Remote Sensing Image Spatiotemporal Fusion Using a Generative Adversarial Network

Hongyan Zhang, *Senior Member, IEEE*, Yiyao Song, *Student Member, IEEE*, Chang Han, *Member, IEEE*, and Liangpei Zhang, *Fellow, IEEE*

***Abstract*— Due to technological limitations and budget constraints, spatiotemporal fusion is considered a promising way to deal with the tradeoff between the temporal and spatial resolutions of remote sensing images. Furthermore, the generative adversarial network (GAN) has shown its capability in a variety of applications. This article presents a remote sensing image spatiotemporal fusion method using a GAN (STFGAN), which adopts a two-stage framework with an end-to-end image fusion GAN (IFGAN) for each stage. The IFGAN contains a generator and a discriminator in competition with each other under the guidance of the optimization function. Considering the huge spatial resolution gap between the high-spatial, low-temporal (HSLT) resolution Landsat imagery and the corresponding low-spatial, high-temporal (LSHT) resolution MODIS imagery, a feature-level fusion strategy is adopted. Specifically, for the generator, we first super-resolve the MODIS images while also extracting the high-frequency features of the Landsat images. Finally, we integrate the features from the MODIS and Landsat images. STFGAN is able to learn an end-to-end mapping between the Landsat–MODIS image pairs and predicts the Landsat-like image for a prediction date by considering all the bands. STFGAN significantly improves the accuracy of phenological change and land-cover-type change prediction with the help of residual blocks and two prior Landsat–MODIS image pairs. To examine the performance of the proposed STFGAN method, experiments were conducted on three representative Landsat–MODIS data sets. The results clearly illustrate the effectiveness of the proposed method.**

***Index Terms*— Generative adversarial network (GAN), multisource satellite data, remote sensing, spatiotemporal fusion.**

I. Introduction

IN AREAS such as dynamic monitoring, change detection, and land-cover classification, high spatial resolution remote sensing images with a dense time series are needed to capture detailed land surface dynamics [1]–[10]. However, due to technological limitations and budget constraints, there is often a tradeoff between the temporal and spatial resolutions of remote sensing images [11]–[14]. In recent years, although great breakthroughs have been made in Earth observation through the availability of remote sensing images with high spatial and temporal resolutions from multiplatform satellites, such as Sentinel-2 and the China High-resolution Earth Observation System (CHEOS), the current availability of remote sensing data is still insufficient in practical applications because of the cloud cover and other disturbances [15], [16]. The insufficient remote sensing data cannot meet the requirements of long-term and detailed land surface dynamics studies, which require dense historical time-series remote sensing images with a high spatial resolution [17]–[20]. Spatiotemporal fusion is a feasible and cost-effective way to promote the applications of the current Earth observation data. Spatiotemporal fusion aims to integrate multisource satellite images to obtain images with both high spatial and high temporal resolutions. For example, MODIS data are characterized by a high temporal resolution and low spatial resolution (LSHT), whereas Landsat Enhanced Thematic Mapper Plus (ETM+) data are characterized by a high spatial resolution and low temporal resolution (HSLT) [21], [22]. Based on one or two Landsat-MODIS image pairs on prior dates and one MODIS image on the prediction date, spatiotemporal fusion models can combine the spatial resolution of Landsat imagery with the temporal frequency of MODIS imagery to generate a Landsat-like image on the prediction date.

In recent years, many spatiotemporal fusion methods have been proposed in an attempt to aggregate remote sensing data from various sensors at different spatial and temporal resolutions. Generally speaking, the current spatiotemporal fusion methods can be classified into three categories: 1) weight function-based methods; 2) unmixing-based methods; and 3) learning-based methods [23], [24]. In the weight function-based methods, the fine pixel values are estimated through combining the information of all the input images by weight functions [24]. Among the weight function-based algorithms, the most representative examples are the spatial and temporal adaptive reflectance fusion model (STARFM) [25] and its enhanced version (ESTARFM) [26]. The classic STARFM builds a simple approximate relationship between the HSLT and LSHT pixels and searches similar neighboring pixels according to the spectral difference, the temporal

Manuscript received January 21, 2020; revised May 27, 2020 and June 24, 2020; accepted July 14, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61871298, in part by the Dragon 5 proposal ID under Grant 58817, in part by the Foundation of Hubei Educational Committee under Grant B2018280, and in part by the Applied Basic Research Programs of Science and Technology Commission Foundation of Wuhan under Grant 2019010701011390. *(Corresponding author: Hongyan Zhang.)*

Hongyan Zhang, Yiyao Song, and Liangpei Zhang are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430072, China (e-mail: zhanghongyan@whu.edu.cn).

Chang Han is with the School of Mechanical and Electrical Engineering, Wuhan Business University, Wuhan 430056, China.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2020.3010530

difference, and the location distance. Considering the existence of complex heterogeneous areas, Zhu *et al.* [26] proposed ESTARFM by assigning different conversion coefficients for homogeneous and heterogeneous areas to modify the weights of the neighboring pixels. However, both algorithms are built under the assumption that the proportion of each land-cover type does not change during the observation period, which does not consider the human activities on the Earth's surface, such as disturbance events (e.g., forest fires) and changes in urban land use. To deal with this problem, Hilker *et al.* [27] proposed a spatial and temporal adaptive algorithm for mapping reflectance change (STAARCH) for sudden disturbance event mapping. Furthermore, Zhao *et al.* [28] developed a robust adaptive spatial and temporal image fusion model (RASTFM) for complex land surface changes. The disadvantage of the weight function-based methods is that adopting neighboring pixels may introduce blur into the predicted image, which, in turn, incurs the loss of high-frequency details.

Based on the linear spectral mixing theory, the unmixing-based methods unmix the coarse pixels to estimate the value of the fine pixels. Zhukov *et al.* [29] first proposed the multisensor multiresolution technique (MMT) to integrate remote sensing images with different spatial resolutions and acquired at different times. However, MMT was confronted with two problems: 1) the large errors caused by spectral unmixing and 2) the lack of endmember spectral variability. In order to address these issues, Zurita-Milla *et al.* [30] proposed the spatial–temporal data fusion approach (STDFA), which obtains the prediction by considering the reflectance change estimated through unmixing the endmember reflectance in a moving window. STDFA has also been enhanced by the use of an adaptive moving window size [31]. However, the unmixing-based methods still have difficulty in land-cover-type change prediction, due to the lack of high spatial resolution land-use databases, which limits their application.

With the development of machine learning, learning-based methods have been proposed in recent years. Huang and Song [32] proposed the sparse representation-based spatiotemporal reflectance fusion model (SPSTFM), which was the first method to bring dictionary-pair learning techniques from natural image super-resolution to spatiotemporal data fusion. Following SPSTFM, to deal with the single prior Landsat-MODIS image pair case, Song and Huang [33] developed another sparse representation-based spatiotemporal satellite image fusion (SSIF) model through one-pair image learning. The sparse representation-based methods aim to extract the mapping relationships between the HSLT Landsat and LSHT MODIS images via learning a dictionary pair. They then predict the fusion image by weighting predictions from the two end dates of the observation period. However, the sparse representation-based methods need to relearn the dictionary for the images of different research areas, which is inefficient. Compared with dictionary learning, deep learning has a better generalization ability over diverse remote sensing scenes. Song *et al.* [34] proposed a spatiotemporal image fusion method using a deep convolutional neural network (STFDCNN). The convolutional neural network (CNN) is adopted to model the relationship between the coarse-resolution (CR) image and fine-resolution (FR) image, and a high-pass fusion model is used for the prediction. Liu *et al.* [35] improved STFDCNN by taking the temporal dependence and temporal consistency into consideration and proposed a two-stream CNN for spatiotemporal image fusion (StfNet). However, there are still several limitations to these two CNN-based methods. First, STFDCNN and StfNet are not end-to-end learning models. The prediction stage is divided into two parts—CNN-based mapping and reconstruction—which increases the complexity of the algorithms. Second, each band needs to be trained separately, which increases the amount of parameters, memory usage, and training time.

In this article, to address the abovementioned problems, we propose a novel spatiotemporal fusion model using a generative adversarial network (STFGAN). As an emerging deep neural network, the generative adversarial network (GAN) [36] shows great potential for exploiting the high-level information and has achieved superior performances in image style transfer, super-resolution, and cloud removal [37]–[40]. In the proposed method, a two-stage framework is developed to improve the accuracy of the fusion results, in which each stage contains an end-to-end image fusion GAN (IFGAN). The generator and discriminator are optimized in an alternating manner to make the generator work as efficiently as possible. Specifically, the generator network consists of three parts: 1) the super-resolution of the MODIS images; 2) the high-frequency feature extraction of the Landsat images; and 3) fusion of the MODIS and Landsat feature maps. The first two parts are achieved via residual blocks.

Compared with the previous learning-based fusion methods, the proposed STFGAN method has the following advantages.

1) To the best of our knowledge, this is the first end-to-end trainable network based on deep learning that can be used to solve the spatiotemporal fusion problem.
2) To generate better spatiotemporal fusion results from the generator, we have developed a residual-blocks architecture for both the Landsat input and MODIS input in the generator network. This approach can capture more textural details and can significantly improve the accuracy of the phenological change and land-cover-type change prediction, with the help of the two prior Landsat-MODIS image pairs.
3) All bands of the remote sensing images are input into the network together, rather than each band separately, which reduces the time and space consumption, especially for large-scale images.

The rest of this article is organized as follows. In Section II, we introduce the proposed STFGAN method in detail. The experimental results and analysis are provided in Section III. Section IV concludes this article.

II. Methodology

A. From GAN to Super-Resolution (SRGAN) to STFGAN

In the natural image processing field, the task of recovering an original high-resolution (HR) image from its corresponding

low-resolution (LR) counterpart is called single-image super-resolution. The development of super-resolution provided motivation for the development of spatiotemporal fusion. For example, Yang *et al.* [41] first proposed a super-resolution algorithm based on sparse representation. Inspired by this, Huang and Song [32], [33] applied the sparse representation to spatiotemporal fusion and proposed the SPSTFM and SSIF algorithms. Subsequently, the application of CNNs in the field of image super-resolution prompted the development of the STFDCNN and StfNet algorithms.

