remote sensing

MDPI

Article

STF-EGFA: A Remote Sensing Spatiotemporal Fusion Network with Edge-Guided Feature Attention

Feifei Cheng [1], Zhitao Fu [1,*], Bohui Tang [1,2], Liang Huang [1], Kun Huang [1] and Xinran Ji [1]

[1] Faculty of Land Resource Engineering, Kunming University of Science and Technology (KUST), Kunming 650093, China; 20192201027@stu.kust.edu.cn (F.C.); tangbh@kust.edu.cn (B.T.); kmhuangliang@kust.edu.cn (L.H.); 20192201033@stu.kust.edu.cn (K.H.); kmjixinran@stu.kust.edu.cn (X.J.)

[2] State Key Laboratory of Resources and Environment Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

* Correspondence: zhitaofu@kust.edu.cn

Abstract: Spatiotemporal fusion in remote sensing plays an important role in Earth science applications by using information complementarity between different remote sensing data to improve image performance. However, several problems still exist, such as edge contour blurring and uneven pixels between the predicted image and the real ground image, in the extraction of salient features by convolutional neural networks (CNNs). We propose a spatiotemporal fusion method with edge-guided feature attention based on remote sensing, called STF-EGFA. First, an edge extraction module is used to maintain edge details, which effectively solves the boundary blurring problem. Second, a feature fusion attention module is used to make adaptive adjustments to the extracted features. Among them, the spatial attention mechanism is used to solve the problem of weight variation in different channels of the network. Additionally, the problem of uneven pixel distribution is addressed with a pixel attention (PA) mechanism to highlight the salient features. We transmit the different features extracted by the edge module and the encoder to the feature attention (FA) module at the same time after the union. Furthermore, the weights of edges, pixels, channels and other features are adaptively learned. Finally, three remote sensing spatiotemporal fusion datasets, Ar Horqin Banner (AHB), Daxing and Tianjin, are used to verify the method. Experiments proved that the proposed method outperformed three typical comparison methods in terms of the overall visual effect and five objective evaluation indexes: spectral angle mapper (SAM), peak signal-to-noise ratio (PSNR), spatial correlation coefficient (SCC), structural similarity (SSIM) and root mean square error (RMSE). Thus, the proposed spatiotemporal fusion algorithm is feasible for remote sensing analysis.

Keywords: convolutional neural network; remote sensing; spatiotemporal; image fusion; feature attention; edge extraction

Citation: Cheng, F.; Fu, Z.; Tang, B.; Huang, L.; Huang, K.; Ji, X. STF-EGFA: A Remote Sensing Spatiotemporal Fusion Network with Edge-Guided Feature Attention. *Remote Sens.* **2022**, *14*, 3057. https://doi.org/10.3390/rs14133057

Academic Editor: Giuseppe Scarpa

Received: 10 May 2022
Accepted: 23 June 2022
Published: 25 June 2022

1. Introduction

The continuous launches of satellites enable a drastic increase in remote sensing data, and new sensors have been developed in the direction of high spectral, spatial and temporal resolution. These sensors include those used by the Landsat, Sentinel and Moderate Resolution Imaging Spectrometer (MODIS) series satellites. However, remote sensing images are limited in spatial, temporal and spectral resolution due to different sensors. Since the 1990s, researchers have been using fusion techniques to combine information from multiple bands or satellites for scientific purposes [1,2]. Through continuous research by scholars, many methods involving spatiotemporal fusion [3,4], spatial-spectral fusion [5] and spatio-temporal-spectral integration [6] have emerged, which have greatly decreased their restrictions and thus greatly improved the application of imagery. Such applications include simulating the soil water content in heterogeneous areas [7], mapping evapotranspiration [8], land use change [9] and extracting abandoned land areas [10].

In studying the spatiotemporal characteristics of land surface features, the spatial and temporal resolutions of satellite sensors have very important effects, but a remote sensing image acquired by a single satellite sensor may be limited by mutual spatial and temporal resolution constraints. Spatiotemporal fusion [11] is a combination of heterogeneous satellite images with high spatial (Landsat) and high temporal (MODIS) resolution to generate dense sequence images with high temporal and spatial resolution using spatiotemporal fusion algorithms. It can be used to effectively generate high spatiotemporal resolution images at low cost and with high efficiency. To date, domestic and foreign scholars have performed extensive research on spatiotemporal fusion algorithms for different data types. However, there are still some problems, such as edge contour blurring and uneven pixels between the predicted image and the real ground image, in the extraction of salient features by convolutional neural networks (CNNs). Therefore, it would be meaningful to develop a method to increase the edge information and focus on the effective information of domains from different remote sensing images to obtain fused images with high spatiotemporal resolution. In this study, we propose such a fusion method for the spatiotemporal fusion of remote sensing images and apply it to three different real remote sensing datasets.

2. Related Work

To date, domestic and foreign scholars have performed extensive research on spatiotemporal fusion algorithms for different data types. In recent years, some scholars [3,4,12,13] have provided detailed reviews of spatiotemporal fusion methods and have sorted and classified the existing methods. At present, most scholars classify them into methods based on weight function, decomposition, model and deep learning.

The spatial and temporal adaptive reflectance fusion model (STARFM) [14] was the first to fuse Landsat and MODIS images with weighted spatiotemporal data and is one of the most widely used algorithms for spatiotemporal fusion. However, the method has some limitations when dealing with complex scenes, and many improved methods are available, such as the enhanced version of STARFM (ESTARFM) [15], the spatiotemporal adaptive algorithm for mapping reflectance change (STAARCH) [16] and spatiotemporal restraint unmixing (STRUM) [17]. ESTARFM [15] is an enhanced spatiotemporal adaptive surface reflectivity fusion model that can increase the accuracy of extraction in different image regions, thus broadening the application of spatiotemporal fusion algorithms for complex scenes. STAARCH [16] is an adaptive spatiotemporal algorithm that maps reflectivity change to detect change points. STRUM [17] is also based on the basic framework of STARFM and predicts changes in ground objects between reference and time phases through mixed-pixel decomposition.

Another scholar proposed a decomposition-based spatiotemporal fusion method, spatial filtering, regression model fitting and residual compensation (FIT-FC) [18] in a new approach. The linear regression (LR)-based fusion algorithm combines spatial filtering and residual compensation to classify spatiotemporal images and has yielded satisfactory results. At present, flexible spatiotemporal data fusion (FSDAF) [19] is also used in a wide range of decomposition methods, and it enables spatiotemporal image prediction with improved spatial detail by combining spectral decomposition and thin-slab sample interpolation concepts to capture the reflectance changes caused by land cover transitions. Additionally, this approach can maintain more spatial detail than can STARFM. Based on this algorithm, scholars also proposed an enhanced FSDAF (EFSDAF) algorithm [20], an improved flexible spatiotemporal data fusion (IFSDAF) approach [21] and other algorithms to improve the image fusion capability of FSDAF. In recent years, several scholars have proposed graphics processing unit (GPU) [22] algorithms for parallel computing, aiming to speed up existing algorithms such as cuSTARFM and cuFSDAF [23].

In the last decade, algorithms that utilize machine models have been increasing and have been heavily researched. Examples include spatiotemporal reflectance fusion via sparse representation (SPSTFM) [24], compression sensing for spatiotemporal fusion (CSSF) [25] and so on. Huang et al. [24] proposed the first spatiotemporal reflectance fusion

via sparse representation (SPSTFM) by using spatiotemporal image pairs, building a priori high temporal resolution images and a priori high spatial resolution images, considering the reflectance variation relationship of the images through dictionary learning, and finally obtaining high temporal resolution time series images. However, there are still some issues with maintaining spectral fidelity and predicting spatial details. Chen et al. [26] proposed a hierarchical spatiotemporal adaptive fusion model (HSTAFM) that adaptively fuses multisensor features to accurately capture seasonal changes and land use/cover changes by enhancing coarse-resolution images with super-resolution information based on sparse representations, followed by preselection for temporal changes, and selecting similar pixels using a two-level strategy. On this basis, Li et al. [27] designed a single-pair learning-based SPSTFM method, combining spatial and temporal expansion models to increase the training set, improve the spatial resolution of high temporal resolution by improving dictionary learning, combine high-pass information obtained by the module fused with high spatial resolution imagery, and successfully improve the accuracy of spatiotemporal prediction. This method improves the prediction effect of SPSTFM.

CNNs have been applied to spatiotemporal image fusion deep learning methods [28,29], and the fusion performance has been improved. The deep CNN for spatiotemporal fusion (STFDCNN) [30] reconstructs spatial resolution images from temporal resolution images using a spatiotemporal fusion method involving deep CNNs; notably, nonlinear mapping super-resolution-based CNNs are less efficient than the sparse representation method. However, the spatial resolution of Landsat and MODIS images is quite different, so using high temporal resolution images to reconstruct high spatial resolution images will bring instability. Li et al. [31] used CNNs to calculate reflectance variations between images to model the heterogeneity of fine pixels from high temporal resolution images, providing a stable and less time-consuming strategy. Li et al. [32] also adopted a spatiotemporal fusion model driven by sensor-bias (BiaSTF) and further used a CNN to learn the bias information between two images, alleviating the spectral and spatial distortion problems in traditional methods. Both of these methods effectively enhance the feature linkages between images and improve image fusion. Spatial information loss is an obvious problem in high temporal resolution images, and scholars have proposed a variety of two-stream CNNs. Liu et al. [33] used a two-stream CNN (StfNet) that considers temporal correlation and temporal consistency between image sequences and can predict fine images not only based on structural similarity (SSIM) but also on texture information in temporally adjacent images to predict fine images and perform spatiotemporal fusion in a CNN-based super-resolution process. Chen et al. [34] designed a combination of a methods using multiscale dual-stream CNNs (STFMCNNs) and atrous spatial pyramid pooling (ASPP) to extract image pair multiscale features, additionally exploiting complementary-based temporal dependencies and temporal consistency information, which can accurately predict differences to obtain predicted images that are similar to real images. Jia et al. [35] also used a mapping method based on temporal variability and spatial information, combining forward and backwards predictions to generate prediction images with high spatial and temporal resolution, with great predictions of vegetation phenology and land use change ability. Tan et al. [36] designed a deep convolutional spatiotemporal fusion network (DCSTFN). It requires an artificial design to build a nonlinear mapping relationship and the extracted features to reconstruct the predicted image. Because the DCSTFN requires a hypothetical equation for mapping between different images, an enhanced DCSTFN (EDCSTFN) was proposed [37], and no hypothetical equation is required in this approach. The difference between reference data and predicted data is learned entirely from the actual data. In addition, a new composite loss function was constructed, and it significantly improved upon the traditional function. Overall, the EDCSTFN model displayed superior performance compared to traditional models, with higher accuracy, visual quality and robustness.

