Multistream STGAN: A Spatiotemporal Image Fusion Model With Improved Temporal Transferability

Fangzheng Lyu, Zijun Yang, Chunyuan Diao, and Shaowen Wang

***Abstract*—Spatiotemporal satellite image fusion aims to generate remote sensing images satisfying both high spatial and temporal resolution by integrating different satellite imagery datasets with distinct spatial and temporal resolutions. Such fusion technique is crucial for numerous applications that require frequent monitoring at fine spatial and temporal scales spanning agriculture, environment, natural resources, and disaster management. However, existing fusion models have difficulty accommodating abrupt spatial changes in land cover among images and dealing with temporally distant image data. This article proposes a novel multistream spatiotemporal fusion generative adversarial network (STGAN) model for spatiotemporal satellite image fusion that can produce accurate fused images and accommodate substantial temporal differences between the input images. The STGAN employs a conditional generative adversarial network architecture with a multistream input design to better learn temporal features. The generator of STGAN comprises convolutional blocks, a spatial transformer module, a channel attention network, and a U-net module designed to better capture spatial and temporal features from the multiresolution input images. Comprehensive evaluations of the proposed STGAN model have been performed on the Coleambally Irrigation Area and Lower Gwydir Catchment datasets, using both visual inspection and spatial and spectral metrics, including root mean square error, relative dimensionless global error synthesis, spectral angle mapping, structural similarity index measure, and local binary pattern. The experiments show that the proposed STGAN model consistently outperforms existing benchmark models and is capable of generating high-quality fused remote sensing data product of high spatial and temporal resolution.**

***Index Terms*—Deep learning, generative adversarial network (GAN), remote sensing, spatiotemporal image fusion.**

Received 2 August 2024; revised 30 September 2024 and 5 November 2024; accepted 11 November 2024. Date of publication 25 November 2024; date of current version 12 December 2024. This work was supported in part by the National Science Foundation (NSF) under Grant 1833225, Grant 2112356, Grant 2118329, and Grant 1951657; in part by the National Aeronautics and Space Administration (NASA) under Grant 80NSSC21K0946; and in part by the United States Department of Agriculture (USDA) under Grant 2021-67021-33446. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF, NASA, and USDA. *(Fangzheng Lyu and Zijun Yang are co-first authors.) (Corresponding authors: Chunyuan Diao; Shaowen Wang.)*

Fangzheng Lyu is with the Department of Geography, Virginia Tech, Blacksburg, VA 24061 USA (e-mail: fangzheng@vt.edu).

Zijun Yang is with the Department of Geography and Geographic Information Science, University of Illinois Urbana-Champaign, Urbana, IL 61801 USA, and also with the Department of Earth and Ocean Sciences, and Center for Marine Science, University of North Carolina Wilmington, Wilmington, NC 28403 USA (e-mail: yangz@uncw.edu).

Chunyuan Diao and Shaowen Wang are with the Department of Geography and Geographic Information Science, University of Illinois Urbana-Champaign, Urbana, IL 61801 USA (e-mail: chunyuan@illinois.edu; shaowen@illinois.edu).

Digital Object Identifier 10.1109/JSTARS.2024.3506879

I. Introduction

SPATIOTEMPORAL satellite image fusion, the process of generating remote sensing images with both high spatial and temporal resolutions, has gained significant attention in recent years. Although an increasing number of satellite missions have been launched, most satellite datasets are still faced with the tradeoff of resolutions, and it is still challenging to provide imagery with both high spatial and temporal resolutions [1]. Spatiotemporal image fusion integrates different remote sensing datasets (i.e., coarse and fine images) with distinct characteristics in their spatial and temporal resolutions. The coarse images are with low spatial resolution but a higher temporal frequency, such as the MODIS data, while the fine images are featured with high spatial resolution but lower temporal frequency, such as the Landsat images. The resulting fusion images with high spatiotemporal resolutions can significantly enhance the application potential of remote sensing data, benefiting a variety of applications, including agriculture, environment, natural resources, disaster managements, etc. [2], [3], [4]. Therefore, spatiotemporal image fusion is a feasible solution to leveraging the diverse collection of satellite images and providing fused satellite datasets with high spatiotemporal resolutions.

Existing developed fusion models can be mainly categorized into four types, namely weight-function-based, unmixing-based, Bayesian-based, and learning-based models. The spatial and temporal adaptive reflectance fusion model represents a classic example of weight-function-based fusion models, which predict reflectance of fine pixels with information from neighboring coarse pixels integrated by a weight function [5]. Pixels with lower spectral difference, temporal difference, and spatial distance are normally assigned with larger weights [6], [7], [8]. Unmixing-based methods first define endmembers from the fine-resolution images, and unmix the low-resolution image with those endmembers in order to predict the reflectance values in the fusion images [9]. Bayesian-based methods are constructed based on the maximum a posterior probability estimation. The predicted image is generated through the process of maximizing the conditional probability given the existing fine and coarse images. With the unprecedented advances in machine learning and deep learning, learning-based models have increasingly

gained more attention in the field of spatiotemporal image fusion [10], [11]. Most traditional methods hold the assumption that land cover types remain mostly unchanged, and their ability to accommodate abrupt changes among images can be limited. In contrast, learning-based models can adapt to complex, nonlinear relationships and sudden changes in data patterns, which provides a solution for spatiotemporal image fusion with dynamic and heterogeneous landscapes.

Deep learning has opened new opportunities for spatiotemporal image fusion. Fusion models based on deep-learning techniques are able to learn complex relationships from large volumes of remote sensing data. Recent advances in deep-learning-based spatiotemporal image fusion models have demonstrated promising performance in providing high-quality fusion imagery with enhanced accuracy. Among various deep-learning modeling architectures, convolutional neural network (CNN) is the most frequently employed deep-learning architecture for spatiotemporal image fusion. For spatiotemporal image fusion, the convolutional operations of CNN allow the models to extract and map spatial features across different sources of images, facilitating accurate construction of the spatial details. Through the design of CNN-based model structures, the relationships between the spatial and textual features in the low- and high-resolution images can be established, enhancing the model accuracy [12], [13], [14]. The spatiotemporal satellite image fusion using deep CNN is the first CNN-based spatiotemporal fusion model, which leverages the super-resolution CNN model to extract fine-scale spatial features and reconstruct the high-resolution images [13]. Recently developed CNN-based models introduce more sophisticated CNN structures in order to extract more representative features and accommodate the discrepancies in spatial, spectral, and temporal aspects between the fine and coarse images [14], [15].

Apart from CNN, generative adversarial network (GAN) models have demonstrated better performances in many remote sensing applications, such as object detection and image classification. GAN models consist of a generator, which generates output data samples, and a discriminator for evaluating them for authenticity by comparing the generated and reference data. Since GAN has the capacity to produce high-resolution images that share similar statistical distributions with the real satellite images, GAN-based have great potential in learning and generating diverse spatial/landscape features when substantial resolution discrepancy exists. Moreover, GAN models can be further enhanced through variants, such as conditional GAN that allows more directed and controlled data generation [16]. Representative GAN-based fusion model includes spatiotemporal fusion method using a GAN [17] which incorporates an image fusion GAN with superior performance in image super resolution. More recently, multilevel feature fusion with GAN incorporates multilevel CNN as the generator in an effort to accommodate the significant resolution difference between low- and high-resolution images [18]. In addition, GAN-based models have demonstrated their potential in lessening the requirement of input data. Compared to traditional models that may require two image pairs (i.e., one coarse and one fine image acquired on the same date), newly proposed GAN-based fusion models may require only one image pair [16], [19]. For instance, the GAN-based spatiotemporal fusion model (GANSTFM) incorporates conditional GAN to reduce the number of model inputs, allowing more flexible deployment of the fusion model [16]. More recently, GAN spatiotemporal fusion model based on multiscale and convolutional block attention module [20] and high-precision remote sensing spatiotemporal fusion method (HPLTS-GAN) [21] both integrate multiscale feature extraction to further enhance model performance with limited one-pair input imagery.

Despite the recent success of deep learning in spatiotemporal image fusion, temporal transferability remains a critical challenge in spatiotemporal fusion. Existing models often require temporally adjacent training samples and need retraining for different temporal periods to obtain accurate fusion results, which limits their generalizability. [18], [22]. The drastic changes that may happen across the temporal dimension pose great challenges to the accurate prediction of reflectance change. Moreover, land cover changes over time may alter the landscape in the images, which brings difficulties for the fusion models to retrieve the spatial structures from the low-resolution images [16]. While selecting temporally close images helps improve model performance, it is not always feasible to acquire images that are acquired near the prediction date. Therefore, a novel modeling structure that can accommodate substantial temporal changes and alleviate the demand for temporally close images is highly desired. The objective of this article is to devise a multistream spatiotemporal fusion GAN (STGAN) model for spatiotemporal satellite image fusion that can accommodate drastic temporal difference for improved temporal transferability. The proposed STGAN model aims to provide accurate prediction of fusion images even without temporally close images, achieving: 1) enhanced temporal transferability that allows the trained model to be applied across different temporal periods without retraining, and 2) a unique random sampling strategy during training that incorporates samples with varying temporal gaps. This approach enables the model to learn and accommodate diverse temporal patterns, making it more robust to different temporal gaps in the input data without requiring temporally close image pairs. To that end, STGAN employs a conditional GAN modeling structure that incorporates multistream input, which allows the model to better learn the temporal changes in surface reflectance from the high-temporal-resolution low-spatial-resolution images. In the meantime, the spatial features could be better retrieved on condition of the image of high spatial resolution, which has great potential in predicting land cover changes over time. By integrating spatial transformer and channel attention modules, STGAN can better capture and emphasize both spatial and temporal features that are important for satellite image fusion, leading to superior fusion results. Through a unique training strategy based on randomly selected reference images, our model is able to effectively accommodate significant temporal gaps, facilitating accurate predictions even when the image pair is temporally distant from the prediction date. Through the unique combination of conditional GAN with

multistream input, spatial transformer, channel attention, U-net, and random training strategy, the proposed STGAN structure possesses enhanced capabilities in establishing the mapping relationships between the low- and high-resolution images, accommodating the temporal changes in both spectral reflectance and spatial structures.