In 2018, Ledig *et al.* [38] first applied an SRGAN and won first place in the CVPR-NTIRE 2018 Image Super-Resolution Championship. SRGAN defines a discriminator network $D_{\theta D}$, which is optimized in an alternating manner along with the generator $G_{\theta G}$ to solve the adversarial min–max problem

$$\min_{\theta_G}\max_{\theta_D}\ \mathbb{E}_{I^{\mathrm{HR}}\sim p_{\mathrm{train}}(I^{\mathrm{HR}})}[\log D_{\theta_D}(I^{\mathrm{HR}})] + \mathbb{E}_{I^{\mathrm{LR}}\sim p_G(I^{\mathrm{LR}})}[\log(1-D_{\theta_D}(G_{\theta_G}(I^{\mathrm{LR}})))] \quad (1)$$

where I^{HR} and I^{LR} represent the reference HR image and its corresponding LR counterpart, respectively. $D(\cdot)$ represents the output probability of the input $\cdot$ coming from the reference data p_{train} rather than the generator p_G. $\mathbb{E}$ represents the mathematical expectation. θ_* denotes the weights and biases of the deep network and is obtained by optimizing the loss function. In the process of solving the objective function, the training generator G generates a super-resolved image to fool the discriminator D, and the training discriminator D distinguishes the super-resolved image from the real image. In this way, the generator can learn to generate super-resolved images that are highly similar to the real images, even when the discriminator has difficulty in distinguishing the real and fake images.

Based on SRGAN, we aim to employ the GAN to complete the task of spatiotemporal fusion. We adapted our generator and discriminator architectures from those in SRGAN. Clearly, there are some differences between super-resolution and spatiotemporal fusion.

1) *Resolution Difference:* Generally speaking, the magnification factor in super-resolution ranges from 2 to 4 and up to 8. However, in spatiotemporal fusion, there can be a huge spatial resolution gap between the two data resources, usually ranging from 8 to 16 (e.g., Landsat with a 250-/500-m spatial resolution and MODIS with a 30-m spatial resolution). In this case, directly applying the super-resolution method to spatiotemporal fusion will result in less accurate results.
2) *Spatial Difference:* It is known that remote sensing images contain more information than natural images. This is because remote sensing images contain more geographic information. The complex feature types and rich textural features increase the difficulty of spatiotemporal fusion.
3) *Temporal Difference:* In single-image super-resolution, there is only one LR image used as input. In spatiotemporal fusion, in contrast, one or two pairs of LR-HR image pairs on prior dates are available. The spatial details of the image pairs, as a supplement information source, can be fully used to improve the spatial resolution of the LR images on the prediction date.
4) *Spectral Difference:* Remote sensing images have multiple bands, unlike natural images containing only three bands of red, green, and blue.

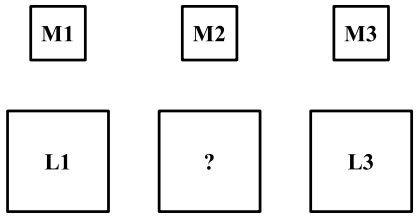

Fig. 1. Target of STFGAN. Five remote sensing images are used to predict the unknown Landsat image at time 2.

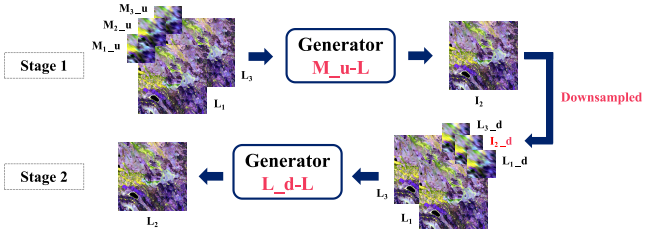

Fig. 2. Flowchart of the two-stage framework. [·]_d and [·]_u indicate the downsampled version (downsampled by four times) or the upsampled version (upsampled by four times) of [·].

We selected Landsat and MODIS images as the HSLT and LSHT images, respectively, to demonstrate the effectiveness of STFGAN. Based on the temporal difference, we considered that there were two pairs of Landsat-MODIS images available, one at time 1 (L_1-M_1) and the other at time 3 (L_3-M_3). Together with the MODIS image at time 2 (M_2), we used these five images (L_1, L_3, M_1, M_2, M_3) to predict the unknown Landsat image at time 2 (L_2), as shown in Fig. 1. Considering the resolution difference and spatial difference, STFGAN makes full use of the prior image pairs to compensate for the great resolution difference between the HR and LR images and provides more textural and structural information for the L_2 prediction. With regard to the spectral difference, and differing from the existing learning-based methods [34], [35], the proposed STFGAN method uses an NIR–red–green composite of the remote sensing image for the network input rather than learning the mapping of each band separately, which reduces the training time and the amount of parameters.

Notably, the resolution difference between the MODIS image and the Landsat image is not an integer multiple and can be considered to be approximately 16 times. Therefore, we first upsample the MODIS image through bicubic interpolation to make the image size difference an integer multiple of 16. One important fact is that the interpolation does not change the resolution of the image but only the pixel size [42]. Considering that the SRGAN model has only a limited ability for super-resolution, we adopt a two-stage framework for the proposed STFGAN method, as shown in Fig. 2, in which each stage is instantiated with an IFGAN with the resolution

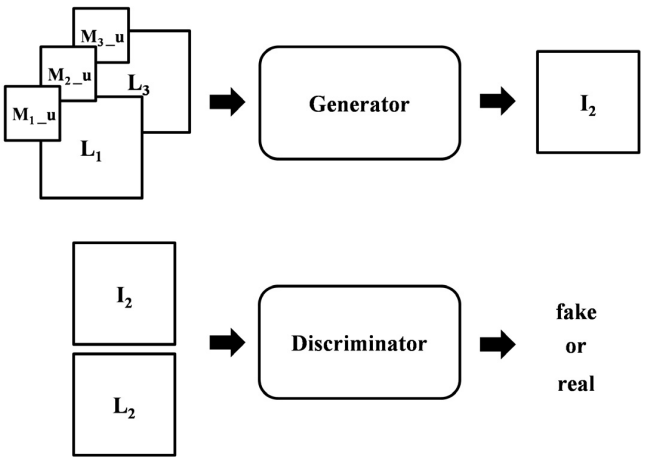

Fig. 3. Overall flowchart of IFGAN. The top and the bottom parts indicate the flowchart of the generator and discriminator, respectively.

enhancement factor of 4. For the first stage, all the MODIS images at times 1–3 are upsampled by four times (M_1_u, M_2_u, and M_3_u) and utilized as the input of the generator G_{M_u-L} with the Landsat images at times 1 and 3, to predict the intermediate image I_2. At this moment, although the image size of the output I_2 is the same as the target image L_2, the resolution of the image obtained by this generator is only improved by four times theoretically. Compared with the target image L_2, the predicted image I_2 may not be optimal. Therefore, the downsampled versions of L_1, I_2, and L_3 by four times (L_1_d, I_2_d, and L_3_d) and L_1 and L_3 are added to the input of the generator G_{L_d-L} of the second stage, to derive the final L_2. It should be noted that in the training phase, the generators of the two stages are not identical. One is trained with upsampled MODIS images (M_u) and Landsat images (L), and the other is trained with downsampled Landsat images (L_d) and Landsat images (L).

B. Network Architecture of IFGAN

In the two-stage framework, each stage contains the same end-to-end IFGAN to generate the spatiotemporal fusion images. Taking the first stage as an example, the overall flowchart of IFGAN is demonstrated in Fig. 3. In the training, the generator G generates a spatiotemporal fusion image with the input of two pairs of prior image pairs and one MODIS image on the prediction date to fool the discriminator D, and the discriminator D distinguishes the spatiotemporal fusion image from the real image L_2. In testing, we can use the trained generator G to derive the predicted I_2.

1) Generator: Considering the spatial difference between super-resolution and spatiotemporal fusion, it is clear that we cannot recover L_2 with high quality only from M_2. Due to the huge resolution gap between Landsat images and MODIS images, there is little complex geographic information contained in the Landsat image remaining in the corresponding MODIS image, which is a barrier to restoring structural and textural details. Due to the temporal dependence of remote sensing images, the L_1–M_1 and L_3–M_3 image pairs are available, which allows the generator to obtain the supplementary information from the prior image pairs to help recover L_2 from M_2. The generator uses an NIR–red–green composite of the remote sensing images as input and output. As shown in Fig. 4(a), the generator network can be divided into three parts: 1) super-resolution of the MODIS images; 2) high-frequency feature extraction of the Landsat images; and 3) fusion of the MODIS and Landsat feature maps.

In the first part, we first upsample M_1–M_3 and concatenate them. In this way, M_1_u and M_3_u provide additional content information for M_2_u. In the existing learning-based spatiotemporal fusion methods [34], [35], a three-layer CNN is used to extract the features from the MODIS images. However, the network architecture of the three-layer CNN is not deep enough to extract high-level abstract features [43]. Thus, we adopt a deeper network to extract the features from the concatenated MODIS images. In theory, the training accuracy of the neural network output increases as the number of layers increases. However, it turns out that for some problems such as vanishing gradients and exploding gradients, a sufficiently deep network can, nevertheless, be difficult to train under the guidance of the optimization function. A residual network is adopted to solve this problem by adding the input of the previous layer to the next layer, which indicates that more information is transmitted to the next layer, and the richness of the information improves the effect of the network training. The deeper the residual network, the better the effect on the training set. Therefore, the core of the first part is the 16 residual blocks with the same layout. Each residual block contains two convolutional layers with a small filter kernel size of 3×3 followed by batch-normalization layers and a rectified linear unit (ReLU) as the activation function. The 16 residual blocks fully learn the features of the MODIS images, which contain the phenological change information and land-cover type change information, laying down the foundation for the next step of super-resolution. Two trained subpixel convolutional layers (PixelShuffler $\times$ 2) are used to increase the spatial resolution of the MODIS images by four times.

