electronics

MDPI

Article

Super-Resolution Reconstruction Model of Spatiotemporal Fusion Remote Sensing Image Based on Double Branch Texture Transformers and Feedback Mechanism

Hui Liu [1], Yurong Qian [2,*], Guangqi Yang [2] and Hao Jiang [2]

[1] College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China; liuhui@stu.xju.edu.cn
[2] School of Software, Xinjiang University, Urumqi 830046, China; ydlam@stu.xju.edu.cn (G.Y.); 107552104334@stu.xju.edu.cn (H.J.)
* Correspondence: qyr@xju.edu.cn; Tel.: +86-189-3594-5059

Abstract: High spatial-temporal resolution plays a vital role in the application of geoscience dynamic observance and prediction. However, thanks to the constraints of technology and budget, it is troublesome for one satellite detector to get high spatial-temporal resolution remote sensing images. Individuals have developed spatiotemporal image fusion technology to resolve this downside, and deep remote sensing images with spatiotemporal resolution have become a possible and efficient answer. Due to the fixed size of the receptive field of convolutional neural networks, the features extracted by convolution operations cannot capture long-range features, so the correlation of global features cannot be modeled in the deep learning process. We propose a spatiotemporal fusion model of remote sensing images to solve these problems based on a dual branch feedback mechanism and texture transformer. The model separates the network from the coarse-fine images with similar structures through the idea of double branches and reduces the dependence of images on time series. It principally merges the benefits of transformer and convolution network and employs feedback mechanism and texture transformer to extract additional spatial and temporal distinction features. The primary function of the transformer module is to learn global temporal correlations and fuse temporal features with spatial features. To completely extract additional elaborated features in several stages, we have a tendency to design a feedback mechanism module. This module chiefly refines the low-level representation through high-level info and obtains additional elaborated features when considering the temporal and spacial characteristics. We have a tendency to receive good results by comparison with four typical spatiotemporal fusion algorithms, proving our model's superiority and robustness.

Keywords: remote sensing images; spatiotemporal image fusion; feedback mechanism; texture tran-sformer; detailed features

Citation: Liu, H.; Qian, Y.; Yang, G.; Jiang, H. Super-Resolution Reconstruction Model of Spatiotemporal Fusion Remote Sensing Image Based on Double Branch Texture Transformers and Feedback Mechanism. *Electronics* **2022**, *11*, 2497. https://doi.org/10.3390/electronics11162497

Academic Editors: Giovanni Ramponi, Raffaela Cefalo and Žiga Kokalj

Received: 27 June 2022
Accepted: 1 August 2022
Published: 10 August 2022

1. Introduction

The advancement and popularization of sensor technology have promoted the wide application of remote sensing image-related research in human activities. For example, the high-quality spatiotemporal remote sensing images we have obtained through satellite remote sensing technology have great research significance in areas [1,2] such as crop monitoring [3], forest monitoring [4], land-cover change monitoring [5], real-time urban disaster monitoring [6], and water resource evaluation [7]. These applications require the spatial resolution of surface details (texture and structure of ground objects) and an intensive time series of remote sensing image data to capture changes in the ground for accurate classification and identification. However, in practical applications, there are still some unavoidable technical and budget restrictions, resulting in trade-offs of time, space, and spectral resolution of remote sensing images of earth observation data,

leading to difficulties in obtaining remote sensing images with high-temporal and high-spatial resolutions [8,9]. In most cases, we select at least two different data sources to generate high-quality fused images:the main features of the Landsat 8 data source are that most spectral bands have a spatial resolution of 30 m and a temporal resolution of 16 days, long-term repeated measurements and the resulting data are coarse images with high temporal-low spatial (HTLS) resolution [10]. The other is the MODerate Resolution Imaging Spectroradiometer (MODISI) image, primarily characterized by covering a large area of our planet, with a spatial resolution of 250 to 1000 m and different wavelengths, and requires being acquired every day. The obtained data are fine images of low spatial and high temporal (LTHS) images [11]. Due to insufficient information from a single data source, researchers have proposed a combination of remote image sensing spatiotemporal fusion algorithms, which merge high spatiotemporal resolution images from multiple data sources to obtain high spatiotemporal information fusion images. We use Landsat 8 and MODIS data sources to synthesize images simultaneously with high spatial-high temporal resolution. In this way, it is proven that multiple information sources will obtain higher and richer characteristic data than only one, and many outstanding research results have been achieved, providing theoretical support for later application research.

Since Tsai proposed [12] image super-resolution reconstruction in 1984, many scholars have researched and discussed the topic. Super-resolution reconstruction technology is a method for obtaining high-resolution images from processing low-pixel images. Due to the advantages of low cost, short consumption period, ample room for improvement, comprehensive coverage, large amount of information, and good durability, this technology has been widely used [13]. High-resolution remote sensing images have been used in various fields, such as environmental monitoring, urban planning, and emergency rescue [14,15], but how to get high-resolution images at a low cost and in a short time has always been a problem that needs to be solved in the field of remote sensing [16]. In this study, high-quality images were obtained by combining the spatiotemporal fusion rules of remote sensing images with super-resolution reconstruction, which laid a good foundation for future applications.

After years of development, researchers have proposed two types of spatial-temporal fusion models for remote sensing images: traditional models and models based on deep learning, as described in Section 2. Although some of these models have achieved exemplary application results, there are still significant theoretical differences between the surface and the collected data. For example, how to select HTLS and low-temporal high-spatial (LTHS) images and reference images is very meaningful because all high-frequency information in the whole modeling process comes from these selected data. If there is a significant difference between the reference image and the predicted image, the final fusion result will not reach an acceptable result. Second, the actual data are often contaminated by penumbra and noise, which is theoretically inconsistent with the processed usable datasets. In order to solve these problems, we must further enhance the quality of the prediction image.

This research proposes a feedback and texture transformer-based spatiotemporal fusion model for remote sensing images, which is obtained by redesigning the network based on the Enhanced Deep Convolutional Spatiotemporal Fusion Network (EDCSTFN) model. Our model contains a total of five characteristics: (1) The model needs at least two pairs of MODIS–Landsat images to get high-quality predicted images, and the predicted image information entirely relies on a continuous time series. (2) The transformer was initially applied in natural language processing [17]. With the development of research, it gradually abandoned convolution and recursion modules and is wholly based on the self-attention mechanism, which has strong parallelism. Our model employs a dual-branch feedback mechanism and a texture converter and considers images' dependence on time series. Therefore, the network not only starts from the structural similarity of the coarse and fine image pairs but also uses the rich texture information in the adjacent images to predict the fine images and improve the quality of image reconstruction. (3) The same

dual-branch texture transformer is used to make the prediction image obtain more detailed information accurately. (4) Both branches of the model use a feedback mechanism to refine low-level representations of high-level information. (5) The model utilizes a composite loss function to analyze the fusion results, which include content and visual losses. The primary purpose is to preserve the high-frequency information to make the generated image clearer. In this study, three datasets, Aruhorqin Banner (AHB), Coleambally Ignition District (CIA), and Lower Gwydir Basin (LGC), were selected for comparative analysis with the classical fusion model. The results display that the model offered in this paper enhances the fusion accuracy, prediction results, and the quality of fused images.

2. Related Works

In machine learning, convolutional neural networks (CNNs) have attracted significant attention [18]. A CNN is a deep feedforward neural network trained and designed using prior knowledge and mainly extracts rich feature information from multiple test paper layers. A classic CNN is composed of five parts: one or more inputs, one or more convolutional layers, and one or more subsampling layers (or pooling layers) [19]. Previous researchers extracted more advanced features by increasing the convolutional layers while avoiding network overfitting caused by the increase in layers. CNNs have become efficient frameworks for addressing the problem of image feature extraction and recognition [20]. With the progress of research, CNNs have been gradually applied to the field of image super-resolution reconstruction and data fusion from their original use for extracting high-level features in image classification and recognition tasks [21,22].

CNN outperforms other computer vision tasks in image super-resolution [23]. Super-Resolution Convolutional Network (SRCNN) first used a three-layer CNN in image SR to learn complex LR-HR mapping. Very Deep Convolutional Networks (VDSR) [24] increase the depth of CNN to 20 layers to use more contextual information in the LR image and adopts the method for jumping connections to overcome the difficulty of optimization when the network is deep. In recent studies, different jump connections have been used to achieve super-resolution image reconstruction. Super-Resolution Generative Adversarial Networks (SRGAN) [25] and Enhanced Deep Residual Networks (EDSR) [26] use the residual jump connection [27], which improves the accuracy of the reconstruction image effect. Super-Resolution Using Dense Skip Connections Networks (SRDenseNet) [28] use the dense skip connection [29] to obtain more characteristic information [30]. A combination of local/global residuals and dense skip connections was incorporated into its RDN. Experiments have revealed that these models perform well in super-resolution image reconstruction. However, the following two problems still exist: The first problem is that due to the use of skip connections or a bottom-up combination of hierarchical features in these network architectures, they only extract low-level features, and the ability of the upper layer to receive information is limited by the small receptive field and lack of sufficient contextual information, which further limits the network's reconstruction ability. The second problem is "space-time contradiction"; that is, there is a problem that the spatial and temporal resolutions of remote sensing images are mutually restricted.

With the continuous development of deep learning models, research based on CNN has gradually been applied to the spatial-temporal fusion of remote sensing images, but it is still at an early stage. By reading a lot of the literature in this research area, we concluded that the existing spatiotemporal fusion algorithms could be divided into five categories: (1) transformation-based, (2) reconstruction-based, (3) Bayesian-based, (4) learning-based, and (5) pan-sharpening-based.

The transform-based methods mainly adopt mathematical transformation technology [31], such as the wavelet transform. Because of multi-source data integration, the original image pixel is represented in another abstract space by mapping. The data are converted from the spatial domain to the frequency domain. This method has two characteristics: the first is that it extracts clear high-frequency information from the transformed LTHS image and fuses it with the HTLS image to obtain high-quality fused images. Second,

it has spatial generality, and different types of features can be extracted from different spatial images using fusion rules [32].

Methods based on reconfiguration are divided into two categories: weight function-based and unmixing-based. The weight function method mainly evaluates HTLS images by setting the weight function and combining the image reflectivity. The classical methods include the spatiotemporal adaptive reflectivity fusion model (STARFM), the spatiotemporal adaptive algorithm for mapping reflectance change (STARCH) [33], and the enhanced STARFM (ESTARFM) [34]. The unmixing method mainly uses spectral unmixing theory and an unmixing algorithm to build the fusion model. It mainly uses HTLS images to reconstruct the corresponding LTHS images. Existing methods include the spatiotemporal reflectivity unmixing model (STRUM) [35], flexible spatiotemporal data fusion (FSDAF) [36], unmixing-based data fusion (UBDF) [37], the spatial attraction model (SAM) [38], and the spatiotemporal data fusion algorithm (STDFA) [39].

