The Spatially Seamless Spatiotemporal Fusion Model Based on Generative Adversarial Networks

ChenYang Weng, Yulin Zhan, Xingfa Gu, Jian Yang, Yan Liu, Hong Guo, Zilong Lian, Shiyuan Zhang, Zhangjie Wang, and Xuechun Zhao

***Abstract*—Spatiotemporal fusion is a method of fusing high spatial resolution low temporal resolution remote sensing images and low spatial resolution high temporal resolution in order to obtain high spatiotemporal resolution remote sensing images, which can provide data support for temporal observation of fine objects, and plays an important role in the fields of Earth sciences, environmental monitoring, and so on. This article reveals an issue that is often overlooked in the field of deep learning-based spatiotemporal fusion: the discontinuity between image blocks and image blocks. This discontinuity may have an impact on the visualization of remote sensing images and subsequent applications. In this regard, this article proposes a spatially seamless stitching approach to optimize the spatiotemporal fusion model based on deep learning. By using this method, we successfully obtain high-quality fused remote sensing images with smoother transitions. The spatiotemporal fusion model used in the experiment is a generative adversarial network-based spatiotemporal fusion model (GAN-STFM) and the data are from the Beijing Gaofen-6 dataset (BJGF6). After our splicing method, the ratio of root-mean-square error (RMSE) at the splicing seam to the overall RMSE is reduced from 1.28 to 0.99, which effectively improves the continuity of the image. This new image splicing method has the potential to improve the utility of deep learning-based spatiotemporal fusion algorithms, which has application value for generating large-scale long time series remote sensing datasets with high temporal and high spatial resolution.**

***Index Terms*—Deep learning, error distribution, Gaofen-6, spatial resolution, spatiotemporal fusion, splicing method.**

I. Introduction

In light of significant progress in remote sensing technology, satellite-acquired image data have become instrumental for Earth observation and environmental monitoring. However, hindered by hardware constraints and observational limitations, individual satellites fall short in delivering image data with extensive spatial and temporal coverage [1]. To fully comprehend and analyze dynamic spatial and temporal patterns across the Earth's surface, and to acquire remote sensing data that are more expansive spatially and consistent temporally, the necessity for spatiotemporal fusion techniques in remote sensing imagery has become paramount.

Spatiotemporal fusion of remote sensing images involves blending data from sources with low temporal resolution and high spatial resolution, like MODIS data, with data from sources featuring high temporal resolution and low spatial resolution, such as Landsat data. This fusion aims to acquire data with both high temporal and spatial resolutions, constituting a technique for blending diverse remote sensing images.

Utilizing spatiotemporal fusion allows for long-term monitoring and change detection at specific locations, providing insights into the temporal evolution of surface processes. It also supports the capture and mapping of high-resolution remote sensing images across extensive geographical areas, aiding in the analysis of spatial distributions of land features.

Spatiotemporal fusion of remote sensing images plays a crucial role in various fields. In recent investigations, researchers have applied spatiotemporal fusion to different domains such as soil carbon emissions monitoring in subtropical forests [2], the retrieval of suspended particulate matter in saline lakes [3], the retrieval of daily surface temperatures, and others [4]. The pronounced utility of spatiotemporal fusion is particularly evident in its pivotal role within the domains of Earth science and environmental monitoring.

Since the release of the spatiotemporal adaptive reflectance fusion model (STARFM) in 2006 [1], the technology of spatiotemporal fusion in remote sensing has continued to evolve. Zhu et al. [5] categorized algorithms for spatiotemporal fusion into five types: methods based on unmixing, methods based on weight functions, Bayesian methods, learning-based methods, and hybrid methods. Unmixing methods [6], [7], based on linear spectral theory, use coarse pixels of the target date to unmix and obtain fine pixels of the target date, such as multisensor multiresolution technology (MMT). Weight function methods combine

Manuscript received 12 December 2023; revised 29 January 2024 and 16 March 2024; accepted 20 March 2024. Date of publication 25 March 2024; date of current version 24 July 2024. This work was supported in part by the National Key R&D Program of China under Grant 2021YFE0117300, in part by Common Application Support Platform for National Civil Space Infrastructure Land Observation Satellites under Grant 2017-000052-73-01-001735, and in part by the Natural Science Foundation of China under Grant 42271358, and in part by the Major Project of High Resolution Earth Observation System under Grant 30-Y60B01-9003-22/23. *(Corresponding authors: Yulin Zhan; Xingfa Gu.)*

ChenYang Weng, Yulin Zhan, Jian Yang, Yan Liu, and Hong Guo are with the National Engineering Laboratory for Satellite Remote Sensing Applications, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, and also with the College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: wengchenyang21@mails.ucas.ac.cn; zhanyl@radi.ac.cn; yangjian@aircas.ac.cn; liuyan@aircas.ac.cn; guohong@radi.ac.cn).

Xingfa Gu is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, also with the University of Chinese Academy of Sciences, Beijing 100049, China, and also with the School of Geography and Remote Sensing, Guangzhou University, GuangZhou 510006, China (e-mail: guxf@aircas.ac.cn).

Zilong Lian, Shiyuan Zhang, Zhangjie Wang, and Xuechun Zhao are with the North China Institute of Aerospace Engineering, Langfang 065000, China (e-mail: kingsleylin@stumail.nciae.edu.cn; 578982408@qq.com; wangzhangjie12@163.com; 2408255020@qq.com).

Digital Object Identifier 10.1109/JSTARS.2024.3381185

input images linearly to obtain fine pixels of the target date, as seen in STARFM and the widely used ESTARFM [8]. Bayesian methods [9], [10] treat spatiotemporal fusion as a maximum a posteriori probability problem, generating fine images of the target date that maximize the probability based on input images. Hybrid methods [11], such as flexible spatiotemporal data fusion (FSDAF) [12], aim to improve model accuracy by combining the strengths of various fusion methods. FSDAF integrates decomposed methods, weighted function-based methods, and spatial interpolation, requiring only a pair of high coarse-resolution images from reference dates and a coarse-resolution image from the target date.

The four methods mentioned above are part of the traditional spatiotemporal fusion of remote sensing images, which involves linear modeling. However, spatiotemporal changes in features cannot be simply described as linear changes. Remote sensing images are affected by hardware performance, physical climate, noise, and other factors [13]. Simple linear modeling results in a significant loss of information. Traditional methods often require numerous manual parameter settings, lack compatibility, and robustness with algorithms, and are not easily scalable [14].

With the development of deep learning, especially the wide use of convolutional neural networks (CNNs) in the field of computer vision [15], [16], [17], [18], [19], an increasing number of researchers are exploring deep learning-based spatiotemporal fusion techniques for remote sensing images. Specifically, deep learning-based methods play an important role in feature extraction of remote sensing images. For example, DC-Net [20], a hyperspectral superresolution framework based on subpixel-level, is built for extracting information from pixel to subpixel-level and from image to feature-level. In addition, a hyperspectral image classification method called GNet [21] is developed using CNN-based feature extraction module to solve the problem of uneven distribution of spectral and spatial features. Furthermore, TDGN [22], a temporal difference-guided hyperspectral image change detection network is devised based on gated recurrent unit (GRU) block to reduce the redundance of features and boost the training efficiency of CNN-based methods. Recently, with the popularity of generative pretrained transformer (GPT) [23], some researchers have started to apply the technique of GPT to the field of remote sensing to fully utilize extensive RS Big Data [24]. We have counted the methodological literature in the field of remote sensing image spatiotemporal fusion over recent years on the Web of Science. As shown in Fig. 1, it can be seen that spatiotemporal fusion techniques based on deep learning are receiving increasing attention, and the proportion of spatiotemporal fusion algorithms based on deep learning is also rising.