However, there are still some problems, such as edge contour blurring and uneven pixels between the predicted image and the real ground image in the extraction of salient features by CNN. Considering the problems above, we designed a remote sensing spa-

tiotemporal fusion network based on edge information–guided feature fusion attention (called EIFA-STF). It can reduce information loss during network layer transfers while focusing on the key available information. The model is validated with three remote sensing spatiotemporal fusion datasets, namely Ar Horqin Banner (AHB), Daxing and Tianjin, which include different types of variations and large differences in MODIS and Landsat. The generalization and robustness of the STF-EGFA model are fully validated by comparing our proposed method with three representative methods. Our main contributions are as follows:

1. We design a spatiotemporal fusion method with edge-guided feature attention based on remote sensing, called STF-EGFA, which strengthens the connections among features in different layers and reduces information loss while utilizing multilayer features.
2. The edge extraction module in STF-EGFA is mainly designed to decrease the boundary information loss in the process of feature extraction and improve the retention of edge details at high spatial resolution to ensure that the predicted spatiotemporal images retain more saliency.
3. The design of the feature attention (FA) module in STF-EGFA focuses on the key available information by using an FA mechanism guided by edge information. Information weighting and pixel heterogeneity are optimized among different channels in the network to provide more accurate predictions of spatiotemporal changes.

The rest of this paper is organized as follows: The second part introduces each module of EIFA-STF. Section 3 introduces the specific implementation process of the method in detail and evaluates the method objectively and subjectively with other typical remote sensing spatial data fusion methods. Section 4 discusses and analyses the method through an ablation test. Section 5 concludes this paper and provides an outlook on future work.

3. Methods

The basic framework of STF-EGFA follows the overall EDCSTFN network architecture [37]. The edge extraction module and FA module are added on the basis of the original network, as shown in Figure 1, with a dual encoder–decoder network structure and a total of five network modules. The first encoder, called the FEncoder, is mainly used to extract features from images with high spatial resolution. The second encoder, called the REncoder, is mainly used to extract the feature differences between high spatial resolution images and high temporal resolution images. The third module is the edge extraction module, which is mainly used to extract the edge features of the input high spatial resolution images. The fourth module is the FA module, which includes a combination of channel attention (CA) and pixel attention (PA) mechanisms and is mainly used to solve problems related to different weights among different channels and heterogeneous pixel distributions in different images. The fifth module is the reconstruction decoder used to generate the prediction image.

First, two groups of images at $t0$ and $t2$ and the MODIS image at $t1$ needing to be predicted were simultaneously input into the encoding modules. As shown in Figure 2, Landsat images at $t1$ were predicted through $t0$ and $t2$ data, and real Landsat images of $t1$ at the predicted time were input into the network as labels to calculate feature losses. Second, the input Landsat data at $t0$ and $t2$ were used to extract edge information through the edge extraction module, and the extracted edge information was added to the encoder information. Next, by inputting the features extracted from the two encoders and the edge feature module into the FA module, the spatiotemporal features to be predicted were generated. Finally, the decoder reconstructed the obtained features to generate the predicted image.

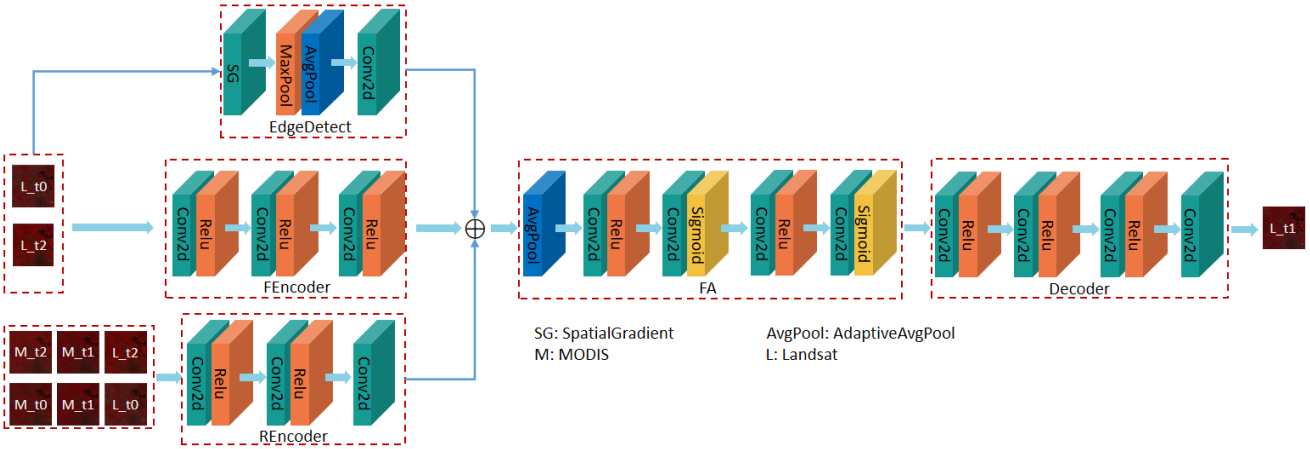

Figure 1. Architecture of the deep convolutional spatiotemporal fusion network (STF-EGFA) model.

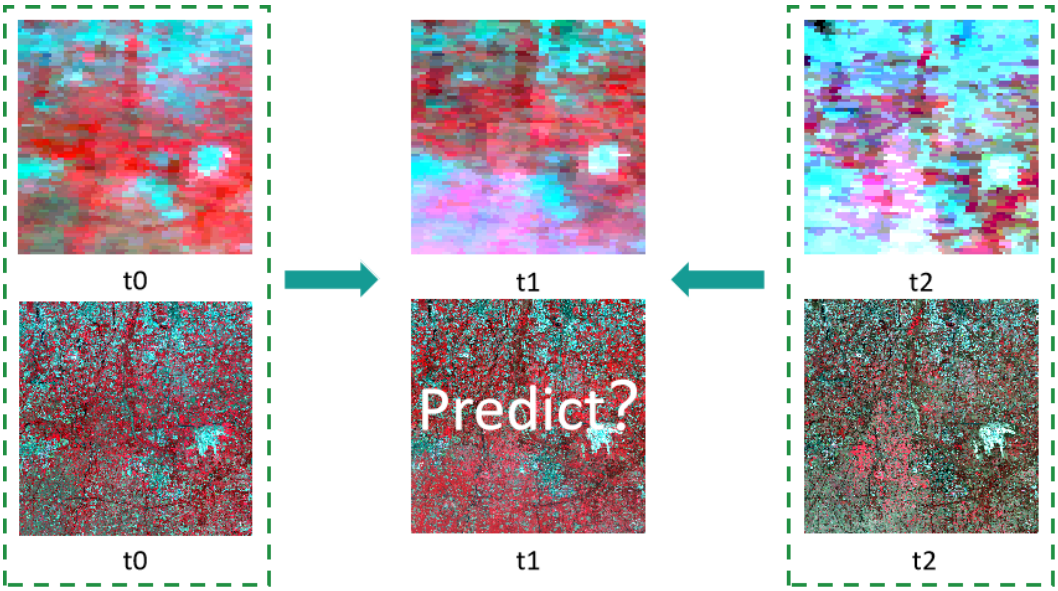

Figure 2. Predicted image based on spatiotemporal fusion, in which all images of $t0$ and $t2$ and the MODIS at time $t1$ are known images, and the Landsat image of $t1$ is the prediction image.

The whole process can be expressed by the following equation:

$$L_{t1} = F(s(f(L_{tk}) + r(L_{tk}, M_{t1}, M_{tk}) + e(L_{tk}))) \ (k \neq 1) \quad (1)$$

where f is the FEncoder encoder, r is the REncoder residual encoder, e is the edge extraction module, s is the FA module, and F is the reconstruction decoder.

3.1. Edge Feature Extraction

Inspired by previous edge feature extraction methods [38], an edge feature extraction module was designed with partial edge information lost, as shown in Figure 3. EdgeDetect contains four parts: the spatial gradient, Maxpool, Avgpool and Conv2d. By using the edge guidance module, the edge features of the image are enhanced to preserve the detail information in the input image and to reduce noise and artefacts while preserving the detail information to achieve a predicted image with abundant detail features. First, the spatial gradient operator is used to identify the horizontal and vertical edge features of the image, and to further enhance the gradient information, the edge enhancement operator S is applied to the available gradient information. Finally, the extracted edge feature

dimensions are converted to feature maps with the same dimension as the encoder output through the convolution layer. The edge enhancement operator S is expressed as follows:

$$S(\nabla L) = \underbrace{\max}_{j=J}(\underbrace{\max}_{i=I}\nabla L(i+1, j+1), \nabla L(i, j)) \quad (2)$$

where ∇L is the gradient feature of the input image, $I = \{1 \ldots, m-1\}$ and $J = \{1 \ldots, m-1\}$. The indexes i and j represent the horizontal and vertical directions of the gradient image, respectively.

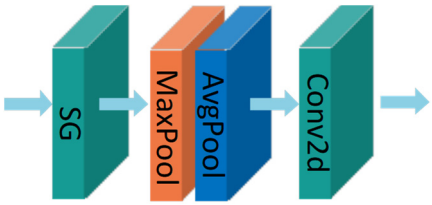

Figure 3. Architecture diagram of the edge-guided extraction module.

3.2. Feature Attention

The retention of edge information can avoid the degradation of details in the process of fusion. Attention mechanisms have been successfully applied in many computer vision tasks due to their ability to capture regions of interest in visual scenes, and the main goal of fusion is to find the appropriate features for each channel. Notably, an FA mechanism [39] can handle images with uneven pixel distributions and channel weights, extend the representation capability of CNNs and achieve good results in restoring the haze-free image. FA includes CA and PA, and Figure 4 shows the network structure of the FA module.

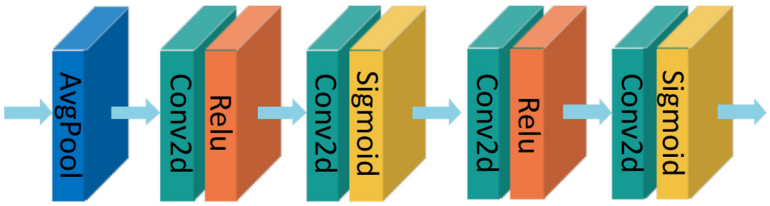

Figure 4. Structural diagram of the feature attention module.

3.2.1. Channel Attention

A CA mechanism [39] learns weights according to the importance of features in different channels, and the weights are different for each channel; however, the weights at different pixel positions for pixels in the same channel are the same, as shown in Figure 5. Considering the complexity of channels in remote sensing spatiotemporal images, a CA module is introduced in this paper. First, we use global average pooling to input the global spatial information from the channel into the channel descriptor, and the feature map size is changed from C × H × W to C × 1 × 1. Subsequently, to obtain the weights of different channels, the features are subsequently convolved through two layers, and sigmoid and rectified linear unit (ReLU) activation functions are used. The size of the obtained feature map is C × 1 × 1. Finally, the input image and the channel weights are multiplied in an element-by-element process to obtain the final CA feature map of size C × H × W.

