2019, Art. no. 2701.

[9] 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.

[10] 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.

[11] X. Zhu, J. Chen, F. Gao, and X. Chen, "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.

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

[13] C. Gevaert and F. 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.

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

[15] 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.

[16] 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.

[17] 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, Jul. 2018, Art. no. 1066.

[18] 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.

[19] W. Li, X. Zhang, Y. Peng, and M. Dong, "DMNet: A network architecture using dilated convolution and multiscale mechanisms for spatiotemporal fusion of remote sensing images," *IEEE Sensors J.*, vol. 20, no. 20, pp. 12190–12202, Oct. 2020.

[20] 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. 2021, Art. no. 5601413.

[21] H. Zhai, F. Huang, and H. Qi, "Generating high resolution LAI based on a modified FSDAF model," *Remote Sens.*, vol. 12, no. 1, Jan. 2020, Art. no. 150.

[22] D. Xie, J. Zhang, P. Sun, Y. Pan, Y. Yun, and Z. Yuan, "Remote sensing data fusion by combining STARFM and downscaling mixed pixel algorithm," *J. Remote Sens.*, vol. 20, no. 1, pp. 62–72, Jan. 2016.

[23] J. Li, Y. Li, and J. Li, "Removing influence of MODIS strip noise in spatiotemporal fusion of remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Jul. 2023, Art. no. 5001805.

[24] C. Dong, C. C., K. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," 2014, *arXiv:1501.00092*.

[25] Y. Zhang, K. Li, Kai Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention network," in *Proc. Comput. Vis.-Eur. Conf. Comput. Vis.*, 2018, pp. 294–310.

[26] M. Zhou, J. Huang, D. Hong, F. Zhao, C. Li, and J. Chanussot, "Rethinking pan-sharpening in closed-loop regularization," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2023.3279931.

[27] M. Zhou, J. Huang, F. Zhao, and D. Hong, "Modality-aware feature integration for pan-sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art. no. 5400312.

[28] K. Zheng et al., "Coupled convolutional neural network with adaptive response function learning for unsupervised hyperspectral super resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2487–2502, Mar. 2021.

[29] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2021, pp. 1833–1844.

[30] Y. Wang, Y. Li, G. Wang, and G. X., "Multi-scale attention network for single image super-resolution," 2022, *arXiv:2209.14145v1*.

[31] Y. Qi et al., "Pulmonary nodule image super-resolution using multiscale deep residual channel attention network with joint optimization," *J. Supercomput.*, vol. 76, no. 2, pp. 1005–1019, Feb. 2020.

[32] J. Masek et al., "North American forest disturbance mapped from a decadal Landsat record," *Remote Sens. Environ.*, vol. 112, no. 6, pp. 2914–2926, Jun. 2008.

[33] J. Li, Y. Li, R. Cai, L. He, J. Chen, and A. Plaza, "Enhanced spatiotemporal fusion via MODIS-like images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2021, Art. no. 5610517.

[34] J. Zhou et al., "Sensitivity of six typical spatiotemporal fusion methods to different influential factors: A comparative study for a normalized difference vegetation index time series reconstruction," *Remote Sens. Environ.*, vol. 252, Jan. 2021, Art. no. 112130.

[35] 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.

Dunyue Cui received the B.S. degree in surveying and mapping engineering from the Hebei Institute of Water Conservancy and Electric Power, Cangzhou, China, in 2021. He is currently working toward the M.S. degree in surveying and mapping engineering with Henan Polytechnic University, Jiaozuo, China.

His research interests include remote sensing image processing, spatiotemporal fusion, and quantitative remote sensing of vegetation.

Shidong Wang received the B.S. degree in land planning and utilization from Henan Polytechnic University, Jiaozuo, China, in 2002, the M.S. degree in cartography and geographic information engineering from Henan Polytechnic University, Jiaozuo, China, in 2005, and the Ph.D. degree in cartography and geographic information systems from Beijing Normal University, Beijing, China, in 2013.

His research interests include (geo) data analysis, image processing, soil moisture estimation, and resource and environmental monitoring.

Cunwei Zhao received the bachelor’s degree in land resource management from Henan Polytechnic University, Jiaozuo, China, in 2006.

His research interests include land resource management, spatial land remediation, and resource and environmental monitoring.

Hebing Zhang received the B.S. degree in surveying engineering from the Jiaozuo Institute of Technology, Jiaozuo, China, in 1999, and the M.S. degree in geodesy and surveying engineering and the Ph.D. degree in surveying engineering from Henan Polytechnic University, Jiaozuo, China, in 2002 and 2016, respectively.

His research interests include land remote sensing monitoring and GIS integration, land spatial planning and informatization, and land restoration and ecological restoration.