II. Data and Material

Two study sites are selected in this article, namely Coleambally Irrigation Area (CIA) and Lower Gwydir Catchment (LGC) with open-source satellite datasets [23]. Located in southern New South Wales, Australia, the CIA site encompasses an agricultural production area with a rice-based irrigation system. The major land cover type in the CIA site is irrigated agriculture, which forms a heterogeneous landscape by a great number of small crop fields with their rapid-changing temporal phenological dynamics during the growing season (Fig. S1). In addition, dryland agriculture and woodlands surround the irrigated agriculture system, with less variable temporal dynamics. Thus, the CIA site is especially appropriate for testing the ability of the fusion models to capture temporal phenological changes. The satellite dataset for the CIA site includes 17 image pairs (Table SI), with the acquisition time ranging from October 2001 to May 2002, corresponding to the summer growing season. However, the temporal gaps between available image pairs vary substantially in the CIA dataset. For instance, temporally proximate pairs may be acquired within a 9-day interval (image pairs available on 8th October 2001 and 17th October 2001), while distant pairs may span over 30 days (12th January–13th February 2022). This inherent variability in temporal sampling requires robust temporal transferability in fusion models. Each image pair in the dataset consists of a Landsat-7 ETM+ image with a 30-m spatial resolution and a MODIS MOD09GA image with a 250-m spatial resolution. Both Landsat and MODIS images are resampled to a 25-m spatial resolution, resulting in an image size of 1720 by 2040.

The LGC site is located in northern New South Wales, Australia. The LGC site encompasses a more natural landscape with a mixture of agriculture and woodlands. Due to a significant flood occurrence in late 2004, the LGC site is known for undergoing abrupt land cover changes before and after the flooding period with excessive wetness and inundated vegetation (Fig. S2). The LGC dataset consists of a total of 14 image pairs (Table SI), acquired between April 2004 and April 2005. Image pairs for the LGC site consist of 30-m Landsat-5 TM and 250-m MODIS MOD09GA products, with both products being resampled to a 25-m resolution. The image size for the LGC site is 3200 by 2720. For both CIA and LGC datasets, the satellite images include six bands, namely blue, green, red, near infrared (NIR), shortwave infrared 1 (SWIR1), and SWIR2.

III. Method

A. Preprocessing

An approach utilizing multiple data streams has been adopted in this article. The prediction of high-resolution images at the target date involves three input streams for generating coarse and fine feature maps as the input to the GAN-based deep-learning model: 1) low-resolution image at the target date, 2) low-resolution image from a randomly selected reference date within the dataset, and 3) high-resolution image from the selected reference date (see Fig. 1). These data streams are integrated to generate both fine and coarse feature maps, which serve as input for the deep learning model (as illustrated in Fig. 1).

The fine feature map and coarse feature map represent information toward spatial and temporal aspects. The fine feature map comprises high-resolution images and the difference between these high-resolution images and their low-resolution correspondent on the reference date. High-resolution images from the reference date capture spatial details of the landscape, whereas the contrast between high- and low-resolution images on the reference date conveys disparities in spatial characteristics between the two sets of images. On the other hand, the coarse feature map encompasses low-resolution images from the target date and the contrast between the low-resolution images on the target date and those from the reference date. The low-resolution image on the target date provides information on the temporal side, indicating the expected state of the landscape at the target time. The difference between the low-resolution image on the target date and the low-resolution image on the reference date represents how the landscape changes over time at a coarse resolution level. Both the fine and coarse feature maps are subsequently input into a GAN-based deep learning model to generate high-resolution images for the target date.

The STGAN model's temporal transferability is achieved through two critical design choices: randomized sampling in reference selection and a multistream input structure. First, the reference image is sampled randomly rather than sequentially, allowing the model to learn from a wide range of temporal intervals without dependence on specific sequences. This randomized sampling helps the model to generalize across temporally distant samples, reducing overfitting to short-term temporal patterns. Additionally, the multistream input design enables the model to capture spatial and temporal differences between the target time and randomly selected reference times, which enhances its adaptability to varied temporal intervals. The multistream input design provides the model with structural information under various levels of temporal changes, making STGAN effective in handling both short and long temporal gaps. Different from existing models, STGAN model does not rely on temporally close training samples and can adapt to distant intervals, achieving enhanced temporal transferability.

The input dataset is initially divided into training, testing, and validation datasets for the proposed deep-learning model. The training and testing datasets are employed in the construction of the machine-learning model, and the validation dataset serves exclusively for result validation. For each image within the training and testing datasets, 50 high-resolution tiles of size 256 pixels $*$ 256 pixels, representing an area of 7.68 km $\times$ 7.68 km, are randomly selected as the output. Each image's acquisition date is deemed as the target date, since the model aims to predict the high-resolution tiles. Their low-resolution correspondents of

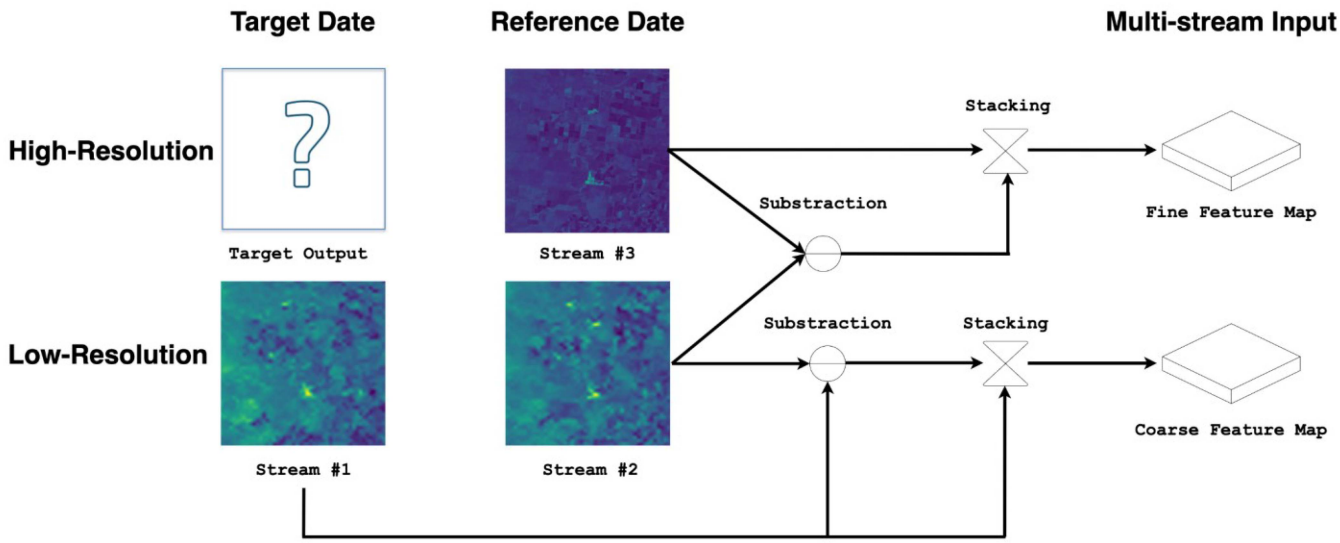


Fig. 1. Multistream Input of the proposed model.

the target date are then used as input, along with 256 pixels × 256 pixels high- and low-resolution tiles located at the same position of the images acquired on a randomly chosen reference date which is different from the target date.

Of the 50 tiles selected from each image, 10 are allocated for testing, while the remaining 40 are employed to train the deep-learning model. For images within the validation dataset, 50 high- and low-resolution tiles are extracted from the images of the validation date and a randomly selected reference date, respectively, following a similar procedure used to generate the training and testing datasets. The randomly selected reference dates ensure that the model encounters various temporal length between the target and reference dates, enhancing the model's capability to predict drastic temporal changes.

As both the CIA and LGC datasets possess six bands representing blue, green, red, near-infrared, short-wave infrared 1, and short-wave infrared 2, each fine and coarse feature map includes 12 bands. In the context of 17 image pairs within the CIA dataset and 14 image pairs within the LGC dataset, 850 images and 700 image tiles will be retrieved CIA and LGC dataset, respectively.

B. Model Structure

1) GAN: In our proposed model, a GAN-based model structure is utilized. This model is based on the deep generative models designed by Goodfellow et al. [24], which are specifically developed to address the generative modeling problem. Unlike traditional deep-learning models, the GAN model employs two neural networks, namely the generator (G) and discriminator (D), which compete with each other in a zero-sum game format. The generator is responsible for generating new data samples that are intended to resemble the target dataset, which in our study is represented by the high-resolution images. The discriminator, on the other hand, evaluates the authenticity of the generated data samples to determine whether they belong to the target dataset [24]. The GAN model represents a zero-sum game, with objective function as

$$\min_D \min_G V\ (D,G) = E_x\left[\ln\left(D\left(x\right)\right)\right] + E_z\left[\ln\left(1-D\left(G\left(x\right)\right)\right)\right] \quad (1)$$

where X is the random variable associated with the input data, while Z represents the random variable associated with the random noise. The generator (D) and discriminator (G) are two players engaged in a minimax game with a value function denoted as *V(D, G)*.

In the field of computer vision, GAN model has been extensively used for various applications, such as object detection [25], [26], image classification [27], image super-resolution [28], [29], and image-to-image translation [30]. Furthermore, this model structure has also been adopted by remote sensing scientists to address challenges in remote sensing image classification [31], hyperspectral image classification [32], [33], and spatiotemporal image fusion [18].

In our proposed study, a GAN model with multistream input is adopted to achieve spatiotemporal image fusion. The proposed structure of the two major components, namely the generator and discriminator, will be illustrated in the following section.