The Bayesian-based method integrates the spatiotemporal spectrum into a unified framework, allowing the input image is not limited to achieving the most realistic prediction results. Existing methods include the unified fusion method [40] and the Bayesian fusion method.

The learning-based method does not require the manual setting of the fusion rules. It mainly uses existing archived data to train the supervised deep learning model. Existing learning-based fusion methods include the sparse-representation-based spatiotemporal reflectance fusion model (SPSTFM) [41], spatiotemporal fusion using a deep convolutional pair neural network (STFDCNN) [42] deep convolutional spatiotemporal fusion network (DCSTFN) [43], enhanced DCSTFN(EDCSTFN) [44], two-stream convolutional neural network for spatiotemporal image fusion (STFNET) [45], and generative adversarial network-based spatiotemporal fusion model (GAN-STFM) [46].

Based on pan-sharpening fusion, the CNN model is applied to panchromatic and multispectral images. With the deepening of remote sensing research, many researchers have proposed various panchromatic sharpening methods; typical methods include intense hue saturation (IHS) [47–49], principal component analysis (PCA) [50,51], Brovey transform (BT) [52], Laplacian pyramid decomposition [53], wavelet transform [54], and curvilinear transformation under different resolutions [55–57].

Although the progress of sensor technology has dramatically improved the accuracy of satellite observation, the following problems remain. First, the absence of technology and budget makes it impossible to obtain high temporal and spatial resolution images directly. Second, there is always a trade-off between the temporal, spatial, and spectral resolution of the observation data we obtain, so it is challenging to continuously get LTHS and HTLS data pairs for research. Therefore, this research model combines the concepts of ConvNet and transformer and uses VGGNet and transformer as the backbone networks, which are mainly reflected in two aspects: on the one hand, members of the team are studying the super-resolution reconstruction technology based on the transformer network and published research results; on the other hand, it borrows the state-of-the-art in spatiotemporal fusion models and super-resolution reconstruction from other papers and adds texture converters and feedback mechanisms to supplement the input data, extract as much helpful information as possible, and at the same time, reduce the model parameters for the best output image quality. We believe this fusion method can provide a reference for future research and has a promising practical application prospect.

3. Methodology

Our proposed method consists of five significant steps: (1) describe the overall framework diagram of the network model in this study, (2) construct a two-branch texture transformer, (3) use a feedback mechanism to obtain more details, (4) use fusion rules for fusion, and (5) use composite loss. The function performs image analysis to generate the final image, the details of which are discussed below.

3.1. Overall Model Architecture

Inspired by the data fusion models of EDCSTFN and DCSTFN, this paper proposes a deep convolution fusion network that can generate high spatial-temporal resolution remote sensing images. This research adopts a similar "encoder merge decoder" architecture with two streams. The overall network architecture is shown in Figure 1. From the figure, it can be seen that the model's architecture has three parts: the first part is the branch in the upper left corner and the branch in the lower left corner, which is called the encoder. The branch in the upper left corner mainly extracts high-frequency information similar to the T0 time image and T1 time image; the branch in the lower left corner mainly extracts high-frequency information similar to the T2 time image and T1 time image. The second part is where the two branches intersect, which is called the fusion part. It mainly uses fusion rules to combine extracted features with the same dimension and size. The third part is the decoder, which is mainly responsible for restoring these advanced features to the original pixel space to get the final high-quality reconstructed image.

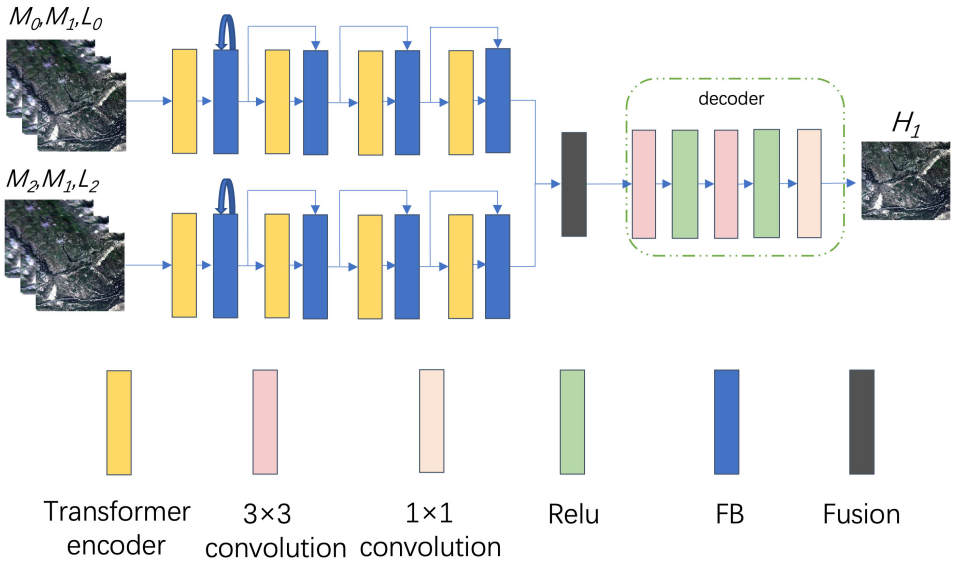

Figure 1. The backbone network is based on two branches. (M_0 represents the coarse image (MODIS Image) of the first pair of reference images, M_1 represents the coarse image of the prediction image, M_2 represents the coarse image of the second pair of reference images, L_0 represents the fine image (Landsat Image) of the first pair of reference images, L_1 represents the fine image of the prediction image, L_2 represents the fine image of the second pair of reference images, and H_1 represents the prediction image).

In this model, the two sub-network branches of the encoder have the same network structure, and its primary purpose is to explain the type of image input. Because the image depends on a continuous time series, the first upper left branch is mainly used to obtain the image at T0 before T1, while the second lower left branch is used to obtain the image at T2 after T1, and finally, a high-quality fused image is obtained through the fusion reconstruction stage. Figure 2 shows the detailed schematic diagram of the model; SFE represents the residual module, which is mainly used to learn the difference between the reference date and the predicted date.

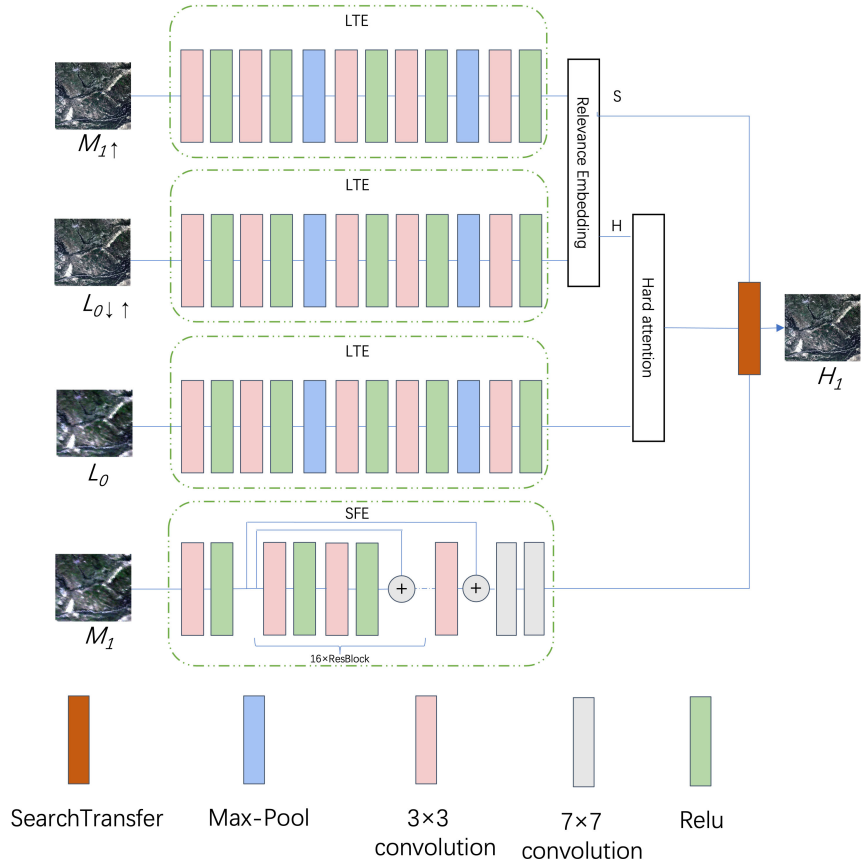

Figure 2. Diagram of backbone network details based on two branches. (S represents soft attention, H represents hard attention, M_1 represents the coarse image of the prediction image, $M_1\uparrow$ represents the coarse image of the up sampled prediction image, L_0 represents the fine image table of the prediction image, $L_0\downarrow\uparrow$ represents the fine image of the up sampled and down sampled prediction image, and H_1 represents the prediction image).

3.2. Texture Transformer

Furthermore, because more texture features can be obtained from LTHS and HTLS while reducing memory and time to achieve the best fusion effect, this study used a texture transformer, which uses high-resolution images (thin images) as reference images (L_0). In this way, the relevant texture is transferred to the LR image [58], transmitting the corresponding and accurate texture features. Texture transformers can be used overlappingly and integrate feature information across scales so that more texture information can be extracted from the reference image and applied to texture restoration at different stages. Our texture transformer model consists of four parts. First, the learnable texture extractor (LTE) is primarily used to update parameters in the end-to-end process while the joint features of LR and the reference image are embedded, ensuring that the attention mechanism has a solid foundation in the super-split reconstruction [59]. Second, the relational embedded (RE) module is mainly used to calculate the correlation between the LR and reference images. In essence, the features extracted from the LR and reference images can serve as a converter to form the pattern of a long beadle and key from which the hard attention (HA) and soft attention (SA) graphs are obtained. The third and fourth modules are the hard and soft attention graphs, which are mainly used to convert high-resolution features from the reference image and fuse them into the LR features extracted from the trunk through an attention graph, as shown in Figure 3.

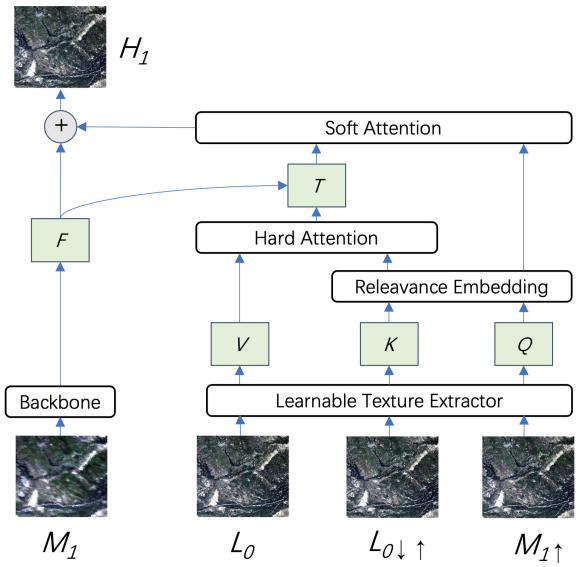

Figure 3. Detail of texture transformer. (M_1 represents the coarse image of the prediction image, $M_1 \uparrow$ represents the coarse image of the up sampled prediction image, L_0 represents the fine image table of the prediction image, $L_0 \downarrow\uparrow$ represents the fine image of the up sampled prediction image, and H1 represents the prediction image).