Initially, researchers directly established the mapping relationship between coarse-resolution images and fine-resolution images [25], [26], [27] by directly enhancing the coarse-resolution images of the target date, and certain fine-resolution images were obtained. Such methods solely rely on coarse-resolution images as input, limiting their ability to effectively utilize images from other reference dates. In addition, the generated images exhibit poor texture details and lower fidelity.

Other methods involve additional reference information [27], [28], [29], [30], [31], [32], [33]. A commonly used approach is to include pairs of high and coarse-resolution images before and after the target date, along with a coarse-resolution remote sensing image of the target date as input to simulate the fine-resolution image for the target date. However, this method can only generate remote sensing images for intermediate dates and is not applicable to fusion requirements for newer dates. Moreover, it imposes limitations on the quantity of input data and time constraints, thereby restricting its application scenarios.

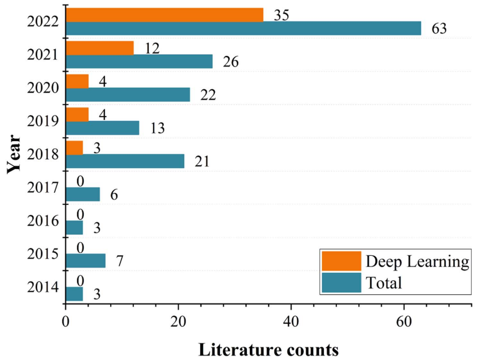


Fig. 1. Yearly literature counts of journal papers introducing spatiotemporal data fusion methods.

Some methods utilize generative adversarial networks (GANs) for image fusion to generate more realistically restored fine-resolution remote sensing images [27], [28], [34], [35]. Approaches like GAN-STFM [28] require only a reference image from another date and the coarse-resolution image of the target date as input, significantly reducing the constraints on the spatiotemporal fusion of remote sensing images. This broadens the application scenarios of the algorithm.

In order to achieve better fusion results and to improve the robustness of the model for more complex mappings, a large number of learnable parameters exist in current end-to-end deep learning models. The number of these parameters is much larger than in traditional spatiotemporal fusion models. Because of this, these parameters require a large amount of computational resources to process massive amounts of data and perform a large number of matrix operations when performing training and tuning. GPUs can execute multiple matrix operations and parallel computations, expediting the training and inference speed of deep learning models. However, due to the limitations of hardware, for relatively large-sized images, using the original image directly as input often leads to memory explosion. A typical approach is to slice the original image, simulate each image block individually, and finally stitch together the simulated image blocks to form a large-scale simulated image [28], [35].

In practical applications, such as spatiotemporal fusion of remote sensing images on a large scale, tessellated splicing breaks are a common problem in the results of block imagery splicing during the processing of analog high-resolution imagery using simulation. This problem may not be prominent in initial observations but can become apparent in detailed analysis and subsequent applications.

Each input image block is a continuous and independent unit of data subdivided from the fused image. However, if these blocks are not properly handled when stitched together after processing by the fusion model to generate a complete simulated image, a noticeable discontinuity from block to block can occur, presenting a phenomenon similar to a tessellated texture. In other words, we expect a smooth and continuous visual effect. However, in reality, there is no continuity between the image blocks, creating many unnatural transitions that degrade the visual effect of the simulated high-resolution image.

The existence of this fault phenomenon not only affects the visual experience but also has a considerable impact on subsequent applications. For instance, in applications like remote sensing image classification, target detection, and environmental monitoring, these abrupt transitions can lead to misinterpretation and affect the accuracy and usability of the results.

Hence, the issue of tessellated splicing sections cannot be overlooked. To address this problem in a targeted manner, we need to conduct an in-depth study on the process of image splicing. The aim is to obtain fused images of higher quality when dealing with spatial and temporal fusion of a wide range of remote sensing images, thus further improving the application value of fused images.

We replicated the GAN-STFM model on a medium-resolution scale (MODIS-Landsat). The spatiotemporal fusion model performs well on this medium-resolution dataset, demonstrating good performance in the errors at the seams. The root-mean-square error (RMSE) between the seams and image blocks differs by only about 2.08%. We reviewed the performance of spatiotemporal fusion models in deep learning at higher resolutions. However, whether in medium-resolution or high-resolution spatiotemporal fusion, the focus is predominantly on the overall spectral and spatial accuracy, rather than the accuracy differences at the seams and across the entire image.

In this article, we have made the following contributions: 1) comparatively tested the phenomenon of splicing gaps in simulated images at different scales and found that obvious splicing gap phenomena appeared in the fused images from 16 to 8 m; 2) proposed an optimized generative adversarial spatiotemporal fusion model based on spatial seamless splicing approach, which can effectively eliminate discontinuous cross-sections due to the splicing of image blocks.

The rest of this article is organized as follows. Section II introduces a spatially seamless stitching method optimized for GAN-based spatiotemporal fusion models. In Section III, we present the data used in the experiments, the discontinuity phenomenon of image block transitions at different scales, and the stitching effects of our proposed method. Finally, Section IV concludes this article.

II. Methods

In this chapter, we introduce our proposed spatially seamless spatiotemporal fusion model. In the first part, we provide an overview of our spatiotemporal fusion model, from data preparation to fusion and final stitching. In the second part, we detail our spatially seamless stitching approach. The last part describes the deep learning model we used and its features.

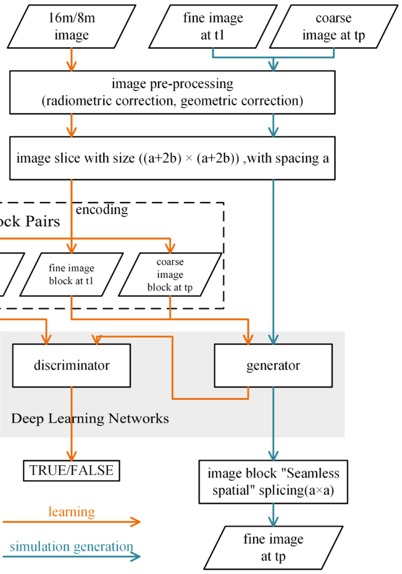


Fig. 2. Spatially seamless spatiotemporal fusion model.

A. Spatially Seamless Spatiotemporal Fusion Model

Our model is divided into three parts: training data preparation, model learning, and model synthesis, as shown in Fig. 2.

In the training data preparation part, we collected the Gaofen-6 16-m resolution and 8-m resolution data for basic remote sensing image preprocessing, including radiometric and geometric corrections. Next, the grid file is generated according to the data range and image block size, image block repetition rate, and the images are blocked according to the grid to get the image block database with high and low resolution for different dates. Finally, the data are encoded in the format of fine image blocks for the target date, fine image blocks for the reference date, and coarse image blocks for the target date to obtain the image pair index table. The image block database and the image pair index table form the spatiotemporal fusion training set for model learning. The image pair index table can be used to read the image block data quickly in the subsequent training stage, while image blocks with a certain repetition rate can improve the model's ability to extract features from neighboring image blocks, thus, improving the edge quality of the image blocks generated by the model.