With respect to the second part, L_1 and L_3 are first concatenated. In super-resolution, there are no extra HR natural images to guide the recovery of the LR natural images. In contrast, in spatiotemporal fusion, although the resolution of the MODIS images is much lower than that of the Landsat images, there are some Landsat images available during the same period, which can guide the recovery of the MODIS images. During a short time period, it can be assumed that the land cover of the study site has not been changed significantly. In other words, the structural and textural information contained in L_1 and L_3 is similar to that in L_2. Based on this, the structural and textural information of L_1 and L_3 can be considered as a complement to M_2_u that provides structural and textural details and aids the recovery of L_2 from M_2_u. We employ eight residual blocks to extract the high-frequency features of the concatenated Landsat images, such as detailed textures and structures, which contributes to the restoration of the field edges. The output size of this subnetwork is the same as the output size of the first part.

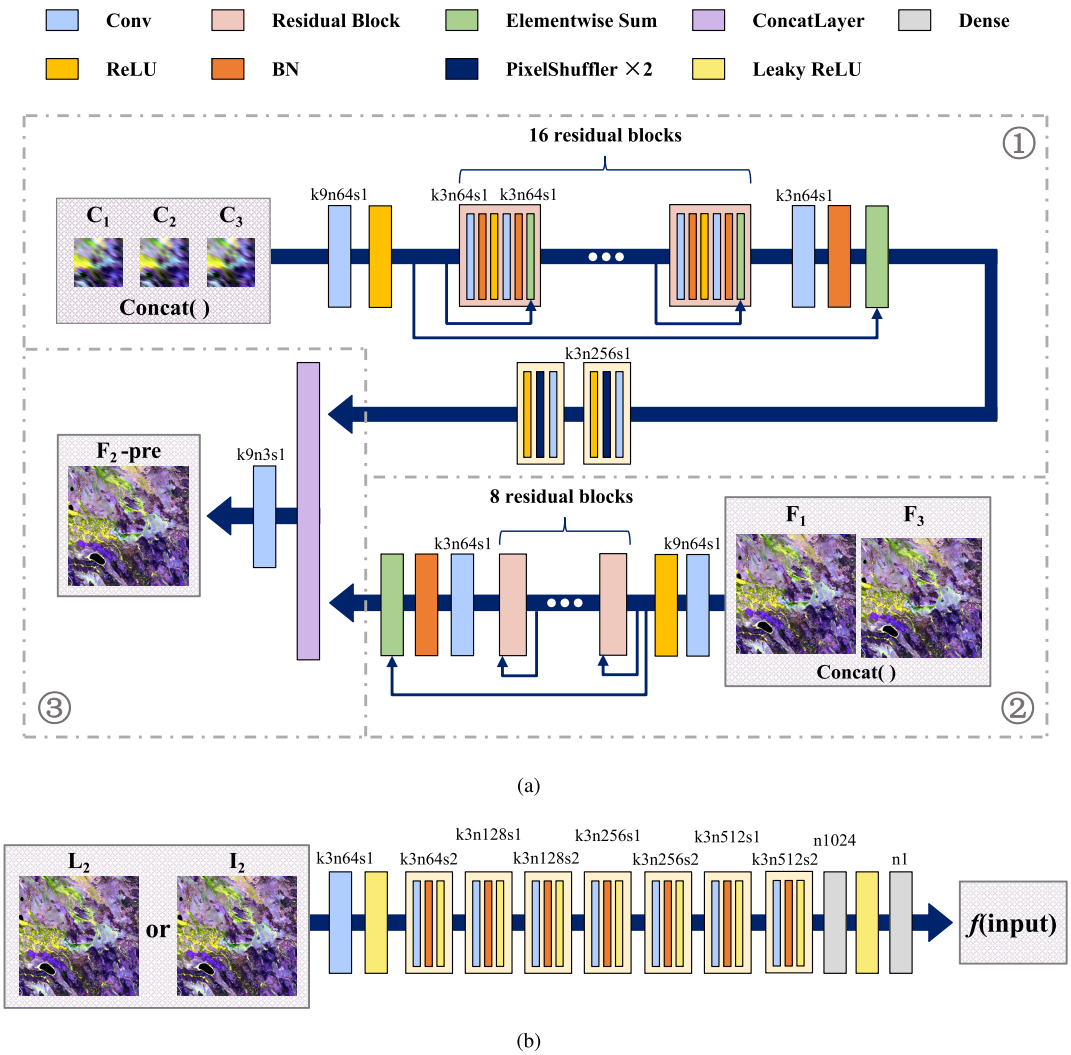

Fig. 4. Network architecture of IFGAN with the corresponding kernel size (k), number of feature maps (n), and stride (s) indicated for each convolutional layer. (a) and (b) Network architecture of the generator and discriminator, respectively.

After these two parts, the feature maps extracted from the FR images and CR images are concatenated by a ConcatLayer. The low-frequency information is obtained from the MODIS images, whereas the high-frequency information is obtained from the Landsat images. By combining these two sources of information, L_2 can be recovered with high-quality content and structural details. Finally, a convolutional layer is used to reduce the output tensor dimension to restore the predicted I_2.

2) Discriminator: The discriminator is used to distinguish the reference images (L_2) and the spatiotemporal fusion images (I_2). As shown in Fig. 4(b), it contains eight convolutional layers with an increasing number of 3×3 filter kernels, increasing by a factor of 2, from 64 to 512 kernels, as in VGGNet [44]. The eight strided convolutional layers fully extract the features of the input and improve the accuracy of the discriminator's classification, which helps the generator to produce more realistic images. The 512 resulting feature maps are followed by two dense layers, and then, the probability of the sample classification is output. Compared with a CNN, the GAN enables the fusion results to include more high-frequency information and finer details, whereas the traditional CNN can only improve the fusion results by selecting the appropriate objective function.

In order to indicate the training process by adversarial loss and obtain the optimal generator, we use the Wasserstein distance to describe the difference between the distributions of the two data sets (the data set of reference images and the data set of spatiotemporal fusion images) [45]. Differing from the discriminator architecture of SRGAN, we remove the sigmoid activation function from the last layer and output the probability of sample classification directly for the Wasserstein distance computation. In addition to this, the loss function and the optimizer are also changed accordingly, which is illustrated in Section II-C.

C. Loss Function of IFGAN

The loss function of IFGAN is formulated as the weighted sum of a content loss and an adversarial loss component. The content loss is formulated as the weighted sum of a mean

square error (mse) loss $\mathcal{L}_{\text{MSE}}^{\text{STF}}$ and a VGG loss $\mathcal{L}_{\text{VGG}/i,j}^{\text{STF}}$. MSE loss $\mathcal{L}_{\text{MSE}}^{\text{STF}}$ is the most widely used optimization target for image restoration, defined as the Euclidean distance between the reconstructed image $G_{\theta_G}(M_1, M_2, M_3, L_1, L_3)$ and the reference image L_2. It is calculated as shown in (2), where W and H describe the dimensions of the reference image and the reconstructed image. The mse loss provides solutions with lower root-mse (RMSE) values that are, however, perceptually rather smooth and less convincing

$$\mathcal{L}_{\text{MSE}}^{\text{STF}} = \frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H}((L_2)_{x,y} - G_{\theta_G}(M_1\_u, M_2\_u, M_3\_u, L_1, L_3)_{x,y})^2 \quad (2)$$

In contrast to mse loss, VGG loss $\mathcal{L}_{\text{VGG}/i,j}^{\text{STF}}$ is closer to perceptual similarity. In other words, VGG loss provides the solutions with higher structural similarity (SSIM) values. The VGG loss is defined as the Euclidean distance between the feature representations of the reconstructed image $G_{\theta_G}(M_1, M_2, M_3, L_1, L_3)$ and the reference image L_2

$$\mathcal{L}_{\text{VGG}/i,j}^{\text{STF}} = \frac{1}{W_{i,j}H_{i,j}}\sum_{x=1}^{W_{i,j}}\sum_{y=1}^{H_{i,j}}(\phi_{i,j}(L_2)_{x,y} - \phi_{i,j}(G_{\theta_G}(M_1\_u, M_2\_u, M_3\_u, L_1, L_3))_{x,y})^2 \quad (3)$$

where $\phi_{i,j}$ indicates the feature map obtained by the jth convolution (after activation) before the ith max-pooling layer within the VGG19 network. $W_{i,j}$ and $H_{i,j}$ describe the dimensions of the corresponding feature map $\phi_{i,j}$ within the VGG network. The content loss combines the advantages of the two loss functions, with which we can recover the high- and low-frequency information of the image at the same time and generate the spatiotemporal fusion image closest to the real image.

Adversarial loss is defined as shown in (4), where N indicates the number of training samples. As mentioned in Section II-B2, in order to indicate the training process by discriminator loss, we use the output probabilities of the discriminator over all the training samples for the Wasserstein distance computing without log function. This encourages the generator network to learn to create solutions that are highly similar to the real image, by trying to fool the discriminator. At the same time, this encourages the discriminator network to optimize itself, to improve the ability to distinguish between spatiotemporal fusion images and real images

$$\mathcal{L}_{\text{Adv}}^{\text{STF}} = \frac{1}{N}\sum_{n=1}^{N} D_{\theta_D}(L_2) - \frac{1}{N}\sum_{n=1}^{N} D_{\theta_D}(G_{\theta_G}(M_1\_u, M_2\_u, M_3\_u, L_1, L_3)). \quad (4)$$

The generator G and discriminator D are trained to solve the min–max optimization problem

$$\min_G \max_D \mathcal{L}^{\text{STF}} = \underbrace{\mathcal{L}_{\text{MSE}}^{\text{STF}} + \alpha\, \mathcal{L}_{\text{VGG}/i,j}^{\text{STF}}}_{\text{content loss}} + \underbrace{\beta\, \mathcal{L}_{\text{Adv}}^{\text{STF}}}_{\text{adversarial loss}} \quad (5)$$

where G tries to minimize this objective against an adversary D that tries to maximize it. α and β are constants and are empirically set as 0.2×10^{-6} and 10^{-3}, respectively, following the success of the work by Ledig *et al.* [38]. This objective function is optimized by adopting RMSProp [45] with standard backpropagation. We empirically set the learning rate to 10^{-4} and decay rate to 0.1.

