Deep-Learning-Based Spatio-Temporal-Spectral Integrated Fusion of Heterogeneous Remote Sensing Images

Menghui Jiang, *Graduate Student Member, IEEE*, Huanfeng Shen, *Senior Member, IEEE*, and Jie Li, *Member, IEEE*

***Abstract*— It is a challenging task to integrate the spatial, temporal, and spectral information of multisource remote sensing images, especially in the case of heterogeneous images. To this end, for the first time, this article proposes a heterogeneous integrated framework based on a novel deep residual cycle generative adversarial network (GAN). The proposed network consists of a forward fusion part and a backward degeneration feedback part. The forward part generates the desired fusion result from the various observations; the backward degeneration feedback part considers the imaging degradation process and regenerates the observations inversely from the fusion result. The heterogeneous integrated fusion framework supported by the proposed network can simultaneously merge the complementary spatial, temporal, and spectral information of multisource heterogeneous observations to achieve heterogeneous spatiospectral fusion, spatiotemporal fusion, and heterogeneous spatiotemporal–spectral fusion. Furthermore, the proposed heterogeneous integrated fusion framework can be leveraged to relieve the two bottlenecks of land-cover change and thick cloud cover. Thus, the inapparent and unobserved variation trends of surface features, which are caused by the low-resolution imaging and cloud contamination, can be detected and reconstructed well. Images from many different remote sensing satellites, i.e., Moderate Resolution Imaging Spectroradiometer (MODIS), Landsat 8, Sentinel-1, and Sentinel-2, were utilized in the experiments conducted in this study, and both the qualitative and quantitative evaluations confirmed the effectiveness of the proposed image fusion method.**

***Index Terms*— Deep residual cycle generative adversarial network (GAN), heterogeneous integrated framework, land-cover change, thick cloud cover.**

I. Introduction

DUE to the hardware limitations, remote sensing system imaging involves a tradeoff between the temporal, spatial, and spectral resolutions [1]. Remote sensing image fusion is an effective way to fuse the complementary information between multisource observations and has been widely considered and developed [2]–[4]. To date, a variety of remote sensing image fusion methods have been proposed. According to the different aims, these methods can be divided into different categories, i.e., spatiospectral fusion, spatiotemporal fusion, and spatiotemporal–spectral fusion [5].

Spatiospectral fusion [6] is aimed at obtaining images of both high spatial and spectral resolutions by fusing the complementary rich spatial and spectral features between two images. Spatiospectral fusion includes panchromatic (PAN)/multispectral (MS) image fusion, PAN/hyperspectral (HS) image fusion, and MS/HS image fusion. The existing spatiospectral fusion methods can be broadly classified into four major branches [7]: component substitution (CS)-based methods [8], [9], multiresolution analysis (MRA)-based methods [10], [11], variational model-based methods [12], [13], and deep-learning-based methods [14], [15]. The abovementioned methods are all aimed at fusing homogeneous optical images. Nevertheless, scholars have proposed some heterogeneous -spectral fusion methods, e.g., synthetic aperture radar (SAR)-optical image fusion, which uses the rich spatial features in SAR images to make up for the spatial deficiencies in optical images with rich spectral information. However, most of the existing SAR-optical image fusion methods [16], [17] have been transferred from the spatiospectral fusion of optical images and are unsuitable for heterogeneous information transformation.

Spatiotemporal fusion [18] is aimed at obtaining images with both high spatial and temporal resolutions by fusing high spatial resolution (HR) images with a long revisit period and low spatial resolution (LR) images with a short revisit period. The spatiotemporal fusion methods can be broadly classified into four main categories [19]: weight function-based methods [20], [21], unmixing-based methods [22], [23], Bayesian-based methods [24], [25], and learning-based methods [26], [27]. Most of these methods can capture phenological changes, but they have a bottleneck in reflecting land-cover changes, especially when the changed land covers are imperceptible in the LR image at the target time. This is a common problem in the case of spatiotemporal fusion under large spatial resolution gaps or severe weather conditions (such as thick cloud cover).

The aforementioned spatiospectral fusion and spatiotemporal fusion methods are dedicated to fusing information from only two of the spatial, temporal, and spectral domains. On this

Manuscript received 21 February 2022; revised 15 May 2022; accepted 9 June 2022. Date of publication 14 July 2022; date of current version 20 July 2022. This work was supported by the National Natural Science Foundation of China under Grant 42130108 and Grant 62071341. *(Corresponding author: Huanfeng Shen.)*

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

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

Jie Li is with the School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China (e-mail: aaronleecool@whu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2022.3188998

basis, Shen [28] first proposed the integrated fusion to integrate the complementary spatial, temporal, and spectral features of multisource remote sensing images to obtain images with the highest spatial–temporal–spectral resolutions. Huang *et al.* [29] then extended the model proposed in [28] by exploring the relationship between the spatiospectral and spatiotemporal fusion methods. Shen *et al.* [5] subsequently proposed an integrated framework that thoroughly analyzes the spatial, spectral, and temporal relationships between the desired image and the multisource remote sensing observations and constructed an integrated relationship model based on maximum *a posteriori* (MAP) theory. Zhao *et al.* [30] exploited the high self-similarity in the spatial domain, the high spectral correlation in the spectral domain, and the temporal changes to develop an integrated sparsity model. However, due to the complex and nonlinear relationships between multisource datasets, the current studies of integrated fusion methods have been limited to homogeneous optical images, and they have not considered the heterogeneous information of different imaging mechanisms, such as SAR and optical images.

In this article, to address these issues, we propose a heterogeneous integrated fusion framework based on a deep residual cycle generative adversarial network (GAN). The main contributions of this article are as follows.

1) We propose a novel deep residual cycle GAN, which consists of a forward fusion part and a backward degeneration feedback part, where a cycle consistency constraint is formed from the inputs of the forward part to the outputs of the backward part.

2) The integrated fusion of multisource heterogeneous remote sensing images is first implemented, which can successfully achieve heterogeneous spatiospectral fusion, spatiotemporal fusion, and heterogeneous spatiotemporal–spectral fusion.

3) The proposed heterogeneous integrated fusion framework can effectively alleviate the two bottlenecks of land-cover change and thick cloud cover to predict, not only, the HR image at the target time but also the changes.

The rest of this article is organized as follows. Section II describes the proposed deep residual cycle GAN in detail. In Section III, the experiments and analyses for two challenging scenarios are presented. Our conclusion and future prospects are reported in Section IV.

II. Proposed Method

Before describing the proposed method in detail, the important notations are introduced. $X \in \mathbb{R}^{M\times N\times B}$ denotes the desired HR MS image at target time $t2$, where M, N, and B represent the width, height, and band number of the image, respectively. $\tilde{X} \in \mathbb{R}^{m\times n\times B}$ denotes the observed LR MS image at $t2$. $S = M/m = N/n$ is the spatial resolution ratio of the LR MS image to the HR MS image. $\hat{X} \in \mathbb{R}^{M\times N\times B}$ is the result of $\tilde{X}$ bicubic upsampling to the same spatial size as X. $Y \in \mathbb{R}^{M\times N\times \mathrm{b}}$ denotes the observed HR SAR image at $t2$, where $b < B$. $Z \in \mathbb{R}^{M\times N\times B}$ denotes the observed HR MS image at auxiliary time $t1$. Note that $t2$ is subsequent to $t1$. The relationships between the observations and the desired image can be formulated as

$$\begin{cases} \tilde{X} = f_{\text{spatial}}(X)\text{=AX+N} \\ Z = f_{\text{temporal}}(X) \\ Y = f_{\text{heterogeneous}}(X) \end{cases} \tag{1}$$

where $f_{\text{spatial}}(\cdot)$ denotes the spatial degradation relationship from X to $\tilde{X}$, usually assumed to be a blurring and downsampling operation [6], which can be expressed by the blurring and downsampling matrix A and the noise N.$f_{\text{temporal}}(\cdot)$ denotes the temporal relationship from X to Z, usually assumed to be a linear model [29], [31]. $f_{\text{heterogeneous}}(\cdot)$ denotes the heterogeneous relationship between X and Y, which is currently difficult to express explicitly.

Fig. 1 displays the flowchart of the proposed method. The proposed deep residual cycle GAN is based on the GAN [32] framework. In the network training depicted in Fig. 1(a), the network can be divided into a forward fusion part and a backward degeneration feedback part. The forward fusion part includes a forward generator and a forward discriminator. The input of the forward generator network consists of observations resized to the same spatial size and concatenated along the spectral dimension. Specifically, for heterogenous spatiotemporal–spectral fusion, the input to the forward generator composes of $t2$ HR SAR, $t1$ HR MS, and resized $t2$ LR MS images, namely, $(Y, Z, \hat{X})$. For heterogeneous spatiospectral fusion, the input to the forward generator is $(Y, \hat{X})$. For spatiotemporal fusion, the input to the forward generator is $(Z, \hat{X})$. That is, three networks are trained to exploit the effect of different fusion strategies. In the following, we describe the proposed network in detail with the example of heterogenous spatiotemporal–spectral fusion. As shown in Fig. 1(a), the output of the forward generator is the fused HR MS image at $t2$. This can be written as

$$X_f = \mathbf{G_F}\big((Y, Z, \hat{X}); \Theta_F\big) \tag{2}$$

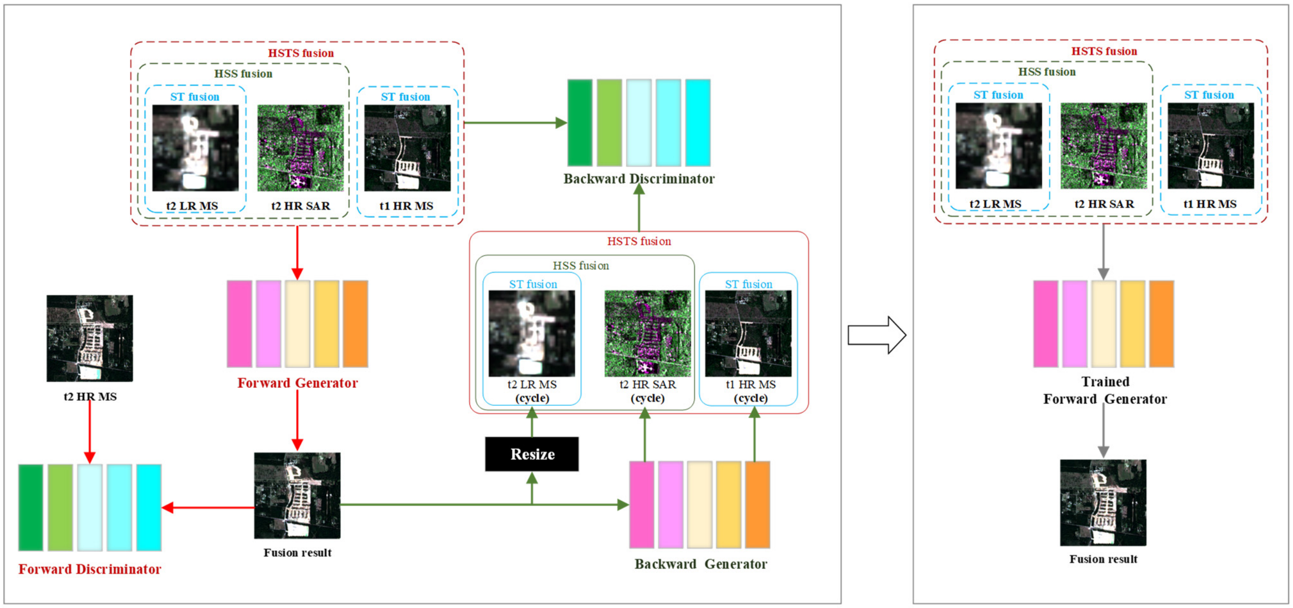
where X_f is the output of the forward generator $\mathbf{G_F}(\cdot)$, and Θ_F represents the trainable parameters. The forward discriminator discriminates X_f and the label data X.