In the texture converter, we obtain texture features such as Q (query), K (key), and V (value) from the up sampled M_1 images, sequential down sampled/up sampled L_0 images and original L_0 images. F is the M_1 feature extracted from the backbone of DNN and further fused with the transmitted texture feature T to generate the SR output. The formula for the texture extraction process is shown in (1)–(3):

$$Q = LTE(M_1 \uparrow) \tag{1}$$

$$K = LTE(L_0 \downarrow\uparrow) \tag{2}$$

$$V = LTE(L_0) \tag{3}$$

Q, K, and V in the formula represent the three essential elements of the internal concern mechanism of the extracted texture feature transformer and will be further used to associate the embedded module. The purpose of correlation embedding is to embed the correlation between M_1 and L_0 images by estimating the similarity between Q and K. The correlation embedding formula is (4):

$$R_{i,j} = < \frac{q_i}{\| q_i \|}, \frac{k_j}{\| k_j \|} > \tag{4}$$

In the formula, the patch expansion of K is shown, and finally, the correlation between the two patches is obtained through normalization processing.

In the hard attention module (H stands for hard attention in the text), an "attention map" H is calculated. We can regard the value of h_i as a hard index, which represents the position in the L_0 image that is most relevant to the ith position in the M_1 image. HR represents the texture feature T obtained from the L_0 image. For H, we can obtain the sum of multiple h_i, where h_i is calculated by correlation and Formulas (5) and (6) are obtained:

$$h_i = \arg\max_j R_{i,j} \tag{5}$$

$$H = \sum_{i=1}^{n} h_i \tag{6}$$

The soft attention (S stands for soft attention in the text) mechanism's formula is determined using Equations (7) and (8):

$$s_i = \arg\max_j R_{i,j} \tag{7}$$

$$S = \sum_{i=1}^{n} s_i \tag{8}$$

The following formula, Formula (9), represents the output of the final image of the texture converter:

$$F_{out} = F + Conv(Concat(F, T)) * W_s + Conv(Concat(F, T)) * W_h \tag{9}$$

where F represents the M_1 feature in the backbone network, T represents the fusion feature, S represents the SA graph, H represents the HA graph, and W represents the training weight. The texture converter can effectively convert HR-related texture features in the L_0 image to M_1 features, which improves the accuracy of the texture generation process.

3.3. Feedback Mechanism

In cognitive theory, feedback connections connecting cortical visual areas can transmit response signals from higher-order areas to lower-order areas [60,61]. Feedback mechanisms work top-down, passing high-level information to the previous level and refining lower-level encoded information. To recover more detailed information from the rough image, that is to say, to recover a better SR image from the LR image, the model uses a feedback mechanism in both branches [62]. Most traditional deep learning-based networks share information in a feedforward manner. However, the feedforward approach prevents the previous layer from accessing helpful information from the lower layer, even if skipped connections are used, and this feedback has strong early reconstruction capabilities with very few parameters. The feedback mechanism can make each network output correct the previous state iteratively. The feedback mechanism in this paper consists of three parts in the iteration process: Provide LR input feature extraction blocks in each iteration—low-resolution feedback block (LRFB) (to ensure the availability of low-level information, which needs to be refined), a feedback block (FB), and a reconstruction block (RB) [62]. Figure 4 shows the structure of the feedback mechanism, in which the weights of each block are shared across time.

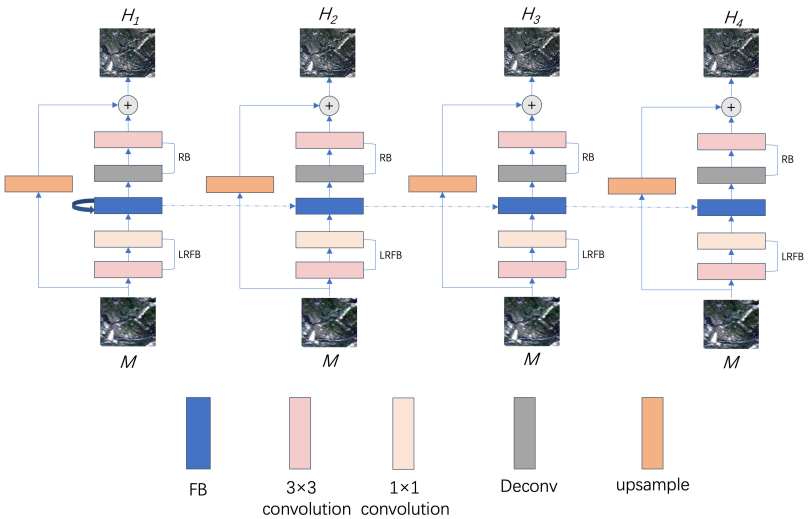

Figure 4. Detail of the feedback. (M represents the original low-resolution image, H_1, H_2, H_3, and H_4 represent the reconstructed images at different stages).

In the process of the M feature extraction block (i.e., feature extraction of a coarse image), the filter is composed of the network layer in the VGG [63] network. Then, we get the shallow feature of the input LR image as F_{in}^{t}, as shown in Formula (10):

$$F_{in}^{t} = f_{LRFB}(I_M) \tag{10}$$

Due to the principle of the feedback mechanism, this shallow feature will then be used as the input to FB. In this paper, F_{in}^{1} is regarded as the initial hidden state F_{out}^{0}. In the t-th iteration, FB receives the hidden state $F_{(t-1)}$ of the previous iteration through the feedback connection and receives the shallow feature F_{in}^{t}. The output of FB is represented by F_{out}^{0} and is shown in Formula (11).

$$F_{out}^{0} = f_{FB}(F_{out}^{t-1}, F_{in}^{1}) \tag{11}$$

In the reconstruction stage, Formulas (12) and (13) after t iterations are as follows:

$$I_{Res}^{t} = f_{RB}(F_{out}^{t}) \tag{12}$$

$$I_{H_1}^{t} = I_{Res}^{t} + f_{up}(I_M) \tag{13}$$

where f_{RB} represents the operation of the reconstruction block, I_{Res}^{t} is a residual block produced according to M, and f_{up} represents the operation of the up sampled kernel.

In the fusion stage, this model is mainly the fusion of the image characteristics of the double branch, the essence of which is in this model from the input to the final stage in the iterative convergence of fusion. That is to say, in the text, the transformer stage fuses to characteristics, the feedback stage fuses to images, and the last are characteristic of the fusion stage finally. Multiple fusions result in optimal image quality.

$L_{compound}$ is a composite loss function, including content loss (content loss is the basic requirement of the image reconstruction process. Its main function is to ensure the integrity of image content, such as texture and tone), feature loss, and visual loss. Its main purpose is to enhance the clarity of the predicted image. The formula is shown in (14):

$$L_{compound} = L_{content} + L_{feature} + L_{vision} + L_{reconstruction} \tag{14}$$

The nonlinear activations used in the network are rectified linear elements. This experiment uses the Landsat image pre-training model, and the feature loss can be expressed by Equation (15):

$$L_{feature} = \frac{1}{N}(\widehat{FL_{t1}} - FL_{t-1})^{b} \tag{15}$$

Visual loss is an auxiliary component designed to improve the overall image quality from the perspective of computer vision. The visual loss of this model is obtained by the combined action of the text transformer and feedback mechanism, which can be written into Equation (16):

$$L_{vision} = I_{H_1}^{t} + \prod_{i=1}^{N} [H_i(t_i - t_{i-1})]^{\alpha_i} [S_i(t_i - t_{i-1})]^{\beta_i} \tag{16}$$

The reconstruction loss is mainly a loss function from low-resolution images to high-resolution images, and the formula is as follows:

$$L_{reconstruction} = \frac{1}{CHW} \parallel I^{HR} - I^{H_1} \parallel_1 + I_{H_1}^{t} \tag{17}$$

4. Experiment and Evaluation

4.1. Study Areas and DataSets

We use three datasets, namely AHB, CIA, and LGC, to test the robustness of the model.

AHB is located in Inner Mongolia, China. The dataset includes 27 pairs of cloud-free Landsat and MODIS images from 30 May 2013 to 6 December 2018, which have lasted for more than 5 years. The Landsat images were obtained by Landsat-8 Operational Land Imager (OLI) sensor, and the MODIS images were obtained by MODIS Terra MOD09GA Collection 5. The AHB dataset has significant phenological changes due to the growth of crops and other vegetation.

CIA is located in the south of New South Wales, Australia. The dataset includes 17 pairs of cloud-free Landsat and MODIS images between October 2001 and May 2002. The Landsat images were obtained by Landsat-7 ETM + sensor, and the MODIS images were obtained by MODIS Terra MOD09GA Collection 5. The CIA dataset includes many changes in phenology but fewer changes in land cover types.

LGC is located in the north of New South Wales, Australia. The dataset consists of 14 pairs of cloud-free Landsat and MODIS images between April 2004 and April 2005. The Landsat images were obtained by Landsat-5 TM sensor. The MODIS images were obtained by MODIS Terra MOD09GA Collection 5. The LGC dataset can be considered to have significant changes in land cover types, and its shape will change regularly due to large floods.

For the three datasets used in this study, there are six bands of Landsat images, of which the size of the AHB dataset Landsat Image is 2480 × 2800, the size of CIA Landsat Image is 1720 × 2040, and the size of LGC dataset Landsat Image is 3200 × 2720, according to relevant research [64]. In the model verification process, we experimented with four bands: red, green, blue, and near-infrared.

4.2. Experiments Settings

Based on previous studies, this study only uses the first four bands for prediction. A group of training data consists of three thick and thin image pairs (low-resolution and high-resolution image pairs), which are the image pairs at T0, T1, and T2. The reference images at T0 and T2 are used to predict a high-resolution image at T1. During the experiment, we used 80% of the images for pre-training, 10% for verification, and 10% for prediction. Because using the entire input image in the training process will lead to insufficient running memory, we used block training data (the entire training requires too much memory) and divided the AHB data into 160 × 160, CIA data into 128 × 128, and LGC data into 128 × 128. In terms of the training details, this model uses the Adam optimization method to update the parameters of the model, with the initial learning rate set to 0.0001, the batch size set to 8, and the epoch set to 40. Python was used to implement the experiments, which were tested on an NVIDIA RTX 3090 device. The specific experimental environment configurations are presented in Table 1.

Table 1. The experimental environment configuration of this study.