III. Experiments

In this section, we first introduce the data sets used in the experiments and the evaluation indices used in the analysis. The experimental results are then presented. We then analyze the results and provide a discussion on the effectiveness and limitations of the proposed method.

A. Study Sites and Data Sets

The three data sets considered in this study were named the Coleambally Irrigation Area (CIA), Lower Gwydir Catchment (LGC), and Shenzhen data sets. The CIA is located in southern New South Wales, Australia, covering an area of 2193 km^2. The data set was made up of 17 cloud-free Landsat-MODIS pairs from October 2001 to May 2002. All the Landsat images were acquired by the Landsat-7 ETM+ sensor and were atmospherically corrected by using MODTRAN4 [46]. The MODIS images were the MODIS Terra MOD09GA Collection 5 data product [47]. The CIA data set includes abundant phenological changes but fewer land-cover-type changes.

The LGC is located in northern New South Wales, Australia, covering an area of 5440 km^2. The data set was made up of 14 cloud-free Landsat-MODIS pairs from April 2004 to April 2005. All the Landsat images were acquired by the Landsat-5 TM sensor and were atmospherically corrected using the algorithm proposed in [48]. The MODIS images were again the MODIS Terra MOD09GA Collection 5 data [47]. The LGC data set can be considered as having significant land-cover-type changes, with regular shape changes due to the occurrence of a large flood.

Shenzhen is located in Guangdong province, China, covering an area of the size of 225 km^2. The Shenzhen data set was made up of three Landsat-MODIS pairs acquired in the same month but in different years, on November 1, 2000, November 7, 2002, and November 8, 2004, respectively. The Landsat images were also acquired by the Landsat-7 ETM+ and were radiometrically and atmospherically corrected using the MODIS 6S approach [49]. The MODIS images were obtained directly from the Land Processes Distributed Active Archive Center (https://lpdaac.usgs.gov/lpdaac). The Shenzhen data set is unique due to its huge time gap between the three image pairs and the intense land-cover-type changes with irregular shape changes over the four years, which increases the difficulty of the prediction.

For the CIA data set, there are six bands with a Landsat image size of 1720×2040; for the LGC data set, there are six bands with a Landsat image size of 3200×2720; and for the Shenzhen data set, there are three bands with a Landsat image size of 500×500. For the three data sets, all the MODIS images were upsampled to a quarter of the image size of

the Landsat image using bicubic interpolation. Differing from STFDCNN, which learns the mapping of each band separately, STFGAN uses an NIR–red–green composite of the remote sensing images as input and output. Thus, bands 4, 3, and 2 (NIR, red, and green bands) were selected for the Landsat images, and the corresponding bands 2, 1, and 4 (NIR, red, and green bands) were selected for the MODIS images. Differing from other deep learning-based methods, STFGAN inputs all three bands into the generator together, rather than training each band separately. In order to remove invalid pixels and facilitate segmentation, the images in the CIA data set were cropped to the size of 1280×1792, the images in the LGC data set were cropped to the size of 3072×2560, and there was no processing for the Shenzhen data set because it was only for testing.

We used an image group as a set of inputs and a reference image. Each image group was made up of three image pairs in a continuous time sequence, as shown in Table I. Each row denotes one image group, and the middle one is the prediction date. For training purposes, we selected nine image groups in the CIA data set and six image groups in the LGC data set. The size of the training subimages was set as 256×256 for all the image groups. In the training data set, there were 1035 subimage groups, including 315 subimage groups from the CIA data set and 720 subimage groups from the LGC data set. For testing purposes, excluding the image groups for training, we selected four image groups from the CIA data set, one image group from the LGC data set, and the image group from the Shenzhen data set.

TABLE I
DATA SET SETTINGS

Dataset		Time 1	Time 2	Time 3
CIA	train	20011007	20011016	20011203
		20011007	20011203	20020104
		20011016	20011203	20020111
		20011203	20020104	20020111
		20011203	20020111	20020212
		20020104	20020111	20020212
		20020104	20020212	20020401
		20020111	20020212	20020401
		20020221	20020401	20020503
	test	20011101	20011108	20011124
		20020221	20020309	20020316
		20020221	20020316	20020417
		20020410	20020417	20020426
LGC	train	20040416	20040502	20040705
		20040502	20040705	20040822
		20040705	20040806	20040822
		20040822	20041025	20041126
		20041025	20041126	20041212
		20050214	20050302	20050403
	test	20041228	20050113	20050129
Shenzhen	test	20001101	20021107	20041108

B. Evaluation Indices

For the evaluation, we compared the fusion results on the prediction date to the real observed Landsat images. In order to evaluate the results quantitatively, we used several evaluation indices to evaluate the fusion effects in terms of content, spectrum, and structure, i.e., the average absolute difference (AAD), the RMSE, the SSIM, the spectral angle mapper (SAM), and the erreur relative global adimensionnelle de synthèse (ERGAS).

AAD and RMSE reflect the reflectance difference between the predicted image and the observed image. AAD and RMSE are defined as follows:

$$\mathrm{AAD} = \frac{1}{P}\sum_{i=1}^{P}|\hat{L}(i) - L(i)| \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{P}\sum_{i=1}^{P}(\hat{L}(i) - L(i))^2} \tag{7}$$

where $\hat{L}$ and L represent the predicted image and the observed image, respectively, the terms $\hat{L}(i)$ and $L(i)$ represent the pixel value of the ith pixel of the predicted image and the observed image, and P is the total number of pixels. The smaller the value, the better the fit of the fusion result to the predicted image.

SSIM measures the similarity of the overall structures between the predicted image and the observed image. It is calculated as

$$\mathrm{SSIM} = \frac{(2\mu_L mu_{\hat{L}} + C_1)(2\sigma_{L\hat{L}} + C_2)}{(\mu_L^2 + mu_{\hat{L}}^2 + C_1)(\sigma_L^2 + \sigma_{\hat{L}}^2 + C_2)} \tag{8}$$

where μ_L and $mu_{\hat{L}}$ represent the mean of the fusion result and the real image, respectively, σ_L and $\sigma_{\hat{L}}$ represent their variances, $\sigma_{L\hat{L}}$ represents the covariance of L and $\hat{L}$, and C_1 and C_2 represent small constants. The larger the value, the higher the similarity between the fusion result and the real image.

SAM measures the spectral distortion of the fusion result. It is defined as follows:

$$\mathrm{SAM} = \frac{1}{P}\sum_{i=1}^{P}\arccos\frac{\sum_{j=1}^{B}(\hat{L}_j(i)L_j(i))}{\sqrt{\sum_{j=1}^{B}(\hat{L}_j(i))^2\sum_{j=1}^{B}(L_j(i))^2}} \tag{9}$$

where B is the total number of bands. A smaller SAM means a closer match to the real image in spectral recovery.

ERGAS evaluates the overall fusion results. It is defined as

$$\mathrm{ERGAS} = 100\frac{h}{l}\sqrt{\frac{1}{B}\sum_{j=1}^{B}\left(\frac{\mathrm{RMSE}(L_j)}{\mu_j}\right)^2} \tag{10}$$

where h is the spatial resolution of the real image, l is the spatial resolution of the fusion image, L_j is the jth band image, and μ_j is the mean of the jth band of the real image.

TABLE II
QUANTITATIVE PERFORMANCE EVALUATION FOR THE TEST DATA SET

Method	AAD			RMSE			SSIM			SAM	ERGAS	Time(s)
	NIR	Red	Green	NIR	Red	Green	NIR	Red	Green			
ESTARFM	0.0864	0.0993	0.1048	0.1370	0.1620	0.1626	0.8790	0.8477	0.8425	0.2865	0.9945	85.71
SPSTFM	0.0786	0.0994	0.1099	0.1100	0.1385	0.1525	0.9114	0.8777	0.8441	0.2398	0.9478	137.14
STFDCNN	0.0814	0.0911	0.1049	0.1095	0.1246	0.1409	0.9118	0.8918	0.8663	0.2347	0.9079	0.66
STFGAN	**0.0681**	**0.0799**	**0.0881**	**0.0914**	**0.1087**	**0.1188**	**0.9357**	**0.9138**	**0.9003**	**0.1964**	**0.7689**	**0.29**

The smaller the value and the closer it is to zero, the higher the fidelity of the overall fusion result.

C. Experimental Results and Analysis

For the testing, we used a test data set made up of four image groups from the CIA data set, one image group from the LGC data set, and the image group from the Shenzhen data set to illustrate the performance of STFGAN in predicting phenological changes and land-cover-type changes. To thoroughly evaluate the performance of the proposed algorithm, ESTARFM [26], SPSTFM [32], and STFDCNN [34] were selected as benchmarks for the comparison.

1) Test Data Set: For all six image groups in the test data set, the quantitative evaluation results of the average of all the bands in terms of AAD, RMSE, SSIM, SAM, ERGAS, and time cost (for every subimage) for the fusion results of ESTARFM, SPSTFM, STFDCNN, and STFGAN are listed in Table II. For the deep learning-based methods (STFDCNN and STFGAN), the model only needed to be trained once and could then be applied to all the test image groups, so we only compare the test time cost. To ensure a fair comparison, all the test experiments were performed on a Windows workstation equipped with an Intel Xeon E3-1220 processor at 3.10 GHz and 8-GB RAM. ESTARFM, SPSTFM, and STFGAN were coded in IDL, MATLAB, and TensorFlow, respectively. STFDCNN was implemented with the support of TensorFlow and MATLAB. It should be noted that the test time computations of all the methods were performed with the CPU. The best results for each evaluation index are labeled in bold, and the second-best results are underlined.