The backward degeneration feedback part takes the degradation process of remote sensing imaging into account and reversely generates observation images from the fusion result X_{f}. Since $f_{\text{spatial}}(\cdot)$ in (1) is relatively clear, but $f_{\text{temporal}}(\cdot)$ and $f_{\text{heterogeneous}}(\cdot)$ in (1) are difficult to accurately model. As shown in Fig. 1(a), the backward part utilizes the "resize" branch to regenerate the $t2$ LR MS image and the backward generator to implement $f_{\text{temporal}}(\cdot)$ and $f_{\text{heterogeneous}}(\cdot)$ implicitly. They can be expressed as

$$\hat{X}^* = \mathbf{resize}\big(X_f\big) \tag{3}$$

$$Y^*, Z^* = \mathbf{G_B}\big(X_f; \Theta_B\big) \tag{4}$$

where $\mathbf{resize}(\cdot)$ represents the blurring and resampling operation. Note that, in the case of thick cloud cover, $\mathbf{resize}(\cdot)$ represents the sequential operation of blurring and resampling and adding the cloud mask. $\hat{X}^*$ is the regenerated $t2$ LR MS image. $\mathbf{G_B}(\cdot)$ and Θ_B denote the backward generator and the corresponding trainable parameters, respectively. Y^* and Z^* denote the regenerated $t2$ HR SAR and $t1$ HR MS images,



(a) (b)

Fig. 1. Flowchart of the proposed method. HSTS fusion represents heterogeneous spatiotemporal–spectral fusion, HSS fusion represents heterogeneous spatiospectral fusion, and ST fusion represents spatiotemporal fusion. (a) Train and (b) test.

respectively. A cycle forms from the inputs of the forward generator to the outputs of the "resize" branch and backward generator. The backward discriminator discriminates $(Y, Z, \hat{X})$ and $(Y^*, Z^*, \hat{X}^*)$. In the network testing depicted in Fig. 1(b), the corresponding observations are input into the trained forward generator of each fusion class, and the output is then the final fusion result.

A. Architecture of the Proposed Network

The proposed deep residual cycle GAN includes two generators and two discriminators. The two generators have the same network structure, and the two discriminators also have the same network structure.

The structure of the adopted generator networks is shown in Fig. 2, which is similar to that of [33] and consists of a feature extraction module, a feature encoding module, a residual learning module, a feature decoding module, and a feature compression module.

1) The feature extraction module extracts features from the inputs. It is a "Conv + BN + ReLU" block that consists of a convolutional layer, a batch normalization layer, and a rectified linear unit (ReLU) activation function layer. The convolutional layer consists of 64 filters of 7 × 7 × InC with stride 1, where InC denotes the channels of the input image.
2) The feature encoding module downsamples the feature maps by convolutional layers with stride 2 [34], which enlarges the receptive field of the features without increasing the convolution kernel size or the network depth. It includes two "Conv + BN + ReLU" blocks, in which the first convolutional layer consists of 128 filters of 3 × 3 × 64 and the second convolutional layer consists of 256 filters of 3 × 3 × 128.
3) The residual learning module utilizes the popular residual learning strategy [35], whose effectiveness has been verified in many tasks [36]–[38]. It consists of six residual blocks. As shown in Fig. 2, a residual block consists of a "Conv + BN + ReLU" structure and a "Conv + BN" structure, in which both convolutional layers include 256 filters of 3 × 3 × 256. $\oplus$ denotes a pixel-by-pixel addition function.
4) The feature decoding module has the opposite function to the feature encoding module, which gradually expands the feature map to the input image size through deconvolution [39]. It is made up of two "DeConv + BN + ReLU" blocks, where the first deconvolutional layer includes 128 filters of 3 × 3 × 256 with stride 2 and the second deconvolutional layer includes 64 filters of 3 × 3 × 128 with stride 2.
5) The feature compression module maps the features back to the image domain. This module consists of a "Conv + Tanh" block, in which the convolutional layer includes OutC filters of 7 × 7 × 64, where OutC denotes the channels of the output image. The tanh activation function layer [40] is empirically used in the last layer of the generator.

For the discriminator networks, we use the popular PatchGAN architecture in [41], which determines whether the image patches are real or fake. Fig. 3 shows the detailed structure of the proposed discriminators, which consist of one "Conv + LeakyReLU" block, three "Conv + BN + LeakyReLU" blocks, and a "Conv + Sigmoid" block. The kernel size of all the convolutional layers is 4 × 4, the stride of the first

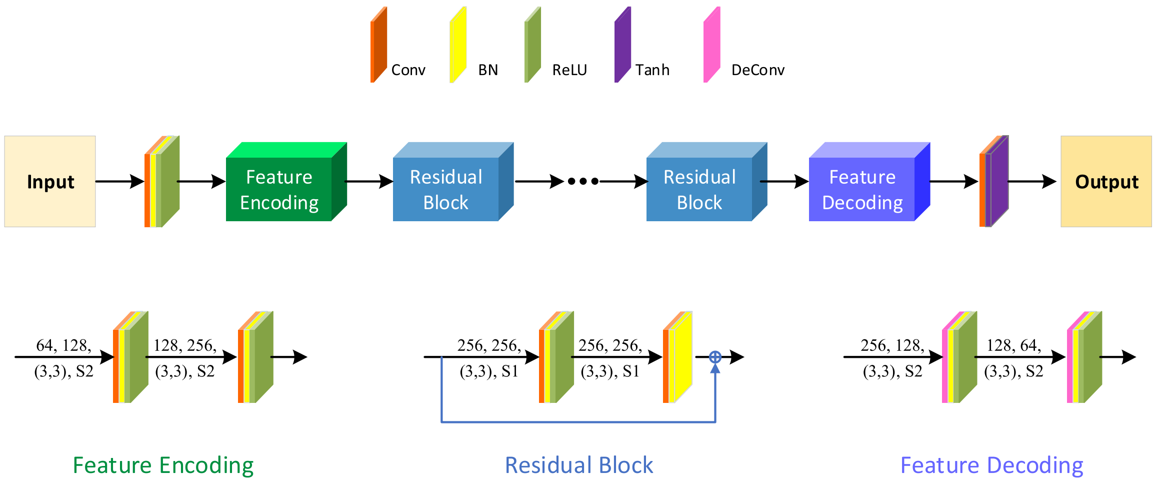


Fig. 2. Structure of the proposed generators.

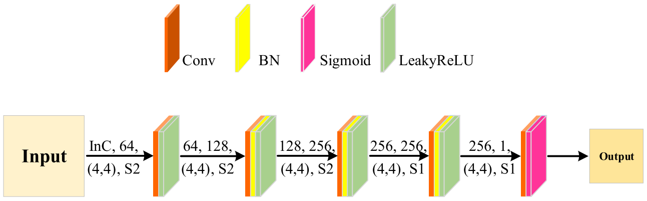


Fig. 3. Structure of the proposed discriminators.

three convolutional layers is 2, and the stride of the last two convolutional layers is 1.

B. Loss Function

With the network structure, the loss function of the proposed network also includes two parts: the loss function of the generator networks and the loss function of the discriminator networks. Two generators are trained together with one loss function, which can be written as follows:

$$L_G = L_{\text{adv}} + L_{\text{con}} \tag{5}$$

where L_G denotes the total loss of the generators and consists of two terms: L_{adv} and L_{con}. L_{adv} is the adversarial loss between the generators and discriminators [42], which is defined as

$$\begin{aligned} L_{\text{adv}} = &\frac{1}{N}\sum_{n=1}^{N}\left\|\mathbf{D_F}\left(X_f\right) - 1\right\|_F^2 \\ &+ \frac{1}{N}\sum_{n=1}^{N}\left\|\mathbf{D_B}\left(\left(Y^*, Z^*, \hat{X}^*\right)\right) - 1\right\|_F^2 \end{aligned} \tag{6}$$

where the first term is the forward discriminator-related adversarial loss and the second term is the backward discriminator-related adversarial loss. N denotes the number of patches in a batch [6]. The mean squared error (MSE) loss [43] is empirically utilized in L_{adv}, and $\|\cdot\|_F$ is the matrix Frobenius norm.

L_{con} in (5) is the content loss to ensure that the outputs of the generators are close to the ground truth. Specifically, the content loss is defined as follows:

$$\begin{aligned} L_{\text{con}} = &\lambda_1 * \frac{1}{N}\sum_{n=1}^{N}\left\|X_f - X\right\|_1 + \lambda_2 * \frac{1}{N}\sum_{n=1}^{N}\| M \odot \\ &\times \left(X_f - X\right)\|_1 \\ &+ \lambda_3 * \frac{1}{N}\sum_{n=1}^{N}\left\|\left(Y^*, Z^*, \hat{X}^*\right) - \left(Y, Z, \hat{X}\right)\right\|_1 \end{aligned} \tag{7}$$

where the first term calculates the global loss between the forward generator output and the ideal fusion result. The second term calculates the local loss between the forward generator output and the ideal fusion result in the cloud-covered areas, where M is the binarized cloud mask, with 1 representing a cloud-covered area and 0 a cloudless area. $\odot$ is the dot product operator. The last term is the cycle consistency loss between the inputs of the forward generator and the outputs of the "resize" branch and the backward generator. λ_1, λ_2, and λ_3 are the adjustable parameters that balance the three terms. Since mean absolute error (MAE) loss [43] is less sensitive to outliers than MSE loss, it is empirically used in the content loss, and $\|\cdot\|_1$is the L1 norm.