Parameter	Numerical Value	Parameter	Numerical Value
operating system	Ubuntu	CUDA	CUDA11.1
CPU	AMD EPYC 7302	cuDNN	cudnn-8.0
GPU	GeForce RTX 3090	Pytorch-GPU	1.9
RAM	63G/DDR4	GPU memory	24G

4.3. Results and Discussion

4.3.1. The Evaluation Index Used in This Experiment

At present, there is no accepted standard that can uniquely evaluate the quality of the fused images [65]. Different fusion indicators can only reflect part of the quality of the fused image [66]. Therefore, this experiment selected six indicators for evaluation. See the discussion below for details.

The peak signal-to-noise ratio (PSNR) [67] is defined as:

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right) = 20 \cdot \log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right) \tag{18}$$

where MAX_I represents the maximum value of the image point color. The higher the PSNR value between the two images, the less distorted the reconstructed image will be with respect to the high-resolution image. The mean square error (MSE) of the two images is defined as:

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1} \| I(i,j) - K(i,j) \|^2 \tag{19}$$

where I and K are two images of size $m \times n$, one is a noisy approximation of the other.

The structural similarity index (SSIM) measures the overall fusion quality by calculating the mean, variance, and covariance of the fused image and the reference image. The specific formula is shown in the following formula (see the quotation for parameter details):

$$\iota(X,Y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \tag{20}$$

$$c(X,Y) = \frac{2\delta_x\delta_y + c_2}{\delta_x^2 + \delta_y^2 + c_2} \tag{21}$$

$$s(X,Y) = \frac{\delta_{xy} + c_3}{\mu_x\mu_y + c_3} \tag{22}$$

$$SSIM(X,Y) = [l(X,Y)]^{\alpha}[c(X,Y)]^{\beta}[s(X,Y)]^{\gamma} \tag{23}$$

Usually, the closer the SSIM value is to 1, the higher the similarity between the two images.

Erreur Relative Globale Adimensionnelle De Synthèse (ERGAS) [66] is defined as:

$$ERGAS = 100\frac{h}{l}\sqrt{\frac{\sum_{i=1}^{N}(RMSE^2(B_i)/M_i^2)}{N}} \tag{24}$$

where h is the resolution of the high-resolution image, l is the resolution of the low-resolution image, N is the number of bands, B_i is the MS image, and M_i is the average of the emissivity values of the MS image. The smaller the value is, the better the spectral quality of the fused image within the spectral range.

The spectral angle mapper (SAM) [68] is defined as:

$$SAM = \arccos\left(\frac{(I_\alpha J_\alpha)}{\| I_\alpha \| \| J_\alpha \|}\right) \tag{25}$$

where I_α and J_α are the pixel vectors of the fused image and the reference image, respectively, at the distance point α. For an ideal fused image, the value of the SAM should be 0.

The spatial correlation coefficient (SCC) [69] needs to extract the high-frequency information of the correlation coefficient (CC) [70] and high pass filter. This paper uses the high Laplace filter, which is defined as:

$$F = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix} \tag{26}$$

$$CC = \frac{\sum_{i=1}^{w}\sum_{j=1}^{h}(X_{i,j} - \mu_X)(Y_{i,j} - \mu_Y)}{\sqrt{\sum_{i=1}^{w}\sum_{j=1}^{h}(X_{i,j} - \mu_X)^2(Y_{i,j} - \mu_Y)^2}} \tag{27}$$

where X is the fused image, Y is the reference image, w and h are the width and height of the image, and μ represents the average value of the image.

4.3.2. Qualitative Analysis of the Three Datasets

In this section, the AHB, CIA, and LGC datasets are used to quantitatively analyze and evaluate the research model.

Table 2 summarizes the results of the ablation experiments in this paper. We selected two backbone network models based on deep learning during the experiment: VGG [63] and RESNET34 [71]. In the ablation experiment in which the backbone network is VGG, the transformer network is combined with double branches and feedback mechanisms to verify it. In the ablation experiment with RESNET34 as the backbone network, the transformer network is combined with double branches and a feedback mechanism to verify it. It can be seen from Table 2 that under the evaluation indicators, such as PSNR, SSIM, RMSE, ERGAS, CC, SAM, and prediction time (TIME), the comprehensive evaluation of our model is more accurate than other models. This shows that applying the idea of double branches and feedback mechanisms to remote sensing image fusion at the same time can improve the quality of the fused image. However, in the parameter comparison of prediction time, our proposed model's time is not optimal due to the reference of double branches. For this problem, we will continue to optimize and verify in the later stage.

Table 2. Ablation experiments based on different backbone network models. (The abscissa represents the index of comparison, and the ordinate represents different methods. In the table, "DB" represents double branch, and "FB" represents feedback).

	PSNR	SAM	SSIM	ERGAS	CC	RMSE	TIME
TTSR	32.2552	0.1010	0.8921	1.8994	0.6023	0.0244	**308.73**
TTSR+DB+FB	32.5730	0.0875	0.8949	1.8624	0.5589	0.0235	336.70
RESNET34+DB+FB	32.3738	**0.0778**	0.9077	1.9480	0.6092	0.0241	340.73
OURS	**32.7311**	0.0889	**0.9091**	**1.5885**	**0.6414**	**0.0231**	398.46

Table 3 presents the quantitative analysis results of different fusion models on the AHB dataset. The evaluation indicators are PSNR, SSIM, RMSE, ERGAS, CC, SAM, and TIME. It can be seen from the table that the model proposed in this paper is at the maximum level except for the CC and TIME indicators. The CC indicator does not surpass other models because the AHB dataset contains rich feature features, and the deep learning model does not thoroughly learn a variety of feature features. The comprehensive evaluation results show that the proposed method can produce better fusion results regarding radiation, spatial structure, and spectrum.

Table 3. Quantitative evaluation of fusion results of the AHB dataset. (The abscissa represents the index of comparison, and the ordinate represents different methods. Bold indicates the best result).

	PSNR	SAM	SSIM	ERGAS	CC	RMSE	TIME
STARFM	25.9854	0.1843	0.7950	5.1480	0.7165	0.0506	1893.60
FSDAF	27.1617	0.1792	0.8166	4.7459	**0.7762**	0.0447	3406.97
DCSTFN	27.8403	0.1270	0.8426	2.1185	0.5622	0.0299	**301.06**
EDCSTFN	32.1307	0.1320	0.8880	2.1874	0.5869	0.0252	353.09
OURS	**32.7311**	**0.0889**	**0.9091**	**1.5885**	0.6414	**0.0231**	398.46

Table 4 presents the quantitative evaluation of the fusion results of the CIA dataset based on PSNR, SSIM, RMSE, ERGAS, CC, and SAM. The table shows that, except for SAM,

the other indicators reached the top level. This demonstrates that the proposed method can produce better radiation, spatial structure, and spectrum fusion results.

Table 4. Quantitative evaluation of fusion results of the CIA dataset. (The abscissa represents the index of comparison, and the ordinate represents different methods. Bold indicates the best result).

	PSNR	SAM	SSIM	ERGAS	CC	RMSE	TIME
STARFM	32.7311	0.0745	0.8914	1.2473	0.8358	0.0233	808.56
FSDAF	32.9512	0.0721	0.8914	1.2251	0.8424	0.0227	1067.51
DCSTFN	30.8206	**0.0638**	0.9040	1.8215	0.7563	0.0294	**25.34**
EDCSTFN	33.2827	0.0678	0.9094	1.1988	0.8580	0.0217	35.40
OURS	**33.5782**	0.0662	**0.9120**	**1.1682**	**0.8713**	**0.0210**	40.30

Table 5 presents the quantitative evaluation of the fusion results of the LGC dataset based on PSNR, SSIM, RMSE, ERGAS, CC, and SAM. As can be seen from the table, our method has reached the leading level in all indicators.

Table 5. Quantitative evaluation of fusion results of LGC dataset. (The abscissa represents the index of comparison, and the ordinate represents different methods. Bold indicates the best result).

	PSNR	SAM	SSIM	ERGAS	CC	RMSE	TIME
STARFM	31.6338	0.0540	0.9336	1.1814	0.7873	0.0266	2410.56
FSDAF	35.5282	0.0456	0.9488	0.7387	0.8984	0.0169	4208.82
DCSTFN	34.2191	0.0435	0.9485	0.8810	0.8949	0.0195	**330.10**
EDCSTFN	35.5021	0.0515	0.9585	0.8180	0.9195	0.0168	340.34
OURS	**37.3654**	**0.0361**	**0.9625**	**0.6405**	**0.9301**	**0.0135**	359.39

The experiments on the three datasets show that our method has achieved acceptable prediction results on the AHB dataset with many irregular regional phenological changes and the CIA dataset with regular regional phenological changes. Similarly, for the LGC dataset that mainly includes land cover type changes, our method is more vital to process temporal and spatial information than traditional methods, and the other two methods are based on deep learning and can achieve better prediction results.

4.3.3. Qualitative Analysis of the Three Datasets

This subsection mainly describes the qualitative analysis of this model and four classical fusion models on the three different datasets. Especially on the AHB dataset, we select other regions to evaluate the fusion results.

Figure 5 displays the results of some pasture areas on 29 August 2017. According to the third column, we can see that our method can better reconstruct the image color and contour details. According to the fourth column, the error of the depth learning method is less than that of traditional methods. The vegetation index detects vegetation growth status, coverage, and so on. Plants absorb red light due to photosynthesis. Therefore, the better-growing plants absorb red light and reflect more near-infrared light. From the fifth column, we can see that the NDVI index of our prediction results shows that the vegetation area is closer to the raw image, which indicates that this method can restore the detailed characteristics of vegetation very well.