2) Generator: The generator component of our proposed model structure comprises of four major components, namely, 1) convolutional block, 2) spatial transformer module, 3) channel attention network, and 4) U-net module, as depicted in Fig. 2. The first component, the convolutional block, is responsible for extracting the spatial and spectral features of the coarse feature map and the fine feature map. Initially, the coarse and fine input streams are fed separately into the convolutional layer. The convolutional layer comprises a sequence of components, including a batch normalization layer, a Leaky Rectified Linear Unit (Leaky ReLU), a two-dimensional convolutional block, a subsequent batch normalization layer, and another Leaky ReLU. The output feature maps from both streams are then added to

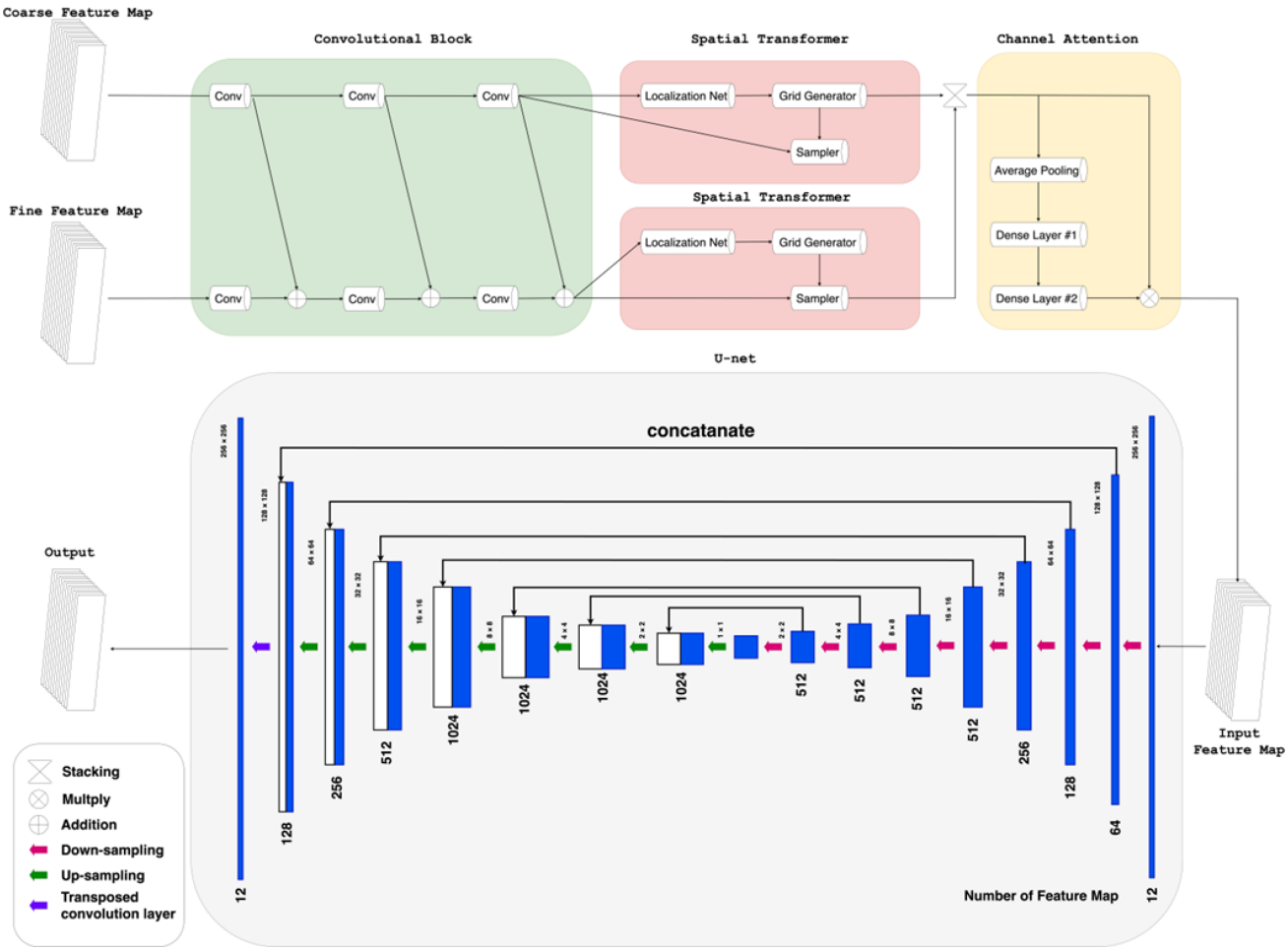


Fig. 2. Generator structure.

generate the new fine feature map to include both spatial details and temporal information generated from coarse feature map. This process is repeated thrice, following which the fine and coarse feature maps are extracted from the multi-stream inputs and fed into the spatial transformer module, separately.

The spatial transformer network is an insertable deep-learning module that helps to increase spatial invariance against various spatial transformations, like translation, scaling, etc. [34]. It has been widely used in object detection [35], image classification [36], image super-resolution [37], and other applications. It is considered a key component in our model, playing a critical role in enhancing model performance by effectively augmenting the spatial features of both coarse and fine feature maps. The spatial transformer mechanism is split into three components—localization net, parameterized sampling grid, and differentiable image sampling involved in the spatial transformer module, as shown in Fig. 2. The first localization network takes the input feature map, and through a number of hidden layers, outputs the parameters of the spatial transformation that should be applied to the feature map. Then, the predicted transformation parameters are used to create a sampling grid, which is a set of points where the input map should be sampled to produce the transformed output. Last, the feature map and the sampling grid are taken as inputs to the sampler, producing the output map sampled from the input at the grid points [34]. To take advantage of the spatial transformer network, we integrated the module in the proposed model structure to increase spatial invariance for fine and coarse feature maps, respectively.

In our proposed model structure, we have also incorporated a channel attention network to capture the temporal information and make full use of the temporal and spectral information in multitemporal remote sensing images [38]. The channel attention module involves a global average pooling layer and followed by two regular densely connected neural network layer. As different features may not contribute to the model output equally, the channel attention mechanism allows the model to highlight informative features that are more relevant to the reconstruction of the fusion imagery. Through dynamically adjusting the weights for different features, the attention block facilitates the model to adapt to different landscapes (e.g., the crop areas which are significant components of the landscape) to enhance the important features in the channel dimension.

The last component of the generator is the U-net module, which takes in the 12-band 256 pixels $\times$ 256 pixels feature map, generated from the convolutional block, spatial transformer module, and channel attention block to produce the

new data sample. The U-net architecture, initially designed by Ronneberger et al. [39], is a fully convolutional network used for biomedical image segmentation. The U-Net architecture serves as the final component, enabling complete information retrieval from the input feature map for image generation and subsequent presentation to the discriminator. Its architecture is shown in the bottom part of Fig. 2. The downsampling process, which represents the contracting path of the U-net model, involves a convolutional network consisting of convolutional layers, followed by batch normalization, ReLU, and a max pooling operation. The upsampling process, which represents the expansive pathway, combines the feature and spatial information through a sequence of upconvolutions and concatenations with high-resolution features from the contracting path. Each upsampling process involves transposed convolution layers, followed by batch normalization and ReLU. The output, as shown in Fig. 2, represents a 12-band image of size 256 pixels × 256 pixels.

Deep-learning models rely on the loss function to obtain the errors between the model prediction and reference data. In the context of spatiotemporal fusion, a simple loss function may not be sufficient for the optimization of all the parameters in deep-learning-based fusion models. For instance, it has been reported that the mean squared error (MSE) loss function tends to generate a comparatively smooth fusion image and often results in the loss of edge information [40], [41]. This is because MSE is highly sensitive to outliers and penalizes large errors. To that end, compound loss functions have been proposed to instruct the fusion models to improve the spatial sharpness in the predicted fusion images. The compound loss functions may integrate error metrics such as MSE, adversarial loss of GAN model, and a perceptual loss which measures loss through comparing the feature maps output from pretrained neural networks [18], [40]. The loss function of the proposed model consists of three main components, namely image loss, perceptual loss, and GAN loss. The total loss function is defined as follows:

$$L(G) = \alpha L_{\text{GAN}} + \beta L_{\text{image}} + \gamma L_{\text{perceptual}} \tag{2}$$

where α, β, and γ are corresponding weight coefficients. The GAN loss follows the form of conventional loss function for GAN model, which helps ensure that statistical distribution of the output image is comparable to that of the reference image. The image loss is designed to measure the difference between the ground truth and the model prediction. Specifically, the image loss consists of three aspects, namely MSE, average difference (AD), and spectrum loss. The MSE and AD are utilized to quantify the difference in reflectance between ground truth images and predicted images. The spectrum loss is constructed based on spectral angle mapper, which measures the cosine similarity between the model prediction and ground truth

$$L_{\text{image}} = L_{\text{mse}} + L_{\text{AD}} + L_{\text{spectrum}} \tag{3}$$

$$L_{\text{mse}} = \frac{1}{N}\sum_{n=1}^{N}(F-R)^2 \tag{4}$$

$$L_{\text{AD}} = \frac{1}{N}\sum_{n=1}^{N}(F-R) \tag{5}$$

$$L_{\text{spectrum}} = I - \frac{F \cdot R}{\| F \| \| R \|} \tag{6}$$

where F is the model prediction, R stands for the ground truth reference image, and N represents the number of pixels in the image.

The perceptual loss is constructed with a pretrained visual geometry group (VGG) model. The VGG network can effectively extract high-level features from the images. Though minimizing the differences between the abstract high feature level, the fusion model can better reconstruct spatial details. The perceptual loss is defined as follows:

$$L_{\text{perceptual}} = \frac{1}{N}\sum_{n=1}^{N}\left[f_{\text{vgg}}(F) - F_{\text{vgg}}(R)\right]^2. \tag{7}$$

3) Discriminator: The discriminator structure employed in this model, commonly known as PatchGAN, was previously proposed in [42] to capture local style statistics. As a popular decimator for GAN model, PatchGAN has been used in GAN-based image-to-image translation models, such as pix2pix [30], and image restoration model [43].

The structure of the discriminator is illustrated in Fig. 3. Initially, the first six bands of the generated image output from the generator are concatenated with the target image, and the resultant image is passed through three consecutive downsampling modules. The downsampling module in the discriminator, similar to that in the generator, consists of a convolutional network comprising convolutional layers, batch normalization, ReLU activation, and max pooling operation. Following downsampling, the module is then fed into two combinations of zero-padding modules and CNNs, which generates a matrix that determines whether the generator output can deceive the discriminator.