From Table II, we can observe that the proposed STFGAN method achieves lower AAD, RMSE, SAM, and ERGAS values and higher SSIM values than ESTARFM, SPSTFM, and STFDCNN for the whole test data set. This indicates that the proposed STFGAN method is capable of producing fusion results with a higher accuracy from the textural detail and spectral aspects. In addition, both deep learning-based methods show the ability to save time, and the proposed STFGAN method takes less time than STFDCNN due to its end-to-end workflow. The weight function-based method and sparse representation-based method take much more time when testing on large-scale images. We show and analyze the experimental results of the three data sets separately in the following sections.

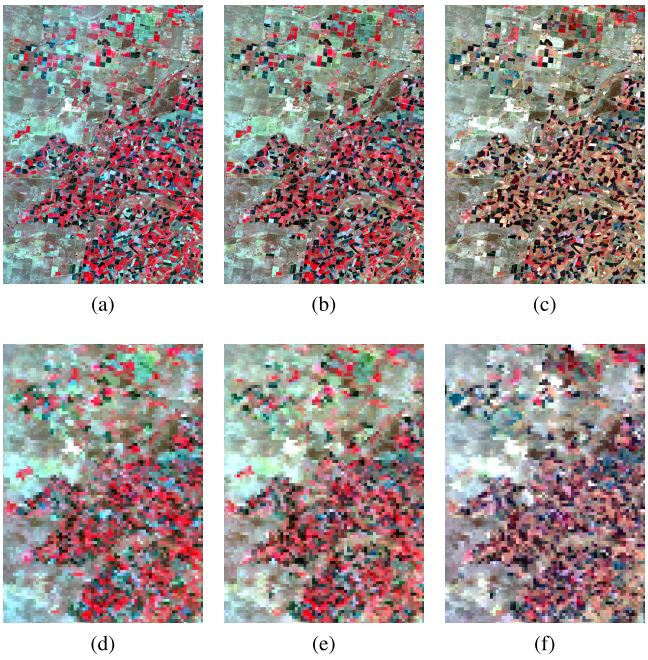

Fig. 5. Illustration of the CIA image group on November 8, 2001. (a)–(c) Landsat images and (d)–(f) corresponding MODIS images. (a) and (d) Pair was acquired on November 1, 2001. (b) and (e) Pair was acquired on November 8, 2001. (c) and (f) Pair was acquired on November 24, 2001.

2) Coleambally Irrigation Area: We selected one of the four image groups for testing from the CIA data set to demonstrate the details of the fusion results. The image pairs in this image group were acquired on three dates: November 1, 2001, November 8, 2001, and November 24, 2001. Scene subsets are shown in Fig. 5 using NIR–red–green as the red–green–blue composite and identical linear stretching. The CIA data set can be considered as a spatially heterogeneous study site for the small field sizes. In this image group, the temporal dynamics are mainly associated with crop phenology, and the land-cover types vary less over time.

The fusion results of ESTARFM, SPSTFM, STFDCNN, and STFGAN are shown in Fig. 6. To facilitate the comparison, we use white rectangles to mark some representative areas in the fusion images. The top row shows the actual observed Landsat image and the fusion results, and the bottom row shows their zoomed details for the parts in the white rectangles. We also used the square error images (i.e., the error images of the fusion results in mse metrics) from the fusion

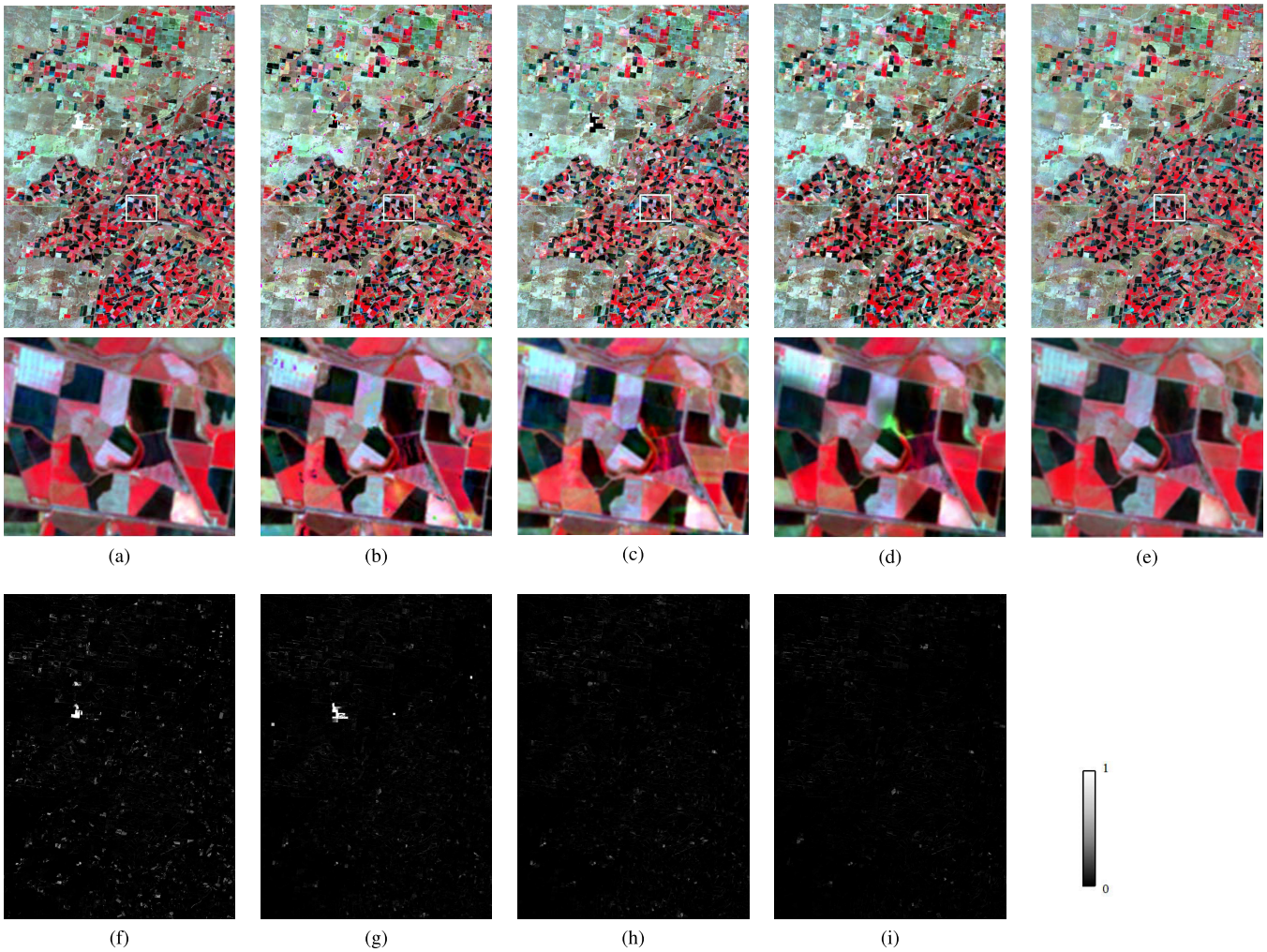

Fig. 6. Comparison between the ground truth and the predicted images for the CIA image group on November 8, 2001. (a) Ground truth. (b) Predicted image obtained by ESTARFM. (c) Predicted image obtained by SPSTFM. (d) Predicted image obtained by STFDCNN. (e) Predicted image obtained by our proposed STFGAN method. (f)–(i) Square error images from the fusion results for ESTARFM, SPSTFM, STFDCNN, and STFGAN.

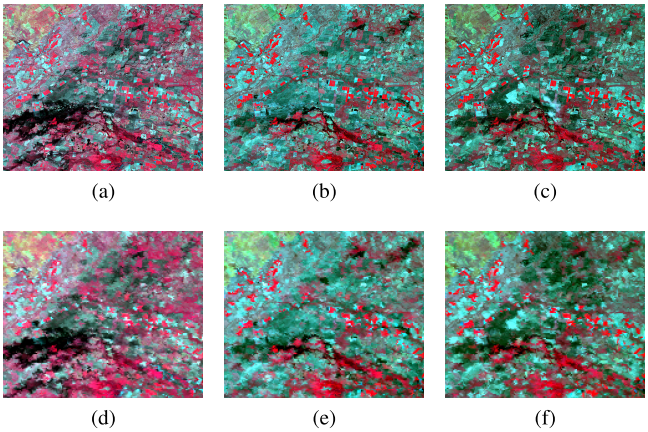

Fig. 7. Illustration of the LGC image group on January 13, 2005. (a)–(c) Landsat images and (d)–(f) corresponding MODIS images. (a) and (d) Pair was acquired on December 28, 2004. (b) and (e) Pair was acquired on January 13, 2005. (c) and (f) Pair was acquired on January 29, 2005.

results to discern the differences between different methods visually. By visually comparing the fusion images with the ground truth, it can be seen that all the methods can capture the phenological changes between the prediction image and the images on the prior dates. However, ESTARFM introduces some unpleasant and disturbing color in some pixels, which may have been caused by the filtering strategy. From the square error images, it can be clearly observed that the deviation between the fusion result of STFGAN and the ground truth is the smallest. For some spatially heterogeneous areas, such as the zoomed areas in the white rectangles, it is clear that STFGAN results in smaller prediction error than ESTRAFM, SPSTFM, and STFDCNN, which reveals that the proposed STFGAN method is more robust than the others in dealing with a spatially heterogeneous case. For the sparse representation-based method, the image features need to be designed manually, which brings instability to the performance, resulting in severe spectral distortion in the fusion result of SPSTFM. In the fusion result of STFDCNN, there is a little spectral distortion at the edge of the field, due to the high-pass modulation [34].