The two discriminators are trained separately with their own loss functions. The forward discriminator distinguishes the forward generator output and the ideal fusion result, i.e., the $t2$ HR MS image. It judges the $t2$ HR MS image to be true with label 1, and the output of the forward generator to be fake with label 0. The loss function can be formulated as follows:

$$L_{D_F} = \frac{1}{2}\left(\frac{1}{N}\sum_{n=1}^{N}\|\mathbf{D_F}\left(X_f\right) - 0\|_F^2 + \frac{1}{N}\sum_{n=1}^{N}\|\mathbf{D_F}(X) - 1\|_F^2\right). \tag{8}$$

Similarly, the backward discriminator distinguishes the inputs of the forward generator and the outputs of the "resize" branch and the backward generator. It judges the former to be true with label 1, and the latter to be fake with label 0. The

TABLE I

DATASETS USED IN THE RESOLUTION IMPROVEMENTS EXPERIMENTS

	Sensor	Time	Resolution	Size (training)	Size (test)	Location (training)	Location (test)
Resolution improvement	MODIS	2017-10-29 (t2)	500 m	128×106×3 60×60×3 124×130×3	60×60×3	(95.62°W, 30.23°N) (95.43°W, 29.86°N) (95.77°W, 29.41°N)	(95.78°W, 29.85°N)
	Sentinel-1	2017-10-28 (t2)	10 m	6400×5300×2 3000×3000×2 6200×6500×2	3000×3000×2		
	Sentinel-2	2016-09-29 (t1)	10 m	6400×5300×3 3000×3000×3 6400×6500×3	3000×3000×3		
	Sentinel-2	2017-10-29 (t2)	10 m	6400×5300×3 3000×3000×3 6400×6500×3	3000×3000×3		

loss function can be formulated as follows:

$$L_{D_B} = \frac{1}{2N}\sum_{n=1}^{N}\left\|\mathbf{D_B}\left(\left(Y, Z, \hat{X}\right)\right) - 1\right\|_F^2 + \frac{1}{2N}\sum_{n=1}^{N}\left\|\mathbf{D_B}\left(\left(Y^*, Z^*, \hat{X}^*\right)\right) - 0\right\|_F^2. \quad (9)$$

In the network training, the trainable parameters of the generators and the discriminators are updated sequentially, according to the corresponding loss functions. Considering the sensitivity of GANs, we learned from [44] and set the hyperparameters accordingly. The learning rate in the first 75 epochs was set to 0.0002, and this was decayed linearly to 0 in the last 75 epochs. In addition, the adjustable weights in (7) were set as: $\lambda_1 = 50$, $\lambda_2 = 100$, and $\lambda_3 = 50$. The Adam optimizer [45] was utilized to optimize the network parameters.

III. EXPERIMENTS

To verify the performance of the proposed method in the two difficult problems of land-cover change and thick cloud cover, three sets of experiments were carried out: resolution improvement experiments, thick cloud removal experiments, and joint processing experiments of thick cloud removal and resolution improvement. Images from multiple sensors were utilized, i.e., the Moderate Resolution Imaging Spectroradiometer (MODIS), the Landsat 8 Operational Land Imager (OLI), the Sentinel-2 Multispectral Instrument (MSI), and the Sentinel-1 C-SAR instrument. All the optical images include the red, green, and blue bands, namely, the B2, B3, and B4 bands of the Sentinel-2 MS image and the Landsat 8 MS image and the B01, B03, and B04 bands of the MODIS MS image. All SAR images include the VH and VV polarization bands. More details of the datasets used in each experiment are provided in the corresponding sections. In this article, five representative indices are used to evaluate the performance of the fusion results quantitatively: the relative dimensionless global error in synthesis (ERGAS) [46], the spectral angle mapper (SAM) [46], the Q metric [46], the peak-signal-to-noise ratio (PSNR), and the structural similarity index (SSIM) [47]. Among the different indices, ERGAS is used for the resolution improvement experiments.

A. Resolution Improvement Experiments

In the resolution improvement experiments, we fused the $t2$ MODIS MS image of a 500-m resolution, the $t2$ Sentinel-1 dual-polarization SAR image of a 10-m resolution, and the $t1$ Sentinel-2 MS image of a 10-m resolution to obtain the $t2$ MS image of a 10-m resolution. Note that due to the different revisit periods of the sensors, the capture time of $t2$ images for different sensors may be slightly different. Table I lists the datasets used for the network training and testing. The $t2$ MODIS MS image is the MODO9GA product captured on October 29, 2017. The $t2$ Sentinel-1 dual-polarization SAR image is the ground range detected (GRD) product of stripmap (SM) mode captured on October 28, 2017. The $t1$ Sentinel-2 MS image is the Level-1C product captured on September 29, 2016. The $t2$ Sentinel-2 MS image captured on October 29, 2017 was used as the label data in the network training and the reference image in the network testing. Three image pairs were utilized to generate the training patches, where the sizes of the MODIS MS images were 128 × 106, 60 × 60, and 124 × 130 and the sizes of the other images were 6400 × 5300, 3000 × 3000, and 6200 × 6500; ×2 and ×3 in the "Size" column represented the spectral bands of images. The center locations of the images were (95.62°W, 30.23°N), (95.43°W, 29.86°N), and (95.77°W, 29.41°N). In total, 1984 patches of size 200 × 200 were randomly generated from these three image pairs. In the network testing, a pair of images was utilized, where the size of the MODIS MS image was 60 × 60 and the size of the other images was 3000 × 3000. The center location of the images was (95.78°W, 29.85°N).

1) Visual Analysis: Fig. 4 displays the fusion results with the size of 3000 × 3000 for the resolution improvement experiments, where the MS images are displayed in the red–green–blue band combination, and the SAR image is displayed in the VV–VH–VV band combination. The first row shows the observations, i.e., the MODIS MS image

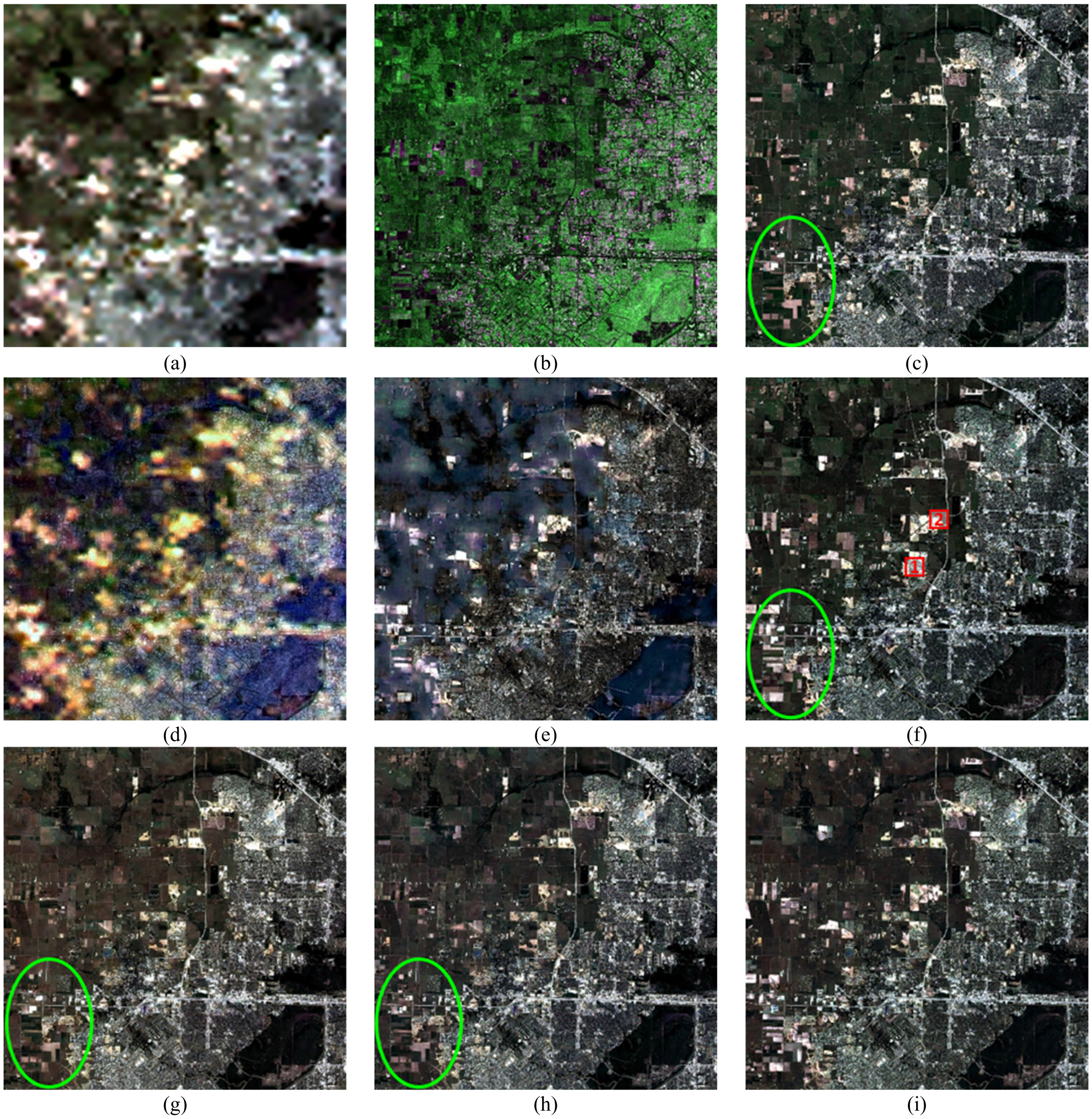


Fig. 4. Fusion results for the resolution improvement experiments. (a) $t2$ MODIS MS, (b) $t2$ Sentinel-1 SAR, (c) $t1$ Sentinel-2 MS, (d) AIHS, (e) Proposed-HSS, (f) $t2$ Sentinel-2 MS (reference), (g) ESRCNN, (h) Proposed-ST, and (i) Proposed-HSTS.

from October 29, 2017 (written as $t2$ MODIS MS), the Sentinel-1 dual-polarization SAR image from October 28, 2017 (written as $t2$ Sentinel-1 SAR), and the Sentinel-2 MS image from September 29, 2016 (written as $t1$ Sentinel-2 MS). Fig. 4(d) and (e) in the second row shows the results of fusing the $t2$ MODIS MS image and the $t2$ Sentinel-1 SAR image. Fig. 4(d) shows the result of adaptive intensity-hue-saturation (AIHS) algorithm [48], which is a comparison algorithm that originated from optical spatiospectral fusion. Fig. 4(e) shows the result of the proposed deep residual cycle GAN adopting the heterogeneous spatiospectral fusion strategy, which is written as the proposed-HSS. Fig. 4(f) is the Sentinel-2 MS image from October 29, 2017 (written as $t2$ Sentinel-2 MS) that acts as the reference image. Fig. 4(g) and (h) in the third row shows the results of fusing the $t2$ MODIS MS image and the $t1$ Sentinel-2 MS image. Fig. 4(g) shows the result of the extended superresolution convolutional neural network (ESRCNN) algorithm [49]. Fig. 4(h) shows the result of the proposed network adopting the spatiotemporal fusion strategy, which is written as the proposed-ST. Fig. 4(i) shows the fusion result of the proposed network adopting the heterogeneous spatiotemporal–spectral fusion strategy of fusing the $t2$ MODIS MS image, the $t2$ SAR image, and the $t1$ Sentinel-2 MS image, which is written as the proposed-HSTS. As shown in Fig. 4, lots of land-cover changes took place between $t1$ and $t2$, as can be seen in the Sentinel-2 MS images at $t1$ and $t2$ (in the green ellipses in Fig. 4(c) and (f), for example). By comparing the fusion results, it can be observed that the AIHS method results in severe global spatial and spectral distortion. The proposed-HSS method effectively increases the spatial information of the $t2$ MODIS MS image, but it results in some local spectral distortion, as can be found in the upper-left corner of Fig. 4(e). The ESRCNN and proposed-ST methods perform much better than the AIHS and proposed-HSS methods, but neither predicts the changed land covers, as shown in the green ellipses in Fig. 4(g) and (h).