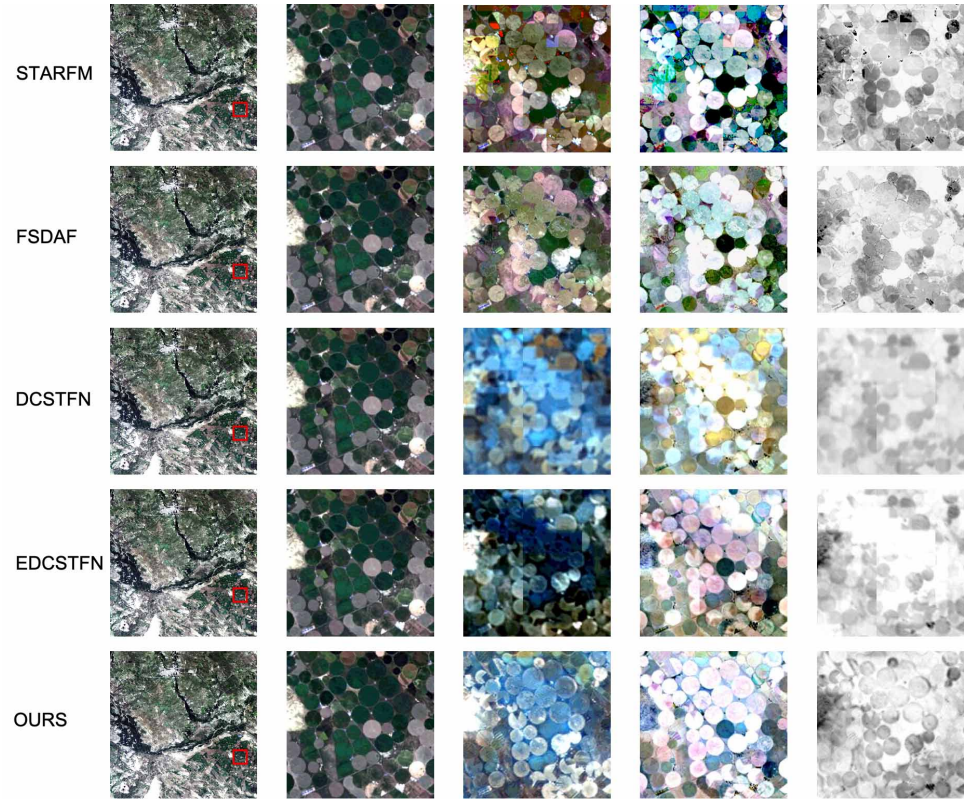

Figure 5. The results in pasture areas of the AHB dataset on 29 August 2007. (The first column shows the original image, the second column exhibits the enlarged part of the red box in the original image, the third column gives the prediction results, and the fourth column displays the difference between the prediction image and the second column of the original image. The fifth column is the calculated normalized differential vegetation index).

Figure 6 shows the results for some cities on 29 August 2017. According to the third column, we can see that this experiment is more reasonable than other models in periods of urban reconstruction, and we can see more urban contours. Although DCSTFN has a practical effect on the reconstruction of the whole image, the urban details are not well reconstructed, probably because the single branch convolution network can not extract more urban details; in the two-branch convolution model of EDCSTFN, this situation has been dramatically improved. The NDVI index of our experiment is close to the actual result.

Figure 7 shows some of the mountain results on 29 August 2017. According to the third and fourth columns, this method is similar to other deep learning and traditional methods. In the vegetation coefficient index, because the mountain area is selected as the study area, there is less green vegetation; we can see that the NDVI index of this experiment is close to the real level.

Figure 8 shows the overall results on 29 August 2017. In the overall results, the subjective effect of this experiment is slightly lower than that of other methods. The preliminary analysis shows that the network based on a transformer is slightly insufficient for the task of high-resolution large-scale images, and it may also be related to the amount of data needed. In addition, there are many different types of feature information in the AHB dataset, which makes the training more difficult.

From the experimental results of Figure 9, it can be seen that the two traditional algorithms, FSDAF and STARFM, have inevitable information loss in the image spectral information from the overall visual effect analysis. It can be seen from NDVI that STARFM has specific information loss in the prediction of green vegetation areas. Deep learning methods can get better prediction results. Figure 10 shows a zoomed-in display of the irrigated area on the CIA dataset. It can be seen that DCSTFN and EDCSTFN are still partially ambiguous, and the predictions obtained by our method are closer to the ground truth.

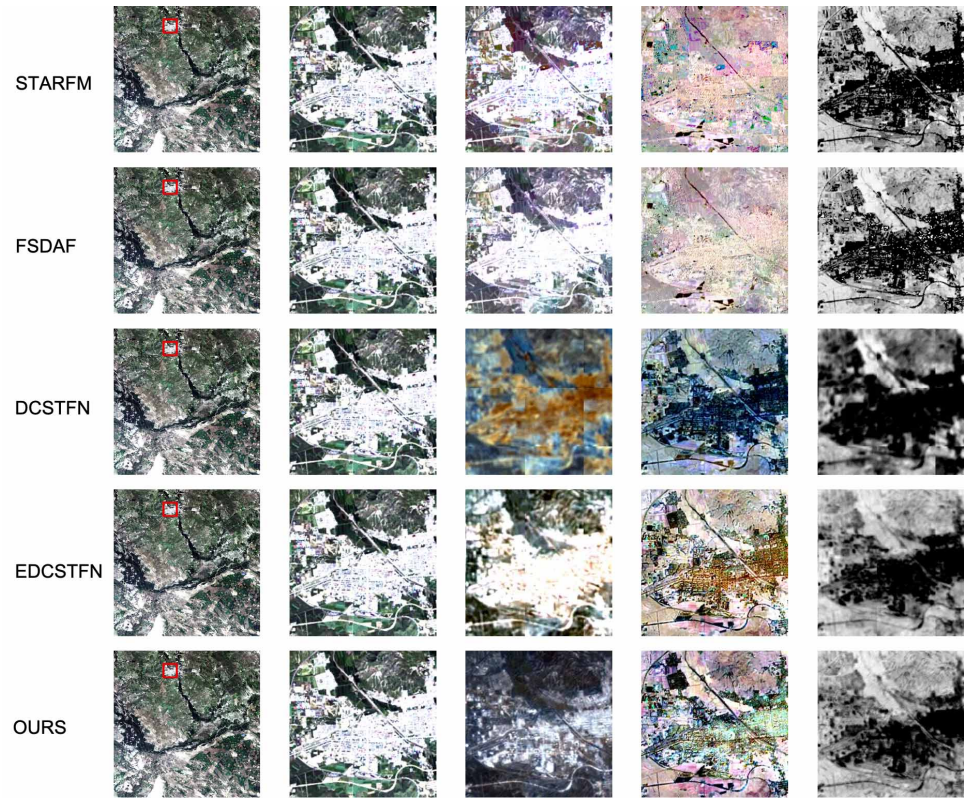

Figure 6. The results in city areas of the AHB dataset on 29 August 2007.

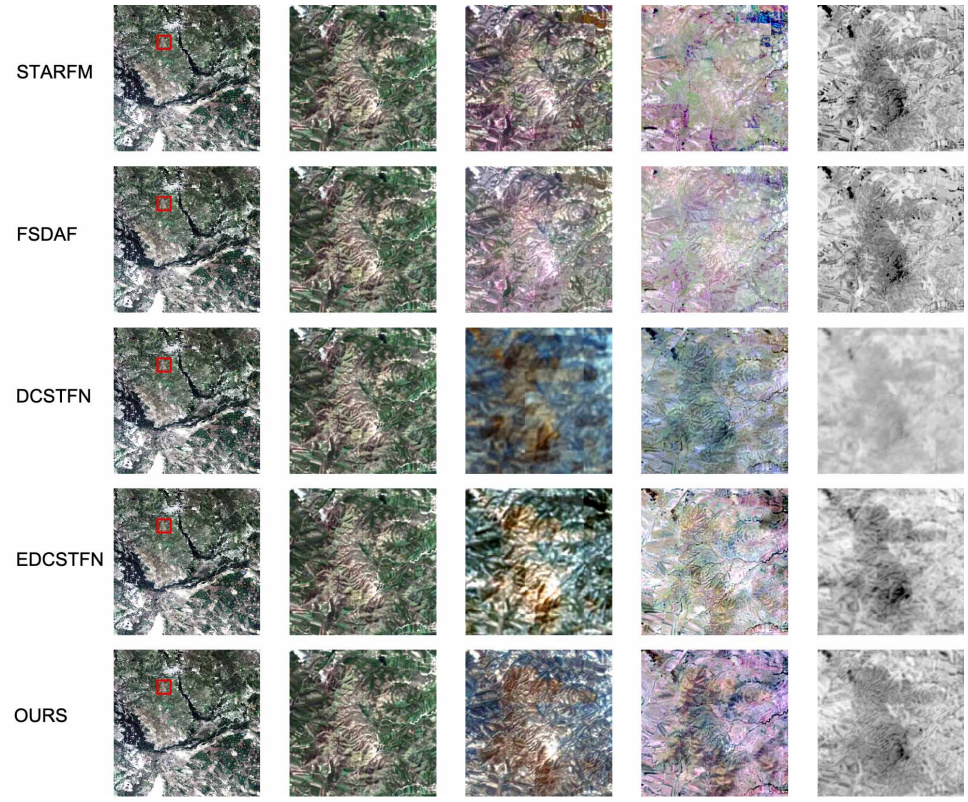

Figure 7. The results in mountain areas of the AHB dataset on 29 August 2007.

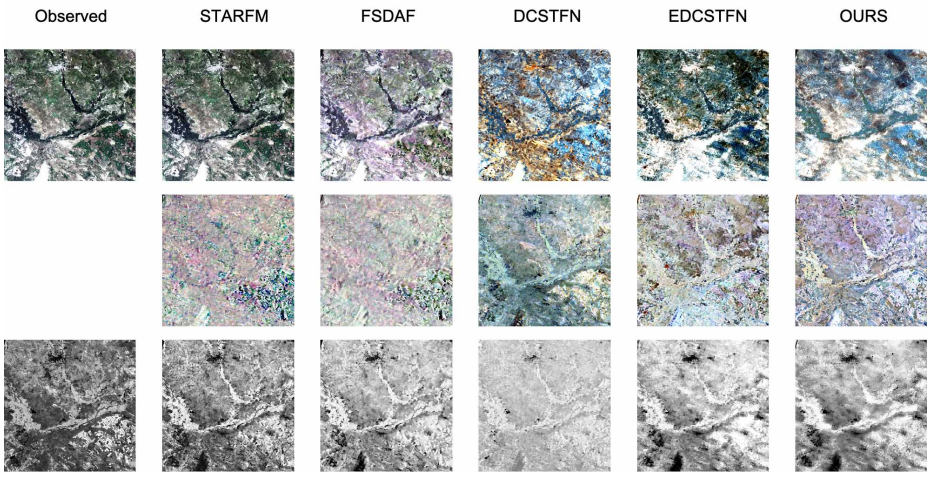

Figure 8. This figure shows the global rendering results on the AHB dataset on 29 August 2007. (In the illustration, "Observed" means the ground truth label, and "Ours" represents the model in this paper. The first row illustrates the actual color image of the individual model, the second row depicts the difference between the predicted image and the original image, and the third row is the model's NDVI prediction of the image).