In the model learning part, taking GAN-STFM as an example, we read the image block pairs from the image pair index file and input them into the generative adversarial model. The continuous confrontation between the generator and the discriminator improves the fusion capability of the generator.

In the simulation generation part of the model, we fuse the fine image of the reference date to be fused with the coarse image of the target date to obtain the fine image of the target date. We first preprocess the input data (this is the same as the preprocessing in the training data preparation part), and then we slice the image blocks, we slice the image blocks with the interval of a pixel, and at the same time, each image block is expanded by b pixels to the surrounding neighboring image blocks, i.e., we slice the image blocks with the size of ((a + 2b) × (a + 2b)) pixels, and the image block is used as the input

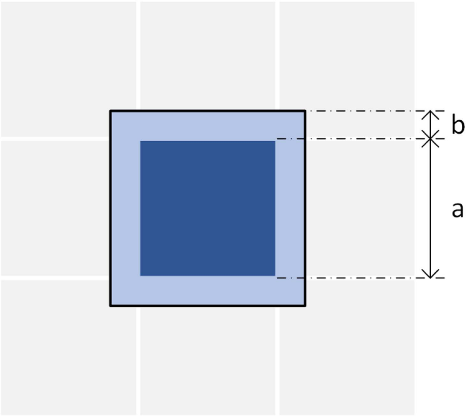


Fig. 3. Spatially seamless stitching.

Algorithm 1: Spatially Seamless Stitching.

```
width = image.width
height = image.height
for x in width, step = a do
    for y in height, step = a do
        input = image[x-b:x+a+b, y-b:y+a+b]
        output = predict(input)
        output = output[b:a+b, b:a+b]
        fusionImage[x:x+a, y:y+a] = output
    end for
end for
```

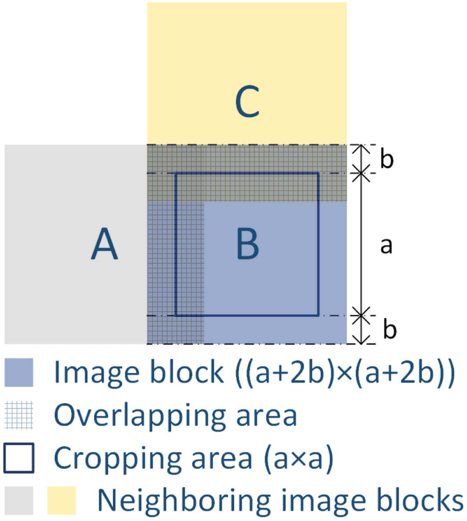


Fig. 4. Overlapping area between image blocks (take three image blocks as an example).

to the GAN-STFM network, and the image block is then used as the input to the GAN-STFM network, and the image block is then used as the input to the GAN-STFM network. The fused image blocks are spatially seamlessly stitched together to obtain a finely simulated image of the target date. (See the next part for details on the splicing method.)

B. Spatially Seamless Stitching

Although deep learning-based fusion models are capable of generating enough fused images for general use, these methods have always had the issue that the splicing between image blocks may result in tessellated cross-sections in the generated fused images. In this article, we propose a new method called "spatial seamless stitching," which provides a more effective means of fusing a wide range of images.

First, we will explain the common splicing method and its issues. The input for the common spatial and temporal fusion model is processed by dividing the input image into nonoverlapping neighboring image blocks. While this approach is simple and straightforward, it can lead to discontinuous transitions in the output. The fusion process is independent from block to block, resulting in a discontinuous transition between neighboring blocks. This compromises the quality of the resulting fused image. The ordinary splicing method does not utilize the transition information between neighboring image blocks, resulting in obvious breaks in the splicing process of the image blocks and degrading the visual effect.

To address this issue, we suggest a new splicing method. Specifically, we focus on the spatially seamless splicing method discussed in this article. In contrast to the conventional splicing method, this technique overlaps image blocks, as illustrated in Fig. 3. The generator of the spatiotemporal fusion model uses the image blocks with (a) as the interval cut and (a + 2b) as the side lengths (the range of the black box line) to predict the target date image. The (a + 2b) × (a + 2b) pixel-sized image blocks output from the deep learning model are stitched together by removing the buffer (light blue area) with radius b around them. The size of the buffer radius b depends on the structure of the chosen deep learning model and is related to the filling radius of the stacked convolutional layers of the spatiotemporal fusion deep learning neural network, as shown in the following:

$$b \geq \sum_{i=1}^{N} p_i. \tag{1}$$

According to (1), the buffer radius b must be greater than or equal to the cumulative sum of the padding radii p_i of all the convolutional layers concatenated in the model. For instance, in the case of the fusion network GAN-STFM, which we are using as an example, there are 17 convolutional layers concatenated on its generator, and the sum of the padding radii is 8. Therefore, b must be an integer greater than or equal to 8.

The input image blocks are processed using spatially seamless stitching, as demonstrated in Fig. 4. In contrast to the standard stitching method, image blocks A and C have a 2b-wide overlap region (shaded area) with image block B during input. It is crucial to note that the overlap region plays a significant role in the model's comprehension of the input image. The deep learning model can extract spatial and spectral features between adjacent image blocks more effectively through this overlap region. This results in a more continuous transition between neighboring image blocks when generating the fused image.

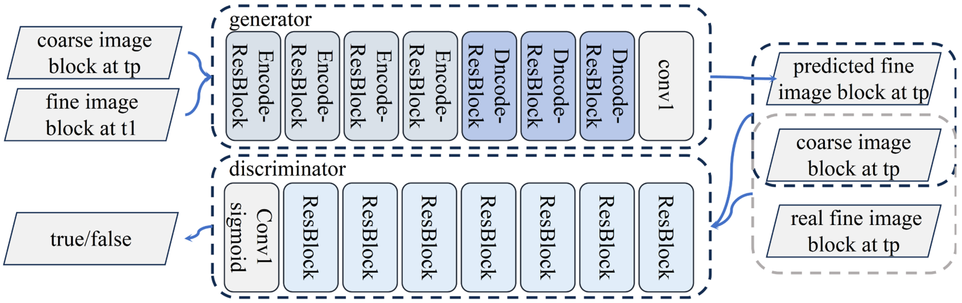


Fig. 5. Whole architecture of the GAN-STFM model for spatiotemporal fusion (GEncoder-ResBlock and GDecoder-ResBlock represent the residual blocks for the generator encoder and decoder, respectively.

The emergence of the spatially seamless stitching method undoubtedly brings a new processing method to the field of deep learning. It greatly reduces the impact of interrupted surfaces of image blocks and enhances the visual effect and quality of the generated fused images. Although this method may affect processing speed, it provides significant help in ensuring overall quality.

C. GAN-STFM Model

The spatiotemporal fusion model used in this article comes from the GAN-STFM proposed by Tan et al. [28]. This model is a spatiotemporal fusion model based on a GAN. As shown in Fig. 5, this model is divided into a generator and a discriminator, both of which are constructed by stacking residual blocks based on convolutional neural networks.

Compared to other models, it has the following features.