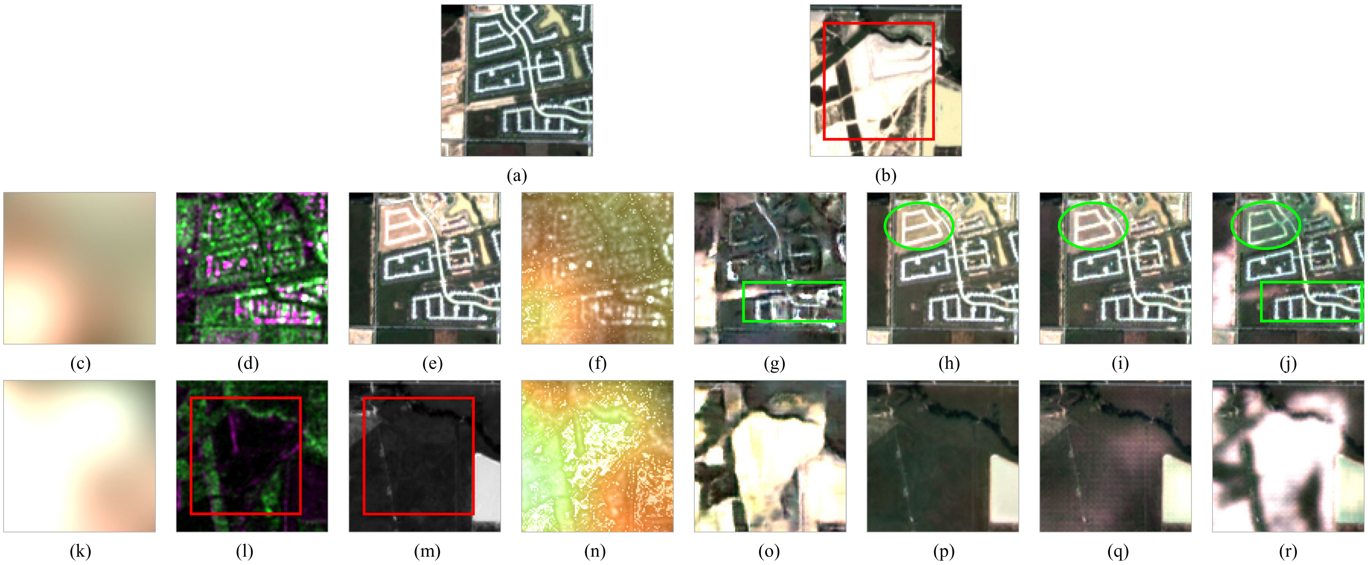


Fig. 5. Sampling area pointed to by red rectangles in Fig. 4(f). S1 and S2 are the abbreviations for Sentinel-1 and Sentinel-2, respectively. (a) $t2$ S2 MS (reference: sampling area 1), (b) $t2$ S2 MS (reference: sampling area 2), (c) $t2$ MODIS MS, (d) $t2$ S1 SAR, (e) $t1$ S2 MS, (f) AIHS, (g) Proposed-HSS, (h) ESRCNN, (i) Proposed-ST, (j) Proposed-HSTS, (k) $t2$ MODIS MS, (l) $t2$ S1 SAR, (m) $t1$ S2 MS, (n) AIHS, (o) Proposed-HSS, (p) ESRCNN, (q) Proposed-ST, and (r) Proposed-HSTS.

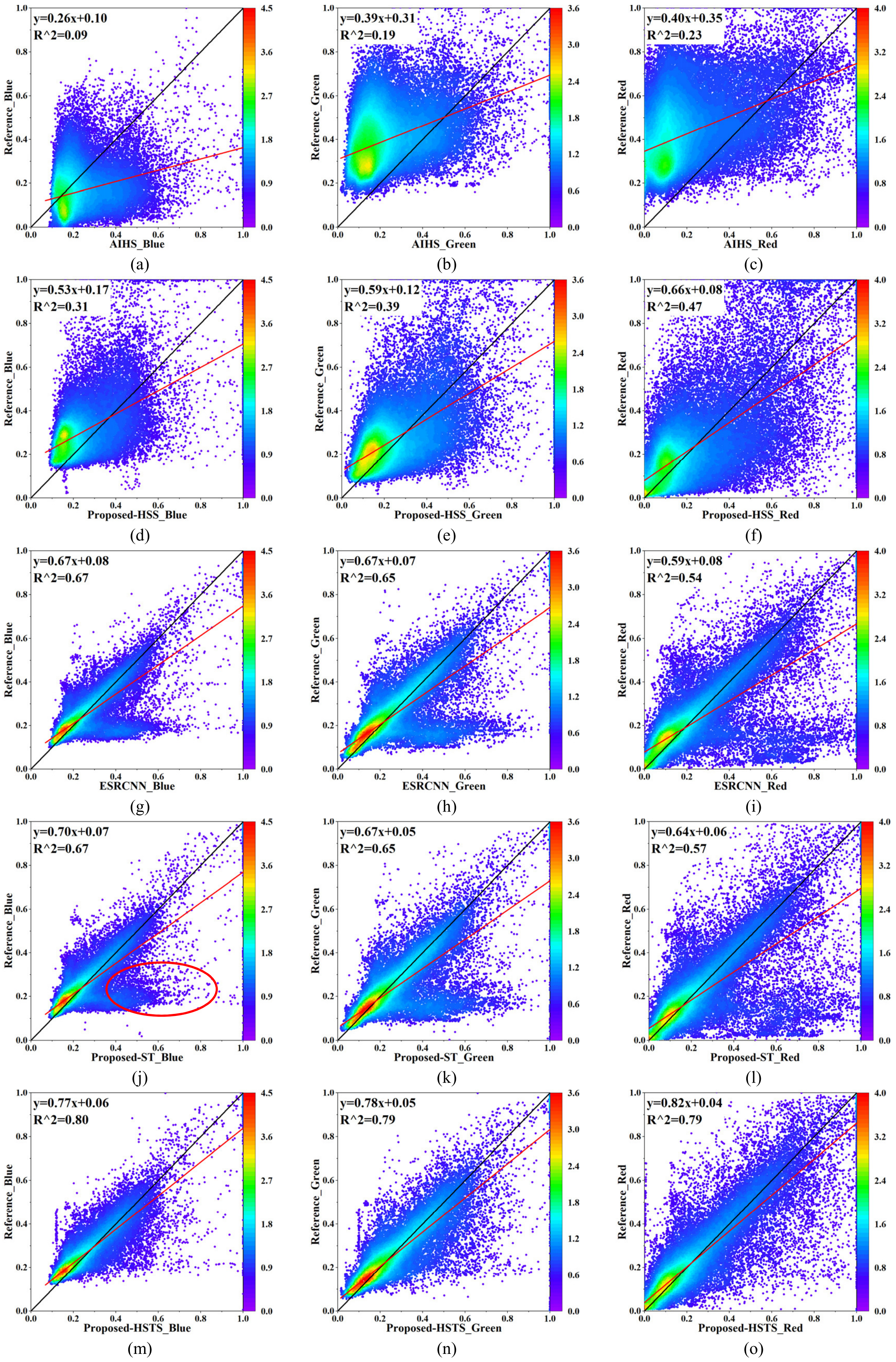
On the whole, the fusion result of the proposed-HSTS method in Fig. 4(i) is visually the closest to the reference.

To further compare the effects of these methods, two representative areas with the size of 200 × 200 in red rectangles in Fig. 4(f) were selected for the analysis in Fig. 5. As shown in Fig. 5, between $t1$ and $t2$, minor land cover changes occurred in sampling area 1; lots of land cover changes occurred in sampling area 2. First, in sampling area 1 shown in the second row of Fig. 5, the AIHS method performs extremely poorly, both spatially and spectrally, which reflects the incompatibility of the migration of the optical image fusion method to SAR-optical image fusion. The proposed-HSS method performs better than the AIHS method, but it produces distorted spatial structure, as shown in the green rectangle in Fig. 5(g). In the two spatiotemporal fusion-based methods, the results of the ESRCNN and proposed-ST methods differ slightly, but neither reflects the land-cover changes, as shown in the green ellipses in Fig. 5(h) and (i). Overall, the proposed-HSTS method combines the advantage of the proposed-ST method to obtain unchanged land covers, as shown in the green rectangle in Fig. 5(j), and the ability of the proposed-HSS method to reflect changed land covers, as shown in the green ellipse in Fig. 5(j).

In sampling area 2 shown in the third row, the white building in the red rectangle is not visible at $t1$ but $t2$, as shown in Fig. 5(b) and (m). Due to the clear structural information of the white building in Fig. 5(l), the proposed-HSS method successfully reconstructs the changed land covers, as shown in Fig. 5(o). In the two spatiotemporal fusion-based methods, the ESRCNN method does not detect any land-cover changes, as in Fig. 5(p), while the proposed-ST method detects some changed land covers but fails to reconstruct them, as shown in Fig. 5(q). The proposed-HSTS method effectively predicts the land-cover changes, as shown in Fig. 5(r), but the result seems not as clear as that of the proposed-HSS method in Fig. 5(o). The reason should be that the proposed-HSTS method obtains HR spatial information by fusing the $t1$ Sentinel-2 MS image and the $t2$ Sentinel-1 SAR image, but the $t1$ Sentinel-2 MS image lacks the corresponding spatial structural information of changed land covers.

2) Quantitative Analysis: Fig. 6 displays the point density between the fusion results and the reference image in Fig. 4. In Fig. 6(a)–(o), the color scheme indicates the point density, the black line refers to the function $y = x$, and the red line represents the band-by-band linear fitting between the fusion result and the reference image. The smaller the angle between the red line and the black line, the closer the slope of the fitted line is to 1. In addition, the narrower the point cloud and the more evenly the points are distributed on both sides of the fitted line, the larger the R^2 value and the more reliable the fitting result. From Fig. 6, it is clear that the proposed-HSS method performs much better than the AIHS method, and the proposed-ST method performs slightly better than the ESRCNN method. Comparing the three proposed methods, the proposed-ST method has a better average slope, a narrower point cloud, and larger R^2 values than the proposed-HSS method, which indicates that the proposed-ST method is generally better than the proposed-HSS method. However, it is noticeable that the proposed-ST method has an uneven distribution of points in all bands, as shown in the red ellipse in Fig. 6(j), which is likely to be caused by land-cover changes. In contrast, the result of the proposed-HSTS method has uniformly distributed points, an average slope closest to 1, and an average R^2 of about 0.80, outperforming other methods by a significant margin.

Table II lists the quantitative evaluation results for the resolution improvement experiments, where the best performance for each index is marked in bold. With the visual results, the proposed-HSS method performs much better than the AIHS method in all the indices, and the proposed-ST method performs slightly better than the ESRCNN method



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

(m) (n) (o)

Fig. 6. Point density between the fusion results and the reference image in Fig. 4. (a) AIHS_Blue. (b) AIHS_Green. (c) AIHS_Red. (d) Proposed-HSS_Blue. (e) Proposed-HSS_Green. (f) Proposed-HSS_Red. (g) ESRCNN_Blue. (h) ESRCNN_Green. (i) ESRCNN_Red. (j) Proposed-ST_Blue. (k) Proposed-ST_Green. (l) Proposed-ST_Red. (m) Proposed-HSTS_Blue. (n) Proposed-HSTS_Green. (o) Proposed-HSTS_Red.

in all the indices. The proposed-HSTS method performs the best in all the quality indices, which verifies the effectiveness of the proposed heterogeneous spatiotemporal–spectral fusion strategy.