The discriminator loss is shown as follows:

$$r_{\text{loss}} = \text{sigmoid\_cross\_entropy}\left(\text{target}_{\text{img}},\ I\right) \tag{8}$$

$$g_{\text{loss}} = \text{sigmoid\_cross\_entropy}\left(\text{generated}_{\text{img}},\ I\right) \tag{9}$$

$$\text{rmse}_{\text{loss}} = \sqrt{\frac{1}{N}\sum_{n\,=\,1}^{N}\left(r_{\text{loss}} - g_{\text{loss}}\right)^2} \tag{10}$$

$$r_{\text{loss}} = r_{\text{loss}} + g_{\text{loss}} + \lambda \text{rmse}_{\text{loss}} \tag{11}$$

where I is an array of ones, r_{loss} and g_{loss} are the sigmoid cross-entropy losses of the target image and the generated images, respectively. LAMBDA (λ), defined as 100, is a hyperparameter that balances the contribution of the generator and the discriminator in the overall loss function. The discriminator outputs are used to distinguish the fake image that comes from the generator.

C. Evaluation Matrices

To evaluate the spatial and spectral accuracy of our model in comparison with benchmark models, we have selected five evaluation metrics for model evaluation and comparison in this article, namely root mean square error (RMSE), local binary patterns (LBP), spectral angle mapping (SAM) [44], relative dimensionless global error synthesis (ERGAS) [45], and structural similarity index measure (SSIM) [46], where RMSE, SAM, and

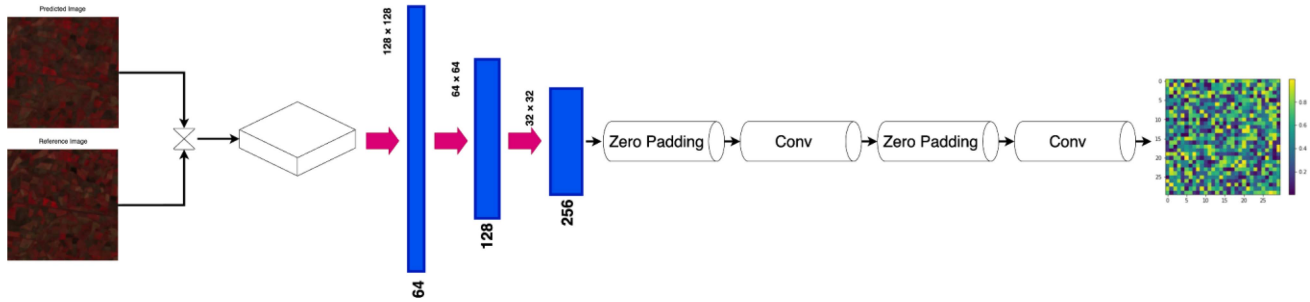


Fig. 3. Discriminator structure.

ERGAS are often used to evaluate the spectral accuracy, while LBP and SSIM are used for spatial accuracy [47]. Among the five evaluation metrics, a larger value indicates better performance for SSIM, while smaller values indicate better performance for RMSE, SAM, ERGAS, and LBP. The formulas for these five metrics are listed as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(F_i - R_i)^2}{N}} \tag{12}$$

where *Fi* is the value of pixel *i* in the generated image, *Ri* is the value of pixel *i* in the target reference image, and *N* is the total number of pixels. And RMSE is one of the most commonly used metrics for evaluating the spectral accuracy for the fused images.

SAM is employed to quantify the spectral distortion between the predicted image and the reference image across multiple bands. The spectral characteristics of each pixel are represented as an *N*-dimensional spectrum vector. The degree of spectral distortion is determined by measuring the angle between the *N*-dimensional vectors of the predicted and reference spectra for each pixel. Lower SAM values indicate that the spectral information of the predicted image is more closely resembles that of the reference image

$$\text{SAM} = \frac{1}{N}\sum_{n=1}^{N}\arccos\frac{F_i \cdot R_i}{\| F_i \| \| R_i \|}. \tag{13}$$

ERGAS evaluates the spectral disparity between the predicted and reference images by utilizing normalized RMSE across bands. It is capable of accommodating variations in reflectance values across different bands and accounting for resolution differences between fine and coarse images

$$\text{ERGAS} = 100\frac{h}{l}\sqrt{\frac{1}{B}\sum_{b=1}^{B}\frac{\sum_{i=1}^{N}(F_{b,i} - R_{b,i})^2}{N\mu_b^2}} \tag{14}$$

where *h* represents the spatial resolution of fine images and *l* represents that of the coarse images. *B* stands for the total number of the bands in the multispectral images and μ_b stands for the mean value of band *b* in the model predicted image.

SSIM evaluates the similarity of the overall spatial structures between the predicted and ground truth images, where a higher SSIM value indicates better spatial quality in the fusion results. The equation for SSIM calculation is listed as follows:

$$\text{SSIM}\ (x,\ y) = \frac{\left(2\mu_x\mu_y + (K_1L)^2\right)\left(2\sigma_{xy} + (K_2L)^2\right)}{\left(\mu_x^2 + \mu_y^2 + (K_1L)^2\right)\left(\sigma_x^2 + \sigma_y^2 + (K_2L)^2\right)} \tag{15}$$

where μ represents the sample mean and σ denotes the sample variance. *L* signifies the dynamic range of the pixel-values, while the default values for k_1 and k_2 are 0.01 and 0.03, respectively.

LBP is a spatial metric that measures the patterns through characterizing local relationships between a central pixel and its neighboring pixels

$$\text{LBP} = \text{decimal}\,(d_1 d_2 \cdots d_8) \tag{16}$$

$$d_i = \begin{cases} 1, & \text{if } D_i > D_c \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

where D_i is the value of pixels surrounding the central pixel in a 3×3 moving window, D_c is the value of the central pixel, and the decimal function is used to convert the binary digits to a decimal number. LBP is a visual descriptor for classification in computer vision by analyzing the local pixels around a point. It has been used as an effective feature for texture classification and is selected as a metric for evaluating the spatial accuracy of the fused images.

D. Computational Support

The training process of the proposed STGAN utilizes high-performance computing (HPC) resources, specifically the Blue Waters supercomputer from the National Center for Supercomputing Applications at the University of Illinois Urbana-Champaign [48]. This efficient use of HPC resources allows us to manage the model's complexity effectively for the most intensive part of our approach, i.e., the training phase, ensuring robust and timely training outcomes. Leveraging these HPC resources, including NVIDIA K20X GPU accelerators with memory of 6 GB, training of the STGAN model with 1000 steps takes approximately 325 s, and our full training process encompasses 20 000 steps, totaling around 108 min. The training time per 1000 steps for GANSTFM, HPLTS-GAN, and enhanced deep convolutional spatiotemporal fusion network (EDCSTFN) is 220, 390, and 140 s, respectively.

Once trained, STGAN's prediction process is designed to be highly efficient, requiring minimal computing resources that

can be easily handled on standard personal computers and laptops. This characteristic enhances the model's practicality for real-world applications, combining the advantages of a sophisticated, pretrained model with low computational demands for prediction. Moreover, due to the model's innovative randomized training strategy and enhanced temporal transferability, it can be used to predict data across different dates within the same location, without the need to retrain for each specific prediction date. This enhanced temporal transferability not only reduces computational costs but also broadens the model's applicability in dynamic and rapidly changing environments.

IV. Result

In this section, we employ the multistream STGAN model proposed in this article to conduct quantitative evaluations and visual inspection on two datasets, namely the CIA dataset and LGC dataset. Subsequently, the performance of our method is assessed by comparing it with three state-of-the-art models, namely 1) flexible spatiotemporal data fusion (FSDAF), which is robust weight-function-based spatiotemporal fusion model [8]; 2) GAN-based spatiotemporal fusion model [16]; 3) HPLTS-GAN [21]; and 4) EDCSTFN, which is built and involves a CNN machine-learning model structure on the aforementioned datasets [15]. Furthermore, we explore the generalizability and scalability of the proposed algorithm by contrasting the use of random reference data with the nearest reference data.

A. Evaluation of CIA Area

1) Quantitative Evaluation: The performance of our proposed model is initially assessed on the CIA dataset. To establish a comprehensive evaluation, we compare our model with three prominent state-of-the-art remote sensing image fusion methods, namely FSDAF, GANSTFM, HPLTS-GAN, and EDCSTFN. To ensure fairness and consistency, we reproduce and fine-tune the best results achieved by these competitive methods, utilizing the publicly available source code provided by the authors.

For the CIA dataset, a set of five validation dates was selected as the last five tiles in the dataset: 04/02/2002, 04/11/2002, 04/18/2002, 04/27/2002, and 05/04/2002. Before partitioning the remaining dataset into training and testing data, the validation tile is separated. The results of the validation model are presented in Table I. The best results are highlighted in bold in Table I for each validation date.

Concerning spectral accuracy, our proposed STGAN model outperforms the benchmark models in RMSE and ERGAS. Specifically, our model consistently achieves an RMSE of approximately 0.03, surpassing the benchmark models by a significant margin across all five selected validation dates in the CIA dataset. Similarly, for ERGAS, our proposed model performs the best among all benchmark models for all validation dates and surpasses the other models by a large margin, except for 4 May where our proposed model performs second to FSDAF. Among the benchmarks, HPLTS-GAN and EDCSTFN demonstrate relatively better performance compared to the other two models. Additionally, with the exception of the validation date of 11 April, where the SAM of the EDCSTFN model is lower than our proposed model, our model outperforms the benchmarks in all four other validation dates for SAM.