3.2.2. Pixel Attention

The PA mechanism [39] learns weights according to the importance of features at different pixel locations, with the same weight for each channel but different weights at different pixel locations for pixels in the same channel, as shown in Figure 6. Considering the complexity and uneven distribution of remotely sensed spatiotemporal images, the PA module is introduced in this paper; notably, the feature information in high-frequency image regions can be prioritized. First, the input image is passed through two convolution

layers, and sigmoid and ReLU activation functions are applied. Then, the feature map is resized from C × H × W to 1 × H × W. Subsequently, the input image and PA features are multiplied in an element-by-element process to obtain the final PA feature map of size C × H × W.

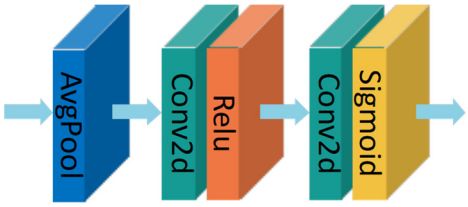

Figure 5. Structural diagram of the channel attention module.

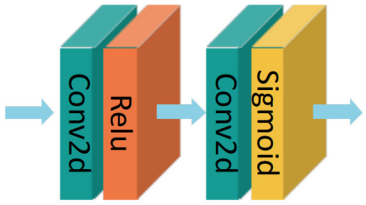

Figure 6. Structural diagram of the pixel attention module.

3.3. Loss Function

To improve the sharpness of the predicted, remotely sensed, spatiotemporally fused images, we applied the loss function concept in EDCSTFN [37] in our experiments to define a new composite loss function. Additionally, in the EDCSTFN, a composite mean square error (MSE) loss function was used to constrain content loss and feature loss; however, Lim et al. [40] showed that in terms of peak signal-to-noise ratio (PSNR) and SSIM metrics, many image recovery models trained with L1 loss functions achieve better performance than those trained with L2 loss functions. To obtain better performance in image prediction, we replaced MSE loss with the more well-constrained L1 loss, and the new composite loss function was constructed as follows:

$$\text{loss} = L_C + L_F + \alpha \cdot L_V \tag{3}$$

where L_C is the content loss, L_F is the feature loss, L_V is the visual loss, α is a scaling factor used to balance the weight of visual loss in the compound loss function, and the value of α is 0.8.

Among them, the content loss and feature loss use the L1 loss function, which can effectively recover image details, and the visual loss function uses MS-SSIM as the loss metric to effectively retain high-frequency feature information. By combining three different loss functions, the loss of features is constrained, and clear predicted images are obtained.

The L1 loss function, a commonly used regression loss function based on the mean of the absolute value of the difference between the target value and the predicted value, reflects the average error magnitude of the predicted value, regardless of the type of input. This function yields a stable gradient and avoids the gradient explosion problem. Without considering the direction of the error, the L1 loss function is stable and robust.

$$L1 = \frac{1}{N}\sum_{i=1}^{N} \|\widehat{FL}_{t1} - FL_{t1}\| \tag{4}$$

where N denotes the number of elements in the feature map, $\widehat{FL}_{t1}$ denotes the features extracted from the STF-EGFA prediction using the pretrained encoder, and FL_{t1} denotes the features extracted from the observed data on the prediction date $t1$.

4. Experiments and Evaluations

The STF-EGFA model has five parts, and the compound loss function is applied in the experiments. We conduct experiments with three different datasets, AHB, Daxing and Tianjin, to verify the generalization and robustness of the STF-EGFA model. We use three different remote sensing spatiotemporal fusion methods for subjective and objective evaluation: STARFM [14], FSDAF [19] and EDCSTFN [37]. The compared methods are tested according to the recommended parameters. Experiments were performed on an Intel(R) Core (TM) i9-10850K central processing unit (CPU) @3.60 GHz, NVIDIA GeForce RTX 3080, and 64 GB RAM configuration, and experiments were performed using PyTorch.

4.1. Datasets

To validate the characteristics of spatiotemporally fused remote sensing images, we use the benchmark dataset proposed by Jun et al. [12] in the field of spatiotemporal fusion. The three datasets in the following experiments are called the AHB dataset, Tianjin dataset and Daxing dataset. Table 1 provides more details. All Landsat images were acquired with the Landsat 8 Operational Land Imager (OLI), with a total of 6 bands. For MODIS images, the AHB dataset is from MOD09GA images, while the Tianjin and Daxing datasets are from MOD02HKM. In the experiment, the spatiotemporal fusion becomes more complicated due to the large obvious strip noise interference between the two shortwave infrared bands, and only the front four bands are used in the experiment. The three datasets are allocated in the same way: the training set is 80% of the datasets (7 groups), and the validation and test sets are 20% of the datasets (2 groups). Three image pairs, namely $t0$, $t1$ and $t2$, constitute a set of training data.

Table 1. The experimental dataset details.

Dataset	Image Size	Experimental Image Size	Experimental Image Pairs	Experimental Image Time Span	Main Change Types
AHB	2480 × 2800 × 6	2432 × 2432 × 6	27	30 May 2013–6 December 2018	Ar Horqin Banner of Inner Mongolia province
Tianjin	2100 × 1970 × 6	1920 × 1920 × 6	27	1 September 2013–18 September 2019	Tianjin city
Daxing	1640 × 1640 × 6	1536 × 1536 × 6	27	1 September 2013–1 August 2019	Daxing district of Beijing

4.1.1. AHB Dataset

The AHB dataset [12] was established to measure phenology changes in rural areas. A total of 27 pairs of images from a cloud-free Landsat-MODIS image set obtained from 30 May 2013 to 6 December 2018, spanning a period of more than 5 years, were used in the experiment. The first and second rows in Figure 7 show the MODIS images (linearly stretched by 2%) of the AHB dataset, while the first and second rows of the AHB dataset in Figure 8 show Landsat images (linearly stretched by 2%). The figures show that there are large differences in the surface features associated with the times the images were acquired, especially in the yellow elliptical areas (Figures 7 and 8), with significant phenology changes.

4.1.2. Tianjin Dataset

The Tianjin dataset [12] was proposed to detect phenological changes in urban areas. A total of 27 image pairs were used in the experiment. The images were cloud-free Landsat-MODIS images obtained from 1 September 2013 to 18 September 2019, spanning over 6 years. The first and second rows in Figure 9 show the MODIS images (linearly stretched by 2%) of the Tianjin dataset, while the first and second rows of the Tianjin dataset in Figure 10 show Landsat images (linearly stretched by 2%). The figures show that the

surface features in the city differ in color and texture over time, with significant urban phenology changes, especially in the yellow elliptical areas (Figures 9 and 10).

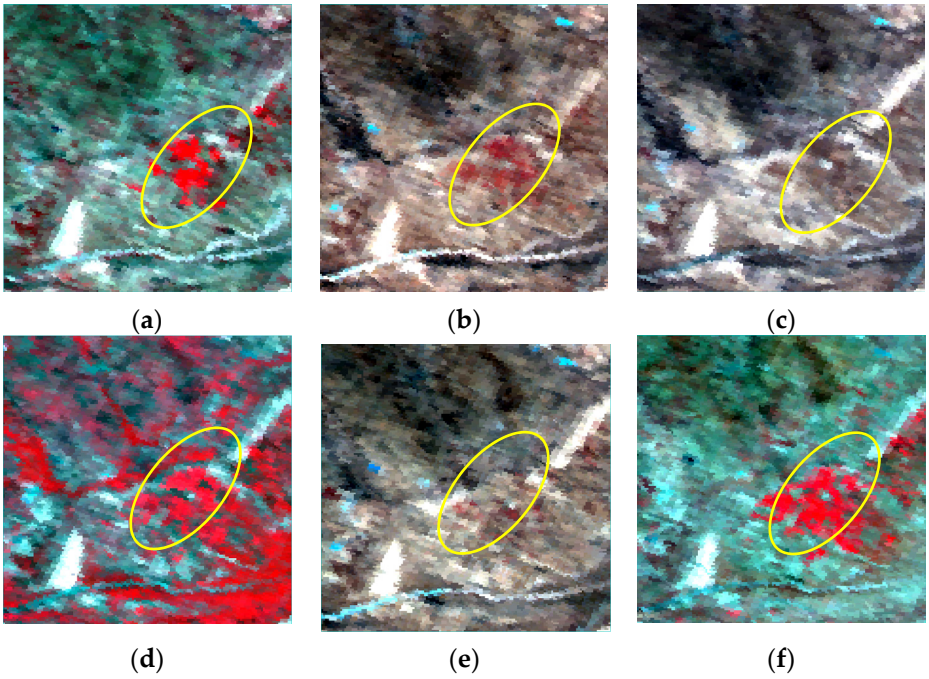

Figure 7. MODIS images from the AHB dataset reflecting changes in phenology in rural and mountainous areas. Dates: (**a**) 30 May 2013; (**b**) 10 February 2014; (**c**) 17 March 2015; (**d**) 26 August 2016; (**e**) 18 February 2017; (**f**) 3 October 2018.

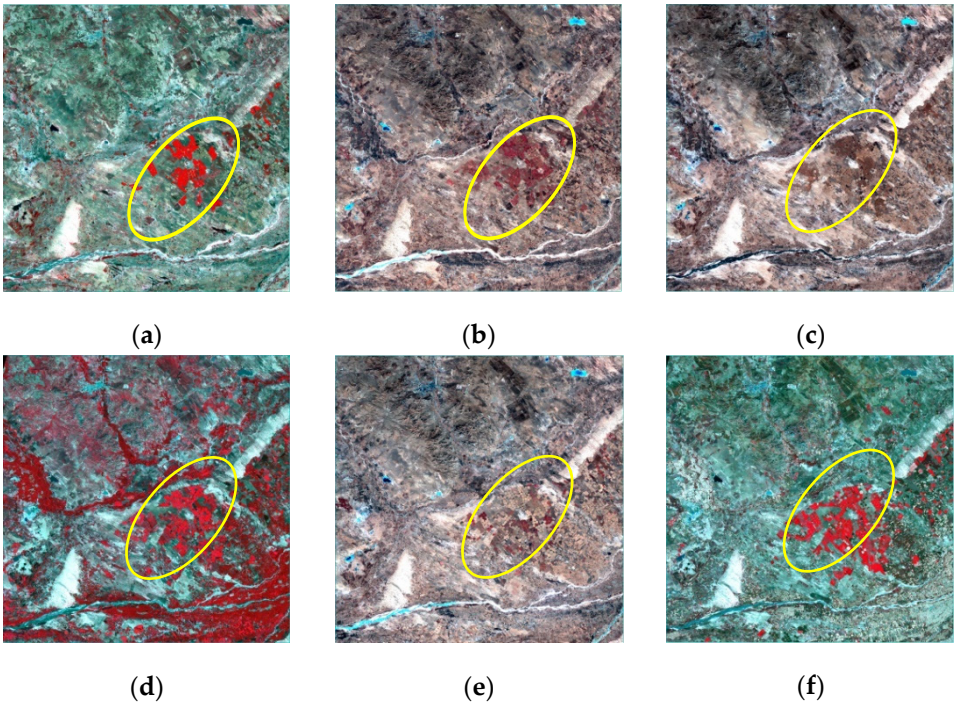

Figure 8. Landsat images from the AHB dataset reflecting changes in phenology in rural and mountainous areas. Dates: (**a**) 30 May 2013; (**b**) 10 February 2014; (**c**) 17 March 2015; (**d**) 26 August 2016; (**e**) 18 February 2017; (**f**) 3 October 2018.