TABLE II
QUANTITATIVE RESULTS FOR THE RESOLUTION IMPROVEMENT EXPERIMENTS

Algorithm	**SAM**	**ERGAS**	**Q**	**PSNR**	**SSIM$_B$**	**SSIM$_G$**	**SSIM$_R$**	**SSIM$_{AVG}$**
Ideal data	0	0	1	$+\infty$	1	1	1	1
AIHS	29.3980	1.4961	−0.0556	12.9123	0.2560	0.1821	0.1046	0.1809
Proposed-HSS	8.1561	1.3627	0.3047	17.6444	0.5955	0.5807	0.5255	0.5672
ESRCNN	6.0370	0.9575	0.6703	21.2852	0.8410	0.8163	0.7307	0.7960
Proposed-ST	5.9166	0.9362	0.6717	21.4413	0.8440	0.8253	0.7478	0.8057
Proposed-HSTS	**5.4886**	**0.7158**	**0.7075**	**23.4775**	**0.8770**	**0.8597**	**0.8008**	**0.8458**

B. Thick Cloud Removal Experiments

Optical remote sensing images are susceptible to clouds, for which lots of related studies have been developed [38], [50]–[54]. To fully demonstrate the effectiveness of the proposed method, in the thick cloud removal experiment, four comparison methods are selected. They are the Simulation-Fusion GAN (Simu-Fus-GAN) method [52], the SAR-Optical-conditional GAN (SAR-opt-cGAN) method [38], the spatiotemporal fusion-based cloud removal (STF-CR) method [53], and the spatial–temporal–spectral deep convolutional neural network (STS-CNN) method [54].

Table III lists the details of the network training and test datasets used for the thick cloud removal experiments. In the experiments, we removed the thick cloud of the Sentinel-2 MS image from December 22, 2019 ($t2$) with the help of the Sentinel-1 dual-polarization SAR image from December 21, 2019 ($t2$) and the cloudless Sentinel-2 MS image from June 10, 2019 ($t1$). Since we cannot simultaneously capture both cloudy and cloudless MS images of the same day, in the network training and simulated experiments, we synthesized the $t2$ cloudy Sentinel-2 MS image by adding a cloud mask to the observed $t2$ cloudless Sentinel-2 MS image. The $t2$ cloudless Sentinel-2 observation was then considered as the label data for the network training and the reference for the simulated experiments. As shown in Table III, in the network training, one image pair of size 5830 × 10 580, with a center location of (88.44°W, 41.96°N), was used to generate the training patches. The cloud coverage for the synthesized $t2$ cloudy Sentinel-2 MS image was 19.72%. In total, 8112 patches of size 128 × 128 were randomly generated. In the simulated experiments, a pair of images of size 5830 × 400, with a center location of (88.43°W, 41.47°N), was utilized to generate 22 small image pairs with the size of 256 × 256. The cloud coverage for the synthesized $t2$ cloudy Sentinel-2 MS image in simulated experiments was 23.15%. In the real-data experiments, the $t2$ cloudy Sentinel-2 MS image was captured on November 22, 2019, its cloud and shadow coverage was 40.11%, and its other parameters were the same as in the simulated experiments.

1) Simulated Thick Cloud Removal Experiments: A group of simulated experiment results with the size of 256 × 256 are displayed in Fig. 7 in the red–green–blue band combination, where the lower-right corner is a magnified display of the image inside the red rectangle. The first row of Fig. 7 displays the observations to be fused. In the second row, the first three are the results of heterogeneous spatiospectral fusion-based methods that fuse the $t2$ cloudy Sentinel-2 MS image and the $t2$ Sentinel-1 SAR image; the fourth is the reference image. In the third row, the first three are the results of spatiotemporal fusion-based methods that fuse the $t2$ cloudy Sentinel-2 MS image and the $t1$ cloudless Sentinel-2 MS image. The last is the result of the proposed-HSTS method that fuses the $t2$ cloudy Sentinel-2 MS image, the $t2$ Sentinel-1 SAR image, and the $t1$ cloudless Sentinel-2 MS image.

First, for the heterogeneous spatiospectral fusion-based methods, the Simu-Fus-GAN method produces severe spectral distortion and the SAR-opt-cGAN method produces severe spatial distortion, as shown in the zoomed areas of Fig. 7(e) and (f). The fusion result of the proposed-HSS method is closer to the reference image, but it has some blurring spatially, as shown in the edges of the bail soil in the yellow ellipse of the zoomed area of Fig. 7(g), which is caused by the unclear structures in the $t2$ SAR images in Fig. 7(b) and (c). Then, among the three spatiotemporal fusion-based methods, the spatial details of the cloud coverage area reconstructed by the STF-CR method are close to the $t1$ Sentinel-2 MS image instead of the reference image, as shown in the zoomed area in Fig. 7(i). The STS-CNN method produces obvious spectral distortion and spatial blurring, as shown in the zoomed area in Fig. 7(j). The result of the proposed-ST method is closer to the reference, both spatially and spectrally, as shown in the zoomed area in Fig. 7(k). Moreover, comparing the results of the three proposed methods, the proposed-HSS method shows some blurring in the zoomed area in Fig. 7(g). The proposed-ST method performs better than the proposed-HSS method, but slightly worse than the proposed-HSTS method, as indicated by the vegetation in the yellow ellipse of the zoomed area in Fig. 7(l). The result of the proposed-HSTS method is the closest to the reference. Table IV lists the quantitative evaluation results for the simulated thick cloud removal experiments, with an average of 22 groups. In Table IV, the best performance for each index is marked in bold. Consistent with the visual results, the proposed-HSS method outperforms the Simu-Fus-GAN and SAR-opt-cGAN methods in all the quality indices. This is also true for the spatiotemporal fusion-based methods, where the proposed-ST method performs better than the STF-CR and STS-CNN methods in all the indices. They show the superiority of the proposed network over the comparison algorithms. Furthermore, comparing the proposed-HSS, proposed-ST, and proposed-HSTS methods, it is clear that the proposed-ST method performs better than the proposed-HSS method in all the indices. This verifies that heterogeneous

TABLE III
DATASETS USED IN THE THICK CLOUD REMOVAL EXPERIMENTS

	Sensor	Time	Resolution	Size (training)	Size (test)	Cloud (training/test: %)	Location (training/test)
Thick cloud removal	Sentinel-2	2019-12-22 (t2)	10 m	5830×10580×3	5830×400×3	19.72/23.15	(88.44°W, 41.96°N)/ (88.43°W, 41.47°N)
	Sentinel-1	2019-12-21 (t2)	10 m	5830×10580×2	5830×400×2	0	
	Sentinel-2	2019-06-10 (t1)	10 m	5830×10580×3	5830×400×3	0	
	Sentinel-2	2019-12-22 (t2)	10 m	5830×10580×3	5830×400×3	0	
	Sentinel-2	2019-11-22 (t2)	10 m	0	5830×400×3	40.11	

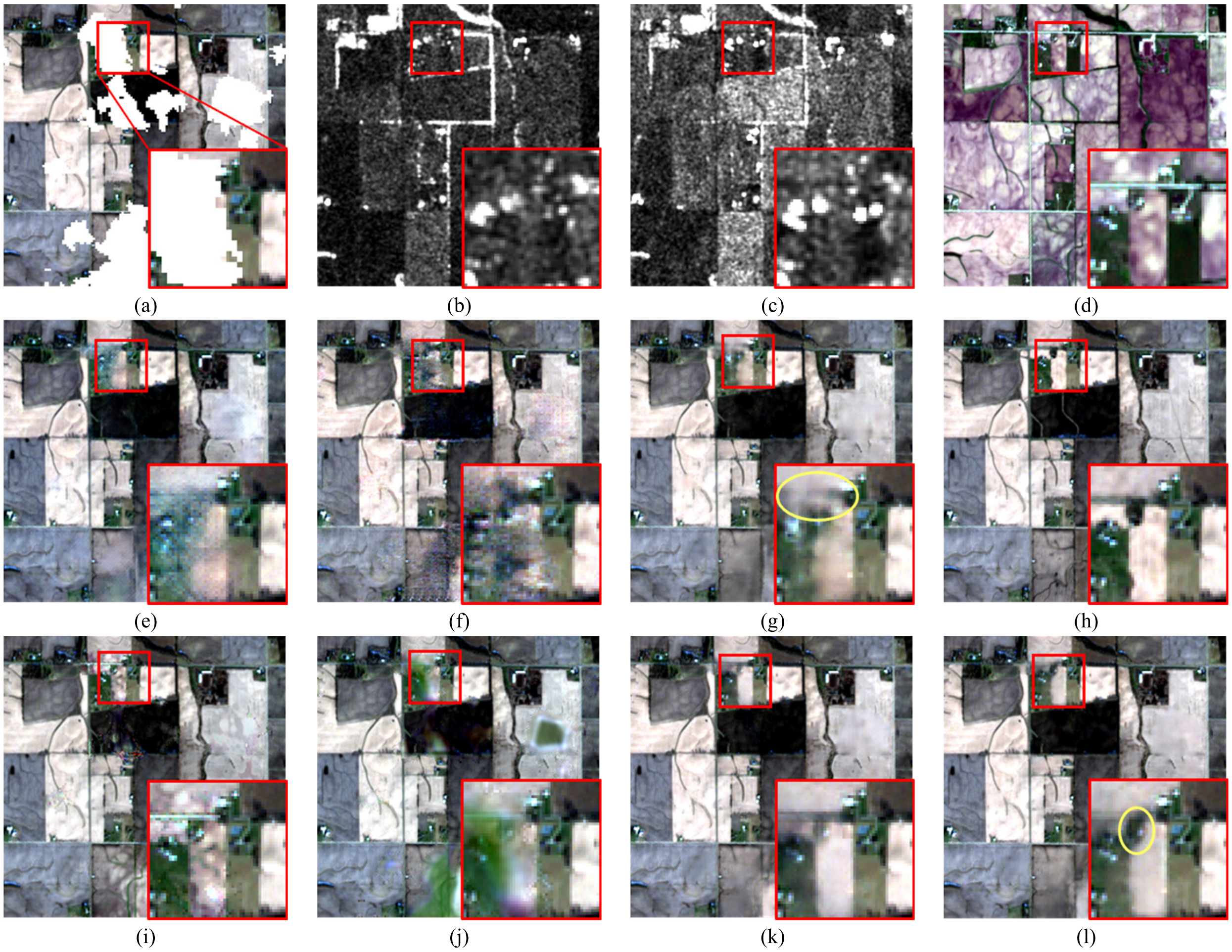


Fig. 7. Results for the simulated thick cloud removal experiment. (a) $t2$ cloudy Sentinel-2 MS, (b) $t2$ Sentinel-1 SAR_VH, (c) $t2$ Sentinel-1 SAR_VV, (d) $t1$ Sentinel-2 MS, (e) Simu-Fus-GAN, (f) SAR-opt-cGAN, (g) Proposed-HSS, (h) $t2$ Sentinel-2 MS (reference), (i) STF-CR, (j) STS-CNN, (k) Proposed-ST, and (l) Proposed-HSTS.

information fusion without a temporal change is still more difficult than homogeneous optical information fusion with a temporal change. The proposed-HSTS method performs the best in all the indices, which confirms the advantage of the heterogeneous spatiotemporal–spectral fusion strategy.