The quantitative evaluation results in terms of AAD, RMSE, SSIM, SAM, and ERGAS are listed in Table III. It can be observed that the deep learning-based methods perform better than the weight function-based method and sparse

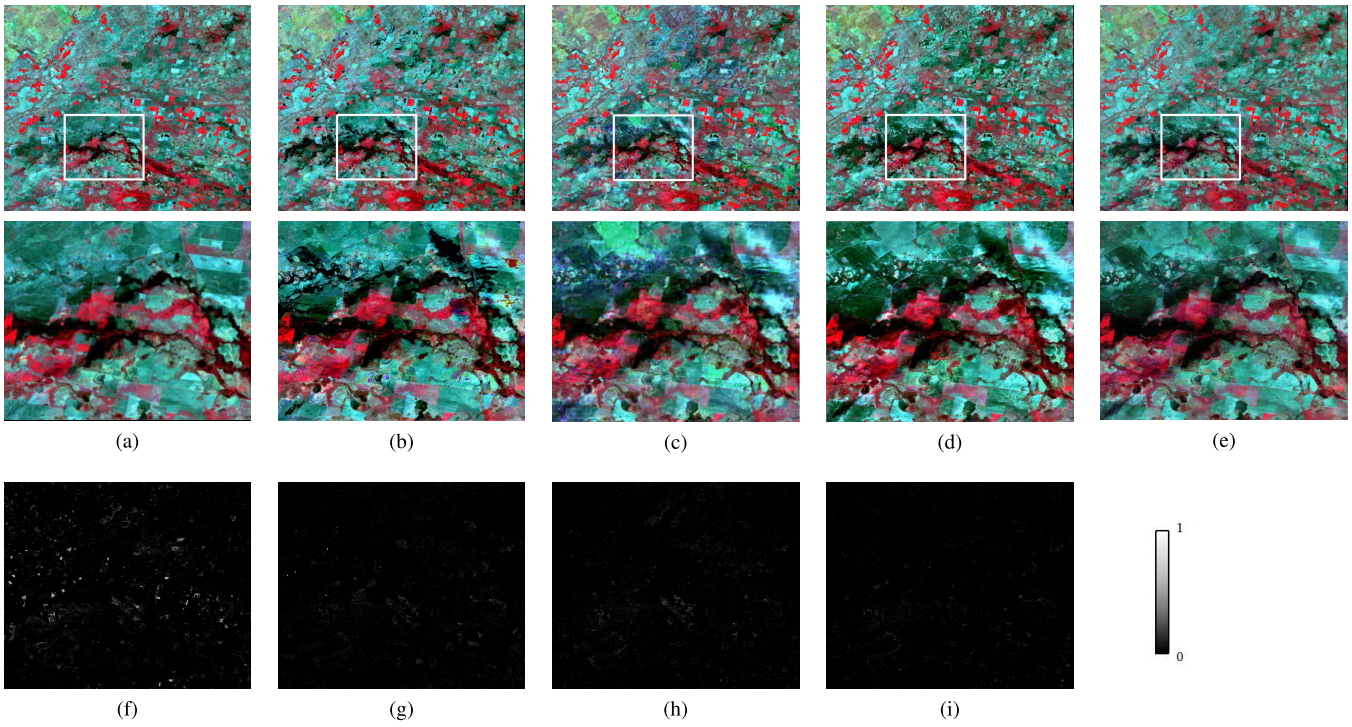

Fig. 8. Comparison between the ground truth and predicted images of the LGC image group on January 13, 2005. (a) Ground truth. (b) Predicted image obtained by ESTARFM. (c) Predicted image obtained by SPSTFM. (d) Predicted image obtained by STFDCNN. (e) Predicted image obtained by our proposed STFGAN method. (f)–(i) Square error images from the fusion results for ESTARFM, SPSTFM, STFDCNN, and STFGAN.

TABLE III
QUANTITATIVE PERFORMANCE EVALUATION FOR THE CIA DATA SET

Method	AAD			RMSE			SSIM			SAM	ERGAS
	NIR	Red	Green	NIR	Red	Green	NIR	Red	Green		
ESTARFM	0.0884	0.0990	0.0974	0.1485	0.1521	0.1467	0.8835	0.8541	0.8584	0.2607	0.9617
SPSTFM	0.0699	0.1012	0.1015	0.1073	0.1423	0.1413	0.9261	0.8684	0.8709	0.2198	0.9071
STFDCNN	0.0720	0.0881	0.0915	0.1071	0.1248	0.1274	0.9262	0.8875	0.8803	0.2148	0.8341
STFGAN	**0.0669**	**0.0837**	**0.0826**	**0.0929**	**0.1129**	**0.1146**	**0.9364**	**0.8995**	**0.8981**	**0.1941**	**0.7646**

TABLE IV
QUANTITATIVE PERFORMANCE EVALUATION FOR THE LGC DATA SET

Method	AAD			RMSE			SSIM			SAM	ERGAS
	NIR	Red	Green	NIR	Red	Green	NIR	Red	Green		
ESTARFM	0.0587	0.0896	0.0934	0.0951	0.1522	0.1529	0.9266	0.8715	0.8689	0.2653	0.8930
SPSTFM	0.0578	0.0940	0.0812	0.0823	0.1293	0.1147	0.9409	0.9011	0.9064	0.2173	0.8242
STFDCNN	0.0582	0.0819	0.0965	0.0836	0.1132	0.1329	0.9333	0.9141	0.8859	0.2226	0.8347
STFGAN	**0.0509**	**0.0753**	**0.0790**	**0.0696**	**0.1024**	**0.1080**	**0.9551**	**0.9272**	**0.9176**	**0.1890**	**0.7177**

representation-based method. The reason for this is that deep learning can map the relationship between the MODIS images and the Landsat images more precisely than the other methods. The proposed STFGAN method achieves the lowest AAD and RMSE values in all three bands, which indicates that STFGAN can reconstruct the Landsat image more precisely than the other three methods. This can be further illustrated through the comparison with ERGAS. STFGAN achieves the highest SSIM values in all three bands, which shows that STFGAN captures more structural details in the surface reflectance than ESTARFM, SPSTFM, and STFDCNN. This might be due to the fact that STFGAN uses residual blocks to extract the features from the MODIS and Landsat images, which can learn more structural information and textural details. Thus, more structural details are accurately predicted. STFGAN achieves the lowest SAM value, which illustrates that STFGAN has also

TABLE V
QUANTITATIVE PERFORMANCE EVALUATION FOR THE SHENZHEN DATA SET

Method	AAD			RMSE			SSIM			SAM	ERGAS
	NIR	Red	Green	NIR	Red	Green	NIR	Red	Green		
ESTARFM	**0.0698**	**0.0776**	0.0885	0.1195	0.1289	0.1391	0.8950	0.8752	0.8588	0.2657	1.0200
SPSTFM	0.0705	0.1003	0.0944	**0.1000**	0.1496	0.1345	0.9212	0.8587	0.8697	0.2501	1.1233
STFDCNN	0.1178	0.0840	0.0825	0.1488	0.1280	0.1221	0.8828	**0.8879**	0.8873	0.2449	1.0575
STFGAN	0.0753	0.0834	**0.0812**	0.1022	**0.1272**	**0.1192**	**0.9249**	0.8822	**0.8921**	**0.2355**	**0.9727**

achieved more accurate results than other methods in spectral recovery. In general, the results of the proposed STFGAN method are better than those of the other benchmark methods. The proposed method can not only predict high-quality spatial information but can also achieve a superior performance in color retention.

3) Lower Gwydir Catchment: The image pairs of the test image group in the LGC data set were acquired on three dates: December 28, 2004, January 13, 2005, and January 29, 2005, as shown in Fig. 7. The LGC data set can be considered as a study site containing both phenological changes and significant land-cover-type changes. In this image group, the land-cover-type changes are mainly associated with receding floodwater, during which time the shape of the flooded area changed regularly. The fusion results and the square error images are shown in Fig. 8. The ground truth and the fusion results are shown in the first row, the zoomed details for the parts in the white rectangles are shown in the second row, and the error images are shown in the third row. By comparing the square error images, it can be clearly observed that the fusion result of STFGAN is the closest to the ground truth. For the areas where flooding occurred (the zoomed areas in the white rectangles), SPSTFM performs the worst, with spectral distortion and severe blurring. The fusion results of both ESTARFM and STFDCNN are affected by the prior image pairs and deviate greatly from the ground truth. The comparison of the fusion results and the square error images demonstrates that the result of STFGAN has minimal color and textural differences, which illustrates the superiority of STFGAN in spectral information retrieval and the prediction of significant land-cover-type changes with regular shapes.

A quantitative evaluation is provided in Table IV. It can be observed that the proposed STFGAN method again outperforms the other three methods, which demonstrates the superiority of STFGAN, both in phenological change prediction and land-cover-type change prediction. Moreover, compared with SPSTFM and STFDCNN, STFGAN achieves lower AAD and RMSE values, higher SSIM values, lower SAM value, and lower ERGAS value over all the bands. This means that STFGAN delivers superior performance in terms of content, structure, spectrum, and comprehensive effects. The improvement of STFGAN over ESTARFM, SPSTFM, and STFDCNN at the LGC study site in terms of AAD, RMSE, SSIM, SAM, and ERGAS illustrates the progress that STFGAN has made in capturing complex change information of land-cover types, as STFGAN incorporates supplementary information provided by the prior Landsat-MODIS image pairs into the prediction.

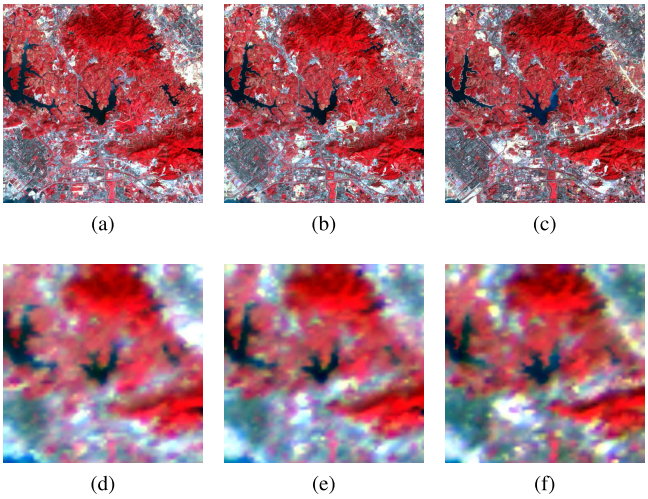

Fig. 9. Illustration of the Shenzhen image group on November 7, 2002. (a)–(c) Landsat images and (d)–(f) corresponding MODIS images. (a) and (d) Pair was acquired on November 1, 2000. (b) and (e) Pair was acquired on November 7, 2002. (c) and (f) Pair was acquired on November 8, 2004.