1) Fewer inputs are needed: Compared with conventional spatiotemporal fusion models, GAN-STFM only needs two images as input. That is, a coarse-resolution image on the prediction date and a fine-resolution reference image from any time in the same area. This reduces the difficulty of data collection and makes practical application more convenient.
2) Lesser time constraints: Traditional spatiotemporal fusion models have strict time constraints on the choice of reference images, while GAN-STFM eliminates this restriction. It can select a fine-resolution reference image at any time, making data preparation more flexible.
3) Better application prospects: GAN-STFM has shown similar to better performance than that of the statistical fusion model in experiments. Its flexibility and simple data requirements make the data preparation for spatiotemporal fusion easier, providing better prospects for its practical application.

D. Model Time Complexity

The runtime of our method is mainly affected by the training and analysis speed of the chosen deep learning-based spatiotemporal fusion network. Different stitching methods can affect the time complexity of the model by affecting the amount of data processed by the network. Assume that the time complexity of the selected deep learning network is $O(f(n)$, then the time complexity of the seamless spatiotemporal fusion model is $O(\frac{(a+2b)^2}{a^2}f(n))$. With our chosen GAN-STFM model as an example, when $a = 256$, $b = 8$, its time complexity is $O(1.129f(n))$. This means that the time complexity of our method is 1.129 times that of conventional methods.

III. Experiments and Results

In this chapter, from the first to the third part, we sequentially introduced the data used in the experiment, including two datasets at 500 m-30 m and 16 m-8 m; the design of the experimental scheme; and the quantitative evaluation indicators of the fusion results. In the fourth part, we compared the discontinuity phenomena of image blocks under different spatial resolutions. In the last part, we compared the fusion effects of the spatially seamless spatiotemporal fusion model and the conventional deep learning model.

A. Data

The Lower Garonne Catchment (LGC)[1] study area data is a commonly used open-source dataset in the field of spatiotemporal fusion [36]. The LGC study area is located in the north of New South Wales, Australia, and covers an area of 5440 square kilometers. This dataset consists of 14 cloud-free Landsat-MODIS pairs from April 2004 to April 2005. The images in the LGC dataset are cropped to a size of 3072 × 2560. The LGC dataset has significant land cover type changes.

Current deep learning-based spatiotemporal fusion models perform well on 500 m-30 m datasets like LGC. We will use this data to test the transition continuity issue between adjacent image blocks in the spatiotemporal fusion model.

The Beijing Gaofen-6 (BJGF6) dataset is a dataset created independently for this study. The Gaofen-6 remote sensing satellite platform provides two payloads: a 16-m multispectral medium resolution wide swath camera (WFV) and a 2-m panchromatic/8-m multispectral high-resolution camera (PMS). The WFV has a swath width of 850 km, while the PMS has only 95 km. As shown in Fig. 6, the middle yellow box is the shooting range of PMS, and the remaining images are shot by WFV. By fusing these two images, the coverage of the PMS image

[1] LGC: http://dx.doi.org/10.4225/08/5111AD2B7FEE6

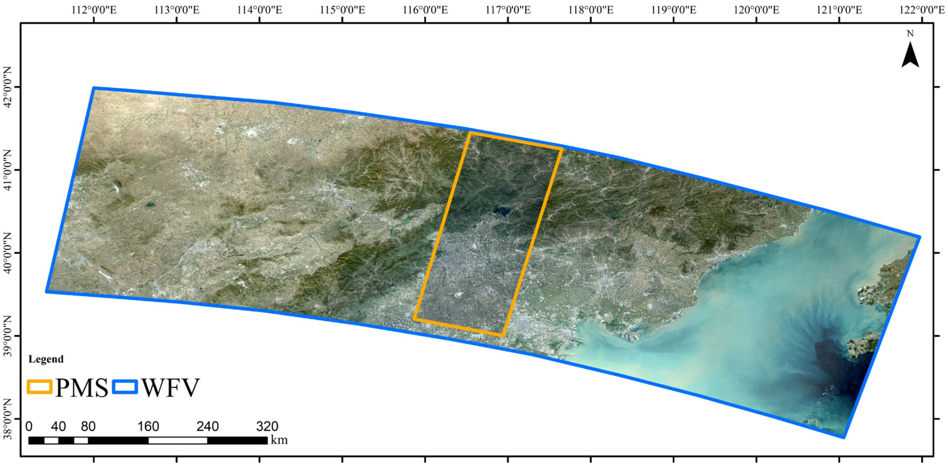


Fig. 6. Comparison of swath width of WFV and PMS.

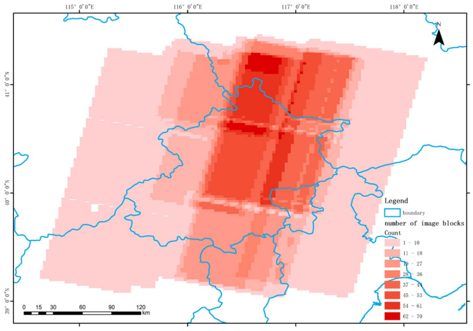


Fig. 7. Spatial distribution of Beijing GF-6 Dataset.

can be expanded significantly, thereby improving the temporal resolution.

The BJGF6 dataset covers a total of 110 cloud-free remote sensing images from 2019 to 2020 of Beijing and its surrounding areas. We perform radiometric and geometric corrections on the images, and register all images to a base map. We then divide the images into image blocks of size 512 × 512 based on the image range, and cut and pair them according to the geographic coordinates of the image blocks. Eventually, we got about 230 000 pairs of images. Each image pair consists of the registered coarse image of the target date (16 m), the fine reference image of other dates (8 m), and the actual image of the target date (8 m). The first and second images are the inputs for the fusion model, and the third image is the reference output when training the deep learning model. Fig. 7 shows the spatial distribution diagram of the dataset, where the shade of red represents the number of image pairs in that position. The darker the color, the more data there are.

The BJGF6 dataset has the following differences compared with the LGC dataset.

1) The BJGF6 dataset is larger in size, more suitable for deep learning model learning.
2) The BJGF6 dataset has a higher resolution, and the ratio difference between high and low spatial resolutions is smaller, which is conducive for model learning.

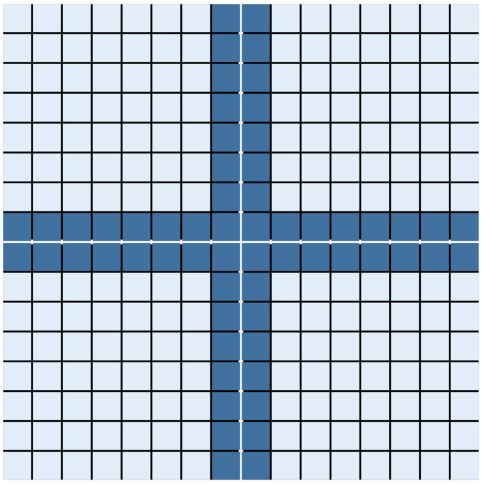


Fig. 8. Schematic diagram of stitching seams.

3) With a higher resolution, the BJGF6 dataset can highlight more easily the issues related to stitching seams when stitching together large remote sensing images in a deep learning-based spatiotemporal fusion model.

B. *Quantitative Evaluation Indicators*

To evaluate the images after spatiotemporal fusion, we divide the test area into stitching seam areas and the overall image. Fig. 8 shows a schematic diagram of the adjacent areas of four image blocks. The white line in the middle is the stitching seam. Each square in the picture represents a pixel. We have analyzed the errors of the stitching seam (dark color) and the overall image (light color + dark color).