2) Real-Data Thick Cloud Removal Experiments: A group of real-data experiment results with the size of 256 × 256 are displayed in Fig. 8 in the red–green–blue band combination. In Fig. 8, the lower-left corner is a magnified display of the image inside the green rectangle. For the real-data experiments, the $t2$ cloudy Sentinel-2 MS image in Fig. 8(a) was obtained on November 22, 2019, and its cloud and shadow coverage was 48.01%. For the heterogeneous spatiospectral fusion-based methods in Fig. 8(e)–(g), the results of the Simu-Fus-GAN and SAR-opt-cGAN methods show obvious cloud-cover boundaries, while the transition from the cloud-covered area to the cloudless area in the result of the proposed-HSS method is more natural. For the spatiotemporal fusion-based methods, the result of the STF-CR method in the cloud-covered areas is similar to the $t1$ Sentinel-2 MS image but is quite different from the reference image, as shown in the zoomed area in Fig. 8(i). The result of the STS-CNN method shows obvious spectral distortion in the cloud-covered areas, as shown in the zoomed area in Fig. 8(j). The result of

TABLE IV

QUANTITATIVE RESULTS FOR THE SIMULATED THICK CLOUD REMOVAL EXPERIMENTS (22 GROUPS)

Algorithm	**SAM**	**Q**	**PSNR**	**SSIM$_B$**	**SSIM$_G$**	**SSIM$_R$**	**SSIM$_{AVG}$**
Ideal data	0	1	$+\infty$	1	1	1	1
Simu-Fus-GAN	4.6584	0.8465	23.2968	0.8746	0.8554	0.7930	0.8410
SAR-opt-cGAN	3.3499	0.8346	24.9575	0.8259	0.8250	0.7857	0.8122
Proposed-HSS	1.3511	0.8940	28.8842	0.9197	0.9015	0.8712	0.8975
STF-CR	1.7946	0.8833	27.7753	0.9018	0.8807	0.8412	0.8746
STS-CNN	2.2743	0.8647	26.3756	0.8937	0.8801	0.8483	0.8740
Proposed-ST	1.3439	0.9120	29.7203	0.9282	0.9123	0.8853	0.9086
Proposed-HSTS	**1.2132**	**0.9200**	**30.2195**	**0.9304**	**0.9169**	**0.8919**	**0.9131**

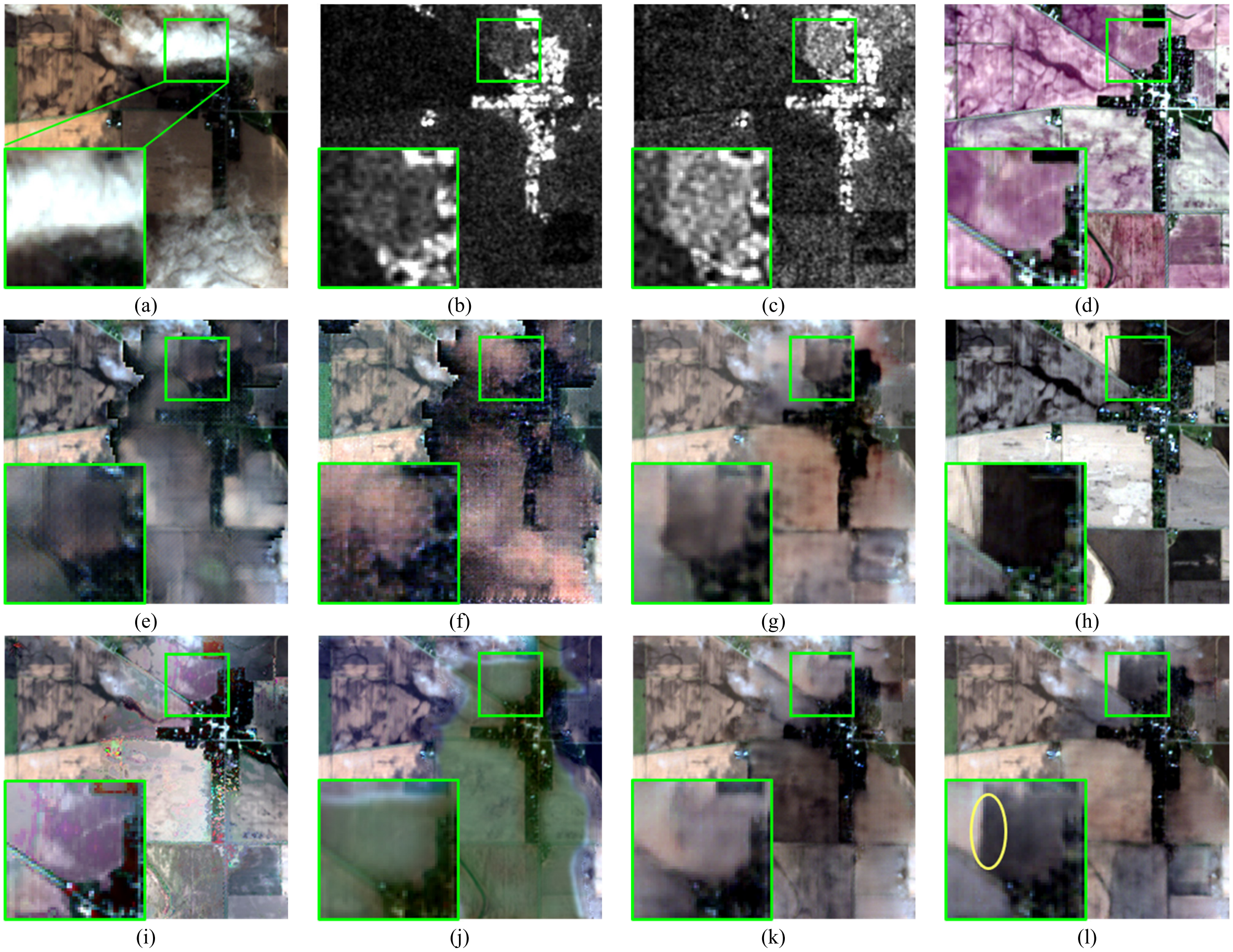


Fig. 8. Results for the real-data thick cloud removal experiments. (a) $t2$ cloudy Sentinel-2 MS, (b) $t2$ Sentinel-1 SAR_VH, (c) $t2$ Sentinel-1 SAR_VV, (d) $t1$ Sentinel-2 MS, (e) Simu-Fus-GAN, (f) SAR-opt-cGAN, (g) Proposed-HSS, (h) $t2$ Sentinel-2 MS (reference), (i) STF-CR, (j) STS-CNN, (k) Proposed-ST, and (l) Proposed-HSTS.

the proposed-ST method is closer to the reference image than the STF-CR and STS-CNN methods, but it is incapable of detecting land-cover changes, as shown in the bare soil of the zoomed area in Fig. 8(k). The proposed-HSTS method reflects the land-cover changes well, as indicated by the bare soil edge in the yellow ellipse of the zoomed area in Fig. 8(l), and it obtains the best results.

C. Joint Processing of Thick Cloud Removal and Resolution Improvement

In the joint processing of thick cloud removal and resolution improvement, we fused the $t2$ cloudy Landsat 8 MS image of a 30-m resolution, the $t2$ cloudless Sentinel-1 dual-polarization SAR image of a 10-m resolution, and the $t1$ cloudless Sentinel-2 MS image of a 10-m resolution to obtain the $t2$ cloudless MS image of a 10-m resolution. Table V gives the details of the network training and test datasets. In the network training and simulated experiments, the $t2$ (October 29, 2017) cloudy Landsat 8 MS image was synthesized by adding a cloud mask to the $t2$ (October 29, 2017) Sentinel-2 MS image, which was previously spatially downsampled to the resolution of the Landsat 8 MS image. The $t2$ Sentinel-1 SAR image was captured on October 28, 2017, and the $t1$ cloudless Sentinel-2 MS image was captured on September 29, 2016. As shown in Table V, one pair of images was utilized to generate

TABLE V

DATASETS USED IN THE JOINT PROCESSING OF THICK CLOUD REMOVAL AND RESOLUTION IMPROVEMENT

	Sensor	Time	Resolution	Size (training)	Size (test)	Cloud (training/test: %)	Location (training/test)
Joint processing of thick cloud removal and resolution improvement	Landsat 8	2017-10-29 (t2)	30 m	1900×2794×3	240×2794×3	32.58/28.28	(95.65°W, 30.09°N)/ (95.32°W, 30.10°N)
	Sentinel-1	2017-10-28 (t2)	10 m	5700×8382×2	720×8382×2	0	
	Sentinel-2	2016-09-29 (t1)	10 m	5700×8382×3	720×8382×3	0	
	Sentinel-2	2017-10-29 (t2)	10 m	5700×8382×3	720×8382×3	0	
	Landsat 8	2017-10-15 (t2)	30 m	0	240×2794×3	34.51	

TABLE VI

QUANTITATIVE RESULT FOR THE SIMULATED EXPERIMENTS IN THE JOINT PROCESSING OF THICK CLOUD REMOVAL AND RESOLUTION IMPROVEMENT (EIGHT GROUPS)

Algorithm	SAM	ERGAS	Q	PSNR	SSIM$_B$	SSIM$_G$	SSIM$_R$	SSIM$_{AVG}$
Ideal data	0	0	1	+∞	1	1	1	1
Proposed-HSS	5.8573	19.8796	0.5741	24.8477	0.8285	0.7845	0.7327	0.7819
Proposed-ST	4.0801	13.4593	0.8375	28.2400	0.9439	**0.9328**	0.8992	0.9253
Proposed-HSTS	**4.0198**	**12.1935**	**0.8396**	**29.0881**	**0.9457**	0.9325	**0.9010**	**0.9264**

the training patches, where the size of the Landsat 8 image was 1900 × 2794, and the size of the other images was 5700 × 8382. The center location was (95.65°W, 30.09°N). The cloud coverage for the synthesized Landsat 8 MS image in the network training was 32.58%. In total, 6336 patches of size 128 × 128 were randomly generated. In the simulated experiments, a pair of images was utilized, where the size of the Landsat 8 MS image was 240 × 2794, with a center location of (95.32°W, 30.10°N). The cloud coverage for the synthesized Landsat 8 MS image in simulated experiments was 28.28%. This pair of images were utilized to generate eight representative image pairs with the size of 256 × 256. For the real-data experiments, the $t2$ cloudy Landsat 8 MS image was captured by the Landsat 8 sensor on October 15, 2017, and the parameters were the same as in the simulated experiments, except that the cloud and shadow coverage was 34.51%.