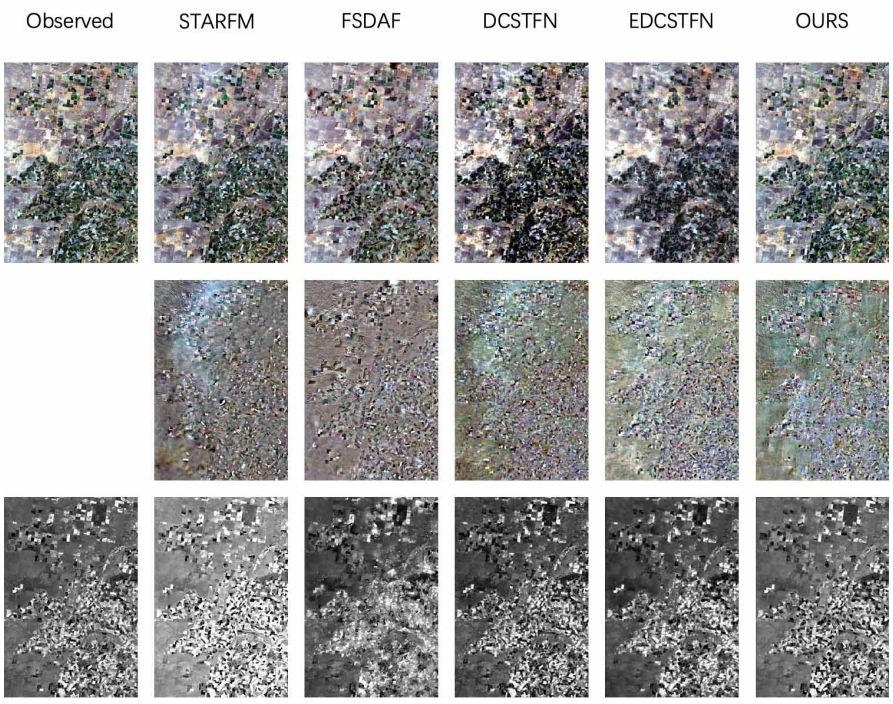

Figure 9. Global renderings showing 9 November 2001 in the CIA dataset.

Figure 11 shows the experimental results we obtained on the LGC dataset on 2 March 2005. From the overall visual effect, the performance of each algorithm is relatively stable, but there are differences in specific boundary processing and spectral information processing. Figure 12 shows an enlarged area of the LGC dataset to show details. We can see from the enlarged display of NDVI and prediction effect that our proposed method can not only better restore vegetation information but also realize the accurate prediction of boundary information and better process spectral information closer to the real value.

In addition, we draw a heat-scattering map based on the obtained experimental data on the three datasets, which mainly shows the distribution between the predicted and actual surface reflections for the four bands. The first to fourth columns in the figure represent four frequency bands, and each row represents a method. Due to the unique advantages of

the transformer and feedback mechanism in this model in extracting global information, it can be seen from Figure 13 that the method proposed in this paper has achieved good results in four bands, especially in the last two bands, which shows that our method can better capture the changes of regional rivers and road boundaries. As can be seen from the comparison chart in Figure 14, the "point cloud" of our proposed method is narrow in each band and has a high correlation, which indicates that our prediction results are closer to the actual observations, and the proposed method is more robust in handling complex changes. As shown in Figure 15, our proposed prediction results significantly impact the LGC dataset, which can capture changes in land cover types and improve the quality of fused images.

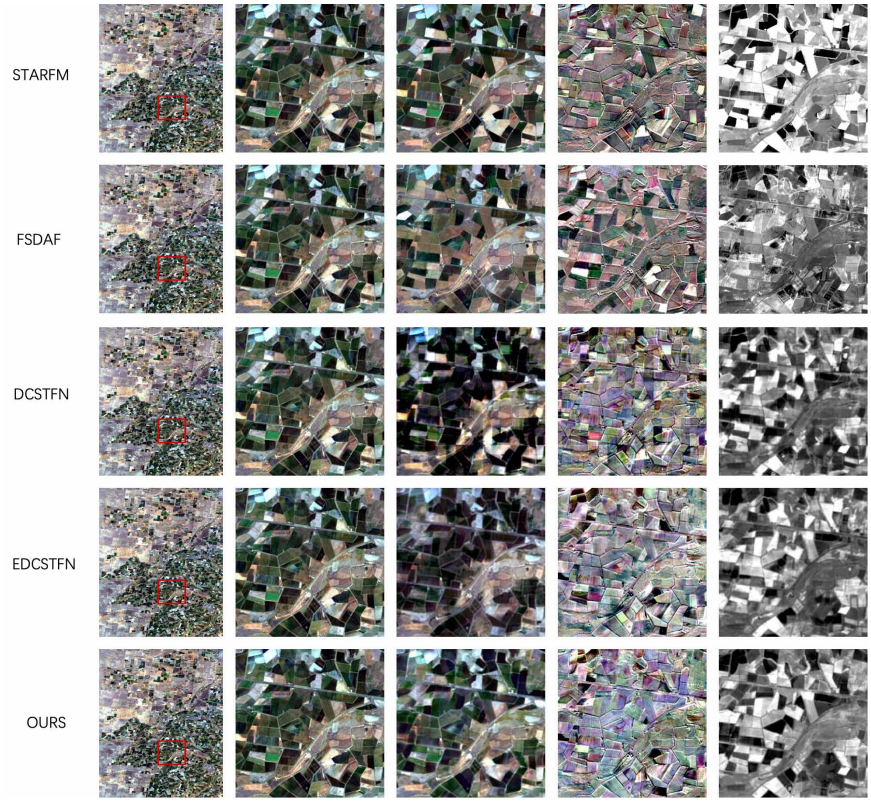

Figure 10. This figure shows the effect of zooming in on the details of the CIA dataset on 9 November 2001.

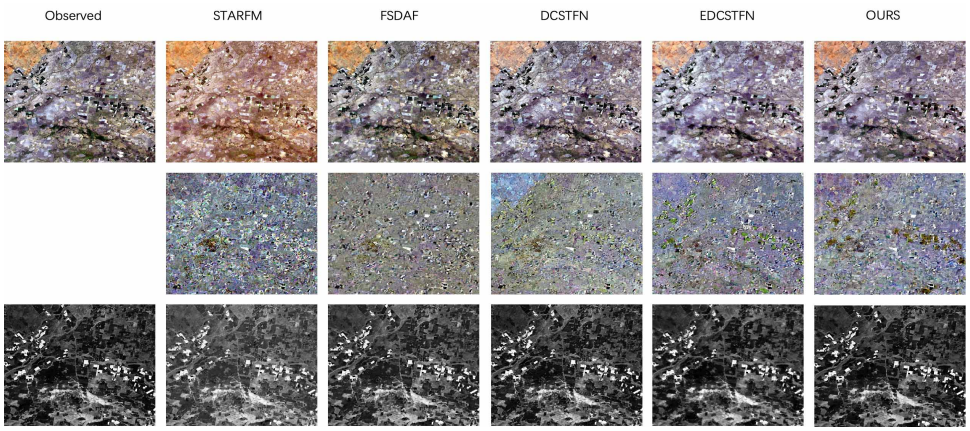

Figure 11. The global presentation results of the LGC dataset on 2 March 2005.

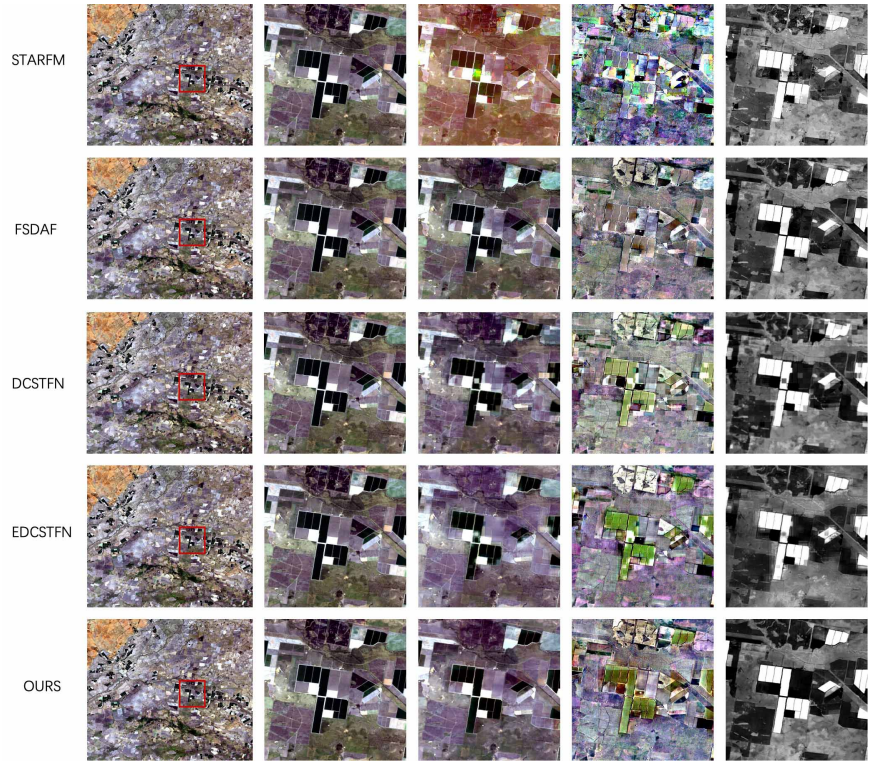

Figure 12. The partially enlarged detail of results of the LGC dataset on 2 March 2005.

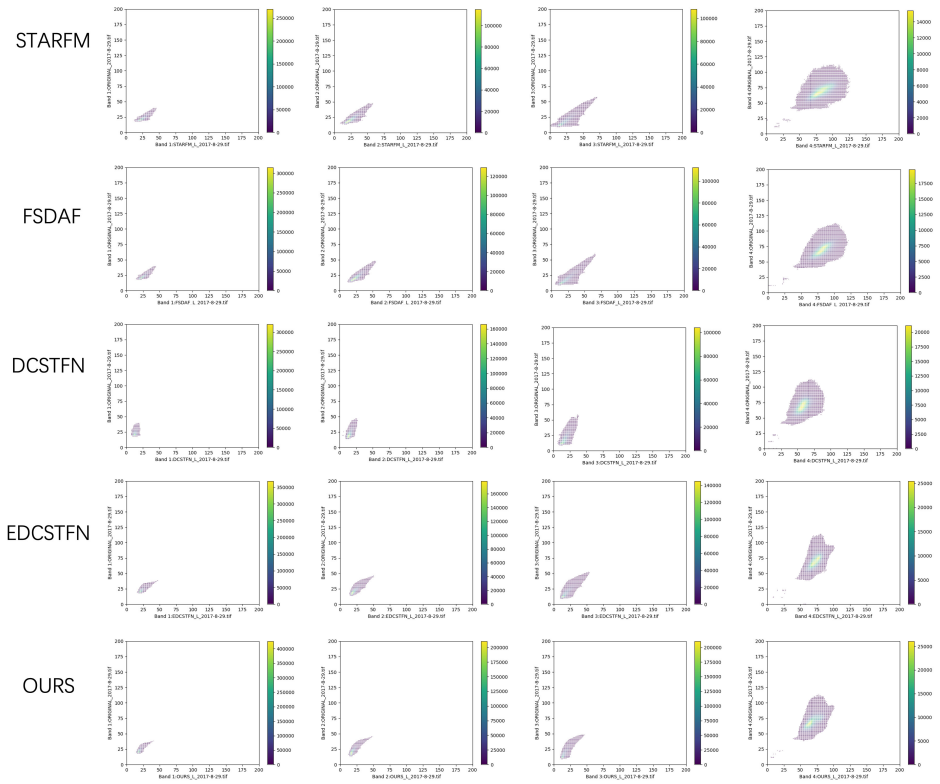

Figure 13. This figure shows a plot of the heat scatter results generated on the AHB dataset. (Zoom in to see that the abscissa represents the band of the predicted image, the ordinate is the band of the actual image, and each column represents the comparison of different models).