In terms of quantitative analysis indicators, we referred to the error evaluation framework by Zhu et al. [37]. Since the stitching seam only consists of a two-pixel width, we did not use spatial accuracy indicators. Instead, we used indicators representing spectral accuracy: RMSE and average difference (AD).

The range of the RMSE value is [0,1]. A value of 0 represents a perfectly fused image; the larger the value, the larger the

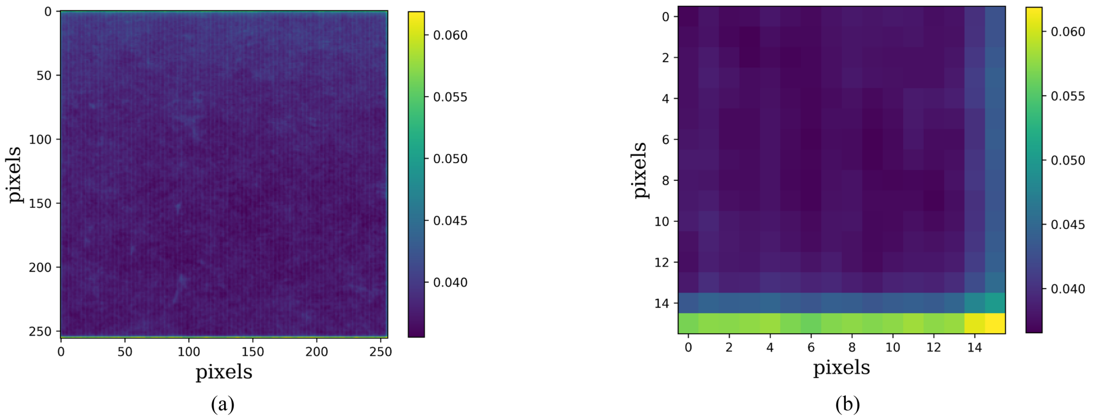


Fig. 9. Error spatial distribution of the fused image block. (a) Error spatial distribution of the entire image block. (b) Error distribution of the 16 × 16 pixels in the lower right corner of the image block.

TABLE I
Quantitative Analysis on the Two Datasets

Dataset	Error position	RMSE	AD	$\frac{\mathrm{RMSE}_{\mathrm{seams}}}{\mathrm{RMSE}_{\mathrm{overall}}}$
1	Stitching seams	0.0123	0.00341	1.26
	Overall image	0.00979	0.00361	
2	Stitching seams	0.0171	0.00333	1.27
	Overall image	0.0135	0.00312	

spectral error in the fused image. The range of the AD value is also $[-1, 1]$. A value of 0 represents a perfectly fused image; more negative values and more positive values indicate larger spectral deviations in the fused image. Negative and positive values represent underestimation and overestimation of spectral information, respectively.

In addition to RMSE and AD we use $\frac{\mathrm{RMSE}_{\mathrm{seams}}}{\mathrm{RMSE}_{\mathrm{overall}}}$ to evaluate the RMSE consistency between the stitching seams and the overall image region, if $\frac{\mathrm{RMSE}_{\mathrm{seams}}}{\mathrm{RMSE}_{\mathrm{overall}}}$ is closer to 1,then the error of the stitching seams is closer to the overall error, and the stitched image is more continuous.

C. Experimental Design

In the experiment, we trained three sets of weights using the GAN-STFM code shared by Tan on GitHub. The parameters were kept the same for all three training sessions, only changing the dataset used and the size of the image blocks. For the first set of weights (Weight I), we used the LGC dataset, and the size of the image block was 256 × 256; for the second set of weights (Weight II), we used our custom BJGF6 dataset, and the size of the image block was 256 × 256; for the third set of weights (Weight III), we also used the BJGF6 dataset, but the size of the image block was 272 × 272. When training the Weight II and Weight III, we made necessary modifications to the code for dataset compatibility.

We used Weight I and Weight II to generate fused image blocks separately and analyzed the generated image blocks to study discontinuities between image blocks at different spatial scales.

We used Weight II and Weight III to generate fused image blocks of different sizes, and we combined these image blocks into whole images using direct stitching and the spatially seamless stitching method proposed in this article. Finally, we performed error analysis on the stitching seams and the overall image of the stitching result.

D. Discontinuity Phenomenon of Image Blocks

1) Error Comparison At Different Resolutions: As shown in Table I, we have carried out a comparative analysis of the errors in the stitching seams and the overall image after directly stitching the image blocks generated by the deep learning-based spatiotemporal fusion model under the LGC dataset and the BJGF6 dataset. It can be seen that, although the error at the stitching seams in the LGC dataset is larger than the overall error, the gap between the two is not large. However, in the BJGF6 dataset, the error at the stitching seams is much larger than the overall error, and the ratio of the two RMSE values has increased from 1.02 to 1.279.

Based on our comparative experiments, we believe that the phenomenon of discontinuity between adjacent image blocks has a significant correlation with the spatial resolution of the fused image. The higher the spatial resolution, the more discontinuous the transition between adjacent image blocks, and the poorer the quality of the fused image.

2) Error Distribution At 8-M Resolution: We continue using Weight II to perform a qualitative analysis on the performance of GAN-STFM on the BJGF6 dataset. Our aim in this analysis is to study the spatial distribution of errors. We carry out an error analysis on the image blocks and the large-scale remote sensing images poststitching.

For the image blocks, we extracted training data from the dataset where fusion times did not exceed 90 days, counting a total of 92 115 image pairs. We calculated the RMSE by subtracting the actual image blocks from the simulated image blocks produced by fusion, and then accumulating these errors. As shown in Fig. 9, one could observe that the model's ability to generate simulated images is not consistent—edges of the image blocks accrue a significantly larger cumulative error compared to that in the center. We accordingly refer to edge errors produced during the generation of image blocks as "edge errors," which

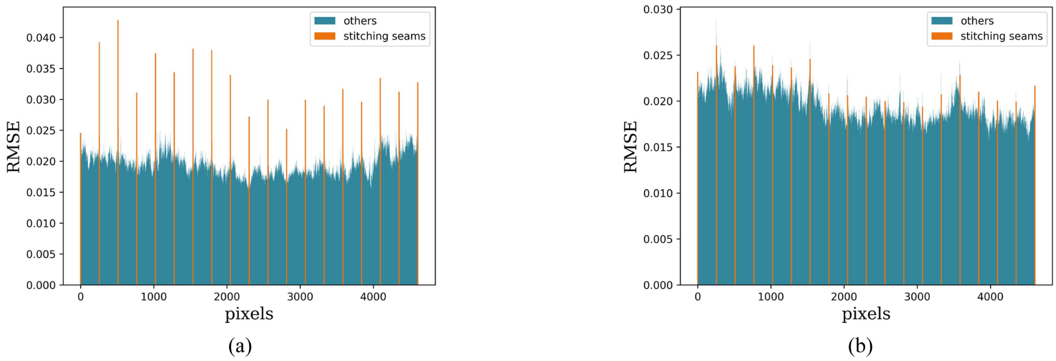


Fig. 10. Average error of the fused image in the horizontal and vertical directions. (a) Average error in the horizontal direction. (b) Average error in vertical direction.