TABLE I
Evaluation Result for CIA Area With Proposed and Benchmark Spatiotemporal Fusion Model in 2002

Validation Date	Model	RMSE	ERGAS	SAM	SSIM	LBP
Apr. 2	FSDAF	0.0336	1.1245	0.0858	0.9450	0.1246
	GANSTFM	0.0511	1.5034	0.1310	0.9084	0.1210
	HPLTS-GAN	0.0405	1.2400	0.1220	0.9370	0.1290
	EDCSTFN	0.0435	1.2216	0.0946	0.9415	0.1202
	STGAN	**0.0316**	**1.0118**	**0.0820**	**0.9567**	**0.1196**
Apr. 11	FSDAF	0.0331	1.3001	0.0995	**0.9390**	0.1259
	GANSTFM	0.0504	1.9117	0.1327	0.8524	0.1235
	HPLTS-GAN	0.0420	1.3100	0.1220	0.9250	**0.1220**
	EDCSTFN	0.0371	1.2896	**0.0897**	0.9357	0.1233
	STGAN	**0.0297**	**1.2743**	0.0958	0.9310	0.1227
Apr. 18	FSDAF	0.0392	1.5012	0.1170	0.9200	0.1330
	GANSTFM	0.0469	1.8808	0.1594	0.8849	0.1241
	HPLTS-GAN	0.0407	1.5440	0.1316	0.9185	0.1265
	EDCSTFN	0.0375	1.4262	0.1088	0.9239	0.1269
	STGAN	**0.0354**	**1.4061**	**0.0936**	**0.9287**	**0.1220**
Apr. 27	FSDAF	0.0398	1.2264	0.1050	0.9480	0.1310
	GANSTFM	0.0470	1.5795	0.1341	0.9074	**0.1296**
	HPLTS-GAN	0.0368	1.2451	0.1192	0.9428	0.1303
	EDCSTFN	0.0382	1.1859	0.1037	0.9517	0.1348
	STGAN	**0.0311**	**1.0377**	**0.0799**	**0.9601**	0.1351
May 4	FSDAF	0.0354	**0.9280**	0.0954	0.9260	0.1390
	GANSTFM	0.0493	1.6259	0.1269	0.8963	**0.1334**
	HPLTS-GAN	0.0380	1.2306	0.1138	0.9400	0.1386
	EDCSTFN	0.0387	1.2217	0.0985	0.9426	0.1384
	STGAN	**0.0304**	1.1479	**0.0784**	**0.9472**	0.1345

In terms of spatial accuracy, our model generates competitive results compared with the benchmark models. For LBP, our model achieves the best performance on April 2 and 18. GANSTFM model performs the best on the last two validation dates, April 27 and May 4. HPLTS-GAN performs the best on Apr. 11 (0.1220), followed by STGAN (0.1227). Regarding SSIM, our proposed model attains values around 0.95 for all validation dates, except for April 11, where the SSIM of the EDCSTFN model (0.9357) marginally outperforms our proposed STGAN model (0.0958). For the remaining validation dates, our proposed model generates superior SSIM values.

Analyzing each one of the five validation dates, we observe that our proposed model consistently outperforms the benchmark models across all five evaluation metrics on 2 April and 18 April for CIA area. However, on the validation date of 11 April, the proposed model falls slightly behind the EDCSTFN model in terms of SSIM and ERGAS, although the differences between the two models are relatively marginal. On 18 April and 27 April, the only metric where our proposed model fails to match the performance of the benchmark models is LBP. Nonetheless, the disparity in LBP values between the best-performing GANSTFM model (0.1296) and our proposed model (0.1351) is minimal on 27 April. Similarly, the LBP values for our proposed model (0.1334 and 0.1345) closely align with those of the GANSTFM model on 4 May. Excluding LBP, our proposed model consistently outperforms the benchmark models across all other indicators on two validation dates.

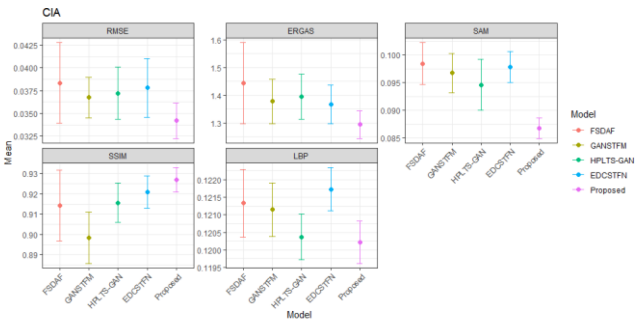


Fig. 4. Mean and standard deviation of the quantitative accuracy evaluations for the fusion images on 11 Apr. 2002 generated from different reference dates in the CIA site. Dots represent the mean value and the error bars represent standard deviations.

An additional advantage observed in our proposed model is its consistent performance across different dates. The model consistently generates good results in both spatial and spectral evaluations, as indicated the following mean values and standard deviations: RMSE (0.03255±0.00285), SSIM (0.94395±0.01615), SAM (1.20895±0.19715), ERGAS (0.0892±0.0076), and LBP (0.12735±0.0075). This consistency highlights the enhanced robustness of our model compared to the benchmark models, particularly when dealing with datasets with varying time intervals between images.

To further evaluate the robustness of our proposed STGAN model, we generated fusion results (prediction date on 11 Apr. 2002 as an example) with different reference dates. Fig. 4 presents the mean and standard deviation of the five accuracy metrics for STGAN and all the benchmark models for the CIA site. Through a comparative analysis involving the benchmark models FSDAF, GANSTFM, HPLTS-GAN, and EDCSTFN, our findings demonstrate a substantial and significant superiority of our proposed model in terms of the model robustness. In all five evaluation metrics, namely RMSE, ERGAS, SAM, SSIM, and LBP, our proposed model consistently outperforms the other benchmark models with more preferable mean values of the metrics. The performance of the remaining four benchmark models remains relatively comparable, except that HPLTS-GAN shows superior performance in LBP compared to the other three benchmark models. The standard deviations (error bars in Fig. 4) also tend to be lower for the proposed STGAN compared to the three benchmark models, suggesting that STGAN is less affected by the images acquired on dates that are far from the prediction dates.

Overall, the observed consistency, robustness, and scalability of our proposed model are valuable attributes, especially when dealing with large datasets spanning multiple decades and varying time intervals between images. Consequently, based on the analysis of the selected validation dates, we argue that our proposed model consistently outperforms the benchmark models in terms of both spatial and spectral accuracy across all the evaluated validation dates in the CIA dataset.

2) Visual Inspection: Fig. 5 displays the synthesized fusion results obtained using the benchmark models FSDAF, GANSTFM, HPLTS-GAN, EDCSTFN, as well as our proposed STGAN model, for the CIA area. The visual demonstration focuses on false color composite composed of the NIR, red, and green bands. The first row provides an overview of the entire CIA area, while the second row presents zoomed-in details of a highlighted region in the dashed yellow box. The third row depicts the error map obtained by comparing the fusion results with the reference. Upon visual inspection, we observe that the error distributions for the FSDAF and EDCSTFN models are noticeably larger compared to both the GANSTFM, HPLTS-GAN, and our proposed STGAN model. And the error distribution of our proposed STGAN model appears to be the best and most consistent among the four models. Notably, our STGAN model exhibits particularly favorable performance in "red" regions, which are agricultural fields during the crop vegetative stages.

Through both quantitative evaluation and visual examination, it is evident that our proposed STGAN model can effectively compete with and even outperform existing benchmark spatiotemporal fusion models.

3) Ablation Study: To validate the architectural design of the proposed STGAN model, we conducted a comprehensive ablation study to quantitatively assess the contribution of each key component—spatial transformer, channel attention, and U-net module. The experiments utilized CIA data from 2nd April 2002, with each architectural variant trained for 2000 steps under identical experimental conditions and RMSE as the evaluation metric.

The baseline STGAN model achieved an RMSE of 0.0355, establishing the benchmark for comparison. Removal of the spatial transformer module, designed to augment the spatial features of both coarse and fine feature maps, increased the RMSE to 0.0371. Similarly, eliminating the channel attention module, which captures the temporal features, resulted in a larger RMSE of 0.0386. To investigate the impact of U-net layers, we conducted experiments removing two and four layers from the U-Net structure, both yielding larger RMSEs of 0.0386 and 0.0361, compared with the baseline STGAN model.

The ablation study demonstrates that each component of the proposed STGAN model, including the spatial transformer module, channel attention module, and the number of U-Net layers, makes a distinct and significant contribution to the model's performance, with the complete STGAN architecture outperforming its ablated variants in spatiotemporal fusion. Component removal generally leads to decreased performance of the model. These findings provide support for our architectural choices and validate the necessity of each component in the proposed STGAN design.

B. Evaluation of LGC Area

1) Quantitative Evaluation: Similar analysis was performed on another location, namely the LGC area, to evaluate the performance of our proposed STGAN model. The same strategy for CIA dataset is applied to the LGC dataset and select the last four times in the dataset: 01/29/2005, 02/14/2005, 03/02/2005, and 04/03/2005 as the validation date. The results of the validation model for the LGC dataset are presented in Table II. As with

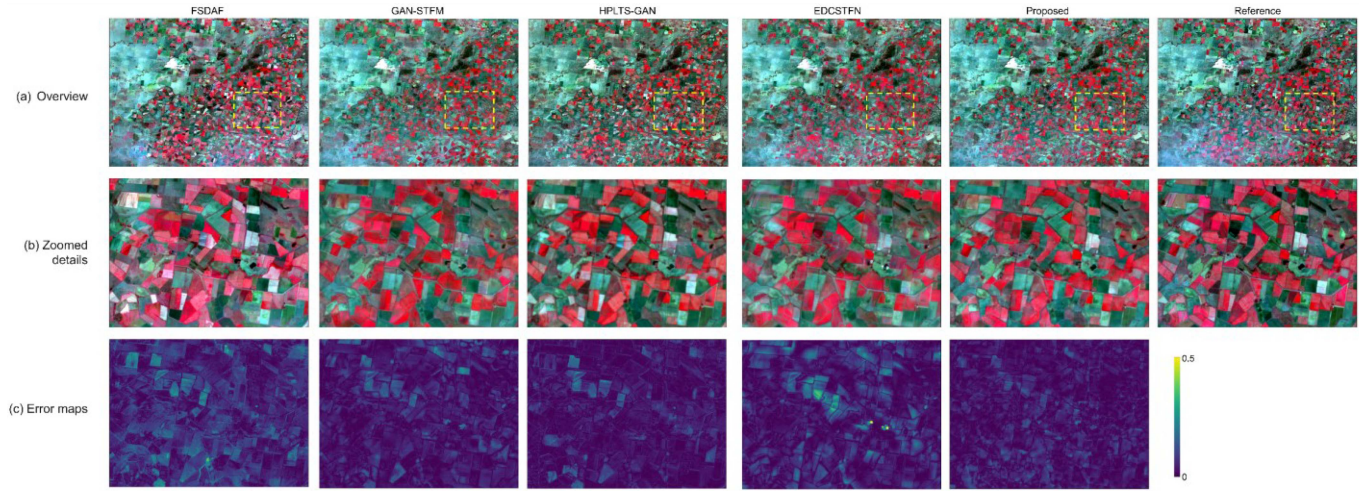


Fig. 5. Fusion results for the CIA area on 11 Apr. with different spatiotemporal fusion methods.