1) Simulated Experiments in the Joint Processing of Thick Cloud Removal and Resolution Improvement: In this section, since there have been few studies of the joint processing of thick cloud removal and resolution improvement, we focus on comparing the three proposed methods, i.e., proposed-HSS, proposed-ST, and proposed-HSTS. A group of simulated experiment results with the size of 256 × 256 are displayed in Fig. 9 in the red–green–blue band combination, where the upper-right corner is a magnified display of the image inside the red rectangle. From the fusion results, it can be seen that all the methods can effectively remove the thick cloud and improve the spatial structure information of the synthesized $t2$ (October 29, 2017) cloudy Landsat 8 MS image in Fig. 9(a). In more detail, as shown in the $t1$ (September 29, 2016) Sentinel-2 MS image in Fig. 9(d) and the reference image in Fig. 9(h), the red rectangle points out one of the most noticeable land-cover changes between $t1$ and $t2$, where the white building next to the road does not exist at $t1$, but it does at $t2$. Unfortunately, in the result of the proposed-ST method obtained fusing the $t2$ cloudy Landsat 8 MS image and the $t1$ cloudless Sentinel-2 MS image, in Fig. 9(f), the white building is not seen, which confirms the inability of the spatiotemporal fusion strategy to predict land-cover changes. In contrast, the $t2$ (October 28, 2017) Sentinel-1 SAR image contains the information of the white building, as shown in Fig. 9(b) and (c). Thus, the proposed-HSS method fusing the $t2$ cloudy Landsat 8 MS image and the $t2$ Sentinel-1 SAR image can effectively reconstruct the white building. However, due to the difference in the imaging mechanisms, the clear road next to the white building in the zoomed area in Fig. 9(h) is unclear in the SAR image in Fig. 9(b) and (c), resulting in an invisible road in the fusion result, as shown in the yellow ellipse in Fig. 9(e). A similar situation can be seen in the white building in the upper-left corner in Fig. 9(e) and (h). The proposed-HSTS method fuses the $t2$ cloudy Landsat 8 MS image, the $t2$ Sentinel-1 SAR image, and the $t1$ cloudless Sentinel-2 MS image. It integrates the respective characteristics of the proposed-HSS and proposed-ST methods, making full use of the complementary information of the $t2$ SAR image and the $t1$ cloudless Sentinel-2 MS image, and effectively reconstructs both the road and the white building in the zoomed area in Fig. 9(g).

Table VI lists the quantitative evaluation results for the simulated experiments in the joint processing f thick cloud removal and resolution improvement, with an average of eight groups, where the best performance for each index is marked in bold. In Table VI, it can be seen that the proposed-HSS method performs the worst in all the indices, due to the difficulty of heterogeneous spatiospectral fusion. The proposed-HSTS method performs slightly better than the proposed-ST method in all the indices, expect SSIM$_G$, which shows the superiority of the heterogeneous spatiotemporal–spectral fusion strategy.

2) Real-Data Experiments in the Joint Processing of Thick Cloud Removal and Resolution Improvement: To further compare the three proposed methods, a group of real-data experiment results with the size of 256 × 256 are displayed in Fig. 10 in the red–green–blue band combination, where the

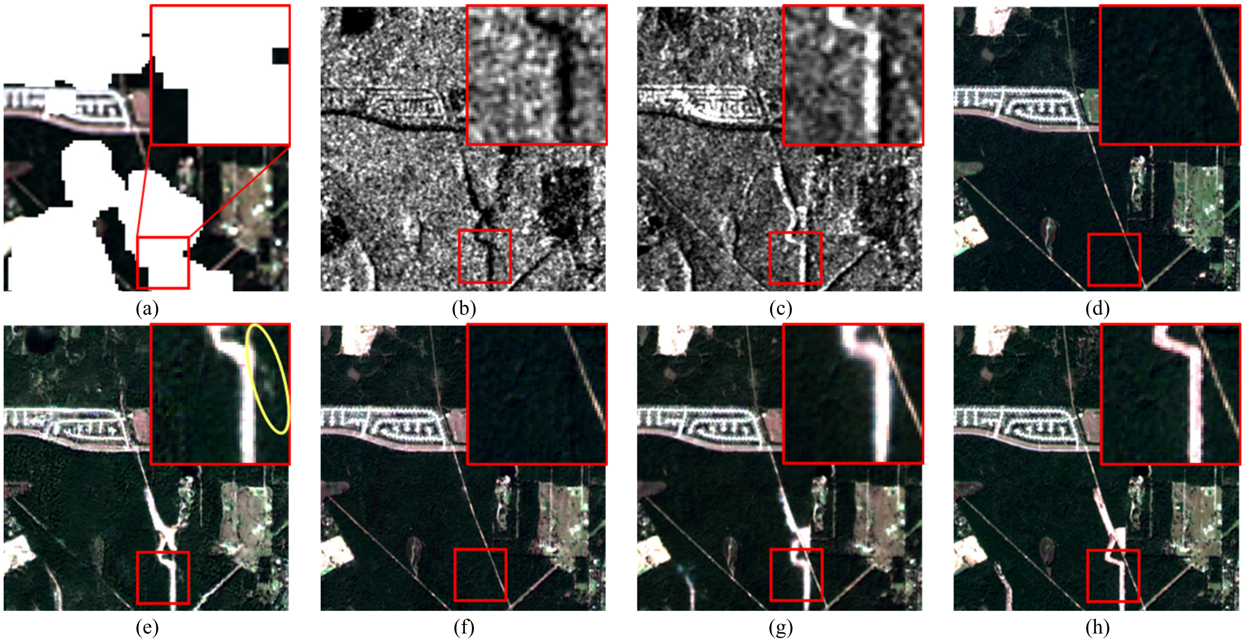


Fig. 9. Results for the simulated experiments in the joint processing of thick cloud removal and resolution improvement. (a) $t2$ cloudy Landsat 8 MS, (b) $t2$ Sentinel-1 SAR_VH, (c) $t2$ Sentinel-1 SAR_VV, (d) $t1$ Sentinel-2 MS, (e) Proposed-HSS, (f) Proposed-ST, (g) Proposed-HSTS, and (h) $t2$ Sentinel-2 MS (reference).

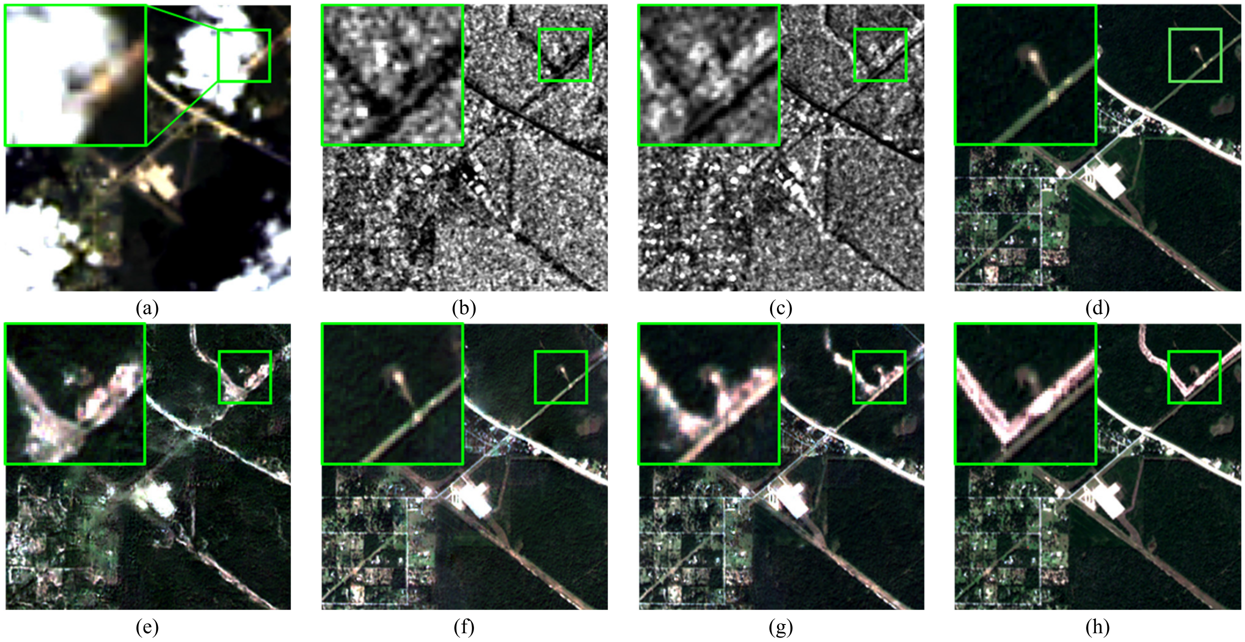


Fig. 10. Results for the real-data experiments in the joint processing of thick cloud removal and resolution improvement. (a) $t2$ Cloudy Landsat 8 MS, (b) $t2$ Sentinel-1 SAR_VH, (c) $t2$ Sentinel-1 SAR_VV, (d) $t1$ Sentinel-2 MS, (e) Proposed-HSS, (f) Proposed-ST, (g) Proposed-HSTS, and (h) $t2$ Sentinel-2 MS (reference).

upper-left corner is a magnified display of the image inside the green rectangle. In the real-data experiments, the cloudy image was not synthetic and was directly obtained from the Landsat 8 sensor on October 15, 2017, and its cloud and shadow coverage was 46.32%. The other dates were the same as in the simulated experiments. In Fig. 10, it can be seen that all the methods can remove all of the thick cloud in the cloudy Landsat 8 MS image in Fig. 10(a) and effectively enhance its spatial structure information. However, due to the difficulty of information fusion in heterogeneous spatiospectral fusion, the result of the proposed-HSS method contains severe global spatial distortion, which is obvious in the lower-left corner of Fig. 10(e). The bottleneck of spatiotemporal fusion in the proposed-ST method means that it is unable to detect the changed land covers between $t1$ and $t2$, as can be seen by comparing the zoomed areas in Fig. 10(f) and (h). The proposed-HSTS method makes full use of the complementarity of the proposed-HSS and proposed-ST methods. It not only obtains a good result with less distortion, but it also effectively reconstructs the changed land covers.

IV. Conclusion and Future Prospects

In this article, for the first time, we have proposed a deep residual cycle GAN-based heterogeneous integrated fusion framework, which can simultaneously fuse the complementary spatial, temporal, and spectral information between multisource heterogeneous observations and can achieve heterogeneous spatiospectral fusion, spatiotemporal fusion, and heterogeneous spatiotemporal–spectral fusion. From the perspective of network design, the proposed method combines a forward fusion part and a backward feedback part to simulate the imaging degradation process. The proposed network preserves the spectral and temporal consistency between the fusion result and the observed $t2$ LR MS image and the spatial consistency between the fusion result and the $t2$ HR SAR image and the $t1$ HR MS image. From the perspective of practical application, the proposed method can effectively relieve the two bottlenecks of land-cover change and thick cloud cover. Three experiments with multisensor images confirmed that when there are no land-cover changes, the proposed-ST method can obtain enough satisfactory results, but when there are land-cover changes, the utilization of the SAR image is essential, that is, when there are sufficient images, the proposed-HSTS method is always the best choice.

In the future, to further improve the structural clarity of the proposed-HSTS method in land-cover changed areas, it will be of great significance to improve the network to adaptively fuse input images. In this way, the network can extract more information from the $t2$ Sentinel-1 SAR image in the land-cover changed areas, while in other areas, the network can extract more information from the $t1$ Sentinel-2 MS image. Moreover, in the backward degradation part of the proposed network, the proposed method only explicitly utilizes the spatial degradation between the $t2$ LR MS and the $t2$ HR MS images. Therefore, in the future, it will be feasible to further explore the heterogeneous relationship model between the SAR and optical images and the temporal relationship model between MS images of different times and to embed this into the network design to strengthen the restraint of the backward part. Last but not least, it will be possible to extend the proposed framework to more satellite images, such as HS satellite images and fully polarized SAR images, and to time-series image reconstruction.

References

[1] Y. Zhang, "Understanding image fusion," *Photogram. Eng. Remote Sens*, vol. 70, no. 6, pp. 657–661, Jun. 2004.