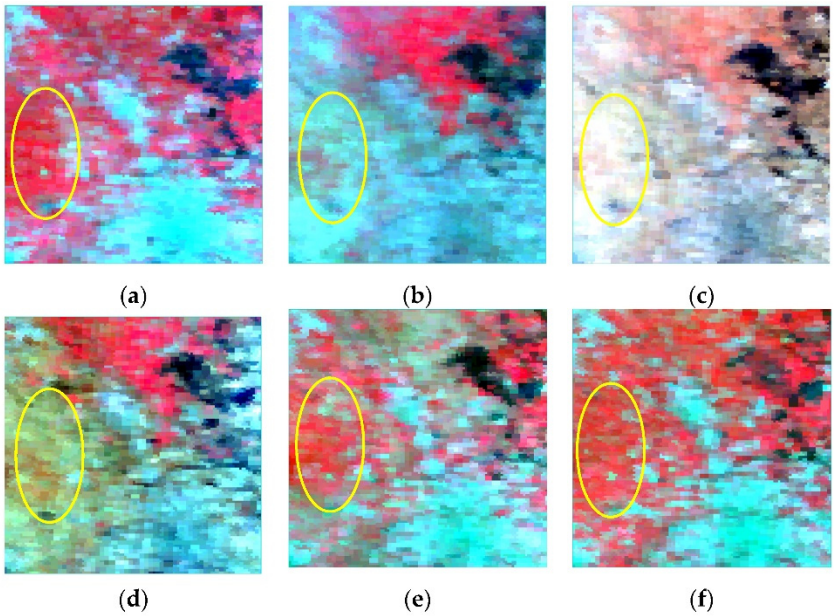

Figure 9. MODIS images from the Tianjin dataset reflecting changes in urban phenology. Dates: (**a**) 1 September 2013; (**b**) 29 April 2014; (**c**) 15 March 2015; (**d**) 14 December 2016; (**e**) 10 July 2017; (**f**) 17 August 2019.

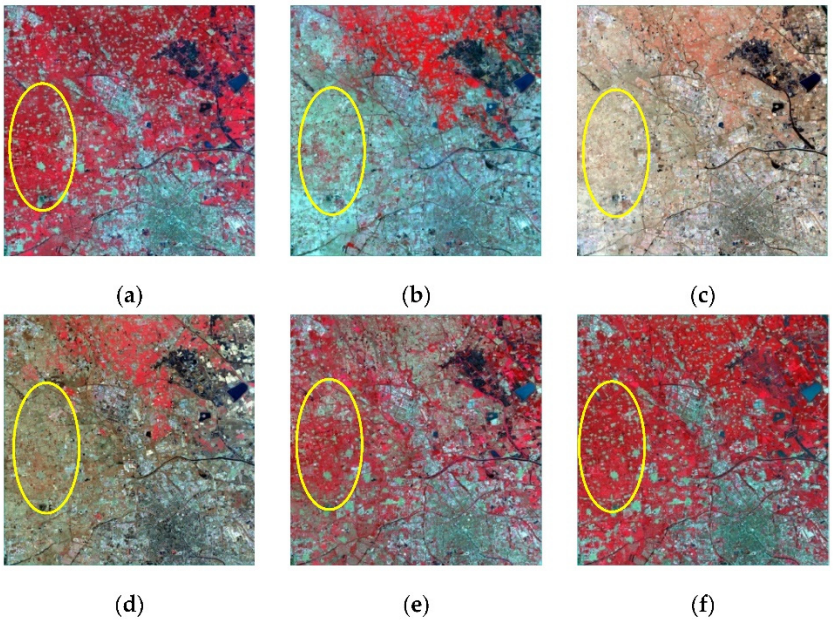

Figure 10. Landsat images from the Tianjin dataset reflecting changes in urban phenology. Dates: (**a**) 1 September 2013; (**b**) 29 April 2014; (**c**) 15 March 2015; (**d**) 14 December 2016; (**e**) 10 July 2017; (**f**) 17 August 2019.

4.1.3. Daxing Dataset

The Daxing dataset [12] was used to assess model performance in detecting land cover change. A total of 27 image pairs were used in the experiment. The images were cloud-free Landsat-MODIS images obtained from 1 September 2013 to 1 August 2019, spanning over 6 years. The first and second rows in Figure 11 show the MODIS images (linearly stretched by 2%) of the Daxing dataset, while the first and second rows of the Daxing dataset in Figure 12 show Landsat images (linearly stretched by 2%). The figures show that the surface land cover in the city changes over time, especially following the construction of Beijing

Daxing International Airport in the yellow elliptical areas (Figures 11 and 12), the ground features have undergone great changes.

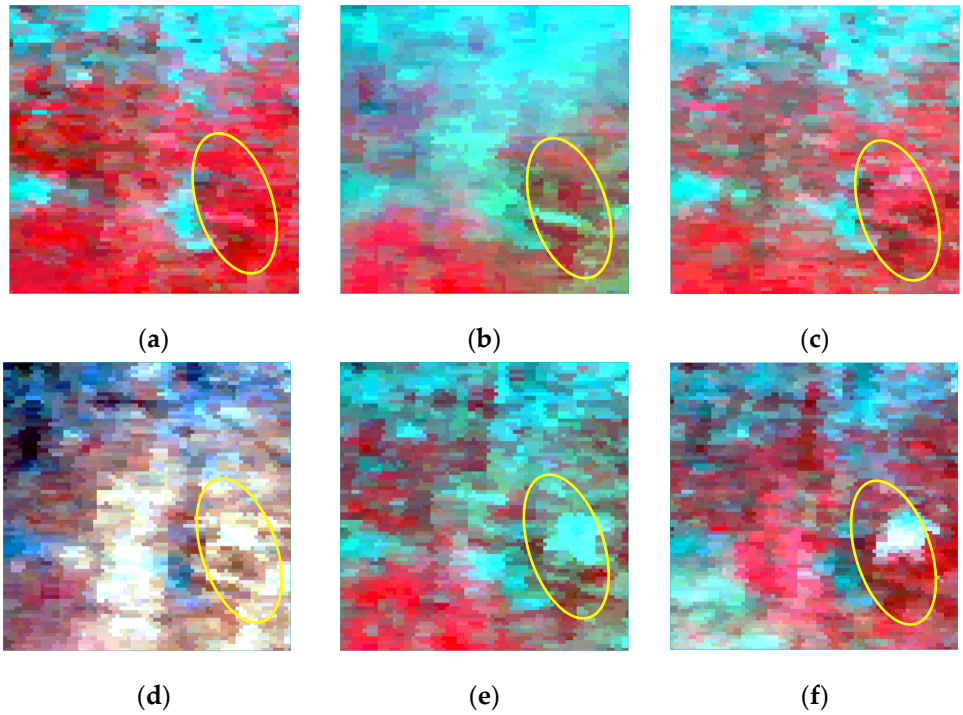

Figure 11. MODIS images from the Daxing dataset reflecting land cover changes. Dates: (**a**) 1 September 2013; (**b**) 29 April 2014; (**c**) 22 August 2015; (**d**) 14 February 2016; (**e**) 7 May 2017; (**f**) 1 October 2018.

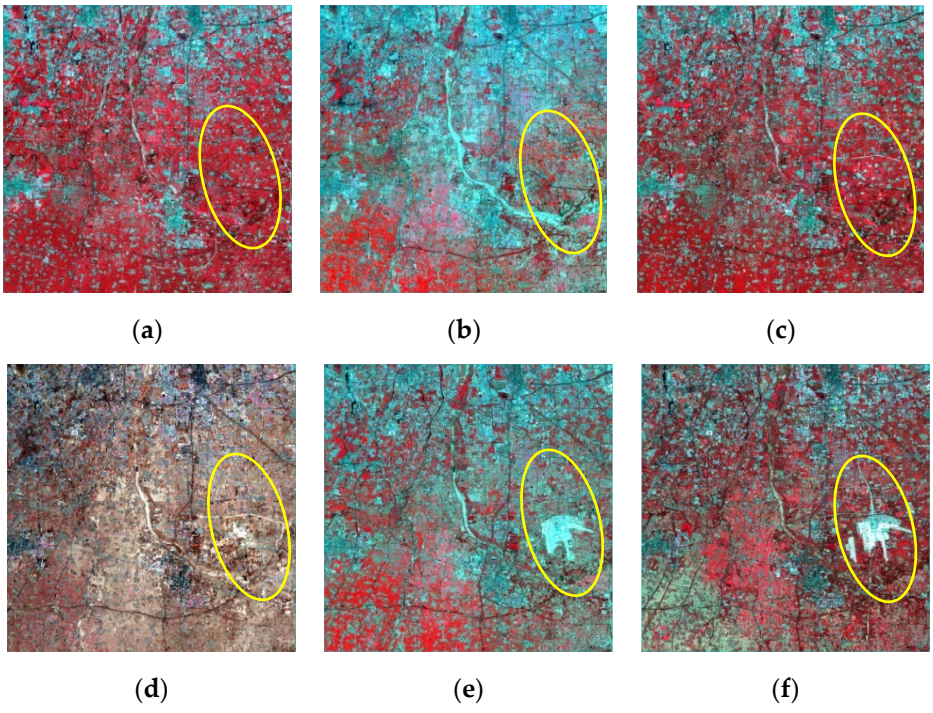

Figure 12. Landsat images from the Daxing dataset reflecting land cover changes. Dates: (**a**) 1 September 2013; (**b**) 29 April 2014; (**c**) 22 August 2015; (**d**) 14 February 2016; (**e**) 7 May 2017; (**f**) 1 October 2018.

4.2. Evaluation

Image quality evaluation has always been an important step in remote sensing spatiotemporal fusion, and to better quantify the quality of a predicted image, objective

evaluation indexes that are representative in mathematics should be applied. Additionally, actual ground data should be compared to the data in predicted and fused images. In this paper, five evaluation metrics—spectral angle mapper (SAM) [41], PSNR [42], spatial correlation coefficient (SCC) [43], SSIM [44] and root mean square error (RMSE) [45]—are used to objectively evaluate and analyze the spatiotemporal fusion results in different datasets. The SAM technique [41] is used to calculate the spectral distortion between the predicted fusion result and the original image; the PSNR [42] reflects the difference between the ground truth image and the predicted fused image based on the statistical mean of the greyscale difference between the corresponding image elements; the SCC [43] is used to assess the similarity of the spatial details in fused and reference images based on high-frequency information; the SSIM [44] measure is implemented between a predicted fused image and a ground truth image; and the RMSE [45] measures the deviation between the predicted and actual reflectance and provides a global description of the radiometric difference between a ground truth image and a predicted fused image. SAM, PSNR and RMSE are spectral quality metrics, and SCC and SSIM are spatial quality metrics.