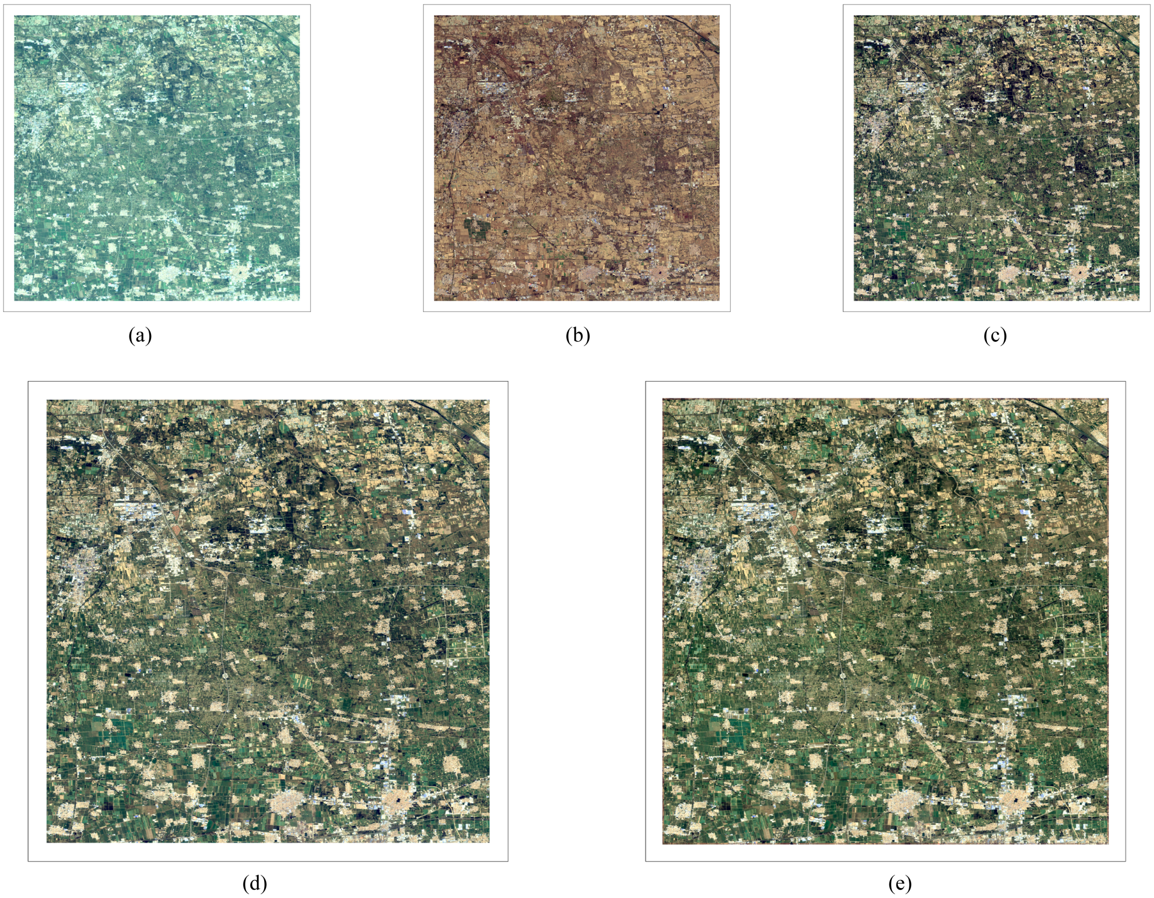


Fig. 11. Comparison of spatiotemporal fusion results of remote sensing images. (a) Input coarse image at target date. (b) Input fine image at other date. (c) Fine image ground truth at target date. (c) Simulated images spliced using direct splicing method. (d) Simulated images spliced by our method.

is one of the causes that leads to discontinuity when different image blocks are combined.

Regarding the stitched images, we analyze a simulated image of 8-m resolution (with a size of 4*4608*4608) generated by a deep learning model. We square the difference between the simulated and actual image, and subsequently average these squared differences in both the horizontal and vertical directions. The resulting statistical diagram, as shown in Fig. 10, includes a Fig. 10(a) average cumulative error diagram in the horizontal direction, and Fig. 10(b) in the vertical direction. It can been seen that the stitching seams' errors (occurring every 256 pixels) are significantly greater than errors found elsewhere in the rows or columns.

E. Comparison of Model Fusion Effects

1) Qualitative Analysis: We perform a qualitative analysis on the simulated images of remote sensing spatiotemporal fusion based on both direct stitching and spatial seamless stitching. To ensure the consistency of the stitching seam spatial position for