[2] P. Sirguey, R. Mathieu, Y. Arnaud, M. M. Khan, and J. Chanussot, "Improving MODIS spatial resolution for snow mapping using wavelet fusion and ARSIS concept," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 1, pp. 78–82, Jan. 2008.

[3] T. R. Martha, N. Kerle, C. J. van Westen, V. Jetten, and K. V. Kumar, "Object-oriented analysis of multi-temporal panchromatic images for creation of historical landslide inventories," *ISPRS J. Photogramm. Remote Sens.*, vol. 67, pp. 105–119, Jan. 2012.

[4] J. He, J. Li, Q. Yuan, H. Shen, and L. Zhang, "Spectral response function-guided deep optimization-driven network for spectral super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Feb. 18, 2021, doi: 10.1109/TNNLS.2021.3056181.

[5] H. F. Shen, X. C. Meng, and L. P. Zhang, "An integrated framework for the spatio-temporal-spectral fusion of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7135–7148, Dec. 2016.

[6] H. Shen, M. Jiang, J. Li, Q. Yuan, W. Wei, and L. Zhang, "Spatial–spectral fusion by combining deep learning and variational model," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 6169–6181, Aug. 2019.

[7] X. Meng, H. Shen, H. Li, L. Zhang, and R. Fu, "Review of the pansharpening methods for remote sensing images based on the idea of meta-analysis: Practical discussion and challenges," *Inf. Fusion*, vol. 46, pp. 102–113, Mar. 2019.

[8] W. Carper, T. Lillesand, and R. Kiefer, "The use of intensity-hue-saturation transformations for merging SPOT panchromatic and multispectral image data," *Photogramm. Eng. Remote Sens.*, vol. 56, no. 4, pp. 459–467, Apr. 2004.

[9] J. Choi, K. Yu, and Y. Kim, "A new adaptive component-substitution-based satellite image fusion by using partial replacement," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 1, pp. 295–309, Jan. 2011.

[10] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "MTF-tailored multiscale fusion of high-resolution MS and Pan imagery," *Photogramm. Eng. Remote Sens.*, vol. 72, no. 5, pp. 591–596, May 2006.

[11] H. R. Shahdoosti and N. Javaheri, "Pansharpening of clustered MS and Pan images considering mixed pixels," *IEEE Trans. Geosci. Remote. Lett.*, vol. 14, no. 6, pp. 826–830, Jun. 2017.

[12] L. Zhang, H. Shen, W. Gong, and H. Zhang, "Adjustable model-based fusion method for multispectral and panchromatic images," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 6, pp. 1693–1704, Dec. 2012.

[13] C. Jiang, H. Zhang, H. Shen, and L. Zhang, "Two-step sparse coding for the Pan-sharpening of remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 5, pp. 1792–1805, May 2014.

[14] Y. Wei, Q. Yuan, H. Shen, and L. Zhang, "Boosting the accuracy of multispectral image pansharpening by learning a deep residual network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1795–1799, Oct. 2017.

[15] M. Jiang, H. Shen, J. Li, Q. Yuan, and L. Zhang, "A differential information residual convolutional neural network for pansharpening," *ISPRS J. Photogramm. Remote Sens.*, vol. 163, pp. 257–271, Jun. 2020.

[16] L. Alparone, S. Baronti, A. Garzelli, and F. Nencini, "Landsat ETM+ and SAR image fusion based on generalized intensity modulation," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 12, pp. 2832–2839, Dec. 2004.

[17] S. Chen, R. Zhang, H. Su, J. Tian, and J. Xia, "SAR and multispectral image fusion using generalized IHS transform based on àtrous wavelet and EMD decompositions," *IEEE Sensors J.*, vol. 10, no. 3, pp. 737–745, Mar. 2010.

[18] J. Wu, Q. Cheng, H. Li, S. Li, X. Guan, and H. Shen, "Spatiotemporal fusion with only two remote sensing images as input," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 6206–6219, Oct. 2020.

[19] 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, Apr. 2020, Art. no. 140301.

[20] 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.

[21] Q. Cheng, H. Liu, H. Shen, P. Wu, and L. Zhang, "A spatial and temporal nonlocal filter-based data fusion method," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4476–4488, Aug. 2017.

[22] 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.

[23] 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.

[24] A. Li, Y. Bo, Y. Zhu, P. Guo, J. Bi, and Y. He, "Blending multi-resolution satellite sea surface temperature (SST) products using Bayesian maximum entropy method," *Remote Sens. Environ.*, vol. 135, no. 4, pp. 52–63, Aug. 2013.

[25] J. Xue, Y. Leung, and T. Fung, "A Bayesian data fusion approach to spatio-temporal fusion of remotely sensed images," *Remote Sens.*, vol. 9, no. 12, p. 1310, Dec. 2017.

[26] 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.

[27] Y. Li, J. Li, L. He, J. Chen, and A. Plaza, “A new sensor bias-driven spatio-temporal fusion model based on convolutional neural networks,” *Sci. China Inf. Sci.*, vol. 63, no. 4, Mar. 2020, Art. no. 140302.

[28] H. Shen, “Integrated fusion method for multiple temporal-spatial–spectral images,” in *Proc. 22nd Congr. ISPRS*, 2012, pp. 407–410.

[29] B. Huang, H. Zhang, H. Song, J. Wang, and C. Song, “Unified fusion of remote-sensing imagery: Generating simultaneously high-resolution synthetic spatial–temporal–spectral Earth observations,” *Remote Sens. Lett.*, vol. 4, no. 6, pp. 561–569, Jun. 2013.

[30] 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.

[31] D. Fasbender, V. Obsomer, J. Radoux, P. Bogaert, and P. Defourny, “Bayesian data fusion: Spatial and temporal applications,” in *Proc. Int. Workshop Anal. Multi-Temporal Remote Sens. Images*, Jul. 2007, pp. 1–6.

[32] I. Goodfellow *et al.*, “Generative adversarial nets,” in *Proc. NIPS*, 2014, pp. 2672–2680.

[33] J. Zhu, T. Park, P. Isola, and A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2223–2232.

[34] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, “Striving for simplicity: The all convolutional net,” in *Proc. Workshop Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[35] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[36] D. Kiku, Y. Monno, M. Tanaka, and M. Okutomi, “Residual interpolation for color image demosaicking,” in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2013, pp. 2304–2308.

[37] R. Timofte, V. De Smet, and L. Van Gool, “A+: Adjusted anchored neighborhood regression for fast super-resolution,” in *Proc. IEEE Asian Conf. Comput. Vis.*, Nov. 2014, pp. 111–126.

[38] J. Gao, Q. Yuan, J. Li, H. Zhang, and X. Su, “Cloud removal with fusion of high resolution optical and SAR images using generative adversarial networks,” *Remote Sens.*, vol. 12, no. 1, p. 191, Jan. 2020.

[39] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[40] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–16.

[41] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[42] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, “FusionGAN: A generative adversarial network for infrared and visible image fusion,” *Inf. Fusion*, vol. 48, pp. 11–26, Aug. 2019.

[43] J. Qi, J. Du, S. M. Siniscalchi, X. Ma, and C.-H. Lee, “On mean absolute error for deep neural network based vector-to-vector regression,” *IEEE Signal Process. Lett.*, vol. 27, pp. 1485–1489, 2020.

[44] J. Pan *et al.*, “Physics-based generative adversarial models for image restoration and beyond,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 7, pp. 2449–2462, Jul. 2021.

[45] D. P. Kingma and J. L. Ba, “Adam: A method for stochastic optimization,” in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–41.

[46] G. Vivone *et al.*, “A critical comparison among pansharpening algorithms,” *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2565–2586, May 2015.

[47] 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.

[48] S. Rahmani, M. Strait, D. Merkurjev, M. Moeller, and T. Wittman, “An adaptive IHS Pan-sharpening method,” *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 4, pp. 746–750, Oct. 2010.

[49] Z. Shao, J. Cai, P. Fu, L. Hu, and T. Liu, “Deep learning-based fusion of Landsat-8 and Sentinel-2 images for a harmonized surface reflectance product,” *Remote Sens. Environ.*, vol. 235, Dec. 2019, Art. no. 111425.

[50] J. Inglada *et al.*, “Assessment of an operational system for crop type map production using high temporal and spatial resolution satellite optical imagery,” *Remote Sens.*, vol. 7, no. 9, pp. 12356–12379, Sep. 2015.

[51] G. Scarpa, M. Gargiulo, A. Mazza, and R. Gaetano, “A CNN-based fusion method for feature extraction from sentinel data,” *Remote Sens.*, vol. 10, no. 2, p. 236, Feb. 2018.

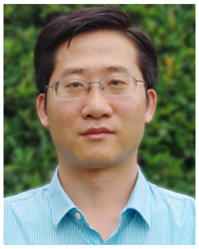
[52] C. Grohnfeldi, M. Schmitt, and X. Zhu, “A conditional generative adversarial network to fuse SAR and multispectral optical data for cloud removal from Sentinel-2 images,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 1726–1729.

[53] H. Shen, J. Wu, Q. Cheng, M. Aihemaiti, C. Zhang, and Z. Li, “A spatiotemporal fusion based cloud removal method for remote sensing images with land cover changes,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 3, pp. 862–874, Mar. 2019.

[54] Q. Zhang, Q. Yuan, C. Zeng, X. Li, and Y. Wei, “Missing data reconstruction in remote sensing image with a unified spatial-temporal-spectral deep convolutional neural network,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4274–4288, Aug. 2018.

Menghui Jiang (Graduate Student Member, IEEE) received the B.S. degree in geographical science and the Ph.D. degree in cartography and geographic information engineering from Wuhan University, Wuhan, China, in 2017 and 2022, respectively.

Her research interests include image data fusion, quality improvement, remote sensing image processing, and deep learning.



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

He is currently a Distinguished Professor with Wuhan University, where he serves as the Dean with the School of Resource and Environmental Sciences. He was or is the Principal Investigator (PI) of two projects supported by the National Key Research and Development Program of China and six projects supported by the National Natural Science Foundation of China. He has authored or coauthored more than 150 peer-reviewed international journal articles, where over 60 appeared in IEEE journals, and published four books as a Chief Editor. His research interests include remote-sensing image processing, multisource data fusion, and intelligent environmental sensing.

Dr. Shen is a fellow of the Institution of Engineering and Technology (IET), an Education Committee Member of the Chinese Society for Geodesy Photogrammetry and Cartography, and a Theory Committee Member of the Chinese Society for Geospatial Information Society. He was a recipient of the First Prize in Natural Science Award of Hubei Province in 2011, the First Prize in Nature Scientific Award of China’s Ministry of Education in 2015, and the First Prize in Scientific and Technological Progress Award of Chinese Society for Geodesy Photogrammetry and Cartography in 2017. He is also a Senior Regional Editor of the *Journal of Applied Remote Sensing* and an Associate Editor of *Geography and Geo-Information Science* and *Journal of Remote Sensing*.

Jie Li (Member, IEEE) received the B.S. degree in sciences and techniques of remote sensing and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2011 and 2016.

He is currently an Associate Professor with the School of Geodesy and Geomatics, Wuhan University. His research interests include image quality improvement, image superresolution reconstruction, data fusion, remote sensing image processing, sparse representation, and deep learning.