TABLE II
EVALUATION RESULT FOR LGC AREA WITH PROPOSED AND BENCHMARK SPATIOTEMPORAL FUSION MODEL IN 2005

Validation Date	Model	RMSE	ERGAS	SAM	SSIM	LBP
Jan. 29	FSDAF	0.0379	1.1770	**0.0516**	0.9454	0.1480
	GANSTFM	0.0299	0.9865	0.0681	**0.9669**	0.1450
	HPLTS-GAN	0.0357	1.2074	0.0528	0.9468	0.1378
	EDCSTFN	0.0356	1.1883	0.0694	0.9496	0.1412
	STGAN	**0.0280**	**0.9352**	0.0665	0.9551	**0.1369**
Feb. 14	FSDAF	0.0344	0.9180	**0.0451**	0.9586	0.1540
	GANSTFM	0.0251	0.8886	0.0628	0.9580	0.1533
	HPLTS-GAN	0.0298	1.1636	0.0575	0.9482	0.1367
	EDCSTFN	0.0270	0.8265	0.0564	0.9698	**0.1501**
	STGAN	**0.0225**	**0.7762**	0.0515	**0.9735**	0.1510
Mar. 2	FSDAF	0.0407	1.1000	**0.0483**	0.9469	0.1610
	GANSTFM	**0.0264**	0.8377	0.0590	**0.9735**	0.1563
	HPLTS-GAN	0.0282	1.1541	0.0553	0.9479	0.1523
	EDCSTFN	0.0344	1.0630	0.0584	0.9465	0.1643
	STGAN	0.0296	**0.7877**	0.0558	0.9703	**0.1504**
Apr. 3	FSDAF	0.0448	1.3010	0.0592	0.9394	0.1720
	GANSTFM	0.0262	0.8359	0.0591	0.9593	0.1742
	HPLTS-GAN	0.0290	1.2015	0.0568	0.9426	**0.1669**
	EDCSTFN	0.0332	0.9995	0.0554	0.9575	0.1797
	STGAN	**0.0262**	**0.8222**	**0.0514**	**0.9692**	0.1673

Table I, the best results obtained on each validation date are highlighted in bold.

Comparing the results obtained for the CIA area, the findings for the LGC area further demonstrate the superiority of our proposed fusion model in both spectral and spatial accuracy (Table II). In terms of spectral accuracy, our proposed model significantly outperforms the other benchmark models both in RMSE and ERGAS. Analyzing the RMSE, our proposed model emerges as the best performing model on the validation dates of 29 January, 14 February, and 3 April. And our model outperforms all other benchmark models for ERGAS. Furthermore, our proposed model exhibits exceptional spatial accuracy in the LGC area. Across all four selected validation dates, our model consistently provides competitive results in terms of the spatial-related evaluation metrics: SSIM and LBP. Specifically, our model achieves the highest performance on 14 February and

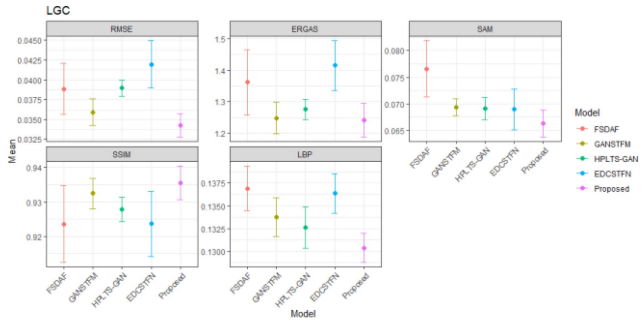


Fig. 6. Mean and standard deviation of the quantitative accuracy evaluations for the fusion images on 14 Feb. 2005 generated from different reference dates in the LGC site. Dots represent the mean value and the error bars represent standard deviations.

3 April based on the SSIM evaluation metric. Our model remains the top-performing model overall across all validation dates when evaluated by LBP, while EDCSTFN and HPLTS-GAN also demonstrate satisfactory performance. Similar to the results obtained for the CIA dataset, all the evaluated indicators on the LGC dataset demonstrate a high level of consistency across each validation date.

In summary, the quantitative evaluation results of the LGC area demonstrate the superior spectral performance of our proposed spatiotemporal fusion model compared to existing benchmark models. Additionally, our model exhibits remarkable spatial accuracy specifically in the LGC area.

Fig. 6 presents the mean and standard deviation of the five accuracy metrics for STGAN and all the benchmark models for the LGC dataset. Similar to the result from CIA dataset, the findings demonstrate consistent superiority of our models over the benchmark models across all indicators in the LGC site. Although the proposed model exhibits comparable performance to GANSTFM in terms of ERGAS, it significantly outperforms the other benchmark models in terms of the mean values of RMSE, SAM, SSIM, and LBP. While FSDAF presents satisfactory evaluation results for SAM (Table II), its performance

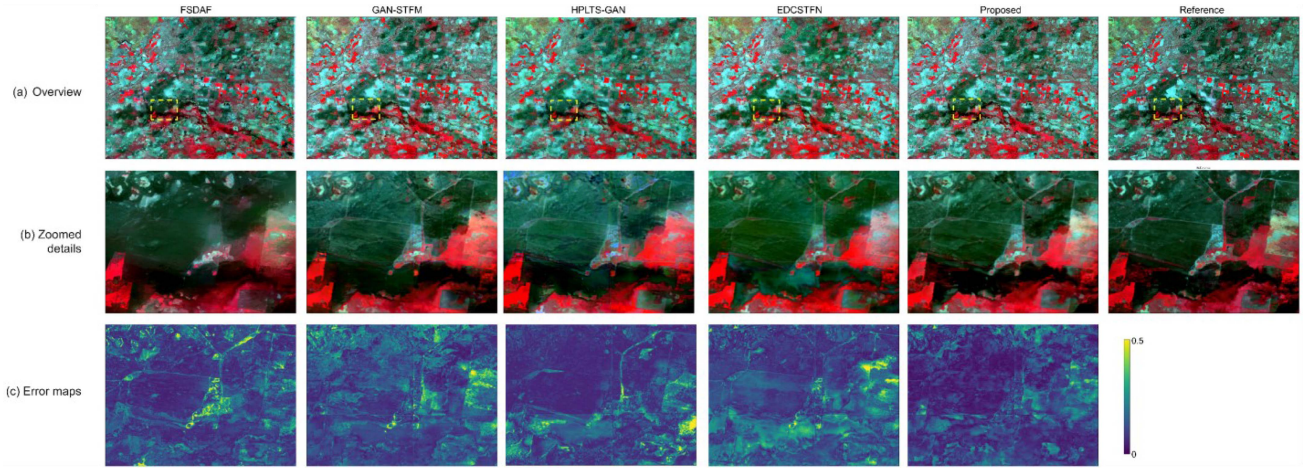


Fig. 7. Fusion results for the LGC area on 14 Feb. with different spatiotemporal fusion methods.

is significantly affected by the distance between the prediction date and the reference that is used to generate the fused images. Notably, the results from our proposed model exhibit enhanced consistency, as indicated by a smaller variance (i.e., error bars in Fig. 6) compared to the benchmark models.

2) Visual Inspection: Similar to the CIA area, a visual examination of the fusion results in the LGC area is presented in Fig. 7, contrasting the predicted images generated by the FSDAF, GANSTFM, HPLTS-GAN, EDCSTFN, and our proposed STGAN models. Upon comparing it with the benchmark model, we observe that our proposed model achieves improved accuracy in terms of error distribution, particularly in edge areas where landscape transitions occur. It is also noticeable that the STGAN model presents lower errors in the flooded areas [the dark strip in Fig. 7(b)], suggesting that STGAN is advantageous in capturing drastic changes that caused substantial landscape and spectral changes among images.

While the weight-function-based FSDAF model may also predict flooded areas well, it may not fully accommodate the edges and fine-scale features in the images. The comparison demonstrates a better ability of STGAN to capture spatial changes in the landscape and effectively minimize spatial inconsistencies. Visually, the general error metrics of our proposed STGAN model exhibit noticeable improvement over the benchmark model.

In conclusion, we conducted experiments on the LGC dataset to evaluate the performance of our proposed model in a different location and compared it with existing benchmark models. The findings from the LGC area align with those from the CIA area, reinforcing the high competitiveness of our model in image fusion compared to the benchmark models. This suggests that our proposed model has the potential for broad applicability across diverse landscapes.

C. Comparison Between Nearest and Random Reference Strategy

The selection of reference images plays a crucial role in determining the prediction accuracy of existing spatiotemporal fusion models. Typically, these models rely on acquiring reference image pairs with acquisition dates that are close to the prediction date. However, in this article, we introduce a key feature: the randomization of reference tile selection. This randomized reference selection becomes essential, particularly for large-scale analysis and image sets without timestamps.

This section aims to evaluate whether our proposed STGAN model can effectively handle randomization reference instead of the traditional nearest reference selection strategy. To achieve this, we compare the performance of our STGAN model using randomized and nearest reference strategies. Specifically, we seek to answer two key questions: 1) How much sacrifice in performance is incurred when transitioning from the nearest reference strategy to the random strategy with proposed STGAN model? 2) Which aspect, spatial or spectral accuracy, is more affected by the randomization strategy, and which evaluation matrices are most impacted? By addressing these questions, we can gain insights into the impact of randomization reference on the performance of our STGAN model and determine whether spatial or spectral accuracy is more influenced, along with identifying the most affected evaluation indicators.

We compare the nearest strategy and random strategy using the following equation:

$$\Delta I = I\,(\text{date},\ \text{nearest}) - I\,(\text{date},\ \text{random}) \tag{18}$$

$$\%I\ = \Delta I / I\,(\text{date}, \text{nearest}) \tag{19}$$

where I is the evaluation matrices, including RMSE, SSIM, SAM, ERGAS, and LBP. ΔI denotes the absolute differences in evaluation metrics between the closest strategy and a random strategy, while $\%I$ represent the percentage change of ΔI relative to the value obtained for the nearest strategy.