4) Shenzhen: In order to illustrate STFGAN's ability for generalization and capacity to predict more complex land-cover-type changes, an image group excluded from the training study sites was considered. As shown in Fig. 9, the image pairs in this image group were acquired on December 28, 2004, January 13, 2005, and January 29, 2005, respectively. Shenzhen can be seen as a study site containing land-cover-type changes with irregular shapes. In addition, the land-cover-type changes are mainly associated with human activities over the four years. The huge time gap means intense and irregular changes in the urban land-cover types and shapes, which increases the difficulty and complexity of the prediction.

The visual comparison and quantitative evaluation are shown in Fig. 10 and Table V. Overall, STFGAN achieves the best or second-best results on all the bands. In the fusion results of ESTARFM and SPSTFM, there are some unpleasant stains and bright spots, respectively, reducing the content similarity of the results with the ground truth. There is also a color difference between the fusion result of STFDCNN and the ground truth. For the region in the black circle, the land-cover types and shapes change irregularly with the construction and demolition of the buildings, and some content

TABLE VI
COMPARISON BETWEEN THE SINGLE- AND TWO-STAGE FRAMEWORKS

Method		AAD			RMSE			SSIM			SAM	ERGAS
		NIR	Red	Green	NIR	Red	Green	NIR	Red	Green		
Testing dataset	Single-Stage	0.0688	0.0802	0.0884	0.0920	0.1090	0.1189	0.9347	**0.9138**	0.9002	0.1970	0.7723
	Two-Stage	**0.0681**	**0.0799**	**0.0881**	**0.0914**	**0.1087**	**0.1188**	**0.9357**	**0.9138**	**0.9003**	**0.1964**	**0.7689**
CIA-20011108	Single-Stage	0.0690	0.0857	0.0830	0.0945	0.1146	0.1143	0.9346	0.8986	**0.8992**	0.2245	0.7742
	Two-Stage	**0.0669**	**0.0837**	**0.0826**	**0.0929**	**0.1129**	**0.1146**	**0.9364**	**0.8995**	0.8981	**0.2225**	**0.7646**
LGC-20050113	Single-Stage	0.0515	0.0754	0.0791	0.0701	0.1026	0.1080	0.9545	**0.9273**	0.9174	0.1898	0.7202
	Two-Stage	**0.0509**	**0.0753**	**0.0790**	**0.0696**	**0.1024**	**0.1080**	**0.9551**	0.9272	**0.9176**	**0.1890**	**0.7177**
Shenzhen-20021107	Single-Stage	**0.0750**	0.0845	0.0817	**0.1021**	0.1283	0.1198	0.9220	0.8806	0.8920	0.2367	0.9808
	Two-Stage	0.0753	**0.0834**	**0.0812**	0.1022	**0.1272**	**0.1192**	**0.9249**	**0.8822**	**0.8921**	**0.2355**	**0.9727**

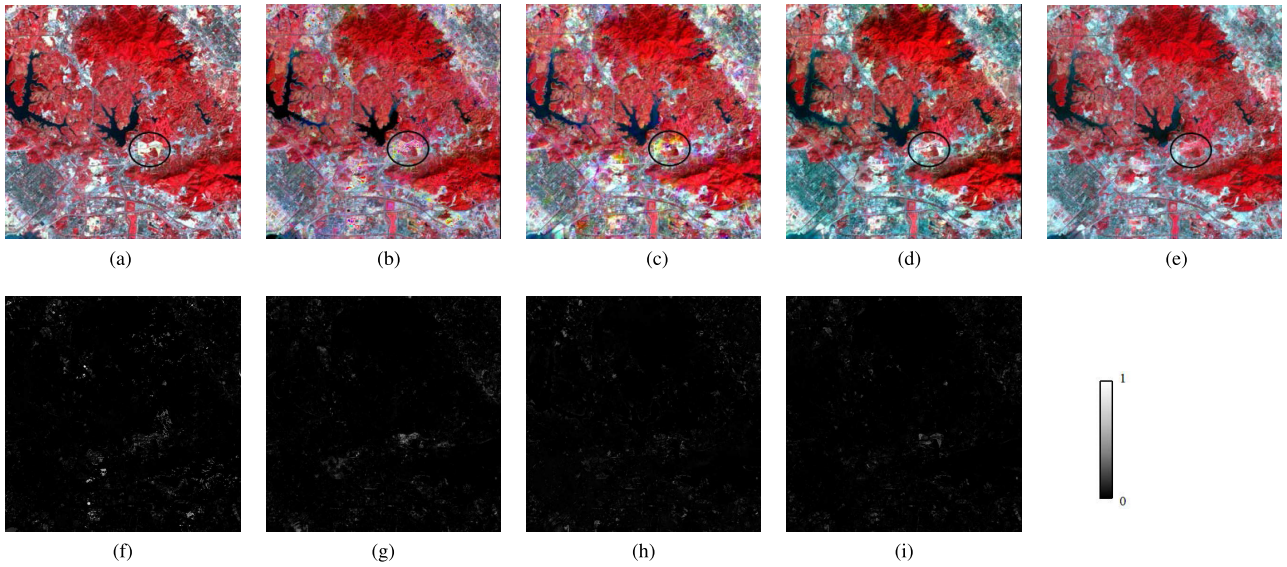

Fig. 10. Comparison between the ground truth and predicted images of the Shenzhen image group on November 7, 2002. (a) Ground truth. (b) Predicted image obtained by ESTARFM. (c) Predicted image obtained by SPSTFM. (d) Predicted image obtained by STFDCNN. (e) Predicted image obtained by our proposed STFGAN method. (f)–(i) Square error images from the fusion results for ESTARFM, SPSTFM, STFDCNN, and STFGAN.

and structural information is only contained in the MODIS image at time 2. In this region, the prediction of STFGAN is not as precise as that of the other methods. This can be attributed to the small weight of the MODIS image at time 2 in the input images. This will be addressed in our future work. Although the other methods may seem to be able to predict such irregular land-cover-type changes, in fact, their errors are also quite large. Furthermore, it can be observed from the square error images that their predictions of the content and structural details of the other regions are not as accurate as those of STFGAN. Moreover, from Table V, we can observe that STFGAN still achieves competitive results, on the whole, which is consistent with the visual analysis. The fusion result of STFGAN in the Shenzhen study site illustrates the excellent generalization ability of the proposed method.

D. Ablation Study for the Two-Stage Framework

To validate the effectiveness of the two-stage framework, we compared it with a single-stage framework. The performance of the single- and two-stage frameworks was evaluated on the whole test data set for all three study sites. The results are listed in Table VI. From Table VI, it can be observed that generally speaking, the two-stage framework performs better than the single-stage framework over the whole test data set. In order to compare the effect of the single- and two-stage frameworks on the different study sites, we selected a set of image groups from the three study sites for a comparative analysis. For the selected image groups in the CIA and LGC study sites, the two-stage framework improves the accuracy by acquiring a more similar content of prediction image and a better characterization and delineation of the type changes. The two-stage framework performs well in the Shenzhen study site, which was not included in the training study sites. Thus, the two-stage framework improves the generalization ability of the proposed method. It is also worth mentioning that the fusion results of the single-stage framework are, in fact, more accurate than those of the other benchmark methods.

IV. Conclusion

In this article, we have proposed a novel remote sensing image STFGAN with a two-stage framework, considering the huge resolution difference between Landsat and MODIS data. For each stage, the generator network takes the coarse MODIS image on the prediction date and two prior Landsat-MODIS image pairs as input and the corresponding fine Landsat image as output. The features super-resolved from the MODIS images and the high-frequency features extracted from the Landsat-like images are fused to generate the predicted Landsat image. In this way, the two prior image pairs can provide the spatial information as the auxiliary information for the MODIS image on the prediction date. Moreover, under the adversarial supervision of the game between the discriminator and the generator, the generator is forced to generate as realistic an image as possible. The proposed STFGAN method was compared with several state-of-the-art spatiotemporal fusion methods by conducting experiments on three data sets, and the experimental results confirmed the effectiveness and generalization ability of STFGAN from both spatial and spectral perspectives. In addition, STFGAN, as a ready-to-use model after the training, requires a much shorter computation time in the inference phase. Our future work will continue along the line of improving the prediction accuracy of land-cover-type changes with irregular shapes.

Acknowledgment

The authors would like to thank Dr. Xiaolin Zhu and Dr. Jin Chen for providing access to the ESTARFM IDL code, Dr. Bo Huang for providing the Shenzhen data and MATLAB code associated with his published article about SPSTFM, and Dr. Irina Emelyanova for providing the CIA and LGC data sets.

References

[1] F. W. Acerbi-Junior, J. G. P. W. Clevers, and M. E. Schaepman, "The assessment of multi-sensor image fusion using wavelet transforms for mapping the Brazilian Savanna," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 8, no. 4, pp. 278–288, Dec. 2006.

[2] C. Senf, P. J. Leitão, D. Pflugmacher, S. van der Linden, and P. Hostert, "Mapping land cover in complex mediterranean landscapes using Landsat: Improved classification accuracies from integrating multi-seasonal and synthetic imagery," *Remote Sens. Environ.*, vol. 156, pp. 527–536, Jan. 2015.

[3] H. Zhang, J. M. Chen, B. Huang, H. Song, and Y. Li, "Reconstructing seasonal variation of Landsat vegetation index related to leaf area index by fusing with MODIS data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 3, pp. 950–960, Mar. 2014.

[4] F. Gao *et al.*, "Fusing Landsat and MODIS data for vegetation monitoring," *IEEE Geosci. Remote Sens. Mag.*, vol. 3, no. 3, pp. 47–60, Sep. 2015.

[5] J. R. Townshend *et al.*, "Global characterization and monitoring of forest cover using Landsat data: Opportunities and challenges," *Int. J. Digit. Earth*, vol. 5, no. 5, pp. 373–397, Sep. 2012.