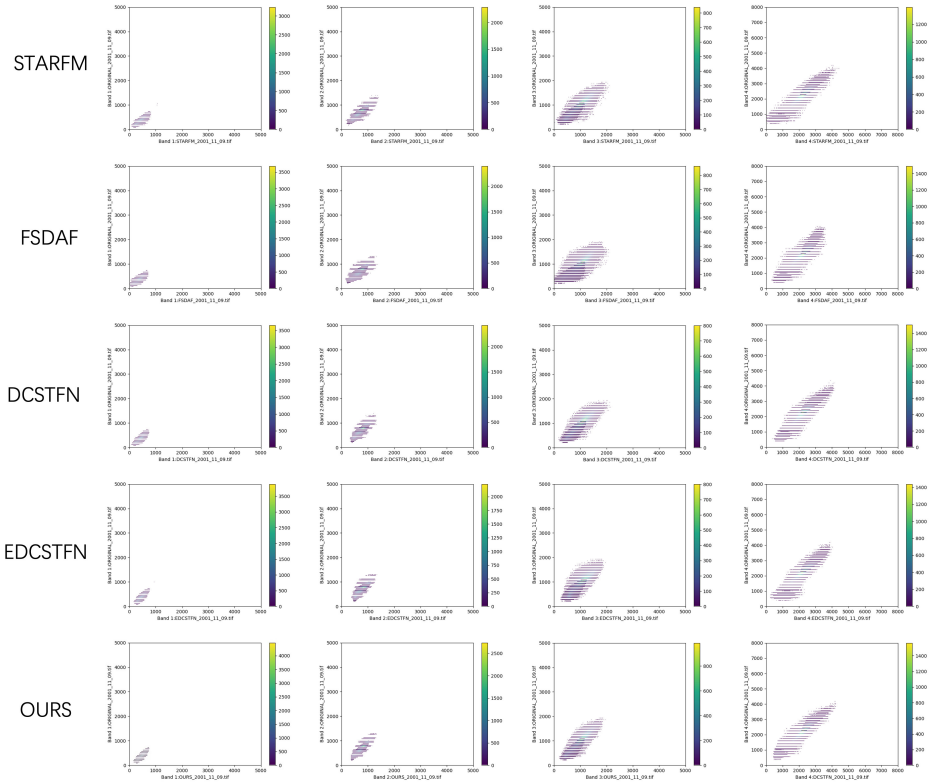

Figure 14. This figure shows the resulting thermal scatter plot on the CIA dataset.

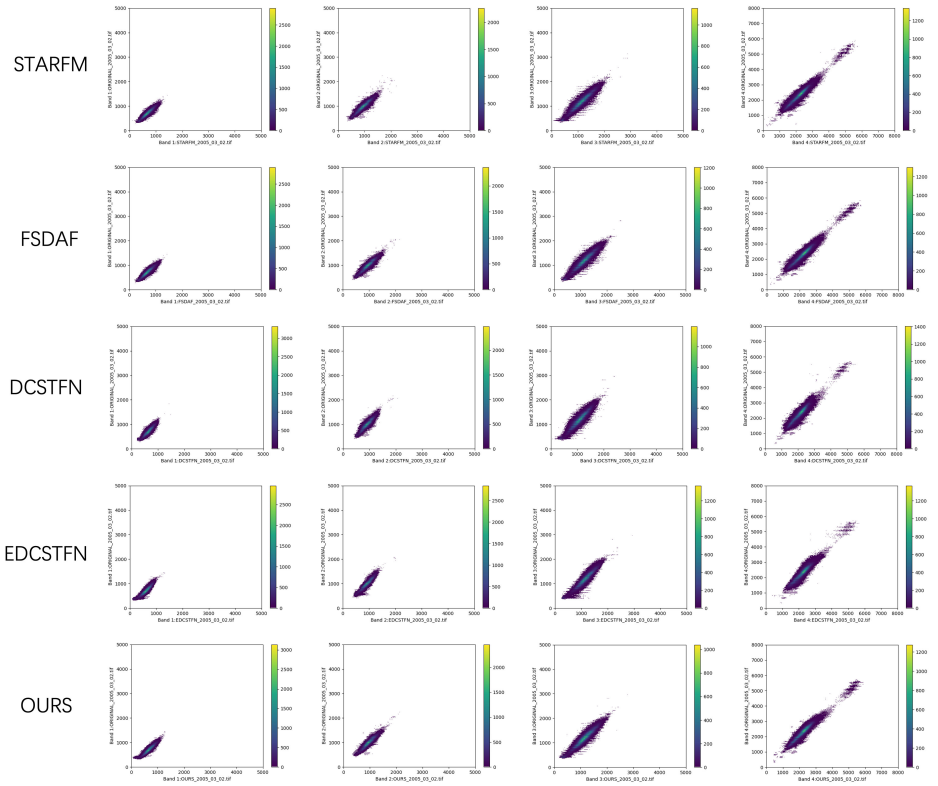

Figure 15. This figure shows the resulting thermal scatter plot on the LGC dataset.

5. Conclusions

This research optimizes and improves the prediction accuracy and image quality while reducing the memory and time consumption of the spatiotemporal fusion model based on

deep learning. This research mainly made the following contributions: first, the texture transformation model is applied to spatiotemporal fusion in the backbone network of this study to provide rich texture features in image prediction. Because the model is a deep neural network model, the interior features can be extracted by adding the self-attention mechanism module so that the model can extract both the overall feature and the local feature from the internal structure information in the image patch. Second, a feedback mechanism with high-level information refinement and low-level representation is used to achieve higher image clarity in the image reconstruction process. Our experiments show that on datasets with significant phenological changes and land cover change such as the LGC dataset and AHB dataset, the proposed model is more stable than other models.

In the future, we are going to study the following points in the field of space-time fusion of remote sensing images: (1) Reducing the model's dependence on reference images and fully takes into account the model's ability to extract complex features without a reference image. (2) The transformer has great potential in the field of remote sensing images. Next, our research will mainly be carried out with transformer and generating countermeasure networks.

Author Contributions: Conceptualization, H.L. and G.Y.; methodology, H.L.; software, H.L. and G.Y.; validation, H.L., G.Y. and H.J.; formal analysis, H.L.; resources, Y.Q.; data curation, H.L. and Y.Q.; writing—original draft preparation, H.L.; writing—review and editing, H.L., G.Y., Y.Q. and H.J.; visualization, H.L. and G.Y.; funding acquisition, Y.Q. 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 (61966035), the National Science Foundation of China under Grant (U1803261), the Xinjiang Uygur Autonomous Region Innovation Team (XJE-DU2017T002), and the Autonomous Region Graduate Innovation Project (XJ2019G069, XJ2021G062 and XJ2020G074).

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

Abbreviations

The following abbreviations are used in this manuscript:

CIA	Coleambally ignition area
LGC	Lower gwydir basin
AHB	Aruhorqin banner
HTLS	High temporal but low spatial resolution
MODIS	Moderate resolution imaging spectroradiometer
LTHS	Low temporal but high spatial resolution
EDCSTFN	Enhanced deep convolutional spatiotemporal fusion network
CNN	Convolutional neural network
SRCNN	Super-resolution convolutional network
VDSR	Very deep convolutional networks
SRGAN	Super-resolution generative adversarial network
EDSR	Enhanced deep residual networks
SRDenseNet	Super-resolution using dense skip connections network
STARFM	Spatial and temporal adaptive reflectance fusion model
STARCH	Spatiotemporal adaptive algorithm for mapping reflectance change
ESTARFM	Enhanced spatial and temporal adaptive reflectance fusion model
STRUM	Spatiotemporal reflectivity unmixing model
FSDAF	Flexible spatiotemporal data fusion
UBDF	Unmixing-based data fusion
SAM	Spatial attraction model
STDFA	Spatiotemporal data fusion algorithm

SPSTFM	Sparse-representation-based spatiotemporal reflectance fusion model
STFDCNN	Spatiotemporal fusion using deep convolutional pair neural network
DCSTFN	Deep convolutional spatiotemporal fusion network
STFNET	Twostream convolutional neural network for spatiotemporal image fusion
GAN-STFM	Generative adversarial network-based spatiotemporal fusion model
CS	Component substitution
MRA	Multiresolution analysis
IHS	Intense-hue-saturation
PCA	Principal component analysis
BT	Brovey transform
LTE	Learnable texture extractor
RE	Relational embedded
HA	Hard attention
SA	Soft attention
FB	Feedback block
RB	Reconstruction block
OLI	Operational land imager
PSNR	Peak signal-to-noise ratio
MSE	Mean squared error
SSIM	Structural similarity
RMSE	Root-mean-square error
SAM	Spectral angular similarity
CC	Correlation coefficient
ERGAS	Erreur relative globale adimensionnelle de synthèse