4.3. Experimental Results and Analysis

In the experiments, we tested the effectiveness of our method with the AHB dataset, Tianjin dataset and Daxing dataset and compared the method with three remote sensing–based spatiotemporal fusion methods: STARFM [14], FSDAF [19] and EDCSTFN [37]. For the comparisons made in the subsequent experiments, we set parameter values in the trained network: the number of epochs is 30 times, the size of batch_size is 16, the cropping size of image blocks is 128 × 128, the cropping step size of image blocks is 128, the value of padding is 16 and the initial learning rate in the network is 0.001.

4.3.1. AHB Dataset

The evaluation of the spatiotemporal fusion prediction of the AHB dataset [12] is based on a combination of subjective and objective evaluation methods. Figure 13 shows the experimental results in the AHB dataset, in which the yellow box area is the enlarged detail area. In the yellow ellipse of the enlarged region, we can clearly see that STARFM predicts better edge details but relatively poor color. FSDAF produces more accurate texture details than STARFM, but block artefacts appear. The EDCSTFN algorithm can also retain texture information better than STARFM. The proposed STF-EGFA model produces complete spatial details compared to EDCSTFN, especially in the yellow elliptical region, indicating that STF-EGFA provides the best edge preservation effect.

To better explain the spatiotemporal fusion prediction effect of the proposed method, the prediction results are evaluated by calculating the average of five evaluation metrics: SAM, PSNR, CC, SSIM and RMSE. Table 2 shows the objective evaluation results of different methods for the AHB dataset, where the bold font represents the optimal value for each evaluation metric. The results show that the two deep learning methods EDCSTFN and EIFA-STF outperform the traditional STARFM and FSDAF methods in various evaluation indicators. The EIFA-STF model achieves the best results among the 5 evaluation metrics, with SAM, PSNR, CC, SSIM and RMSE improving by 8.91%, 2.15%, 12.01%, 3.45% and 15.26%, respectively, compared with the EDCSTFN method. The method uses FA and edge extraction modules to effectively enhance the spatial information, capture interchannel and pixel information, and enhance the fusion results. In summary, the method in this paper demonstrates reliable applicability to the AHB dataset.

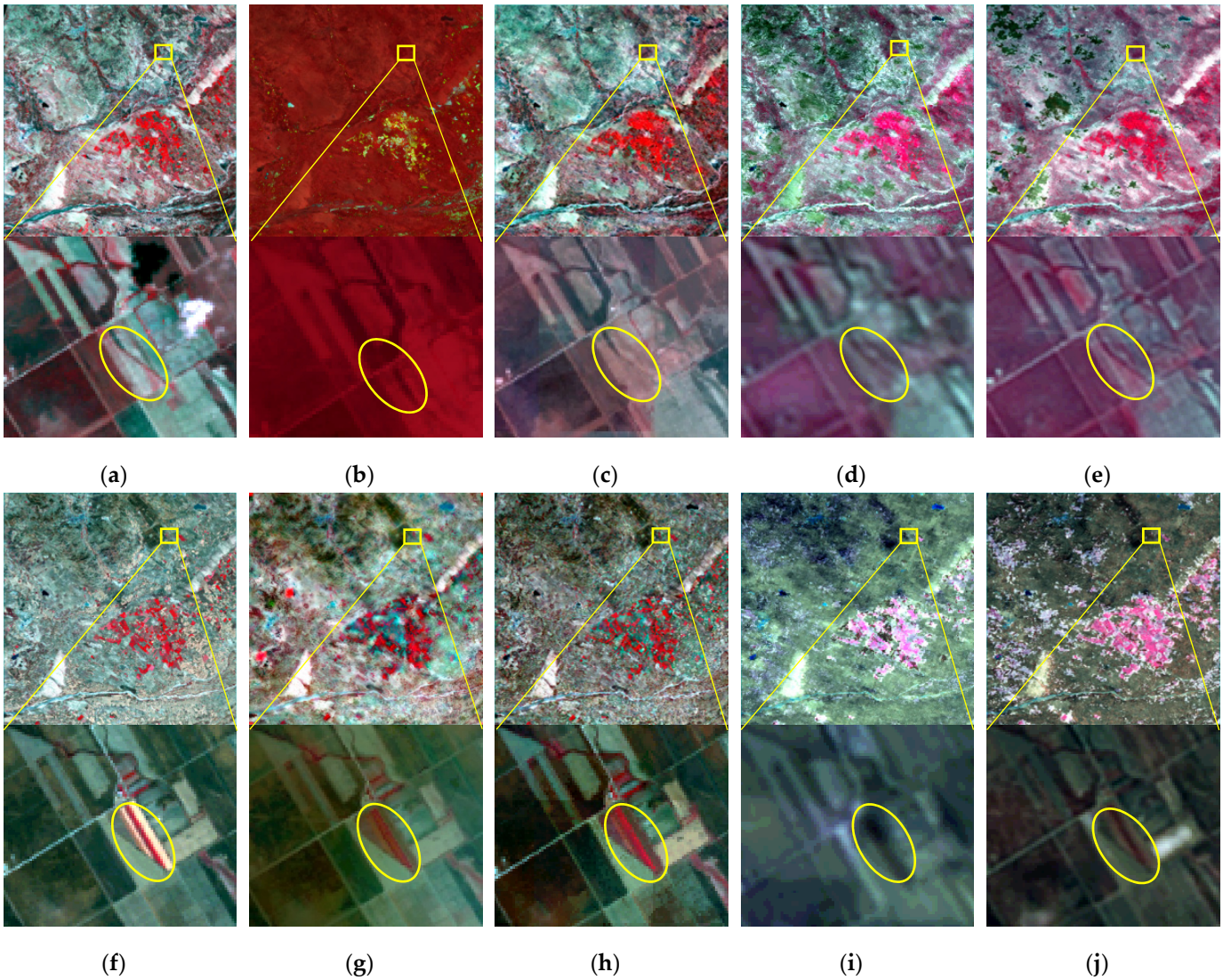

Figure 13. Experimental results based on the AHB dataset for spatiotemporal fusion: (**a**,**f**) Ground truth image; (**b**,**g**) STARFM method results; (**c**,**h**) FSDAF method results; (**d**,**i**) EDCSTFN method results; and (**e**,**j**) STF-EGFA method results.

Table 2. Objective evaluation results for the AHB test set.

Method	SAM ↓ *	PSNR ↑	CC ↑	SSIM ↑	RMSE ↓
STARFM	0.339	22.220	0.351	0.687	22.946
FSDAF	0.324	22.412	0.524	0.681	20.444
EDCSTFN	0.101	28.147	0.433	0.840	11.368
STF-EGFA	**0.092**	**28.751**	**0.485**	**0.869**	**9.633**

* ↑ indicates that larger is better, ↓ indicates that smaller is better.

4.3.2. Tianjin Dataset

We also test the proposed method on the urban phenological change dataset, as shown in Figure 14, the result of the Tianjin dataset [12]. The yellow boxed area is the area of detail amplification. In the yellow ellipse of the enlarged region, STARFM yields the worst prediction effect and displays some variability for different images. FSDAF and STF-EGFA produce more accurate texture details and spectral information than STARFM and retain the edge information from the image. In addition, the EDCSTFN method also has a good effect in maintaining texture information, and the simulation results are quite different from

the original image. Therefore, the STF-EGFA algorithm proposed in this paper achieves good results while maintaining the image edge and spectral information.

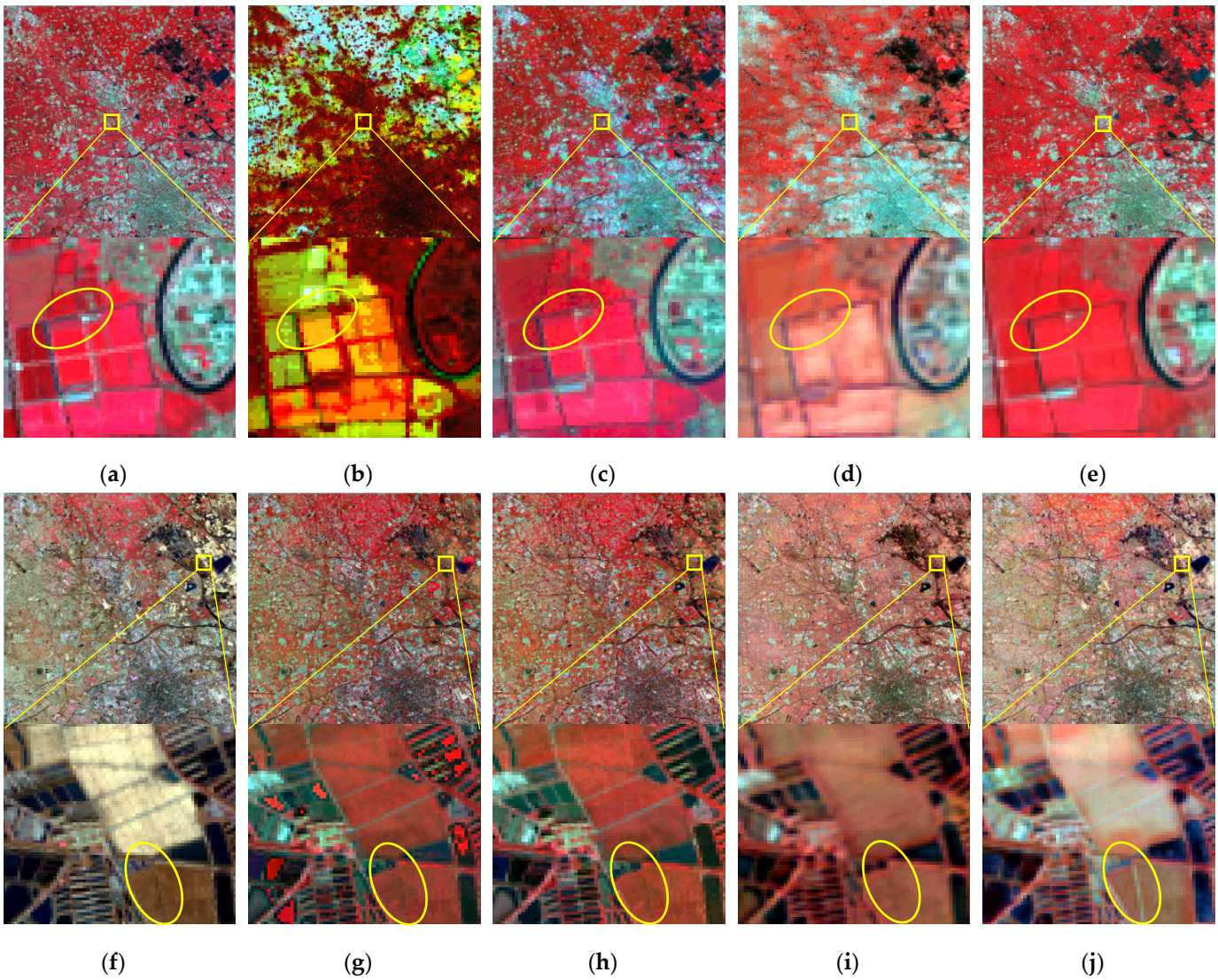

Figure 14. Experimental results based on the Tianjin dataset for spatiotemporal fusion: (**a**,**f**) Ground truth image; (**b**,**g**) STARFM method results; (**c**,**h**) FSDAF method results; (**d**,**i**) EDCSTFN method results; and (**e**,**j**) STF-EGFA method results.