Tables III and IV present the results of performance variation observed when comparing the random reference strategy with the nearest reference strategy for the proposed STGAN model in the CIA and LGC areas, respectively. Among the five evaluation metrics, SSIM and LBP are minimally affected by the transition from the nearest strategy to the random strategy, with percentage changes of less than 0.05 for both the CIA and LGC areas. Notably, the transition from the nearest strategy to

TABLE III
PERFORMANCE CHANGE USING RANDOM REFERENCE STRATEGY COMPARED WITH THE NEAREST REFERENCE STRATEGY FOR PROPOSED STGAN MODEL IN CIA AREA IN 2002

Validation Date	ΔRMSE	ΔERGAS	ΔSAM	ΔSSIM	ΔLBP
Apr. 2	-0.0038	0.0041	-0.0128	0.0029	0.0033
Apr. 11	-0.0013	-0.0035	-0.2789	0.0079	-0.0001
Apr. 18	-0.0026	0.0025	-0.2009	-0.0514	-0.0009
Apr. 27	-0.0052	-0.0034	-0.062	0.0039	0.0015
May 4	-0.0052	-0.0051	-0.175	0.0154	0.0075
Average	-0.00362	-0.00108	-0.14592	-0.00426	0.00262
%Average	-0.115145148	-0.013485851	-0.117318653	-0.004701781	0.020203799

TABLE IV
PERFORMANCE CHANGE USING RANDOM REFERENCE STRATEGY COMPARED WITH THE NEAREST REFERENCE STRATEGY FOR PROPOSED STGAN MODEL IN LGC AREA IN 2005

Validation Date	ΔRMSE	ΔERGAS	ΔSAM	ΔSSIM	ΔLBP
Jan. 29	-0.0024	-0.0032	-0.0119	0.0092	-0.0012
Feb. 14	-0.0023	-0.006	-0.0739	0.0033	-0.0111
Mar. 2	-0.0022	-0.0154	-0.0546	0.0036	0.0016
Apr. 3	0.0028	0.0051	-0.0327	0.0003	0.0003
Average	-0.001025	-0.004875	-0.043275	0.0041	-0.0026
%Average	-0.038847651	-0.085347257	-0.054254762	0.004260514	-0.017460993

a random strategy surprisingly yields a slight positive effect on the evaluation outcomes pertaining to LBP and SSIM in the CIA dataset. Conversely, the remaining three evaluation metrics, namely RMSE, SAM, and ERGAS, are relatively more influenced by the adoption of the random reference strategy.

The influence of transitioning from the nearest reference strategy to a random one on RMSE, ERGAS, and SAM exhibits location-dependent characteristics. In the case of the LGC dataset, the impact on RMSE and SAM is relatively insignificant, with percentage changes of -0.0388 and -0.0543, respectively. However, for the CIA dataset, the impact is relatively substantial, with absolute percentage changes exceeding 0.1. In terms of ERGAS, the impact on the CIA dataset is relatively small (-0.013) compared to that observed for the LGC dataset (-0.085). Hence, it can be inferred that the effects on RMSE, ERGAS, and SAM are contingent upon the specific study location. Additionally, the supplementary materials also indicate that the adoption of the nearest strategy in our proposed model does not generate a substantial negative impact especially when compared to other benchmark models. Overall, we maintain that the transition from the nearest to the random reference strategy does not exert a significant influence on the STGAN model's evaluation metrics, suggesting that the associated errors can be managed effectively.

V. DISCUSSION

This article introduces a multistream STGAN that is specifically designed to enhance the temporal transferability in an effort to accommodate the drastic changes among available satellite images. The innovative integration of multistream inputs, spatial transformer modules, channel attention networks, and U-net architecture collectively enhances the model's ability to learn spatial and temporal patterns, while relaxing the amount of image pairs required for accurate spatiotemporal image fusion in previous fusion models. Compared to the four benchmark models, our proposed STGAN model demonstrates superior performance when evaluated by a comprehensive collection of spectral and spatial accuracy metrics. Parameter analysis further confirms that all the model components contribute significantly to the superior model performance.

STGAN not only achieves lower errors but also performs more stably (i.e., lower standard deviations) when using different image pairs to predict the same date, as shown in Figs. 4 and 6. Such stable performance suggests that STGAN is equipped with enhanced ability to accommodate varying levels of temporal changes. This advantage of enhanced temporal transferability is further reflected in the comparison between nearest and random reference strategy (Tables III and IV). Through the randomized reference strategy, we aim to introduce varying levels of temporal changes between images in the model training process. The comparable performance of STGAN under both randomized and nearest reference strategies highlights its robust capability to accommodate substantial temporal variations.

The enhanced robustness and temporal transferability can be attributed to the randomized design of our model during training with the reference dataset. The randomized design of our model offers several advantages. First, it enables the model to learn and accommodate various levels of temporal changes, resulting in consistently high-quality fused images across different prediction dates. Second, it demonstrates robustness by enabling the model to handle different edge cases and generate reasonable results for such scenarios. Last, the scalability of our model is noteworthy, as the machine-learning algorithm only needs to be run once per location to generate a pretrained model applicable for different times. Subsequently, batch image generation in the same location for different prediction dates can be performed without requiring further training.

STGAN has great potential to enhance environmental monitoring applications with its substantial improvements in temporal transferability. For instance, temporally dense imagery generated by STGAN can make significant contributions to monitoring land cover changes, vegetation health, agriculture productivity, and water bodies over time. For instance, literature has found that accurate fusion images are conducive to better characterization of vegetation phenology and early-season mapping of crop species [2], [49], [50]. Furthermore, disaster management could leverage STGAN for assessing areas impacted by disturbance events, such as floods, wildfires, or hurricanes, by providing accurate fusion images with high spatial and temporal resolutions before and after such events to support rapid response to natural disasters [51], [52].

With the rapid growth of satellite missions and the increasing availability of remote sensing datasets, there is a growing demand for innovative models to integrate and synthesize multisource satellite data. The proposed STGAN model, with its novel architecture, exemplifies this trend by effectively capturing complex spatiotemporal patterns from multisource satellite images, with a great potential to incorporate more

recent satellite missions, such as Sentinel-2 and PlanetScope. While STGAN exhibits promising performance with enhanced temporal transferability, the training process is designed to accommodate specific satellite image scenes. Thus, applying STGAN across extended geographic regions may present uncertainties given the diverse spatial complexity and various temporal change patterns in different regions. Recent advances in deep learning present compelling opportunities to further enhance the scalability of spatiotemporal fusion models [41], [53]. New deep-learning model structures, such as transformers, have emerged as powerful tools with enhanced feature extraction and sequence modeling capabilities, particularly in capturing complex spatiotemporal patterns from multisource satellite images [49], [54]. Recent discoveries in self-supervised learning strategies, such as masked autoencoders and contrastive learning approaches, offer significant potential for enabling models to learn robust representations from multi-source imagery across not only prolonged temporal periods but also extended geographic regions [55], [56]. Built upon these ideas, geospatial foundation models leverage large-scale training on diverse and complex spatiotemporal patterns, enabling them to learn generalized and transferable features that span wide geographic areas and extended temporal scales [57], [58], [59]. Future efforts may be devoted to leveraging geospatial foundation models to build scalable frameworks that further enhance the scalability of spatiotemporal fusion models across space and time.

VI. Conclusion

In summary, this article presents a novel and resilient deep-learning model, named STGAN, which utilizes a GAN-based approach for spatiotemporal satellite image fusion, emphasizing improved transferability. The proposed GAN-based fusion model incorporates a multistream input architecture, enabling it to learn temporal variations in surface reflectance from high-temporal-resolution MODIS images, while leveraging the spatial features obtained from high-spatial-resolution Landsat image correspondents. The model incorporates a GAN-based structure that integrates a spatial transformer, a channel attention module, and a U-net structure. Additionally, STGAN adopts a unique random reference strategy, allowing for the handling of multidecade data with varying temporal gaps and data lacking timestamps.

Experiments including quantitively evaluations and visual inspections conducted on the CIA and LGC datasets demonstrate that the proposed STGAN model exhibits high robustness and competitiveness, producing fused images that are spatially and spectrally accurate. Comparative analysis against existing benchmark models, including FSDAF, GANSTFM, HPLTS-GAN, and EDCSTFN, reveals that STGAN outperforms these benchmarks across multiple dimensions, as assessed by evaluation metrics including RMSE, SSIM, SAM, ERGAS, and LBP, as well as visual inspection.

Acknowledgment

Our computational work used ROGER, which is a geospatial supercomputer supported by the CyberGIS Center for Advanced Digital and Spatial Studies and the School of Earth, Society, and Environment at University of Illinois Urbana-Champaign.

References

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

[2] F. Gao et al., "Toward mapping crop progress at field scales through fusion of Landsat and MODIS imagery," *Remote Sens. Environ.*, vol. 188, pp. 9–25, Jan. 2017.

[3] H. Kimm et al., "Deriving high-spatiotemporal-resolution leaf area index for agroecosystems in the U.S. corn belt using planet labs CubeSat and STAIR fusion data," *Remote Sens. Environ.*, vol. 239, Mar. 2020, Art. no. 111615.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] Z. Shao and J. Cai, "Remote sensing image fusion with deep convolutional neural network," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1656–1669, May 2018.

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

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

[15] Z. Tan, L. Di, M. Zhang, L. Guo, and M. Gao, "An enhanced deep convolutional model for spatiotemporal image fusion," *Remote Sens.*, vol. 11, no. 24, 2019, Art. no. 2898.

[16] Z. Tan, M. Gao, X. Li, and L. Jiang, "A flexible reference-insensitive spatiotemporal fusion model for remote sensing images using conditional generative adversarial network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5601413.

[17] H. Zhang, Y. Song, C. Han, and L. Zhang, "Remote sensing image spatiotemporal fusion using a generative adversarial network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4273–4286, May 2021.

[18] B. Song et al., "MLFF-GAN: A multilevel feature fusion with GAN for spatiotemporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4410816.