References

1. Tong, X.; Zhao, W.; Xing, J.; Fu, W. Status and development of china high-resolution earth observation system and application. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 3738–3741.
2. Li, D.; Wang, M.; Jiang, J. China's high-resolution optical remote sensing satellites and their mapping applications. *Geo-Spat. Inf. Sci.* **2021**, *24*, 85–94. [CrossRef]
3. Yu, B.; Shang, S. Multi-Year Mapping of Maize and Sunflower in Hetao Irrigation District of China with High Spatial and Temporal Resolution Vegetation Index Series. *Remote Sens.* **2017**, *9*, 855. [CrossRef]
4. Walker, J.; De Beurs, K.; Wynne, R.; Gao, F. Evaluation of Landsat and MODIS data fusion products for analysis of dryland forest phenology. *Remote Sens. Environ.* **2012**, *117*, 381–393. doi: 10.1016/j.rse.2011.10.014. [CrossRef]
5. Hansen, M.C.; Loveland, T.R. A review of large area monitoring of land cover change using Landsat data. *Remote Sens. Environ.* **2012**, *122*, 66–74. doi: 10.1016/j.rse.2011.08.024. [CrossRef]
6. Kyrkou, C.; Theocharides, T. EmergencyNet: Efficient Aerial Image Classification for Drone-Based Emergency Monitoring using Atrous Convolutional Feature Fusion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 1687–1699. [CrossRef]
7. Nair, H.C.; Padmalal, D.; Joseph, A.; Vinod, P. Delineation of groundwater potential zones in river basins using geospatial tools—An example from southern western Ghats, Kerala, India. *J. Geovisualization Spat. Anal.* **2017**, *1*, 5. [CrossRef]
8. Patanè, G.; Spagnuolo, M. Heterogeneous Spatial Data: Fusion, Modeling, and Analysis for GIS Applications. *Synth. Lect. Vis. Comput. Comput. Graph. Animat. Comput. Photogr. Imaging* **2016**, *8*, 1–155.
9. Shen, H.; Meng, X.; Zhang, L. An integrated framework for the spatio–temporal–spectral fusion of remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7135–7148. [CrossRef]
10. Zhu, X.; Cai, F.; Tian, J.; Williams, T. Spatiotemporal Fusion of Multisource Remote Sensing Data: Literature Survey, Taxonomy, Principles, Applications, and Future Directions. *Remote Sens.* **2018**, *10*, 527. [CrossRef]
11. 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.
12. Tsai, R. Multiframe image restoration and registration. *Adv. Comput. Vis. Image Process.* **1984**, *1*, 317–339.
13. Zhang, K.; Tao, D.; Gao, X.; Li, X.; Xiong, Z. Learning multiple linear mappings for efficient single image super-resolution. *IEEE Trans. Image Process.* **2015**, *24*, 846–861. [CrossRef] [PubMed]
14. Zhang, Y.; Wu, W.; Dai, Y.; Yang, X.; Yan, B.; Lu, W. Remote sensing images super-resolution based on sparse dictionaries and residual dictionaries. In Proceedings of the 2013 IEEE 11th International Conference on Dependable, Autonomic and Secure Computing, Chengdu, China, 21–22 December 2013; pp. 318–323.
15. Wu, W.; Yang, X.; Liu, K.; Liu, Y.; Yan, B.; Hua, H. A new framework for remote sensing image super-resolution: Sparse representation-based method by processing dictionaries with multi-type features. *J. Syst. Archit.* **2016**, *64*, 63–75. [CrossRef]
16. Yang, J.; Wright, J.; Huang, T.S.; Ma, Y. Image super-resolution via sparse representation. *IEEE Trans. Image Process.* **2010**, *19*, 2861–2873. [CrossRef]

17. Singh, S.; Mahmood, A. The NLP Cookbook: Modern Recipes for Transformer Based Deep Learning Architectures. *IEEE Access* **2021**, *9*, 68675–68702. [CrossRef]
18. Guo, Y.; Liu, Y.; Lao, S.; Bakker, E.M.; Bai, L.; Lew, M.S. Bag of Surrogate Parts Feature for Visual Recognition. *IEEE Trans. Multimed.* **2017**, *20*, 1525–1536. [CrossRef]
19. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
20. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef]
21. 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]
22. Liu, Y.; Chen, X.; Peng, H.; Wang, Z. Multi-focus image fusion with a deep convolutional neural network. *Inf. Fusion* **2017**, *36*, 191–207. [CrossRef]
23. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 295–307. [CrossRef]
24. Zhang, X.; Zou, J.; He, K.; Jian, S. Accelerating Very Deep Convolutional Networks for Classification and Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1943-1955. [CrossRef]
25. Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. *arXiv* **2016**, arXiv:1609.04802.
26. Lim, B.; Son, S.; Kim, H.; Nah, S.; Mu Lee, K. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017 ; pp. 136–144.
27. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
28. Tong, T.; Li, G.; Liu, X.; Gao, Q. Image super-resolution using dense skip connections. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4799–4807.
29. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
30. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y. Residual dense network for image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2472–2481.
31. Acerbi-Junior, F.; Clevers, J.; Schaepman, M.E. The assessment of multi-sensor image fusion using wavelet transforms for mapping the Brazilian Savanna. *Int. J. Appl. Earth Obs. Geoinf.* **2006**, *8*, 278–288. [CrossRef]
32. Chen, B.; Huang, B.; Xu, B. Comparison of spatiotemporal fusion models: A review. *Remote Sens.* **2015**, *7*, 1798–1835. [CrossRef]
33. 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]
34. 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]
35. Gevaert, C.M.; García-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]
36. 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.* **2016**, *172*, 165–177. [CrossRef]
37. Zurita-Milla, R.; Clevers, J.G.; Schaepman, M.E. Unmixing-based Landsat TM and MERIS FR data fusion. *IEEE Geosci. Remote Sens. Lett.* **2008**, *5*, 453–457. doi: 10.1109/LGRS.2008.919685. [CrossRef]
38. Lu, L.; Huang, Y.; Di, L.; Hang, D. A new spatial attraction model for improving subpixel land cover classification. *Remote Sens.* **2017**, *9*, 360. [CrossRef]
39. Wu, M.; Niu, Z.; Wang, C.; Wu, C.; Wang, L. Use of MODIS and Landsat time series data to generate high-resolution temporal synthetic Landsat data using a spatial and temporal reflectance fusion model. *J. Appl. Remote Sens.* **2012**, *6*, 063507.
40. Belgiu, M.; Stein, A. Spatiotemporal image fusion in remote sensing. *Remote Sens.* **2019**, *11*, 818. [CrossRef]
41. Huang, B.; Zhang, H.; Song, H.; Wang, J.; Song, C. Unified fusion of remote-sensing imagery: Generating simultaneously high-resolution synthetic spatial–temporal–spectral earth observations. *Remote Sens. Lett.* **2013**, *4*, 561–569. [CrossRef]
42. Xue, J.; Leung, Y.; Fung, T. A Bayesian data fusion approach to spatio-temporal fusion of remotely sensed images. *Remote Sens.* **2017**, *9*, 1310. [CrossRef]
43. Tan, Z.; Yue, P.; Di, L.; Tang, J. Deriving high spatiotemporal remote sensing images using deep convolutional network. *Remote Sens.* **2018**, *10*, 1066. [CrossRef]
44. 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]
45. 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]
46. Tan, Z.; Gao, M.; Li, X.; Jiang, L. A flexible reference-insensitive spatiotemporal fusion model for remote sensing images using conditional generative adversarial network. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*. [CrossRef]

47. CARPER, W.; LILLESAND, T.; KIEFER, R. The use of intensity-hue-saturation transformations for merging SPOT panchromatic and multispectral image data. *Photogramm. Eng. Remote Sens.* **1990**, *56*, 459–467.
48. Tu, T.M.; Su, S.C.; Shyu, H.C.; Huang, P.S. A new look at IHS-like image fusion methods. *Inf. Fusion* **2001**, *2*, 177–186. [CrossRef]
49. González-Audícana, M.; Saleta, J.L.; Catalán, R.G.; García, R. Fusion of multispectral and panchromatic images using improved IHS and PCA mergers based on wavelet decomposition. *IEEE Trans. Geosci. Remote Sens.* **2004**, *42*, 1291–1299. [CrossRef]
50. Pohl, C.; Van Genderen, J.L. Review article multisensor image fusion in remote sensing: Concepts, methods and applications. *Int. J. Remote Sens.* **1998**, *19*, 823–854. [CrossRef]
51. Shahdoosti, H.R.; Ghassemian, H. Combining the spectral PCA and spatial PCA fusion methods by an optimal filter. *Inf. Fusion* **2016**, *27*, 150–160. [CrossRef]
52. Choi, J.; Yu, K.; Kim, Y. A new adaptive component-substitution-based satellite image fusion by using partial replacement. *IEEE Trans. Geosci. Remote Sens.* **2010**, *49*, 295–309. doi: 10.1109/TGRS.2010.2051674. [CrossRef]
53. Burt, P.J.; Adelson, E.H. The Laplacian pyramid as a compact image code. In *Readings in Computer Vision*; Elsevier: Amsterdam, The Netherlands, 1987; pp. 671–679.
54. Shensa, M.J. The discrete wavelet transform: Wedding the a trous and Mallat algorithms. *IEEE Trans. Signal Process.* **1992**, *40*, 2464–2482. [CrossRef]
55. Choi, M.; Kim, R.Y.; Nam, M.R.; Kim, H.O. Fusion of multispectral and panchromatic satellite images using the curvelet transform. *IEEE Geosci. Remote Sens. Lett.* **2005**, *2*, 136–140. [CrossRef]
56. Ghahremani, M.; Ghassemian, H. Remote-sensing image fusion based on curvelets and ICA. *Int. J. Remote Sens.* **2015**, *36*, 4131–4143. [CrossRef]
57. Ji, X.; Zhang, G. Image fusion method of SAR and infrared image based on Curvelet transform with adaptive weighting. *Multimed. Tools Appl.* **2017**, *76*, 17633–17649. [CrossRef]
58. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 6000–6010.
59. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
60. Gilbert, C.D.; Sigman, M. Brain states: Top-down influences in sensory processing. *Neuron* **2007**, *54*, 677–696. [CrossRef]
61. Hupé, J.; James, A.; Payne, B.; Lomber, S.; Girard, P.; Bullier, J. Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. *Nature* **1998**, *394*, 784–787. [CrossRef]
62. Li, Z.; Yang, J.; Liu, Z.; Yang, X.; Jeon, G.; Wu, W. Feedback network for image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3867–3876.
63. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
64. Li, J.; Li, Y.; He, L.; Chen, J.; Plaza, A. Spatio-temporal fusion for remote sensing data: An overview and new benchmark. *Sci. China Inf. Sci.* **2020**, *63*, 140301. [CrossRef]
65. Chen, Z.; Pu, H.; Wang, B.; Jiang, G.M. Fusion of hyperspectral and multispectral images: A novel framework based on generalization of pan-sharpening methods. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 1418–1422. [CrossRef]
66. Wald, L. Quality of high resolution synthesised images: Is there a simple criterion? In Proceedings of the Third Conference "Fusion of Earth Data: Merging Point Measurements, Raster Maps and Remotely Sensed Images", Sophia Antipolis, France, 28–30 January 2000; pp. 99–103.
67. Huynh-Thu, Q.; Ghanbari, M. Scope of validity of PSNR in image/video quality assessment. *Electron. Lett.* **2008**, *44*, 800–801. [CrossRef]
68. Yuhas, R.H.; Goetz, A.F.; Boardman, J.W. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm. In Proceedings of the Summaries 3rd Annual JPL Airborne Geoscience Workshop (AVIRIS Workshop), Pasadena, CA, USA, 1–5 June 1992; Volume 1, pp. 147–149.
69. Zhou, J.; Civco, D.; Silander, J. A wavelet transform method to merge Landsat TM and SPOT panchromatic data. *Int. J. Remote Sens.* **1998**, *19*, 743–757. [CrossRef]
70. 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]
71. Lau, S.; Wang, X.; Yang, X.; Chong, E. Automated Pavement Crack Segmentation Using Fully Convolutional U-Net with a Pretrained ResNet-34 Encoder. *IEEE Access* **2020**, *8*, 114892–114899. [CrossRef]