On the basis of the Tianjin test set, five evaluation indicators—SAM, PSNR, CC, SSIM and RMSE—are calculated and evaluated. Table 3 shows the objective evaluation results of different methods on the Tianjin dataset. It is worth noting that our proposed STF-EGFA method outperforms the traditional STARFM and FSDAF methods on every metric. Compared with EDCSTFN, the proposed EIFA-STF approach is superior, with a 19.09% improvement in SAM, a 6.21% improvement in PSNR, a 15.11% improvement in CC, a 9.33% improvement in SSIM, and an 18.27% improvement in RMSE. Compared with other algorithms, the use of FA and edge extraction technology can better preserve more edge and detail information and make more accurate predictions for spatiotemporal fusion images. In summary, the method in this paper also demonstrates reliable applicability on the Tianjin dataset. The best values of objective evaluation indexes are shown in bold.

Table 3. Objective evaluation of the Tianjin test set.

Method	SAM ↓	PSNR ↑	CC ↑	SSIM ↑	RMSE ↓
STARFM	0.375	17.462	0.268	0.589	53.019
FSDAF	0.240	20.261	0.438	0.632	32.500
EDCSTFN	0.110	28.554	0.761	0.772	9.536
STF-EGFA	**0.089**	**30.327**	**0.876**	**0.844**	**7.794**

4.3.3. Daxing Dataset

We also test the applicability of the network using a land use change dataset. Figure 15 below shows the experimental results of various methods for the Daxing dataset [12]; the yellow boxed area is the area of detail amplification. In the yellow ellipse of the enlarged region, the images obtained by STARFM and FSDAF are similar in color, and the image details obtained by the FSDAF algorithm are better than those obtained by the STARFM method. In addition, the STARFM results display partial pixel loss. Comparatively, the EDCSTFN and STF-EGFA methods can better retain edge information. The proposed STF-EGFA method achieves the best results in preserving edge information, especially in the yellow elliptical regions, which indicates that STF-EGFA has advantages in edge preservation; however, improvements in color prediction are needed. The proposed STF-EGFA method can better preserve the edge information in the image, especially in the yellow elliptical regions, which indicates that STF-EGFA has advantages in edge preservation; however, improvements in color prediction are needed.

On the basis of the Daxing test set, five evaluation indicators—SAM, PSNR, CC, SSIM and RMSE—are calculated and evaluated. The objective evaluation of the Daxing dataset is shown in Table 4. Obviously, EDCSTFN and EIFA-STF perform better than traditional STARFM and FSDAF. The EIFA-STF method performs better than the EDCSTFN deep learning algorithm, and the optimal values of all 5 evaluation indexes are obtained. Compared with EDCSTFN, the proposed EIFA-STF approach is superior, with a 10.96% improvement in SAM, a 2.87% improvement in PSNR, a 4.88% improvement in CC, a 4.12% improvement in SSIM, and a 9.92% improvement in RMSE. The subjective and objective evaluations show that the STF-EGFA method can better preserve salient information such as edges and can better predict the effect of the image. In summary, the method in this paper also demonstrates reliable applicability for the Daxing dataset. The best values of objective evaluation indexes are shown in bold.

Table 4. Objective evaluation of the Daxing test set.

Method	SAM ↓	PSNR ↑	CC ↑	SSIM ↑	RMSE ↓
STARFM	0.090	27.942	0.731	0.805	10.221
FSDAF	0.088	28.941	0.776	0.811	9.144
EDCSTFN	0.073	30.766	0.841	0.826	7.426
STF-EGFA	**0.065**	**31.650**	**0.882**	**0.860**	**6.689**

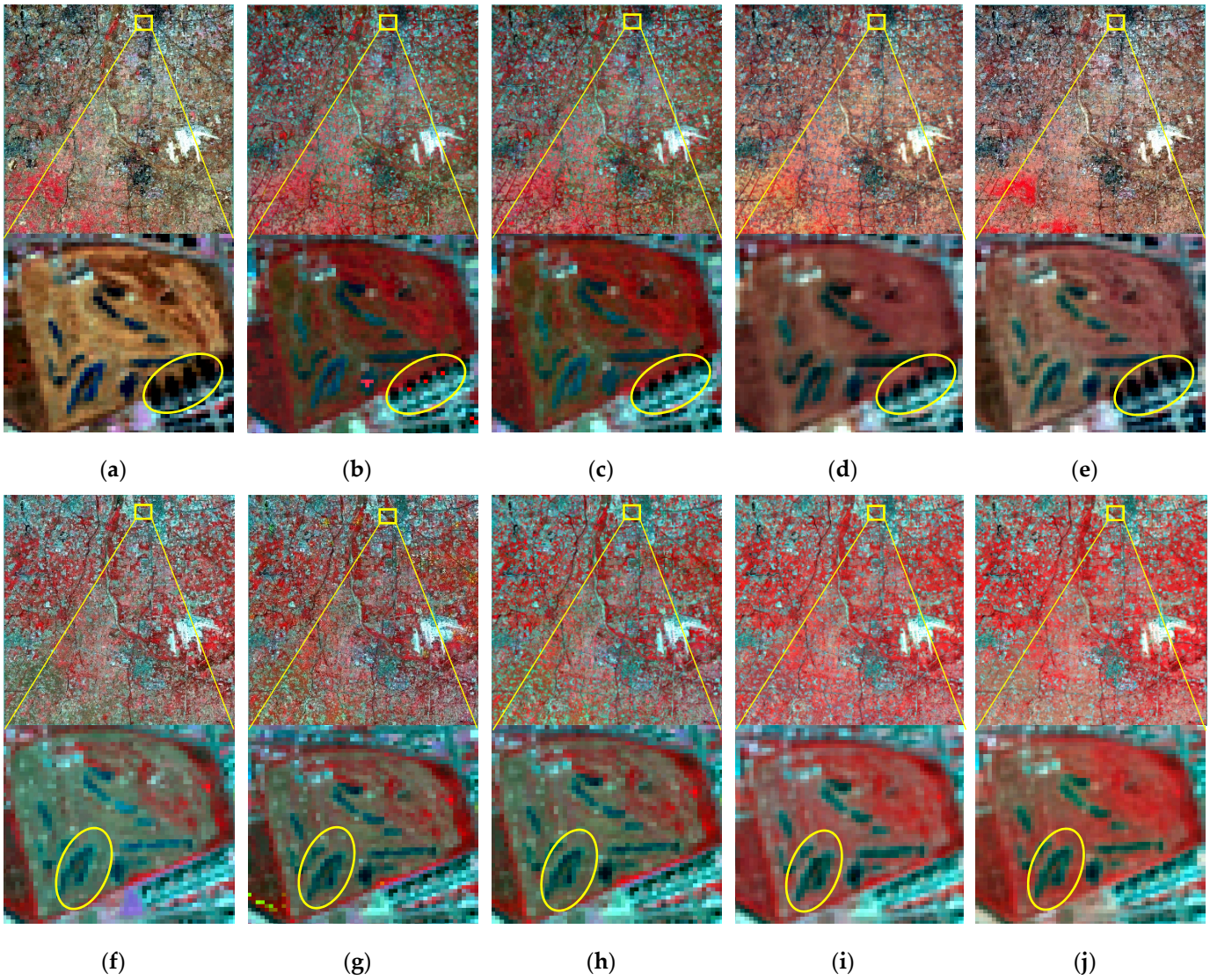

Figure 15. Experimental results based on the Daxing dataset for spatiotemporal fusion: (**a**,**f**) Ground truth image; (**b**,**g**) STARFM method results; (**c**,**h**) FSDAF method results; (**d**,**i**) EDCSTFN method results; and (**e**,**j**) STF-EGFA method results.

5. Discussion

5.1. Ablation Experiment

To better verify the added module effectiveness of the STF-EGFA, we performed ablation experiments to discuss and verify the added modules. We divided the test into three types: (a) only an edge feature extraction module is used in the model, called Only edge; (b) while adding the edge feature module, add FA (after the FEncoder) at the same time, called Edge-FA-encoder; and (c) while adding the edge feature module, add FA (before the decoder) at the same time, called Edge-FA-decoder. The three models were tested and validated with the Daxing dataset. The schematic diagrams of the three model structures are shown in Figure 16.

The three kinds of network training results are shown in Figure 17, which illustrates that adding modules results in a substantial improvement in the effectiveness of spatiotemporal fusion. First, edge information is better retained in the network when the edge module is added, thus preserving the edge features of the predicted images. Additionally, the model with an edge extraction module and an FA module (after the FEncoder) better optimizes the predicted image channels and the pixel features in the network. In addition,

while adding the edge feature module, adding FA (before the decoder) effectively retains both edge information and color information to obtain a better fusion effect, displaying certain stability. Through the ablation experiments of the above three modules, it can be intuitively seen from the figure that the STF-EGFA method can better preserve the edge structure of the image and extract detailed features.

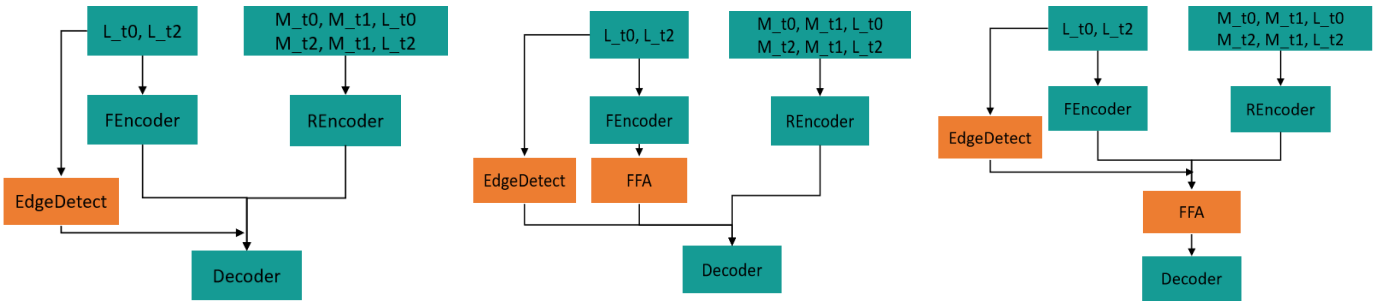

Figure 16. Network models used in the ablation experiment: (**a**) Only edge; (**b**) Edge-FA-encoder; and (**c**) Edge-FA-decoder.