[6] G. Xian and M. Crane, "Assessments of urban growth in the Tampa bay watershed using remote sensing data," *Remote Sens. Environ.*, vol. 97, no. 2, pp. 203–215, Jul. 2005.

[7] J. Huang *et al.*, "Assimilating a synthetic Kalman filter leaf area index series into the Wofost model to estimate regional winter wheat yield," *Agricult. Forest Meteorol.*, vol. 216, pp. 188–202, Jan. 2016.

[8] J. Zhang, Q. Zhou, X. Shen, and Y. Li, "Cloud detection in high-resolution remote sensing images using multi-features of ground objects," *J. Geovis. Spatial Anal.*, vol. 3, no. 2, Dec. 2019.

[9] H. Zhai, H. Zhang, L. Zhang, and P. Li, "Total variation regularized collaborative representation clustering with a locally adaptive dictionary for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 166–180, Jan. 2019.

[10] H. Zhang, J. Kang, X. Xu, and L. Zhang, "Accessing the temporal and spectral features in crop type mapping using multi-temporal Sentinel-2 imagery: A case study of Yi'an county, Heilongjiang province, China," *Comput. Electron. Agricult.*, vol. 176, Sep. 2020, Art. no. 105618, doi: 10.1016/j.compag.2020.105618.

[11] Q. Weng, P. Fu, and F. Gao, "Generating daily land surface temperature at Landsat resolution by fusing Landsat and MODIS data," *Remote Sens. Environ.*, vol. 145, pp. 55–67, Apr. 2014.

[12] C. Cammalleri, M. C. Anderson, F. Gao, C. R. Hain, and W. P. Kustas, "A data fusion approach for mapping daily evapotranspiration at field scale," *Water Resour. Res.*, vol. 49, no. 8, pp. 4672–4686, Aug. 2013.

[13] C. M. Gevaert and F. J. García-Haro, "A comparison of STARFM and an unmixing-based algorithm for Landsat and MODIS data fusion," *Remote Sens. Environ.*, vol. 156, pp. 34–44, Jan. 2015.

[14] 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.

[15] J. Ju and D. P. Roy, "The availability of cloud-free Landsat ETM+ data over the conterminous united states and globally," *Remote Sens. Environ.*, vol. 112, no. 3, pp. 1196–1211, Mar. 2008.

[16] H. Zhai, H. Zhang, L. Zhang, and P. Li, "Cloud/shadow detection based on spectral indices for multi/hyperspectral optical remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 144, pp. 235–253, Oct. 2018.

[17] J. G. 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.

[18] R. Michishita, Z. Jiang, and B. Xu, "Monitoring two decades of urbanization in the Poyang lake area, China through spectral unmixing," *Remote Sens. Environ.*, vol. 117, pp. 3–18, Feb. 2012.

[19] H. Zhai, H. Zhang, L. Zhang, and P. Li, "Laplacian-regularized low-rank subspace clustering for hyperspectral image band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1723–1740, Mar. 2019.

[20] H. Zhang, L. Liu, W. He, and L. Zhang, "Hyperspectral image denoising with total variation regularization and nonlocal low-rank tensor decomposition," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3071–3084, May 2020.

[21] C. O. Justice *et al.*, "An overview of MODIS land data processing and product status," *Remote Sens. Environ.*, vol. 83, nos. 1–2, pp. 3–15, Nov. 2002.

[22] D. P. Roy *et al.*, "Landsat-8: Science and product vision for terrestrial global change research," *Remote Sens. Environ.*, vol. 145, pp. 154–172, Apr. 2014.

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

[24] X. Zhu, F. Cai, J. Tian, and T. Williams, "Spatiotemporal fusion of multisource remote sensing data: Literature survey, taxonomy, principles, applications, and future directions," *Remote Sens.*, vol. 10, no. 4, p. 527, Mar. 2018.

[25] 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.

[26] X. Zhu, J. Chen, F. Gao, X. Chen, and J. G. Masek, "An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions," *Remote Sens. Environ.*, vol. 114, no. 11, pp. 2610–2623, Nov. 2010.

[27] T. Hilker *et al.*, "A new data fusion model for high spatial- and temporal-resolution mapping of forest disturbance based on Landsat and MODIS," *Remote Sens. Environ.*, vol. 113, no. 8, pp. 1613–1627, Aug. 2009.

[28] Y. Zhao, B. Huang, and H. Song, "A robust adaptive spatial and temporal image fusion model for complex land surface changes," *Remote Sens. Environ.*, vol. 208, pp. 42–62, Apr. 2018.

[29] 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.

[30] R. Zurita-Milla, J. Clevers, and M. E. Schaepman, "Unmixing-based Landsat TM and MERIS FR data fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 3, pp. 453–457, Jul. 2008.

[31] M. Wu, H. Wenjiang, N. Zheng, and W. Changyao, "Generating daily synthetic Landsat imagery by combining Landsat and MODIS data," *Sensors*, vol. 15, no. 9, pp. 24002–24025, 2015.
[32] B. Huang and H. Song, "Spatiotemporal reflectance fusion via sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 10, pp. 3707–3716, Oct. 2012.
[33] H. Song and B. Huang, "Spatiotemporal satellite image fusion through one-pair image learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 4, pp. 1883–1896, Apr. 2013.
[34] 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.
[35] X. Liu, C. Deng, J. Chanussot, D. Hong, and B. Zhang, "StfNet: A two-stream convolutional neural network for spatiotemporal image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6552–6564, Sep. 2019.
[36] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[37] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1125–1134.
[38] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4681–4690.
[39] N. C. Rakotonirina and A. Rasoanaivo, "ESRGAN+ : Further improving enhanced super-resolution generative adversarial network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 63–79.
[40] H. Zhang and V. M. Patel, "Densely connected pyramid dehazing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3194–3203.
[41] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.
[42] Z. Lin and H.-Y. Shum, "Fundamental limits of reconstruction-based superresolution algorithms under local translation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 1, pp. 83–97, Jan. 2004.
[43] W. Yang, X. Zhang, Y. Tian, W. Wang, and J. Xue, "Deep learning for single image super-resolution: A brief review," *IEEE Trans. Multimedia*, vol. 21, no. 12, pp. 3106–3121, 2019.
[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556
[45] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 214–223.
[46] A. Berk *et al.*, "MODTRAN4 radiative transfer modeling for atmospheric correction," *Proc. SPIE*, vol. 3756, pp. 348–353, Oct. 1999.
[47] I. V. Emelyanova, T. R. McVicar, T. G. Van Niel, L. T. Li, and A. I. J. M. van Dijk, "Assessing the accuracy of blending Landsat–MODIS surface reflectances in two landscapes with contrasting spatial and temporal dynamics: A framework for algorithm selection," *Remote Sens. Environ.*, vol. 133, pp. 193–209, Jun. 2013.
[48] F. Li *et al.*, "An evaluation of the use of atmospheric and BRDF correction to standardize Landsat data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 3, no. 3, pp. 257–270, Sep. 2010.
[49] E. F. Vermote, D. Tanre, J. L. Deuze, M. Herman, and J.-J. Morcette, "Second simulation of the satellite signal in the solar spectrum, 6S: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 35, no. 3, pp. 675–686, May 1997.

Hongyan Zhang (Senior Member, IEEE) received the B.S. degree in geographic information system and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2005 and 2010, respectively.

Since 2016, he has been a Full Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University. He is a Young Chang-Jiang Scholar appointed by the Ministry of Education of China. He scored first in the Pairwise Semantic Stereo Challenge of the 2019 Data Fusion Contest organized by the IEEE Image Analysis and Data Fusion Technical Committee. He has authored or coauthored more than 90 research articles. His research interests include image reconstruction for quality improvement, hyperspectral information processing, and agricultural remote sensing.

Dr. Zhang was the Session Chair of the 2016 IEEE IGARSS Conference and the 2015 IEEE WHISPERS Conference. He is a reviewer for more than 30 international academic journals, including the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, and the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS. He serves as an Associate Editor for the *Photogrammetric Engineering and Remote Sensing and Computers and Geosciences*.

Yiyao Song (Student Member, IEEE) received the B.S. degree from the School of Electronic Information, Wuhan University, Wuhan, China, in 2018, where she is pursuing the M.S. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing.

Her research interests include remote sensing image spatiotemporal fusion, super-resolution, and deep learning.

Chang Han (Member, IEEE) received the B.S. degree in electronic science and technology from Nanchang Hangkong University, Nanchang, China, in 2006, the M.S. degree in physical electronics from Fuzhou University, Fuzhou, China, in 2009, and the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2016.

He is a Lecturer with the School of Mechanical and Electrical Engineering, Wuhan Business University, Wuhan. His research interests include image reconstruction, sparse representation, and computer vision.

Liangpei Zhang (Fellow, IEEE) received the B.S. degree in physics from Hunan Normal University, Changsha, China, in 1982, the M.S. degree in optics from the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China, in 1988, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1998.

He was a Principal Scientist with the China State Key Basic Research Project from 2011 to 2016 appointed by the Ministry of National Science and Technology of China to lead the Remote Sensing Program in China. He is a Chang-Jiang Scholar' Chair Professor appointed by the Ministry of Education of China with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University. He has published more than 700 research articles and five books. He is a highly cited author of the Institute for Scientific Information (ISI). He is the holder of 30 patents. His research interests include hyperspectral remote sensing, high-resolution remote sensing, image processing, and artificial intelligence.

Dr. Zhang is a fellow of the Institution of Engineering and Technology (IET). He was a recipient of the 2010 Best Paper Boeing Award, the 2013 Best Paper ERDAS Award from the American Society of Photogrammetry and Remote Sensing (ASPRS), and the 2016 Best Paper Theoretical Innovation Award from the International Society for Optics and Photonics (SPIE). His research teams won the top three prizes of the IEEE Geoscience and Remote Sensing Society (GRSS) 2014 Data Fusion Contest and his students have been selected as the winners or finalists of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS) Student Paper Contest in recent years. He is the Founding Chair of the IEEE GRSS Wuhan Chapter. He also serves as an Associate Editor or an Editor for more than ten international journals. He is serving as an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.