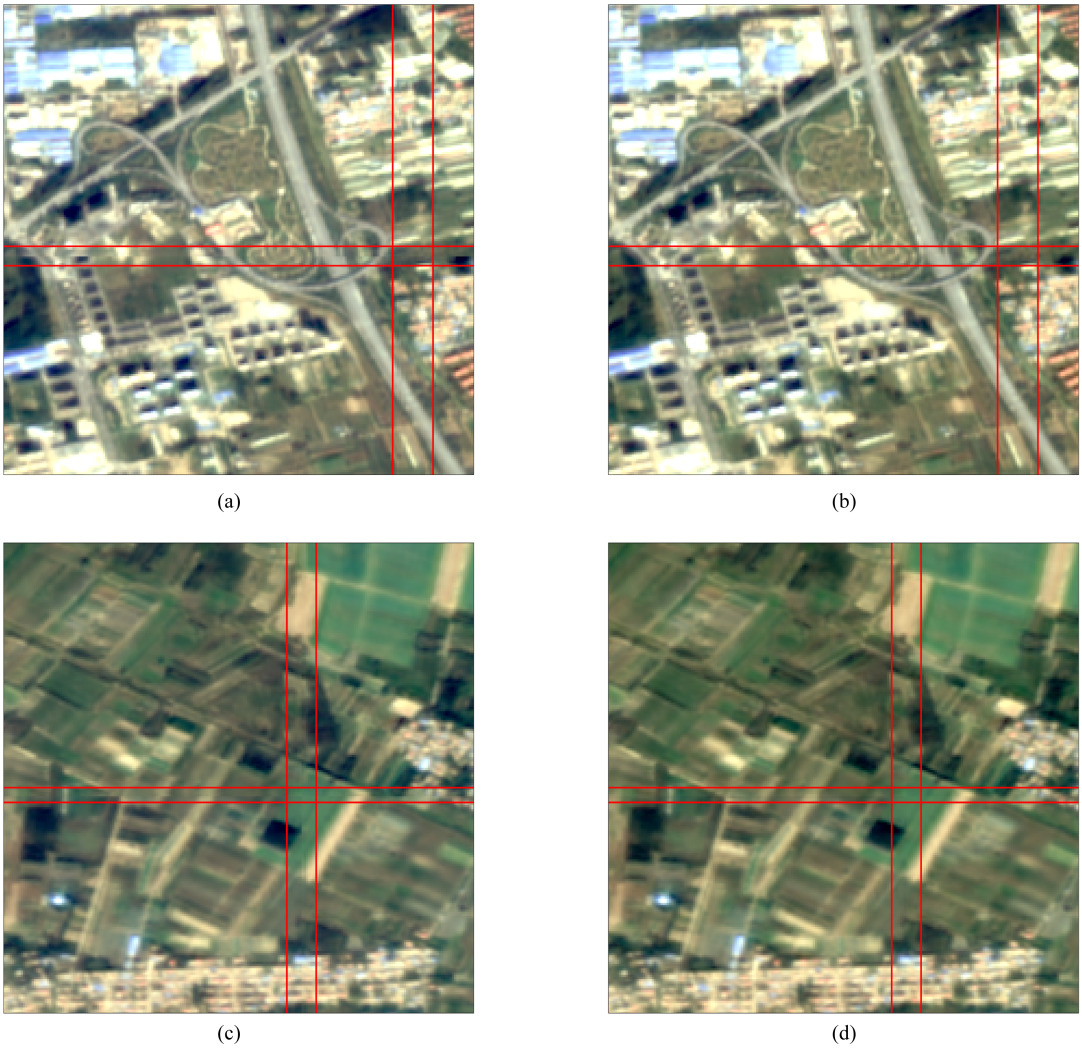


Fig. 12. Detailed comparison of the fusion results, the red lines are the splicing gaps. (a) Details of our method in urban areas. (b) Details of direct splicing method in urban area. (c) Details of our method in farmland. (d) Details of direct splicing method in farmland.

both, we use the weight II for direct stitching and the weight III for spatial seamless stitching. This way, both models use image blocks of 256 × 256 in size when stitching. We stitch 16 × 16 image blocks into a remote sensing image of 4096 × 4096 pixels. The first row of Fig. 11(a)–(c) shows the input image and actual image used for quantitative analysis; the second row of Fig. 11(d) shows the result of direct stitching, and Fig. 11(e) shows the result obtained using our spatial seamless stitching method.

Since Fig. 11 is a thumbnail, the stitching seams cannot be seen clearly. We have selected urban and farmland areas in the test region for a detailed comparison. The first row of Fig. 12(a) and (b) shows the detailed comparisons of the two methods in the urban area, and Fig. 12(c) and (d) shows the detailed comparisons of the two methods in the farmland. The part within the red solid line is the stitching area. It can be clearly seen that there are obvious stitching seams in the first column, which represents the direct stitching method, particularly noticeable on the horizontal direction in terms of the road stitching cross-section and the color difference in the farmland stitching. However, the second column, which represents the spatial seamless stitching method proposed in this article, can effectively eliminate these stitching cross-sections and make the transition between image blocks smoother.

2) Quantitative Analysis: We quantitatively evaluate the fusion effects of the two methods. As shown in Table II, the error values at different positions of image blocks for the two methods are presented, revealing that the spatial seamless stitching method has relatively consistent errors both at the stitching seams and in the overall image, while the direct stitching method produces larger errors at the stitching seams. Looking at the fourth column of the table for the $\frac{\text{RMSE}_{\text{seams}}}{\text{RMSE}_{\text{overall}}}$ metrics reveals that the space seamless method has almost the same RMSE at the spliced seams as the whole, while the direct splicing method will have an overall 28% higher RMSE at the spliced seams than the whole. Hence, the spatial seamless stitching method can effectively diminish errors located at the stitching seams of the algorithm, thereby increasing the practicality of the spatiotemporal fusion algorithm based on deep learning.

TABLE II
Quantitative Analysis of the Two Stitching Methods

Band	Splicing method	Error position	RMSE	AD	$\frac{RMSE_{seams}}{RMSE_{overall}}$
1	Direct splicing	Stitching seams	0.0123	**0.00341**	1.26
		Overall image	0.00979	0.00361	
	Space seamless	Stitching seams	**0.00965**	0.00353	**1.00**
		Overall image	**0.00967**	**0.00353**	
2	Direct splicing	Stitching seams	0.0171	0.00333	1.27
		Overall image	0.0135	0.00312	
	Space seamless	Stitching seams	**0.0133**	**0.0031**	**1.00**
		Overall image	**0.0133**	**0.00306**	
3	Direct splicing	Stitching seams	0.0245	0.00365	1.28
		Overall image	0.0192	0.00358	
	Space seamless	Stitching seams	**0.0189**	**0.00361**	**0.99**
		Overall image	**0.019**	**0.00346**	
4	Direct splicing	Stitching seams	0.039	0.000953	1.30
		Overall image	0.0301	**0.000564**	
	Space seamless	Stitching seams	**0.0295**	**0.000795**	**0.99**
		Overall image	**0.0298**	0.000699	
Total	Direct splicing	Stitching seams	0.0253	0.00283	1.28
		Overall image	0.0197	0.00272	
	Space seamless	Stitching seams	**0.0193**	**0.00276**	**0.99**
		Overall image	**0.0195**	**0.00269**	

IV. Conclusion

The generation of discontinuity in transitions between adjacent image blocks primarily arises from two aspects. First, the edge-padding operations on feature maps during large stacks of convolutional layers upscaling and downscaling of convolution-based generative models lead to a degradation in the generation quality at the edges of the image blocks. Second, the artificial segmentation between input image blocks results in spatial discontinuity, thereby affecting the continuity among generated results. We address this issue of discontinuity between two neighboring image blocks by expanding the input range and cropping the output image blocks.

In this article, we comparatively analyze the fusion effects of deep learning-based spatiotemporal fusion techniques at different resolutions. We find that as the resolution of the fused images increases, the stitching seam error between image blocks is larger, and the RMSE at the seam reaches 1.279 times of the overall RMSE at 8-m resolution, which reveals the problem of discontinuities in the image stitching process.

In response to the above. We offer an innovative solution: a spatially seamless stitching approach. This method makes the input image blocks continuous with each other and cover each other, thus making the transition between the generated image blocks more continuous and smooth. The experiments verify that our approach can significantly reduce transitional discontinuity and improve visual effects noticeably, and the RMSE at the seam is kept between 0.99 1 with the overall RMSE.

In summary, this study proposes and validates a new image stitching method that effectively improves the quality of spatiotemporal fusion of remote sensing images under the burden of complexity that can be afforded. This is of great significance for remote sensing applications such as remote sensing mapping based on fused images. We expect this new stitching method to be more widely used in the future deep learning-based spatiotemporal remote sensing image fusion. In addition, the factors affecting the quality of spatiotemporal remote sensing image fusion include the quality of spatial alignment of remote sensing images and image imaging time differences, in addition to the spatial resolution of the images. In the future, we will study the sensitivity analysis of spatiotemporal fusion techniques by combining these influencing factors.

References

[1] 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.

[2] X. Ablat, C. Huang, G. Tang, N. Erkin, and R. Sawut, “Modeling soil co2 efflux in a subtropical forest by combining fused remote sensing images with linear mixed effect models,” *Remote Sens.*, vol. 15, no. 5, 2023, Art. no. 1415.

[3] F. Zhang et al., “An advanced spatiotemporal fusion model for suspended particulate matter monitoring in an intermontane lake,” *Remote Sens.*, vol. 15, no. 5, 2023, Art. no. 1204.

[4] M. Li et al., “Stability analysis of unmixing-based spatiotemporal fusion model: A case of land surface temperature product downscaling,” *Remote Sens.*, vol. 15, no. 4, 2023, Art. no. 901.

[5] 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, 2018, Art. no. 527.

[6] 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.

[7] M. Lu, J. Chen, H. Tang, Y. Rao, P. Yang, and W. Wu, “Land cover change detection by integrating object-based data blending model of landsat and modis,” *Remote Sens. Environ.*, vol. 184, pp. 374–386, 2016.