In addition, we performed an objective evaluation of two groups of images from the Daxing dataset in three structural experiments. The experimental results obtained with the EIFA-STF network under different conditions are shown in Table 5; five evaluation metrics, namely SAM, PSNR, CC, SSIM, and RMSE, were used to assess the results. Notably, the addition of the edge extraction module improved all five metrics compared to those for the original network, indicating that the addition of the edge extraction module in the network positively influenced the prediction results of spatiotemporal fusion. In another set of experiments, we used edge extraction while adding FA (after the FEncoder). Adding the second module yielded a certain improvement in the experimental effect compared to that when only the edge feature extraction module was considered. On this basis, we evaluated whether the position of the feature module influences the fused image result and placed the FA module before the decoder. The experimental results showed that the network achieved an improved fusion effect, and four evaluation indexes reached optimal values among those observed. Additionally, the SAM index was improved by 10.96% compared with that for the EDCSTFN method. In summary, the proposed STF-EGFA network not only provides the best spatiotemporally fused images but also best retains spectral information and edge features. It is obvious that the simultaneous use of the edge feature extraction module and the FA module can not only improve the fusion effect of the network but also increase the stability of network fusion. We have carefully checked the funding information and the funding information provided is accurate. The best values of objective evaluation indexes are shown in bold.

Table 5. Objective evaluation of the Daxing dataset in ablation experiments.

Method	SAM ↓	PSNR ↑	CC ↑	SSIM ↑	RMSE ↓
EDCSTFN	0.073	30.766	0.841	0.826	7.426
Only edge	**0.062**	31.561	0.877	0.858	6.752
Edge-FA-encoder	0.063	31.585	**0.882**	**0.860**	6.748
Edge-FA-decoder	0.065	**31.650**	**0.882**	**0.860**	**6.689**

5.2. Discussion

The experiments on the above three datasets indicate that our method achieves good prediction results on the AHB dataset, the Tianjin dataset and the Daxing dataset of rural and urban phenology changes and obtains edge features similar to real ground images. In this experiment, we use the newly proposed datasets to break through the limitations of

the Coleambally Irrigation Area (CIA) and Lower Gwydir Catchment (LGC) in this field and provide reference values for further pertinent research.

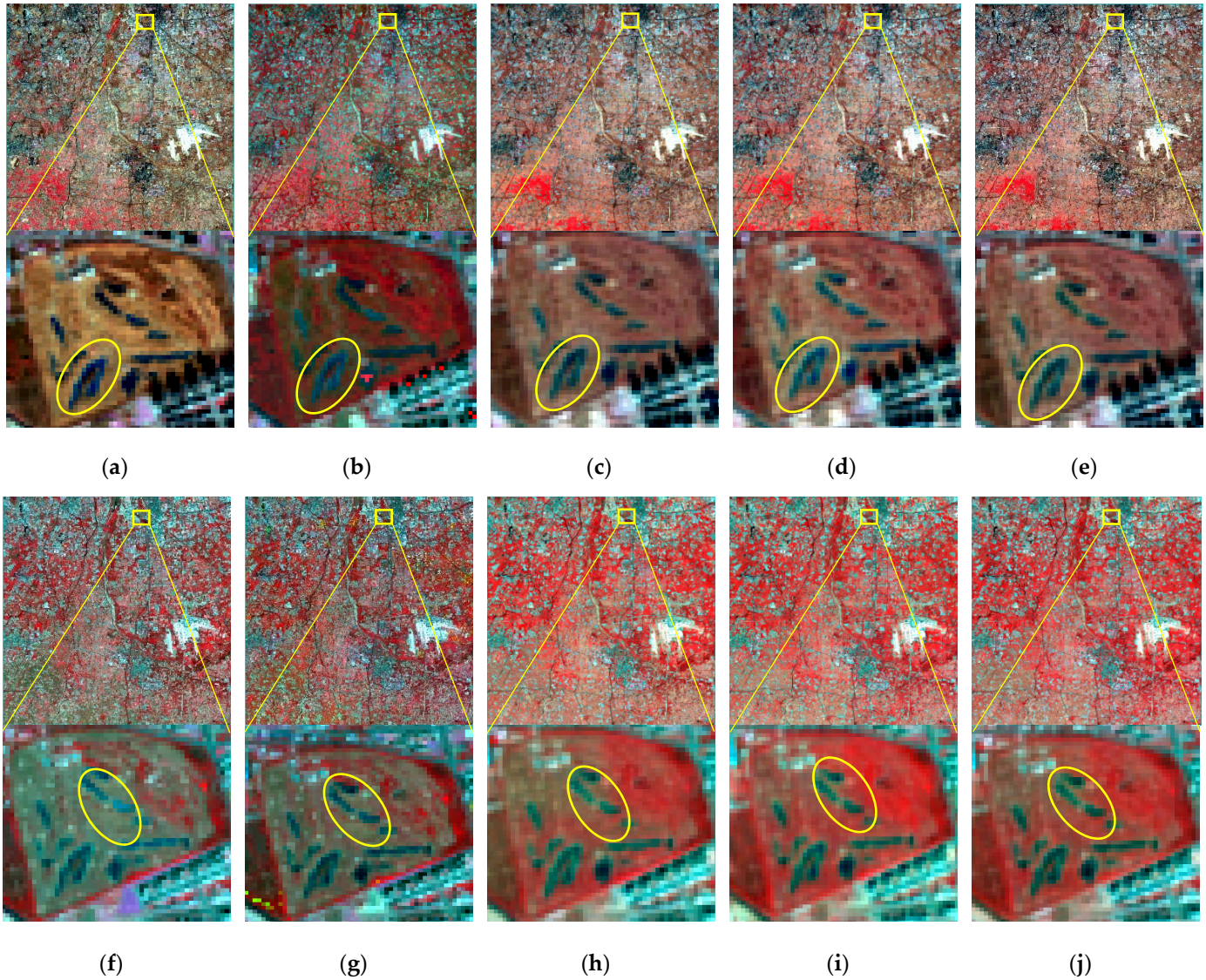

Figure 17. Experimental results based on the Daxing dataset for spatiotemporal fusion with different module structures: (**a**,**f**) Ground truth image; (**b**,**g**) EDCSTFN method results; (**c**,**h**) Only edge method results; (**d**,**i**) Edge-FA-encoder method results; and (**e**,**j**) Edge-FA-decoder method results.

Second, we ablate the proposed method through experimental verification and comprehensive analysis of the method of different modules and modules for the location of the influence of the method. First, the influence of the edge module on the experimental results is analyzed with the Daxing dataset. The experimental analysis shows that the increase in the edge module greatly improves the image fusion effect and edge features, similar to the ground real image edge features. Moreover, for the comprehensive analysis of the FA module, especially for the FA module, we show different experimental results obtained in different positions with the addition of the AM and with the addition of the experimental analysis module in different positions in the network. Therefore, we can study the influence of the modules on the experimental results by considering the different positions of the modules to propose more efficient and suitable methods.

However, our model has some shortcomings that can be improved. Compared to the EDCSTFN method, which has 281,764 parameters, our proposed algorithm has 342,101 parameters, which is an increase of 60,337 parameters. In future research, we

will consider using a lighter network structure and additional modules to improve the prediction accuracy of spatiotemporal fusion.

6. Conclusions

In this paper, we propose a spatiotemporal fusion method with edge-guided feature attention based on remote sensing, called STF-EGFA, which is designed to enhance the expression of salient features and improve the quality of predicted images. The added edge-guided extraction module enhances the retention of predicted image edge details. The combined feature fusion attention module with channels and pixels achieves the adaptive adjustment of features to highlight salient features in images, thus solving the information weighting and pixel heterogeneity problems among different channels in the network. Using a combination of subjective and objective evaluations, the proposed model is shown to achieve good performance, as verified for three datasets with different types of variations and large discrepancies, indicating the robustness of the established method. The proposed STF-EGFA model can capture edge, spectral and channel information more effectively than other methods and is useful for remote sensing spatiotemporal image fusion tasks.

Author Contributions: Design of the method and experiment, F.C.; writing—original draft preparation, F.C., K.H. and Z.F.; data preprocessing, F.C., K.H. and X.J.; writing—review and editing, Z.F., L.H. and B.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under grant Nos. 41961053 and 41871244. It was also supported Yunnan Fundamental Research Project under grant Nos. 202101AT070102 and 202201AT070164.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

STF-EGFA	Spatiotemporal fusion network with edge-guided feature attention
CNN	Convolutional neural network
MODIS	Moderate Resolution Imaging Spectrometer
STARFM	Spatial and temporal adaptive reflectance fusion model
STRUM	Spatiotemporal restraint unmixing
STAARCH	Spatiotemporal adaptive algorithm for mapping reflectance change
FIT-FC	Fitting, and residual compensation
LR	Linear regression
FSDAF	Flexible spatiotemporal data fusion
GPU	Graphics Processing Units
CSSF	Compression sensing for spatiotemporal fusion
SPSTFM	Spatiotemporal reflectance fusion via sparse representation
HSTAFM	Hierarchical spatiotemporal adaptive fusion model
BiaSTF	Spatiotemporal fusion model driven by sensor bias
ASPP	Atrous spatial pyramid pooling
DCSTFN	Deep convolutional spatiotemporal fusion network
FA	Feature attention
PA	Pixel attention
CA	Channel attention
SAM	Spectral angle mapper
PSNR	Peak signal-to-noise ratio
CC	Correlation coefficient
SSIM	Structural similarity
RMSE	Root mean square error