[19] Y. Song, H. Zhang, H. Huang, and L. Zhang, "Remote sensing image spatiotemporal fusion via a generative adversarial network with one prior image pair," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5528117.

[20] H. Liu, G. Yang, F. Deng, Y. Qian, and Y. Fan, "MCBAM-GAN: The GAN spatiotemporal fusion model based on multiscale and CBAM for remote sensing images," *Remote Sens.*, vol. 15, no. 6, Mar. 2023, Art. no. 1583.

[21] D. Lei et al., "HPLTS-GAN: A high-precision remote sensing spatiotemporal fusion method based on low temporal sensitivity," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5407416.
[22] Z. Yang, C. Diao, and B. Li, "A robust hybrid deep learning model for spatiotemporal image fusion," *Remote Sens.*, vol. 13, no. 24, Dec. 2021, Art. no. 5005.
[23] 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.
[24] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020.
[25] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1222–1230.
[26] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "SOD-MTGAN: Small object detection via multi-task generative adversarial network," in *Proc. Comput. Vis.*, 2018, pp. 210–226.
[27] Q. Kong, B. Tong, M. Klinkigt, Y. Watanabe, N. Akira, and T. Murakami, "Active generative adversarial network for image classification," *AAAI*, vol. 33, no. 01, pp. 4090–4097, Jul. 2019.
[28] X. Wang et al., "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Lecture Notes in Computer Science*. Berlin, Germany: Springer-Verlag, 2019, pp. 63–79.
[29] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4681–4690.
[30] 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.*, 2017, pp. 1125–1134.
[31] X. Wang, H. Yan, C. Huo, J. Yu, and C. Pant, "Enhancing Pix2Pix for Remote sensing image classification," in *Proc. 24th Int. Conf. Pattern Recognit.*, 2018, pp. 2332–2336.
[32] L. Zhu, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Generative adversarial networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5046–5063, Sep. 2018.
[33] Y. Zhan, D. Hu, Y. Wang, and X. Yu, "Semisupervised hyperspectral image classification based on generative adversarial networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 212–216, Feb. 2018.
[34] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," *Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.
[35] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.
[36] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural Comput.*, vol. 29, no. 9, pp. 2352–2449, Sep. 2017.
[37] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 286–301.
[38] P. Tang, P. Du, J. Xia, P. Zhang, and W. Zhang, "Channel attention-based temporal convolutional network for satellite image time series classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8016505.
[39] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.
[40] D. Lei, G. Ran, L. Zhang, and W. Li, "A spatiotemporal fusion method based on multiscale feature extraction and spatial channel attention mechanism," *Remote Sens.*, vol. 14, no. 3, Jan. 2022, Art. no. 461.
[41] J. Xiao, A. K. Aggarwal, N. H. Duc, A. Arya, U. K. Rage, and R. Avtar, "A review of remote sensing image spatiotemporal fusion: Challenges, applications and recent trends," *Remote Sens. Appl. Soc. Environ.*, vol. 32, 2023, Art. no. 101005.
[42] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.
[43] X. Yu, Y. Qu, and M. Hong, "Underwater-GAN: Underwater image restoration via conditional generative adversarial network," in *Pattern Recognition and Information Forensics*. Berlin, Germany: Springer-Verlag, 2019, pp. 66–75.
[44] F. A. Kruse et al., "The spectral image processing system (SIPS)—Interactive visualization and analysis of imaging spectrometer data," *Remote Sens. Environ.*, vol. 44, no. 2, pp. 145–163, May 1993.
[45] L. Wald, "Quality of high resolution synthesised images: Is there a simple criterion?," in *Proc. 3rd Conf. "Fusion Earth Data: Merging Point Meas., Raster Maps Remotely Sensed Images"*, 2000, pp. 99–103.
[46] 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.
[47] X. Zhu et al., "A novel framework to assess all-round performances of spatiotemporal fusion models," *Remote Sens. Environ.*, vol. 274, 2022, Art. no. 113002.
[48] S. Wang, "A CyberGIS framework for the synthesis of cyberinfrastructure, GIS, and spatial analysis," *Ann. Assoc. Amer. Geographers*, vol. 100, no. 3, pp. 535–557, Jun. 2010.
[49] Z. Yang, C. Diao, F. Gao, and B. Li, "EMET: An emergence-based thermal phenological framework for near real-time crop type mapping," *ISPRS J. Photogrammetry Remote Sens.*, vol. 215, pp. 271–291, Sep. 2024.
[50] Y. Zhao, C. Diao, C. K. Augspurger, and Z. Yang, "Monitoring spring leaf phenology of individual trees in a temperate forest fragment with multi-scale satellite time series," *Remote Sens. Environ.*, vol. 297, 2023, Art. no. 113790.
[51] Y. Kim, "Applicability assessment of a spatiotemporal geostatistical fusion model for disaster monitoring: Two cases of flood and wildfire," *Remote Sens.*, vol. 14, no. 24, Dec. 2022, Art. no. 6204.
[52] Z. Xu, H. Sun, T. Zhang, H. Xu, D. Wu, and J. Gao, "The high spatial resolution drought response index (HiDRI): An integrated framework for monitoring vegetation drought with remote sensing, deep learning, and spatiotemporal fusion," *Remote Sens. Environ.*, vol. 312, 2024, Art. no. 114324.
[53] T. Zhao et al., "Artificial intelligence for geoscience: Progress, challenges, and perspectives," *Innovation*, vol. 5, no. 5, Sep. 2024, Art. no. 100691.
[54] P. Li, Y. Wang, T. Si, K. Ullah, W. Han, and L. Wang, "MFFSP: Multi-scale feature fusion scene parsing network for landslides detection based on high-resolution satellite images," *Eng. Appl. Artif. Intell.*, vol. 127, 2024, Art. no. 107337.
[55] M. Wang, F. Gao, J. Dong, H.-C. Li, and Q. Du, "Nearest neighbor-based contrastive learning for hyperspectral and LiDAR data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5501816.
[56] J. Lin, F. Gao, X. Shi, J. Dong, and Q. Du, "SS-MAE: Spatial–spectral masked autoencoder for multisource remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5531614.
[57] C.-Y. Hsu, W. Li, and S. Wang, "Geospatial foundation models for image analysis: Evaluating and enhancing NASA-IBM Prithvi's domain adaptability," *Int. J. Geographical Inf. Sci.*, pp. 1–30, 2024.
[58] G. Tseng, R. Cartuyvels, I. Zvonkov, M. Purohit, D. Rolnick, and H. Kerner, "Lightweight, pre-trained transformers for remote sensing timeseries," 2023, *arXiv:2304.14065*.
[59] S. Lu et al., "AI foundation models in remote sensing: A survey," 2024, *arXiv:2408.03464*.

Fangzheng Lyu received the B.E. degree in computer engineering from the University of Hong Kong, Hong Kong, in 2018, and the M.S. and Ph.D. degrees in geography from the University of Illinois Urbana-Champaign, Champaign, IL, USA, in 2021 and 2024, respectively.

He is currently an Assistant Professor with the Department of Geography, Virginia Polytechnic Institute and State University (Virginia Tech), Blacksburg, VA, USA. His research interests include advancing GIScience, geospatial data science and computational science to tackle complex geospatial problems and understand multiscale dynamics using heterogeneous geospatial Big Data.

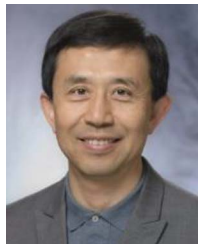
Zijun Yang received the B.S. degree in geographic information systems from the Sun Yat-sen University, Guangzhou, China, in 2016, the M.S. degree in natural resources and environment from the University of Michigan, Ann Arbor, MI, USA, in 2018, and the Ph.D. degree in geography from the University of Illinois Urbana-Champaign, Champaign, IL, USA, in 2024.

He is currently an Assistant Professor with the Department of Earth and Ocean Sciences, University of North Carolina Wilmington, Wilmington, NC, USA. His research interests include remote sensing, machine learning, agriculture, and environmental sustainability. His work aims to better understand ecosystem dynamics under the changing climate through an integration of multiscale remote sensing, field data collection, and GeoAI modeling.

Chunyuan Diao received the B.S. degree in resources science and engineering from the Beijing Normal University, Beijing, China, in 2010, and the M.A. degree in biostatistics and the Ph.D. degree in geography from the State University of New York at Buffalo, Buffalo, NY, USA, in 2014 and 2017, respectively.

She is currently an Associate Professor with the Department of Geography and Geographic Information Science, University of Illinois Urbana-Champaign, Champaign, IL, USA. Her research interests include confluence of remote sensing, geographic information science and biogeography, with current research focusing on computational remote sensing in characterizing land surface patterns and processes, underlying mechanisms, and responses to climate change and human activities. Her research is funded by the National Center for Supercomputing Applications, National Science Foundation, United States Department of Agriculture, etc.

Dr. Diao was the recipient of several awards for her research, including the NSF CAREER award, the Early Career Scholars in Remote Sensing Award from the AAG Remote Sensing Specialty Group, the Microsoft AI for Earth Award, etc.



Shaowen Wang received the B.S. degree in computer engineering from the Tianjin University, Tianjin, China, in 1995, the M.S. degree in geography from the Peking University, Beijing, China, in 1998, and the M.S. degree in computer science and the Ph.D. degree in geography from the University of Iowa, Iowa City, IA, USA, in 2002 and 2004, respectively.

He is currently a Professor of Geography and Geographic Information Science (primary), Computing and Data Science, and Urban and Regional Planning with the University of Illinois Urbana-Champaign (UIUC), Champaign, IL, USA, where he also serves as an Associate Dean for Life and Physical Sciences with the College of Liberal Arts and Sciences. He directs the UIUC's CyberGIS Center for Advanced Digital & Spatial Studies and leads the Institute for Geospatial Understanding through an Integrative Discovery Environment supported by the U.S. National Science Foundation. His research interests include advancing cyberGIS and geospatial data science for scalable solutions to complex geospatial problems and sustainability challenges.

Dr. Wang is a Fellow of the American Association for the Advancement of Science, the American Association of Geographers, and the University Consortium for Geographic Information Science.