[8] 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, 2010.

[9] A. Li, Y. Bo, Y. Zhu, P. Guo, J. Bi, and Y. He, “Blending multi-resolution satellite sea surface temperature (SST) products using Bayesian maximum entropy method,” *Remote Sens. Environ.*, vol. 135, pp. 52–63, 2013.

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

[11] Q. Wang and P. M. Atkinson, “Spatio-temporal fusion for daily sentinel-2 images,” *Remote Sens. Environ.*, vol. 204, pp. 31–42, 2018.

[12] 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, 2016.

[13] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, “An augmented linear mixing model to address spectral variability for hyperspectral unmixing,” *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

[14] G. Chen, P. Jiao, Q. Hu, L. Xiao, and Z. Ye, “SwinSTFM: Remote sensing spatiotemporal fusion using swin transformer,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5410618.

[15] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[16] R. Girshick, “Fast r-cnn,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[17] 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.

[18] D. Hong et al., “Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks,” *Remote Sens. Environ.*, vol. 299, 2023, Art. no. 113856.

[19] Z. Li, K. Zheng, L. Ni, and L. Gao, “Level merging attention based on dense network for remote sensing image scene classification,” in *Proc. Int. Conf. Remote Sens., Mapping, Geographic Syst.*, SPIE, 2023, vol. 12815, pp. 88–93.

[20] D. Hong, J. Yao, C. Li, D. Meng, N. Yokoya, and J. Chanussot, “Decoupled-and-coupled networks: Self-supervised hyperspectral image super-resolution with subpixel fusion,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5527812.

[21] Z. Chen, D. Hong, and H. Gao, “Grid network: Feature extraction in anisotropic perspective for hyperspectral image classification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 5507105.

[22] Z. Chen et al., “Temporal difference-guided network for hyperspectral image change detection,” *Int. J. Remote Sens.*, vol. 44, no. 19, pp. 6033–6059, 2023.

[23] A. Radford et al., “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, 2019, Art. no. 9.

[24] D. Hong et al., “Spectralgpt: Spectral foundation model,” 2023, *arXiv:2311.07113*.

[25] 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.

[26] D. Jia, C. Cheng, C. Song, S. Shen, L. Ning, and T. Zhang, “A hybrid deep learning-based spatiotemporal fusion method for combining satellite images with different resolutions,” *Remote Sens.*, vol. 13, no. 4, 2021, Art. no. 645.

[27] 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.

[28] 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.

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

[30] 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.

[31] Y. Li, J. Li, L. He, J. Chen, and A. Plaza, “A new sensor bias-driven spatio-temporal fusion model based on convolutional neural networks,” *Sci. China Inf. Sci.*, vol. 63, pp. 1–16, 2020.

[32] W. Li, C. Yang, Y. Peng, and J. Du, “A pseudo-Siamese deep convolutional neural network for spatiotemporal satellite image fusion,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1205–1220, 2022.

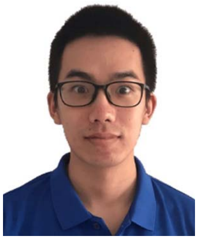
[33] G. Yang et al., “Msfusion: Multistage for remote sensing image spatiotemporal fusion based on texture transformer and convolutional neural network,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4653–4666, 2022.

[34] C. Shang et al., “Spatiotemporal reflectance fusion using a generative adversarial network,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5400915.

[35] J. Chen, L. Wang, R. Feng, P. Liu, W. Han, and X. Chen, “CycleGAN-STF: Spatiotemporal fusion via cyclegan-based image generation,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5851–5865, Jul. 2021.

[36] I. V. Emelyanova, T. R. McVicar, T. G. Van Niel, L. T. Li, and A. I. 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, 2013.

[37] 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.



ChenYang Weng received the bachelor of engineering degree from the China University of Geosciences, Wuhan, China, in 2021. He is currently working toward the master’s degree in cartography and geographic information systems with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

His research interests include data fusion, deep learning, and quantitative remote sensing.

Yulin Zhan received the B.S. and M.S. degrees from Nanchang University, Nanchang, China, in 1998 and 2002, respectively, and the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2005.

He is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His current research interests include data fusion, crop classification, and remote sensing of the urban environment.

Xingfa Gu received the B.S. degree in surveying and mapping from Wuhan University, Wuhan, China, in 1982, and the Ph.D. degree in remote sensing of physics from the University of Paris VII, Paris, France, in 1991.

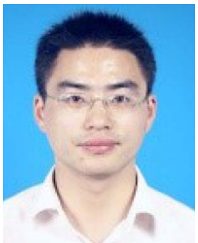
He is currently a Professor with Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China, and an Academician of the International Academy of Astronautics. His research interests include radiometric calibration and atmospheric remote sensing.

Jian Yang received the Ph.D. degree in cartography and geography information system from the Institute of Remote Sensing Application, Chinese Academy of Sciences, Beijing, China, in 2008.

He is currently a Professor with Aerospace Information Research Institute, Chinese Academy of Sciences. His main research interests include remotely sensing imagery processing and analyzing algorithms on high spatial resolution RS image, including multifeature extraction, multisource data fusion, land cover classification, and change detection of urban region.

Yan Liu received the B.S. degree from Lanzhou University, Lanzhou, China, in 2009, the M.S. degree in remote sensing from Beijing Normal University, Beijing, China, in 2012, and the Ph.D. degree in environmental science from the University of Massachusetts, Boston, Boston, MA, USA, in 2016.

She is currently an Assistant Research Fellow with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing. Her research interests include data fusion, BRDF inversion and application, and phenology monitoring using multiply scales data.



Hong Guo received the Ph.D. degree in cartography and geography information system from the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China, in 2014.

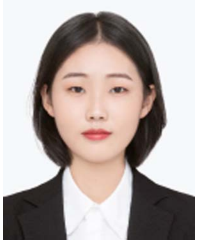
He is currently an Associate Professor with Aerospace Information Research Institute, Chinese Academy of Sciences. His research interest includes aerosol parameters inversion algorithms.

Zilong Lian is currently working toward the master's degree in aeronautical and astronautical science and technology from the School of Remote Sensing Information Engineering, North China Institute of Aerospace Engineering, Langfang, China.

His research interest includes the spatiotemporal fusion method of remote sensing.

Shiyuan Zhang received the bachelor's degree in network engineering from Jinan University, Guangzhou, China, in 2018. He is currently working toward the master's degree in aeronautical and astronautical science and technology with the School of Remote Sensing Information Engineering, North China Institute of Aerospace Engineering, Langfang, China.

His research interest includes leaf area index.



Zhangjie Wang received the bachelor of engineering degree from Yunnan Police College, Kunming, Yunnan, in 2021. She is currently working toward the master's degree in aeronautical and astronautical science and technology with the School of Remote Sensing Information Engineering, North China Institute of Aerospace Engineering, Langfang, China.

Her research interests include quantitative remote sensing and vegetation coverage.

Xuechun Zhao received the bachelor of engineering degree in 2021 from the North China Institute of Aerospace Engineering, Langfang, China, where she is currently working toward the master's degree in aeronautical and astronautical science and technology with the School of Remote Sensing Information Engineering.

Her research interests include quantitative remote sensing and vegetation coverage.