References

1. Ghamisi, P.; Rasti, B.; Yokoya, N.; Wang, Q.; Hofle, B.; Bruzzone, L.; Bovolo, F.; Chi, M.; Anders, K.; Gloaguen, R.; et al. Multisource and Multitemporal Data Fusion in Remote Sensing: A Comprehensive Review of the State of the Art. *IEEE Geosci. Remote Sens. Mag.* **2019**, *7*, 6–39. [CrossRef]
2. Liu, S.; Marinelli, D.; Bruzzone, L.; Bovolo, F. A Review of Change Detection in Multitemporal Hyperspectral Images: Current Techniques, Applications, and Challenges. *IEEE Geosci. Remote Sens. Mag.* **2019**, *7*, 140–158. [CrossRef]
3. Chen, B.; Huang, B.; Xu, B. Comparison of Spatiotemporal Fusion Models: A Review. *Remote Sens.* **2015**, *7*, 1798–1835. [CrossRef]
4. Zhu, X.; Cai, F.; Tian, J.; Williams, T.K.-A. Spatiotemporal Fusion of Multisource Remote Sensing Data: Literature Survey, Taxonomy, Principles, Applications, and Future Directions. *Remote Sens.* **2018**, *10*, 527. [CrossRef]
5. Javan, F.D.; Samadzadegan, F.; Mehravar, S.; Toosi, A.; Khatami, R.; Stein, A. A review of image fusion techniques for pan-sharpening of high-resolution satellite imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *171*, 101–117. [CrossRef]
6. Peng, Y.; Li, W.; Luo, X.; Du, J.; Gan, Y.; Gao, X. Integrated fusion framework based on semicoupled sparse tensor factorization for spatio-temporal–spectral fusion of remote sensing images. *Inf. Fusion* **2021**, *65*, 21–36. [CrossRef]
7. Chiesi, M.; Battista, P.; Fibbi, L.; Gardin, L.; Pieri, M.; Rapi, B.; Romani, M.; Sabatini, F.; Maselli, F. Spatio-temporal fusion of NDVI data for simulating soil water content in heterogeneous Mediterranean areas. *Eur. J. Remote Sens.* **2019**, *52*, 88–95. [CrossRef]
8. Wang, T.; Tang, R.; Li, Z.-L.; Tang, B.; Wu, H.; Jiang, Y.; Liu, M. A Comparison of Two Spatio-Temporal Data Fusion Schemes to Increase the Spatial Resolution of Mapping Actual Evapotranspiration. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 7023–7026.
9. Yang, X.; Lo, C.P. Using a time series of satellite imagery to detect land use and land cover changes in the Atlanta, Georgia metropolitan area. *Int. J. Remote Sens.* **2002**, *23*, 1775–1798. [CrossRef]
10. He, S.; Shao, H.; Xian, X.; Zhang, S.; Zhong, J.; Qi, J. Extraction of Abandoned Land in Hilly Areas Based on the Spatio-Temporal Fusion of Multi-Source Remote Sensing Images. *Photogramm. Eng. Remote Sens.* **2021**, *13*, 3956. [CrossRef]
11. Huang, B.; Zhao, Y. Research Status and Prospect of Spatiotemporal Fusion of Multi-source Satellite Remote Sensing Imagery. *Acta Geod. Cartogr. Sin.* **2017**, *46*, 1492–1499.
12. Li, J.; Li, Y.; He, L.; Chen, J.; Plaza, A. Spatio-temporal fusion for remote sensing data: An overview and new benchmark. *China Inf.* **2020**, *63*, 140301. [CrossRef]
13. Yang, G.Q.; Liu, H.; Zhong, X.W.; Chen, L.; Qian, Y.R. Temporal and spatial fusion of remote sensing images: A comprehensive revie. *Comput. Eng. Appl.* **2022**, *58*, 27–40.
14. Gao, F.; Masek, J.; Schwaller, M.; Hall, F. On the Blending of the Landsat and MODIS Surface Reflectance: Predicting Daily Landsat Surface Reflectance. *IEEE Trans. Geosci. Remote Sens.* **2006**, *44*, 2207–2218.
15. Zhu, X.; Chen, J.; Gao, F.; Chen, X.; Masek, J.G. An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions. *Remote Sens. Environ.* **2010**, *114*, 2610–2623. [CrossRef]
16. Hilker, T.; Wulder, M.A.; Coops, N.C.; Linke, J.; McDermid, G.; Masek, J.G.; Gao, F.; White, J.C. A new data fusion model for high spatial- and temporal-resolution mapping of forest disturbance based on Landsat and MODIS. *Remote Sens. Environ.* **2009**, *113*, 1613–1627. [CrossRef]
17. Gevaert, C.M.; Garcia-Haro, F.J. A comparison of STARFM and an unmixing-based algorithm for Landsat and MODIS data fusion. *Remote Sens. Environ.* **2015**, *156*, 34–44. [CrossRef]
18. Wang, Q.; Atkinson, P.M. Spatio-temporal fusion for daily Sentinel-2 images. *Remote Sens. Environ.* **2018**, *204*, 31–42. [CrossRef]
19. Zhu, X.; Helmer, E.H.; Gao, F.; Liu, D.; Chen, J.; Lefsky, M.A. A flexible spatiotemporal method for fusing satellite images with different resolutions. *Remote Sens. Environ. Interdiscip. J.* **2016**, *172*, 165–177. [CrossRef]
20. Shi, C.; Wang, X.; Zhang, M.; Liang, X.; Niu, L.; Han, H.; Zhu, X. A Comprehensive and Automated Fusion Method: The Enhanced Flexible Spatiotemporal DAta Fusion Model for Monitoring Dynamic Changes of Land Surface. *Appl. Sci.* **2019**, *9*, 3693. [CrossRef]
21. Liu, M.; Yang, W.; Zhu, X.; Chen, J.; Chen, X.; Yang, L.; Helmer, E.H. An Improved Flexible Spatiotemporal DAta Fusion (IFSDAF) method for producing high spatiotemporal resolution normalized difference vegetation index time series. *Remote Sens. Environ.* **2019**, *227*, 74–89. [CrossRef]
22. Bernabe, S.; Martin, G.; Nascimento, J.M.P.; Bioucas-Dias, J.M.; Plaza, A.; Silva, V. Parallel Hyperspectral Coded Aperture for Compressive Sensing on GPUs. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 932–944. [CrossRef]
23. Gao, H.; Zhu, X.; Guan, Q.; Yang, X.; Yao, Y.; Zeng, W.; Peng, X. cuFSDAF: An Enhanced Flexible Spatiotemporal Data Fusion Algorithm Parallelized Using Graphics Processing Units. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–16. [CrossRef]
24. Huang, B.; Song, H. Spatiotemporal Reflectance Fusion via Sparse Representation. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 3707–3716. [CrossRef]
25. Li, L.; Liu, P.; Wu, J.; Wang, L.; He, G. Spatiotemporal Remote-Sensing Image Fusion with Patch-Group Compressed Sensing. *IEEE Access* **2020**, *8*, 209199–209211. [CrossRef]
26. Chen, B.; Huang, B.; Xu, B. A hierarchical spatiotemporal adaptive fusion model using one image pair. *Int. J. Digit. Earth* **2016**, *10*, 639–655. [CrossRef]
27. Li, D.; Li, Y.; Yang, W.; Ge, Y.; Han, Q.; Ma, L.; Chen, Y.; Li, X. An Enhanced Single-Pair Learning-Based Reflectance Fusion Algorithm with Spatiotemporally Extended Training Samples. *Remote Sens.* **2018**, *10*, 1207. [CrossRef]

28. Lei, D.; Ran, G.; Zhang, L.; Li, W. A spatiotemporal fusion method based on multiscale feature extraction and spatial channel attention mechanism. *Remote Sens.* **2022**, *14*, 461. [CrossRef]
29. Li, W.; Zhang, X.; Peng, Y.; Dong, M. DMNet: A network architecture using dilated convolution and multiscale mechanisms for spatiotemporal fusion of remote sensing images. *IEEE Sens. J.* **2020**, *20*, 12190–12202. [CrossRef]
30. Song, H.; Liu, Q.; Wang, G.; Hang, R.; Huang, B. Spatiotemporal Satellite Image Fusion Using Deep Convolutional Neural Networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 821–829. [CrossRef]
31. Li, Y.; Liu, C.; Yan, L.; Li, J.; Plaza, A.; Li, B. A New Spatio-Temporal Fusion Method for Remotely Sensed Data Based on Convolutional Neural Networks. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–5 August 2019; p. 835.
32. Li, Y.; Li, J.; He, L.; Chen, J.; Plaza, A. A new sensor bias-driven spatio-temporal fusion model based on convolutional neural networks. *China Inf.* **2020**, *63*, 140302. [CrossRef]
33. Liu, X.; Deng, C.; Chanussot, J.; Hong, D.; Zhao, B. StfNet: A Two-Stream Convolutional Neural Network for Spatiotemporal Image Fusion. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 6552–6564. [CrossRef]
34. Chen, Y.; Shi, K.; Ge, Y.; Zhou, Y. Spatiotemporal Remote Sensing Image Fusion Using Multiscale Two-Stream Convolutional Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 4402112. [CrossRef]
35. Jia, D.; Song, C.; Cheng, C.; Shen, S.; Ning, L.; Hui, C. A Novel Deep Learning-Based Spatiotemporal Fusion Method for Combining Satellite Images with Different Resolutions Using a Two-Stream Convolutional Neural Network. *Remote Sens.* **2020**, *12*, 698. [CrossRef]
36. Tan, Z.; Yue, P.; Di, L.; Tang, J. Deriving High Spatiotemporal Remote Sensing Images Using Deep Convolutional Network. *Remote Sens.* **2018**, *10*, 1066. [CrossRef]
37. Tan, Z.; Di, L.; Zhang, M.; Guo, L.; Gao, M. An Enhanced Deep Convolutional Model for Spatiotemporal Image Fusion. *Remote Sens.* **2019**, *11*, 2898. [CrossRef]
38. Liu, J.; Fan, X.; Jiang, J.; Liu, R.; Luo, Z. Learning a Deep Multi-scale Feature Ensemble and an Edge-attention Guidance for Image Fusion. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 105–119. [CrossRef]
39. Qin, X.; Wang, Z.; Bai, Y.; Xie, X.; Jia, H. FFA-Net: Feature Fusion Attention Network for Single Image Dehazing. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 11908–11915. [CrossRef]
40. Lim, B.; Son, S.; Kim, H.; Nah, S.; Lee, K.M. Enhanced Deep Residual Networks for Single Image Super-Resolution. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 1132–1140.
41. Yuhas, R.H.; Goetz, A.F.H.; Boardman, J.W. Discrimination among semi-arid landscape endmembers using the Spectral Angle Mapper (SAM) algorithm. In Proceedings of the Twenty-Fourth Lunar and Planetary Science Conference, Pasadena, CA, USA, 1–5 June 1992; pp. 147–149.
42. Huynh-Thu, Q.; Ghanbari, M. Scope of validity of PSNR in image/video quality assessment. *Electron. Lett.* **2008**, *44*, 800–801. [CrossRef]
43. Alparone, L.; Wald, L.; Chanussot, J.; Thomas, C.; Gamba, P.; Bruce, L.M. Comparison of Pansharpening Algorithms: Outcome of the 2006 GRS-S Data Fusion Contest. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 3012–3021. [CrossRef]
44. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef]
45. Li, W.; Cao, D.; Peng, Y.; Yang, C. MSNet: A Multi-Stream Fusion Network for Remote Sensing Spatiotemporal Fusion Based on Transformer and Convolution. *Remote Sens.* **2021**, *13*, 3724. [CrossRef]